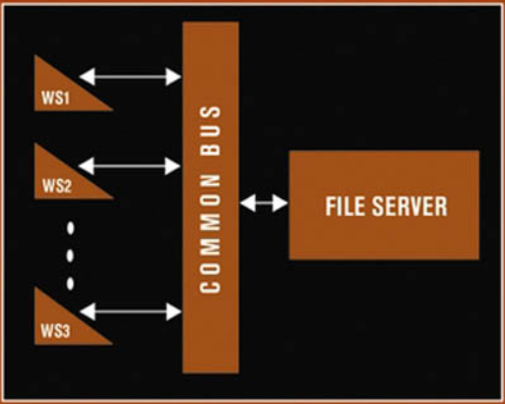
WILEY

Modeling and Analysis of Telecommunications Networks



JEREMIAH F. HAYES

THIMMA V. J. GANESH BABU

ftp://

MODELING AND ANALYSIS OF TELECOMMUNICATIONS NETWORKS

MODELING AND ANALYSIS OF TELECOMMUNICATIONS NETWORKS

JEREMIAH F. HAYES
THIMMA V. J. GANESH BABU

WILEY-
INTERSCIENCE

A JOHN WILEY & SONS, INC., PUBLICATION

Published by John Wiley & Sons, Inc., Hoboken, New Jersey.
Published simultaneously in Canada.

Limit of Liability/Disclaimer of Warranty: While the publisher and author have used their best efforts in preparing this book, they make no representations or warranties with respect to the accuracy or completeness of the contents of this book and specifically disclaim any implied warranties of merchantability or fitness for a particular purpose. No warranty may be created or extended by sales representatives or written sales materials. The advice and strategies contained herein may not be suitable for your situation. You should consult with a professional where appropriate. Neither the publisher nor author shall be liable for any loss of profit or any other commercial damages, including but not limited to special, incidental, consequential, or other damages.

For general information on our other products and services please contact our Customer Care Department within the U.S. at 877-762-2974, outside the U.S. at 317-572-4002.

Wiley also published its books in a variety of electronic formats. Some content that appears in print, however, may not be available in electronic format.

Library of Congress Cataloging-in-Publication Data:

Hayes, Jeremiah F., 1934-
Modeling and analysis of telecommunications networks/Jeremiah F.
Hayes & Thimma V.J. Ganesh Babu.
p. cm.
"A Wiley-Interscience Publication."
ISBN 0-471-34845-7 (Cloth)
1. Telecommunication systems. I. Babu, Thimma V.J. Ganesh. II. Title.

TK5101 .H39 2004
621.382′1–dc22

2003020806

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

To the Older Generation

T. V. Janarthanan, T. J. Kanchanamala,
Yvette Sacoutis, and . . .

To the Younger Generation

Katie, Liam, Alex, Taian, Sophie, Oliver, Stephanie,
Alexa, and . . .

CONTENTS

PREFACE

BACKGROUND

The insinuation of telecommunications into the daily fabric of our lives has been arguably the most important and surprising development of the last 25 years. Before this revolution, telephone service and its place in our lives had been largely stable for more than a generation. The growth was, so to speak, lateral, as the global reach of telecommunications extended and more people got telephone service. The distinction between oversea and domestic calls blurred with the advances in switching and transmission, undersea cable, and communication satellites. Traffic on the network remained overwhelmingly voice, largely in analog format with facsimile (Fax) beginning to make inroads. A relatively small amount of data traffic was carried by modems operating at rates up to 9600 bits per second over voice connections. Multiplexing of signals was rudimentary—most connections were point-to-point business applications.

The contrast with today's network is overwhelming. The conversion from analog to digital has long since been completed. A wide range of services, each with its unique set of traffic characteristics and performance requirements, are available. At the core of the change is the Internet, which is becoming accepted to handle all telecommunications traffic and functions.

In order to effect such a far-reaching change, many streams converged. At the most basic level was the explosive growth of the technology. The digital switching and processing that are intrinsic to the modern network are possible only through integrated-circuit technology. There are any number of examples of this technology at work. For one who has worked in data communications, the most striking is the voice-band modem. At the beginning of the era the rule of thumb for cost was a dollar per bit per second. Modems were typically the size of a VCR. Integrated-circuit technology and other factors, which we consider next, contributed to the several-fold increase in performance, with 56 kilobits per second now routine with a modem costing pennies.

A second key technological development was optical fiber, which is virtually the perfect transmission medium. Using techniques learned in the manufacture of semiconductors the transparency of fiber has been brought down to fractions of a decibel (dB) loss per kilometer. The deployment of fiber led to a quantum leap in transmission rates. We have had to learn new prefixes to quantify the rates: *giga*bits per second (10^9) and *tera*bits per second (10^{12}). On the horizon are *peta*bits per second (10^{15}) and *exa*bits per second (10^{18}) rates. Moreover, the deployment of fiber has decreased the error rate in the telephone network by orders of magnitudes.

Important advances in modulation and coding also contributed to the revolution. This stream of knowledge has been particularly important in the mobile phone business, where a dB increase in performance can increase profit by millions. It is ironic and instructive that, at one point, error-correcting codes were written off as only being of academic interest. Recently, turbocodes have been shown to achieve Shannon capacity, which had been the Holy Grail of transmission performance.

A significant contribution to the advance was made by what may be broadly, but perhaps inaccurately, called "software." Under this heading, we place the new ways of organizing the information flow in the network. As we mentioned, at the beginning of the period, the telephone network, oriented as it was toward voice traffic, was not hospitable to data traffic. Impairments, which were no problem for voice, would kill data signals. Further, the charging structure was totally oriented toward voice traffic. The response to these problem was *packet switching*, in which data are encapsulated into packets and passed through the network in hops between intermediate points. Error can be controlled each at each hop or end-to-end, as appropriate to the error conditions. The user pays only for the amount of information transmitted, not for the duration of a 3-min voice call.

Packet switching is at the core of the system that epitomized the telecommunications revolution, the ARPANET, which went into service in 1971. The network-linked computers at universities that receive research contracts, from the Advanced Research Projects Agency (ARPA), an entity within the U.S. Department of Defense. So successful was this network that it rapidly, and in an unprecedented fashion, evolved into the worldwide network we have today: the Internet.

The design objective was to provide reliable communications, suited to traffic between computers, over an unreliable network. The TCP/IP (Transmission Control Protocol/Internet Protocol) protocols did the job so effectively that they are now the ascendant way of organizing information flow in the telecommunications network, in spite of the fact that the telephone network has improved greatly in reliability.

One can make the case that the first major use of the ARPANET, e-mail (electronic mail), was almost accidental. The original purpose of the ARPANET was to provide a resource that would allow all the university computers on the network to be linked into one giant computational engine. Simple mail protocols were written to allow the technicians on the project to talk to one another in the course of their work. Faculty members associated with the project also found the service useful to keep in touch with their peers. And so on it went, from faculty member on the project to members of electrical engineering and computer science departments, to the rest of the university and on to commercial operation. The miracle was that it was all wholly unplanned. The basic system was so powerful and flexible that human ingenuity found a way to use and evolve it.

Another important development at the beginning of the Internet era was ALOHA, which was developed at the University of Hawaii with the objective of allowing a large number of low-volume users to share a wireless data network among the Hawaiian islands. The technique is packet-oriented and employs a dispersed system of traffic control*, random access.* ALOHA was the ancestor of the Ethernet, which is the ubiquitous, de facto standard for forming networks of computers in industry, government, university, and home settings. An Ethernet, perhaps including wireless access points, can link together all the appliances in your home and all the processors in your car.

We should also mention important legal developments that played a part in the communications revolution. Previously, the telephone network was operated as a public utility, like water and electricity. Innovations were brought on line in accordance with the conservative economics governing utilities. The telephone administrations had a monopoly on the equipment deployed. The system served society very well, but was slow to respond to the opportunities and demands of advancing technology, as illustrated by difficulties with data communication, involving some degree of networking. The 1971 “Carterfone” decision in the United States seems to have broken the logjam. The decision was to allow nontelephone equipment to be connected to the network subject only to safety considerations. This was followed by antitrust action by the U.S. government, which broke American Telephone and Telegraph into a number of smaller companies. Then, the present intensely competitive environment ensued, with both benefits and losses for the public. Only history will be able to say if this entire process was progress.

From our perspective, the most important change is in the nature of the traffic carried on the network. Voice and simple data traffic have given way to a range of traffic of daunting complexity. Each of the many applications has its own characteristic and requirements. At one end of the spectrum is voice, which is robust to error, but which had stringent delay requirements. In contrast, the bulk transfer of records must be totally accurate, but the timeframe of delivery is relaxed. Furthermore, the real-time interactions on the network give rise to the entirely new phenomenon of fractal-like traffic patterns. Such behavior has, for example, been observed in Ethernets.

In the face of the complex nature of telecommunications traffic, mathematics, basically in the form of queuing theory, has a role in evaluating performance. We present an introduction to the subject of modeling and analysis starting from first principles.

TEXT OVERVIEW

The text is composed of nine chapters, outlined as follows:

Chapter 1—a survey of queuing theory from a historical perspective.

Chapter 2—a review of the mathematics of probability with particular emphasis on material used later in the text.

Chapter 3—a study of pure birth and birth and death processes. In the former, we stress the Poisson process, which is basic to elementary traffic analysis. The latter processes are applied directly to queuing models. Even at this level widely useful results are derived.

Chapter 4—networks made up a number of queues. These queues are treated as multidimensional birth and death processes, and the networks can be used to model a wide range of telecommunications systems.

Chapter 5—Markov chains, which were introduced in Chapter 2. They are developed here and applied to multiplexing for synchronous, asynchronous, and random access.

Chapter 6—extention of the concept of the Markov chaine to the embedded Markov chain. The important application is the analysis of queues with a general distribution of service times.

Chapter 7—the fluid flow approximation. This is a new technique that allows the modeling of traffic that is engendered by the new applications in the telephone network. The Poisson process is not appropriate for much of this traffic.

Chapter 8—the matrix analytic technique, introduced at a basic level. This is an alternative, more powerful approach to modeling a wide range of traffic.

Chapter 9—simulation, which complements the analytic techniques in the rest of the text. This chapter presents the basic concepts.

A unique feature of the text is the accompanying software that appears on the Website for the text. The motivation for this material is to generate concrete numbers from the formulas that we derive. In order to compute these numbers, we have used three tools, each with its own capabilities. *Excel* does simple calculations quickly and efficiently. Further, it has a nice interactive capability. *Matlab* is a more powerful tool, which is particularly well suited to matrix operations. It also provides the ability to make very attractive graphics. The third tool that we use is *Maple*, which allows symbolic manipulations. The text material has been used in final-year undergraduate courses and in first-year graduate courses. There are several sequences that can be followed in teaching the text material. Presumably, the students have a background in probability; consequently, the material in Chapter 2 is meant for review and reference. Chapter 3 is needed for all the subsequent chapters, except 9. Chapter 4 stands on its own and is not required for the remaining chapters. Chapters 5 and 6 should be covered in tandem. The advanced material in Chapters 7 and 8 could be covered with a prerequisite of only Chapter 3 and some general background. Finally, Chapter 9, which covers the basics of simulations, requires only basic probability.

In thinking about the book and its subject, the words of Oscar Wilde in his play *The Importance of Being Ernest* seem appropriate: “Oh, yes Dr. Chasuble is a most learned man. He has never written a book, so you can imagine how much he knows.”

The quotation is meant to portray our trepidation in approaching the vast and complex task of applying mathematics to the performance of the telecommunications networks. Our goal is suitably modest—to furnish a guide for beginners, some of whom will go far beyond the bounds of our text. Hopefully, all who use this book will appreciate the place of mathematics in the work of understanding one of humankind’s most complex works—the telecommunications network.

We would like to thank Steve Weinstein, Peder Emstad, and Sanjeev Verma for their help. Finally, one of the authors expresses his appreciation to his wife, Niki Sacoutis, for all her support, in particular the comfortable home atmosphere that is the result of her work.

Jeremiah F. Hayes
Thimma V. J. Ganesh Babu

Concordia University
Montreal, Quebec, Canada

RETRIEVING FILES FROM THE WILEY FTP AND INTERNET SITES

To download software referred to in the examples and used to generate figures in the book, use an ftp program or a Web browser.

FTP ACCESS

If you are using an ftp program, type the following at your ftp prompt:

```
ftp://ftp.wiley.com
```

Some programs may provide the first “ftp” for you, in which case type:

```
ftp.wiley.com
```

Log in as anonymous (e.g., User ID: anonymous). Leave the password blank. After you have connected to the Wiley ftp site, navigate through the directory path of:

```
ftp://ftp.wiley.com/public/sci_tech_med/
telecommunications_networks
```

WEB ACCESS

If you are using a standard Web browser, type the URL address of:

```
ftp://ftp.wiley.com
```

Navigate through the directory path of:

```
ftp://ftp.wiley.com/public/sci_tech_med/
telecommunications_networks
```

If you need further information about downloading the files, you can call Wiley Technical Support at 201-748-6753.

1

PERFORMANCE EVALUATION IN TELECOMMUNICATIONS

1.1 INTRODUCTION: THE TELEPHONE NETWORK

The performance of a telecommunications system is a subject of considerable importance. Before going further, we must first describe the system under study.[1] The model of a generic telephone system as depicted in Figure 1.1 shows four basic components: *customer premises equipment*, the *local network*, the *switching plant*, and the *long-haul network*. As we look in detail at each of the categories, we see increasing complexity reflecting the explosive growth of the telecommunications industry.

1.1.1 Customer Premises Equipment

The most common piece of equipment on the customer's premises is, of course, the ordinary telephone, the handset. Today, a number of other items can also be found. There is scarcely a telephone without an answering service, which is provided by a box next to the phone. Up until the early 1990s the Internet was the domain of the technical types. Today the personal computer with the attendant *modem* is an appliance found in many homes. Higher-capacity modems convey data from more

[1] The Further Reading section at the end of the chapter includes several telecommunication sources that we have found to be useful. The history of telecommunications is itself an interesting study. We have found two Websites on the subject: *http://www-stall.rz.fht-esslingen.de/telehistory/* and *http://www.webbconsult.com/timeline.html*.

Modeling and Analysis of Telecommunications Networks, by Jeremiah F. Hayes and
Thimma V. J. Ganesh Babu
ISBN 0-471-34845-7 © 2004 John Wiley & Sons, Inc.

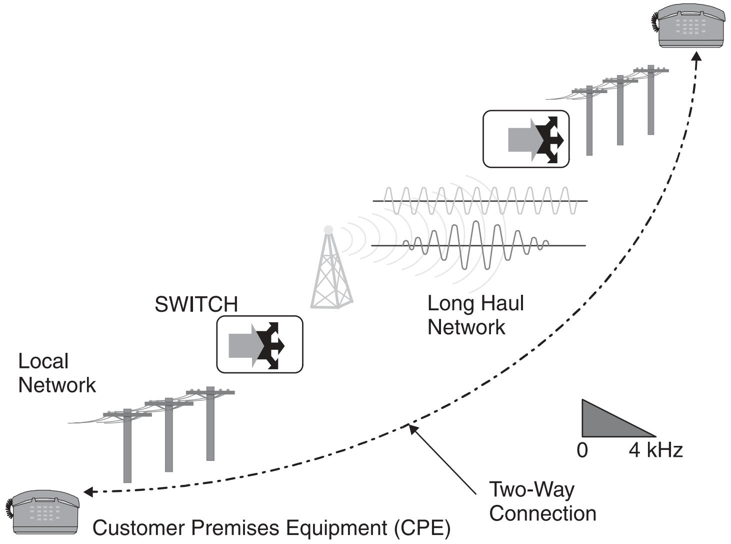


Figure 1.1 Telephone system.

complex processing center industrial sites (see discussion below). *Facsimile* (Fax) *machines* have also experienced explosive growth in deployment, and the private branch exchange (PBX), which is simply a local switchboard for connecting the phones in an office building to the outside world, has in many instances been replaced by central national switching hubs.

In industrial, commercial, and scientific sites an ubiquitous installation is the *local-area network* (LAN). A LAN may range from a simple bus with a few computers to a complex of switches and routers linking together hundreds of computers and specialized pieces of equipment. A LAN may or may not be connected to the telephone network, but it is a very important part of modern telecommunication. A great deal of effort has been devoted to a study of its performance.[2]

The complexity and growth of the network precludes easy categorization. Where do we place the ubiquitous cell phone? Considering its function, we can put it in the same category as the ordinary wired phone as a piece of customer service equipment.

1.1.2 The Local Network

The *local network* is the means by which the customer premises equipment is connected to the telephone network. All of the ordinary telephones are connected to the local telephone end office through pairs of 22- or 26-gauge wire twisted together

[2] Three papers from which modern LANs have evolved are listed in the Further Reading section. For example, the ubiquitous Ethernet may be seen as a lineal descendant of ALOHA.

in order to minimize crosstalk. Hundreds of these twisted pairs are bundled in a cable. In order to improve the quality of the analog voice signal, *loading coils* are attached to the twisted pairs. These same coils are removed for digital transmission of modem signals. Since they are by far the most common means of access to the premises, a great deal of work is going into increasing the information-carrying capability of twisted pairs. This would allow the classic telephone network to play a role in what might be called the Internet revolution with a relatively modest increase in investment. The generic term for these techniques is *digital subscriber links* (DSLs). Asynchronous *digital subscriber links* (ADSLs) recognize that most Internet users receive far more information than they transmit; accordingly, the rates provided are 1.536 megabits per second (Mbps) on the downlink and 400 kilobits per second (kbps) on the uplink. These datastreams coexist with the ordinary voice signal on the same line. The latest development in this area is the very high-data-rate digital subscriber line (VDSL), which provides uplink and downlink rate of 52 Mbps over optical links close to the subscriber premises.

Blurring of Distinctions Until recently, the telephone network and the relatively primitive cable television network (CTV) have been physically separate and distinct in function; however, the dizzying pace of regulatory as well as technical development has led to the CTV and the telephone companies competing for the same business. The immediate point of contention is Internet access. The *cable modem* operation over the coaxial cable entering customer premises has the potential to allow rapid access to Internet material. The same technology would allow the same cable to carry a wide range of services. There is a potential topological problem. The CTV network is a *fanout*, where one point feeds many; accordingly, techniques for controlling uplink traffic, specifically, many to one, are required.

CTV companies are certainly not secure in their base market. Direct-access satellites and ground radio channels leapfrog the local cable network. Although these services now focus on distribution of video images, it is only a matter of time until there is an uplink channel and access to the Internet for a range of services.

Wireless Transmission An area of explosive growth is wireless transmission. It seems that it is not possible to walk down the street without seeing someone, possibly on a bicycle, chatting away with someone on a cell phone. The handset receives and transmits from a base station that serves a limited geographic area, the cell. The base station serves the same function as the end office in the wire network, connecting to the system at large. The connection could be to a mobile user in the same cell or, through switching, to the long haul. In developing countries, wireless services are a great economic benefit since they avoid the installation of local (landline) wire distribution, which is the most expensive part of the network. Up to this point, wireless services have emphasized voice and low speed data of the order of 19.2 kbps; however, the future growth is in high speed multimedia data services of the order of Mbps. The current deployment of 3G wireless network strives to provide 2 Mbps per user and the leading research on 4G wireless networking technologies aims at providing QoS based services at a rate of the order of 20 Mbps.

1.1.3 Long-Haul Network

The long-haul network carries traffic from one telephone end office to another end office. In general, the long-haul network is a mesh of interconnected links (see Fig. 1.2). There may be a number of links in this path. The switch, which will be discussed in the next subsection, serves to route the flow of traffic between links.

A number of different kinds of transmission media may be used to implement the links in the long-haul network: twisted pairs with repeaters, coaxial cable, microwave radio, satellite, and optical fiber. Increasingly, optical fiber is replacing the metallic media and microwave radio as the transmission medium. This is the case for transoceanic cable as well. The dominance of optical fiber is not difficult to understand since it provides virtually error-free transmission at gigabit rates. The only challenge to the hegemony of fiber is the satellite system in its area of application. Satellites allow direct access to any point in their footprint; thus, new networks can be set up quickly. Further, earthly impediments and distances constitute no barrier. Satellites are unsurpassed in linking to remote areas, for example.

1.1.4 Switching

It is clear that a telephone system of any size consisting of hard connections between all pairs of subscribers would be impossibly large; accordingly, there is a need for trunks that are transmission lines that can be used by different pairs of users at different times (see Fig. 1.2). The switching onto these trunks is carried out at exchanges. The first manual exchange was developed in 1878. The first automatic exchange was put into operation in Illinois in 1892.

Until the late 1960s the traffic was routed using *circuit switching* in which an electrical path with a nominal 4 kHz bandwidth is established between transmitter and receiver (see Fig. 1.3). This technique works well for voice traffic since the time required to set up the path, which is approximately 0.5 s, is small compared to the duration of a typical call whose average is about 3 minutes.

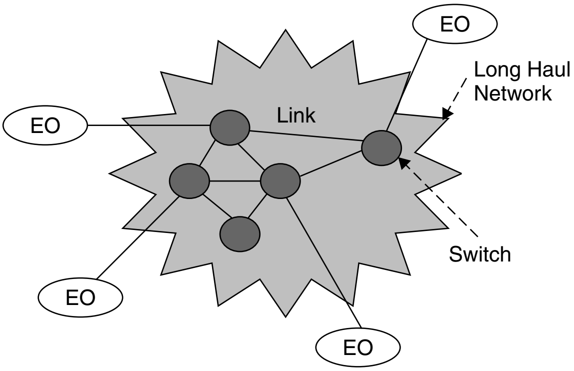


Figure 1.2 Long-haul network.

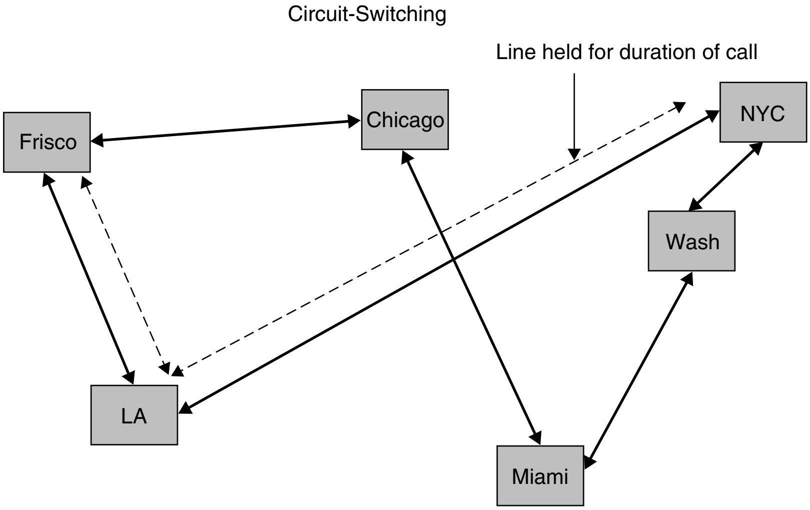


Figure 1.3 Circuit switching.

The call setup time is one of the problems encountered by data traffic connections. A typical data message may only contain 1000 bits; accordingly, even at the low rate of 4.8 kbps, the transmission time is of the same order as the setup time. The remedy *is packet switching*, whereby information is segmented into fixed-size blocks. Source and destination addresses are appended to the data payload in addition to control fields for recovering from errors and other desired functions to form a packet. Transmission facilities are dedicated only for the time required to transmit the packet.

The addition of parity check bits combats the high bit error rate, which was a common impairment of the old voice network. The packets are routed through the network in *store-and-forward* fashion over links connecting packet switches (see Fig. 1.4). At each switch along the way the packet can be checked for errors before being routed to the destination. Implementation of early packet-switched networks, notably the ARPANET, have evolved to the present Internet. Seminal papers in the development of the Internet are listed at the end of the chapter.

There is a form of packet switching, which emulates circuit switching. In *asynchronous transfer mode* (ATM) the packets, called cells in this context contain 424 bits including 40 overhead bits. The small, constant cell length allows processing associated with routing and multiplexing to be done very quickly with current technology. The basic objective is to have a degree of flexibility in handling a wide range of traffic types. The cells making up a call follow the same path for its duration, which is called a *virtual circuit.*

In the early implementation of ARPANET-based IP networks, the focus was on handling data traffic in a best-effort manner. Since the IP network provided a connectionless network service and then the applications such as ftp (File Transfer

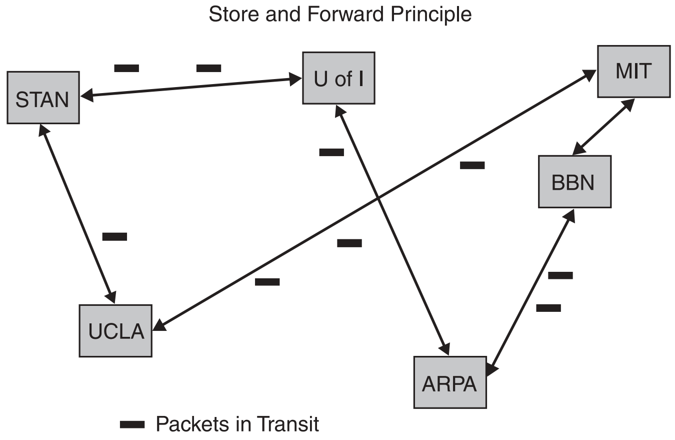


Figure 1.4 Packet-switched networks.

Protocol) and Telnet (a terminal application program for TCP/IP networks) were the only applications widely used, the requirements of service from the network for such applications were not so stringent. Although users would have preferred better response times, for their real-time sessions, such as Telnet, or chat (UNIX-based), the focus of the network implementers was on improving the ad hoc techniques developed from the point of implementation. With World Wide Web (WWW) traffic and the explosive growth of new (in particular e-commerce) applications, an IP network with just best-effort service was not sufficient. This lead to the definition of *differentiated* services and *integrated* services in the context of next-generation IP networks.

1.1.5 The Functional Organization of Network Protocols

A useful view of a telecommunications system is by the functions that it must perform in order to convey information from one point to another. The Open Systems Interconnection (OSI)[3] protocol structure provides a delineation of these functions (see Fig. 1.5). In OSI closely related functions are grouped into one of seven layers. This grouping allows simple interface between the layers. Although the seven-layer structure can be fitted to any telecommunications network, it is most suited to packet-switched networks. Fortunately, these are the networks that we are most interested in.

At the lowest level, the *physical layer*, the signal that is to be conveyed from one point to another is transformed into a signal suitable for transmission over the medium at hand. In digital communication systems, our main concern is that zeros

[3] The original purpose of OSI was to provide a common framework for all telecommunications. Because of its complexity and the rapid advance of the technology, it never functioned in this way.

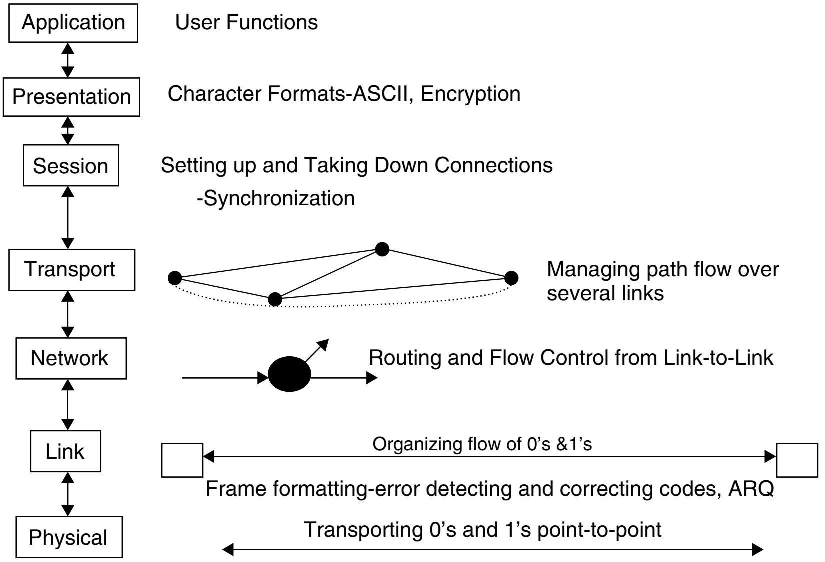

Application
User Functions
Presentation
Character Formats-ASCII, Encryption
Session
Setting up and Taking Down Connections
-Synchronization
Transport
Managing path flow over several links
Network
Routing and Flow Control from Link-to-Link
Link
Organizing flow of 0's &1's
Frame formatting-error detecting and correcting codes, ARQ
Physical
Transporting 0's and 1's point-to-point

Figure 1.5 OSI seven-layer reference model.

and ones of a digital stream modulate electrical or optical pulses that are transmitted over the respective media. The implementation of functions such as signal filtering and phase and timing recovery and tracking are handled at this level. The performance issue is to perform these functions in such a way that the probability of error is minimized within the implementation constraints.

The physical layer delivers raw bits to the *link level*, whose function is to organize the flow of bits over a segment of the path called a *link*. A link could be, for example, an optical fiber between buildings on a university campus or an intercontinental satellite channel. The basic format of bits at the link level is the *frame*, which consists of user information bits together with a number of overhead bits. The overhead bits perform several functions. In general, frames are not of fixed length; accordingly, framing bits indicating the beginning and end of a frame must be provided. Parity bits can be included in a frame in order to detect and/or correct errors that have occurred in transmission. In order to ensure that frames are delivered in order, sequencing bits must also be included in the frame. The primary performance issue at the link level is the efficacy of error detection and correction.

The frames, which are supposed to be error-free, are passed from the link layer to the *network layer*, which is responsible for routing and flow control of the links in the path between the source and the destination. At this level, a packet is formed by adding addresses as well as other overhead to the frame.

In packet-switched networks buffering is required to smooth flow between links. The *probability of buffer overflow* is a function of the rate of traffic flow and the information-carrying capacity of the link. A routing strategy could be based on minimizing this probability.

At the *transport layer* the end-to-end flow over all of the links in a path is managed. Flow control can be used for efficient operation between systems whose speeds are mismatched and as a type of congestion control. In some systems, errors are detected and corrected at this level. Further, there is a requirement to provide enough buffering to bring probability of buffer overflow to an acceptable level.

Calls are established at the *session layer.* In a circuit-switched network, the task is to find a suitable path through the network. The criterion that is relevant here is the *probability of call blocking*, specifically, the probability that a suitable path is not available to an arriving call. In packet-switched networks where resources are assigned on demand, the criterion is whether there will be enough resources available throughout the duration of a call. This is the *admission control* problem. The other functions of the session layer include data transfer control using tokens, in the case of request-response type dialogs (i.e. half duplex), synchronization in case of data transfer failures.

The performance issues that we will be considering do not arise in the two upper layers. At the *presentation layer* the information is formatted for purposes other than communication. Encryption is a form of formatting, for example. The *application layer* deals with the particular function that the user is exercising, such as mail or image transfer, for example. Since many functionalities of different applications have a common structure for the communication, the application layer supports these common functionalities in terms of protocols called application service elements (ASEs) called common ASEs (CASEs) and those functionalities which are specific to the application are modeled as specific ASEs (SASEs). The overall control of these ASEs to implement a particular application is done by using Control Functions (CFs). For modeling applications OSI recommends the concept called abstract service definition conventions (ASDC).

1.2 APPROACHES TO PERFORMANCE EVALUATION

There are three general approaches to evaluating the performance of a network. The one with which we are most familiar through course work is analysis. The calculation of the probability of error for a time-invariant channel disturbed by additive white Gaussian noise should be a familiar example. Calculation of the probability of overflow for Poisson arrival of packet to a buffer is another. At the switch level the probability of blocking, that is, the probability that an output line is not available, is of interest.

Analysis is certainly the best approach for simple models since it is fast and accurate. The problem is that mathematical approaches are tractable for only a limited number of models. There is a real art in finding an analyzable model, which approximates a given real system to some degree of accuracy. Analysis is most useful in the early stages of a project when it is necessary to do a rough assessment of available options.

After an analysis, the next step of refinement in the development of a system would be *computer simulation*, in which a more detailed model of the system under study is emulated in software. In most cases of interest, one wishes to assess the effect of random inputs in telecommunications systems. Random path delays in wireless channels and random arrival patterns to a switch are examples. The technique for dealing with a system with random stimuli is called *Monte Carlo simulation*.[4] The essence of this approach is repeated trials to obtain a set of responses to random inputs. A statistical analysis is performed on the output set in order to estimate a performance measure. Two examples will serve as illustration:

1. The number of bit errors that occur when a channel is disturbed by non-Gaussian noise is counted. This number divided by the total number of trials serves as an estimate of the probability of bit error.
2. The time of arrival to and departure from a system of a sequence of messages is recorded. The average message delay in the system can be estimated from these data.

In no sense is simulation a substitution for analysis—both techniques have their own place. If the program is even moderately complicated, it is difficult to be certain that there are no errors despite thorough checking. An analytic model is a valuable authentication tool. It can ascertain whether the simulation outputs are in the ballpark. Also, it may be possible to run the simulation for cases that can be analyzed. Mathematics is necessary to analyze the results of simulation. For example, how many data points are required for one to judge the accuracy of an estimate? Finally, in addition to verification and analysis of results, mathematics can play a role in implementation. Analytically tractable models may be used for subsystems of a larger system, thereby simplifying the program. Furthermore, in many cases the event being measured is rare. For example, the probability of error on an optical link may be on the order of 10^{-9}; to obtain valid estimates, 10^{11} samples may be required, implying a prohibitively long runtime for the program. Mathematical techniques can be used to obtain accurate results with reasonable run times. The variance reduction technique treated in Chapter 9 are examples.

With enough time and effort expended the most accurate approach to evaluating a system is building a *prototype*. It is the kind of exercise that is best suited to the final stages of a project; it is really too inflexible for early to middle stages of design. Certainly, one would not want to build a prototype to test the large number of alternatives that arise in the early stages of a project. Modern telecommunication systems have a large software component. In this respect the distinction between prototype and simulation could be blurred. A working simulation of certain system components could be directly converted into silicon, for example.

1.3 QUEUEING MODELS

1.3.1 Basic Form

The primary analytic tool that we will use to evaluate systems is *queueing theory*, which is the application of stochastic processes to the study of

[4] The fundamentals are presented in Chapter 9.

waiting lines.[5] The generic model for queueing systems is illustrated in Figure 1.6. It consists of three basic components: (1) an arrival process, (2) a storage facility, and (3) a server. A bakery is an everyday example of a queueing system. Customers arrive at rates that vary according to the season and time of day. If the sales personnel are occupied, the customer queues, that is, stands in line. The time required to fill an order varies according to the customer's demands. Of course, the number of servers is the number of sales clerks.

The *Poisson arrival process* is the most widely applied model of the arrival process and is the most tractable mathematically. In the models that we shall study, the service time is a random variable, which is independent of the arrival process in our study. We develop a general model for the service time. Storage facilities hold customers until servers are available. There are two important cases. The storage facilities can be so large that they may be considered infinite. On the other extreme, we have facilities that hold only customers who are in the process of being served.

Queueing models are widely applicable. A search of a university library for books applying queueing theory showed 96 books on the following topics: service and manufacturing, storage facilities with special emphasis on dams, inventory and maintenance, construction and mining, insurance risk, and social organization. This wide applicability notwithstanding, the most successful and the most important application of queueing models has been and continues to be telecommunications. This is the application that is the subject of this book. The generic model for telephony is shown in Figure 1.7. The arrival process consists of the generation of calls or messages. (The distinction will be made clear later.) Calls and messages are stored in a buffer prior to transmission over telephone lines. The call- and message-handling capacity of these lines is a key element in determination of the performance of the system.

In terms of performance requirements, telephone systems fall into two basic categories: *loss and delay systems.* In the former, calls leave the system if transmission facilities are not available. An everyday example is trying to place a telephone call at a later time when the line is busy. In telephone applications handling data, it is frequently the case that delay is not as critical as loss and messages can be stored until transmission facilities are available.

1.3.2 A Brief Historical Sketch

The Classical Period—Erlang and Others Before getting on with the detailed mathematical models in the rest of the text, we begin with an historical sketch of the field of queueing models. Virtually all the results that we cite will be derived later in the text. We can regard the development queueing theory as falling into three

[5] At the end of the chapter several queueing theoretic texts are listed, each shedding light on a different aspect of the subject. While all cover the fundamentals, the first three of these give an interesting historical perspective. We have found that the next four have special tutorial merit. Finally, the last two delve more deeply into the subject.

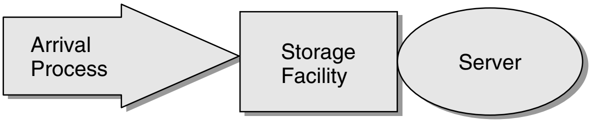


Figure 1.6 Generic queueing model.

distinct phases, which, for purposes of explanation, we call the *classical*, the *romantic*, and the *modern eras*.

As we all know, the genesis of the telecommunications industry was the invention of the telephone by Alexander Graham Bell in 1876. Bell subsequently exploited his invention for practical commercial purposes. The fact that the name Bell and telephony are inextricably linked bears witness to his energy and foresight. It is an often repeated anecdote that the telegraph company declined the opportunity to deploy the telephone because it could see no practical use for speech separated from physical contact.

It is clear that a telephone system of any size consisting of hard connections between all pairs of subscribers would be impossibly large; accordingly, there is a need for trunks, which are transmission lines that can be used by different pairs of users at different times (see Fig. 1.2). The switching onto these trunks is carried out at exchanges. The first manual exchange was developed in 1878. The first automatic exchange was put into operation in Illinois in 1892. These advances gave rise to questions of performance. For example, how many trunks would be required between a pair of exchanges so that subscribers could be connected to one another a high percentage of the time?

Questions of performance were compounded as telephone networks grew in size and complexity. The basic form of the telephone system was the *circuit-switched network* shown in Figure 1.3. For the duration of a call, transmission facilities are dedicated to a call over a fixed path or circuit. The time required to set up this path is an important parameter of service.

The first analysis of telephone traffic was reported by G. T. Blood in an unpublished memorandum in 1898. Although there are other early contributors—Rory, Joannen, and Grinsted—the father of queueing theory is undoubtedly the Danish mathematician, Agner Krarup Erlang (1878–1929). Erlang's models were based on the Poisson arrival process and an exponentially distributed duration of

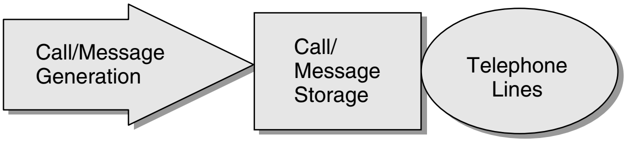


Figure 1.7 Generic telephony model.

calls.[6] During the period 1909–1917 he obtained steady-state solutions leading to formulas still used by telephone engineers. The *Erlang B* formula gives the probability that a trunk is not available as a function of demand or load and the number of trunks in a *loss system* in which calls cannot be stored. The same result for a *delay system* in which calls can be held until a line becomes available is given by the *Erlang C* formula.

In 1939, Erlang's analysis and those of subsequent contributors were unified by Feller by the application of the theory of birth and death processes. Among his many insights into the analytical model, Erlang had conjectured that the probability of a lost call (Erlang B) is insensitive to the probability distribution of the call duration. This result was subsequently proved by Kosten in 1948. Extensions and refinements of Erlang's work were carried out by Molina, Engset, and O'Dell, all of whom were on the staff of telephone administrations. This later work was concerned with such questions as retries when a call is initially blocked. The last result of the classical era that we'll cite is the *Pollaczek–Khinchin* formula,which gives the average delay in completing service in a system with Poisson arrivals, a general distribution of service time, and infinite storage. This formula was subsequently derived by Kendall in 1951 using the theory of *imbedded Markov chains.* We cover this material in Chapters 3, 5, and 6 of the text.

The Romantic Era—Packet Switching As discussed above, deficiencies in the operation of the voice-oriented telephone network led to the development of packet switching. The pioneering work in the analysis of this new kind of network was Kleinrock's doctoral thesis. This work has as its basis *Jackson networks.* This concept and its extensions are covered in Chapter 4.

A significant category of packet network developed later has been the local-area network. Much of the queuing theoretic work in this area focused on techniques for sharing a common transmission medium among a number of users. The simplest approach is dedicated capacity, as in *frequency-division multiple access* (FDMA) and *time-division multiple access* (TDMA). An alternative is the *random access technique* introduced in the *ALOHA* network. Other techniques, such as token passing, can be analyzed as variations on *polling.* These media access techniques are analyzed in Chapters 5 and 6.

The Modern Era Two developments in the technology, optical fiber and very large-scale integration (VLSI), have given rise to the modern era in telecommunications. The first of these provided orders of magnitude increases in the capacity available on transmission facilities, with rates of the order of Gbps and very low error rates. The second allows the implementation of modern digital processing techniques for switching and multiplexing. Further, VLSI is at the heart of modern computer technology, which in itself stimulates applications. The result of these developments is the development of many new services, each with its own traffic

[6] In the subsequent chapters of the text, we shall derive the results mentioned in the remainder of this paragraph and in the next subsection, on the romantic era.

characteristic and performance requirements. From the point of view of queueing models the traffic may be divided into two categories: real-time and delay-insensitive, corresponding roughly to the loss and delay systems, respectively, discussed above. In the real-time class we would find teleconferencing and high-definition television (HDTV) as well as conventional voice traffic. Representative of delay-insensitive traffic are forms of voice traffic medical images and TV on demand. Among the new services there is emphasis on visual information since it consumes so much bandwidth.

With huge available bandwidth and multiple types of applications, the emphasis on required quality of service (QoS) and congestion control techniques has grown. Although initially strict QoS-based ATM became an almost sure concept, there was still a great debate on the wastage of bandwidth of ATM cells of almost 10% in terms of cell header. Because of the ease of implementation, Internet Engineering Task Force (IETF) proposals based on IntServ and DiffServ have become widely accepted standards for QoS-based IP networks. While Integrated Services architecture is based on individual flow-level QoS provision using the Resource Reservation Protocol (RSVP), Differentiated Services architecture is based on the QoS provision at the aggregate of flows. DiffServ architecture identifies four different types of aggregates, and defines their per hop behavior at each node, apart from metering, access rate control (e.g., using the "leaky bucket" mechanism, which we discuss in Chapter 7), and/or smoothing functions at the entry of the Differentiated Services network based on the service-level agreement between the user and the Differentiated Services (DS) network, prior to using the DS network. The typical mechanisms used in sharing bandwidth among the aggregates, are based on packetized generalised processor sharing (GPS) schemes and the active queue management to implement different levels of loss probabilities can be based on the random early detection (RED) scheme, which was originally proposed to improve congestion control mechanism associated with TCP (Transmission Control Protocol) dynamics. Most of the mechanisms studied use mathematical models for evaluating the performance for which our book presents the fundamentals.

It has been amply demonstrated that the Poisson model for the new classes of traffic yields misleading results; accordingly, a great deal of effort has been expended on models that capture the salient characteristics of the traffic generated by the new services. In Chapter 7, the *fluid flow model* approximates the discrete flow of digital traffic is by a fluidlike model. In Chapter 8, a more general tool, the matrix analytic technique, is studied.

1.4 COMPUTATIONAL TOOLS

The emphasis in the text is on getting numbers in order to evaluate the performance of a system. We will extensively use three different software tools to do this. The simplest is the *Excel* spreadsheet. The virtue of Excel is that it is easy to use and allows interaction. Simple formulas can be quickly evaluated, and curves can be drawn. The effect of changing parameters can be seen immediately.

The second tool we use is *Matlab*. It is more suited than Excel to complex computations, and it has much better graphical capability. We use Matlab extensively to simulate communications systems and techniques by means of Monte Carlo simulation. As its name may indicate, Matlab is particularly well suited to matrix operations. This tool was very valuable in the last three chapters of the text in handling complex matrix operations and in doing simulation.

The last tool that is used in the course is *Maple*. The particular strength of this tool is symbolic manipulation. In Chapter 6, for example, differentiation of complex functions was carried out. Maple was used in Chapter 8 to simplify very complex equations.

FURTHER READING

Material on Telecommunications Systems

Freeman, R. L., *Fundamentals of Telecommunications*, Wiley, New York, 1999.

"100 years of communications progress," *IEEE Commun. Soc. Mag.*, **22** (5), (May 1984).

Leon-Garcia, A., and I. Widjaja, *Communication Networks: Fundamental Concepts and Key Architectures*, McGraw-Hill, New York, 2000.

Genesis of LANs

Abramson, N., "The ALOHA system—another alternative for computer communications," *1970 Fall Joint Computer Conf., AFIPS Conference Proceedings*, Vol. 37, 1970, pp. 281–285.

Farmer, W. D., and E. E. Newhall, "An experimental distributed switching system to handle bursty computer traffic," *Proc. ACM Symp., Problems in Optimization Data Communication Systems*, DP. 1–34, Pine Mountain, GA, 1969.

Metcalf, R. M., and D. R. Boggs, "Ethernet: Distributed packet switching for local computer networks," *Commun. ACM* **19**: 395–404 (July 1976).

Pierce, J., "How far can data loops go?" *IEEE Trans. Commun.* **COM-20**, 527–530 (June 1972).

Queuing Texts of Historical Interest

Bear, D., *Principles of Telecommunication Traffic Engineering*, Peter Peregrinus Ltd, 1976

Saaty, T. L., *Elements of Queueing Theory*, McGraw-Hill, New York, 1961.

Syski, R., *Congestion Theory in Telephone Systems*, Oliver & Boyd, 1960.

Queuing Texts with Tutorial Value

Allen, A. O., *Probability, Statistics and Queuing Theory*, Academic Press, New York, 1978.

Cox, D. R., and W. L. Smith, *Queues*, Methuen, London, 1961.

Gross, D., and C. M. Harris, *Fundamentals of Queueing Theory*, Wiley, New York, 1998.

Kleinrock, L., *Queueing Systems*, Vol. 1: *Theory*, Wiley, New York, 1975.

Comprehensive Theoretical Treatments of Queuing Theory

Cohen, J. W., *The Single Server Queue*, North-Holland, 1968.

Takagi, H., *Queueing Analysis*: *A Foundation of Performance Evaluation*, Vols. 1–3, Elsevier Science Publishers B.V, 1993.

Milestones in the Development of Queueing Theory

Brockmeyer, E., H. L. Halstrom, and A. Jensen, "The life and works of A. K. Erlang," *Trans. Danish Acad. Tech. Sci. ATS*, (2) (1948).

Erlang, A. K., "The theory of probabilities and telephone conversations," *Nyt Tidsskrift Matematik B*, **20**: 33–39 (1909).

Erlang, A. K., "Solution of some problems in the theory of probabilities of significance in automatic telephone exchanges," *Electroteknikeren* **13**: 5–13 (1917) [in English: *PO Electric. Eng. J.*, **10**: 189–197 (1917–1918)].

Feller, W., *An Introduction to Probability Theory and Its Applications*, Vol. 1, 3rd ed., Wiley, New York, 1968.

Kendall, D. G., "Some problems in the theory of queues," *J. Roy. Stat. Ser. B* **13**: 151–185, (1951).

Kendall, D. G., "Stochastic process occurring in the theory of queues and their analysis by the method of the imbedded Markov chain," *Ann. Math. Stat.*, **24**: 338–354 (1953).

Khinchin, A. Y., "Mathematical theory of stationary queues," *Mat. Sbornik* **39**: 73–84 (1932).

Kosten, L. *On Loss and Queueing Problems* (Dutch), thesis, Technicological Univ., Delft, The Netherlands, 1942.

Kosten, L., "On the validity of the Erlang and Engset loss formulae," *Het P. T. T. Bedjijf* (*Netherlands Post Office Journal*).

Pollaczek, F., "Uber eine Aufgab der Wahrscheinlichkeitstheorie," *I–II Math. Zeitschrift.* **32**: 64–100, 729–750 (1903).

Packet Switching and the Internet

Baran, P., "On Distributed Communications Networks," RAND paper P-2626, Sept. 1962; also, *IEEE Trans. Commun. Sys.* **CS-12**(1): 1–9 (March 1964).

Cerf, V. G. and R. E. Kahn, "A Protocol for Packet Network Interconnection," *IEEE Trans. Comm. Techn.*, Vol. Comm-22, No. 5, 627–641, May 1974.

IEEE Communications Magazine, Issues of March and May 2002.

Kahn, R. E., "Resource-sharing computer networks," *in Computer Networks, a Tutorial*, M. Abrams, R. P. Blanc, and I. W. Cotton, eds., IEEE Press, 1978, pp. 5-8–5-18.

Kleinrock, L., *Communication Nets: Stochastic Message Flow and Delay*, McGraw-Hill, New York, 1964.

Leiner, B. M. et al., "A brief history of the Internet," http://www.isoc.org/internet-history/brief.html.

Roberts, L. "Multiple Computer Networks and Intercomputer Communications," ACM, Gatlinburg Conf., October 1967.

Roberts, L. G., "Data by the packet," *IEEE Spectrum* **11**(2): 46–51 (Feb. 1974).

Proceedings of the IEEE Special Issue on Packet Communications, Vol. 66, No. 11, Nov. 1978.

The Modern Network

Leland, W. E. et. al., “On the self-similar nature of Ethernet traffic (extended version),” *IEEE/ACM Trans. Network.*, 1–15 (Feb. 1994).

On Modern Internet Services

Blake, S., D. Black, M. Carlson, E. Davies, Z. Wang, and W. Weiss, *An Architecture for Differentiated Services*, RFC 2475, Dec. 1998.

Braden, R., D. Clark, and S. Shenker, *Integrated Services in the Internet Architecture: An Overview*, RFC 1633, June 1994.

Braden, R., L. Zhang, S. Berson, S. Herzog, and S. Jamin, *Resource Reservation Protocol* (*RSVP*)—*Version 1 Functional Specification*, RFC 2205, Sept. 1997.

2

PROBABILITY AND RANDOM PROCESSES REVIEW

2.1 BASIC RELATIONS

The material presented in this text is based on queuing theoretic models of systems. In a broad sense, queuing models deal with systems of waiting lines with *random* arrivals of customers with *random* service requirements. The bases for the study of random phenomena are probability and stochastic processes, which are the subject of this chapter. The treatment is that of a review; accordingly, a number of concepts will simply be introduced with the idea of placing them in a general context.[1] Proofs will be omitted. Where appropriate they will be given in subsequent chapters.

2.1.1 Set Functions and the Axioms of Probability

Three set-theoretic concepts form the basis for probability theory: the set of all possible outcomes, the set of events, and the probability of an event. We define Ω as the set of all possible outcomes of an experiment, such as 1, 2, 3, 4, 5, 6 in the toss of a die. We define the set of events **A**. An *event* is defined to be a subset of the set of all experimental outcomes, for example, odd numbers in the toss of a die. On the set of

[1] The reader is referred to any one of a number of excellent texts on the subject, several of which are cited at the end of the chapter.

Modeling and Analysis of Telecommunications Networks, by Jeremiah F. Hayes and
Thimma V. J. Ganesh Babu
ISBN 0-471-34845-7 © 2004 John Wiley & Sons, Inc.

events we define the following set operations. The Venn diagrams in Figure 2.1 illustrate these operations:

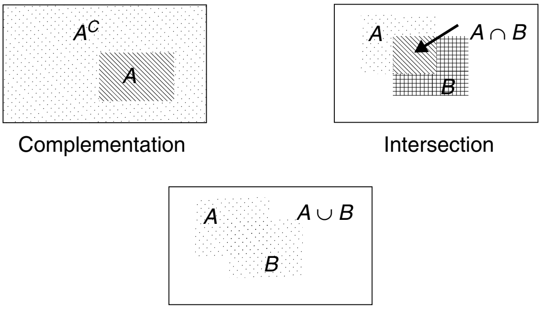
1. *Complementation.* A^C is the event that event A does not occur. This is the set of all experiments in Ω that are not in A. A^C is called the *complement of* A.
2. *Intersection.* $A \cap B$ is the event that events A and B occur. This is the set of experimental outcomes common to both A and B.
3. *Union.* $A \cup B$ is the event that events A or B occur. This is the set of experimental outcomes in A or B or both.
4. *Inclusion.* $A \subset B$ An event A occurring implies event B occurs.

The set of events **A** is closed under these operations, meaning that performing them on included members results in no new members.

Thus, if A and B are events then A^C, B^C, $A \cap B$, $A \cup B, \ldots$, are also events. For example, if in the toss of a die, $\{1, 4, 5\}$ and $\{1\}$ are events, so are $\{2, 3, 6\}$, $\{1, 2, 3, 6\}$, $\{2, 3, 4, 5, 6\}$, $\{4, 5\}$, $\varnothing$ (empty or null set), and Ω. Two sets of events are said to be disjoint if they have no common elements; that is, events A and B are disjoint if $A \cap B = \varnothing$.

Let $P(A)$ denote the *probability of event* A. The probability of event A is a function on the set of events that satisfies the following five axioms:

1. There is a nonempty set of experimental outcomes Ω, a set of events **A** defined over the set of outcomes.
2. For any event $A \in \mathbf{A}$, $P(A) \geq 0$.
3. For the set of all possible experimental outcomes, Ω, $P(\Omega) = 1$.
4. If the events and A and B are disjoint or mutually exclusive, $A \cap B = \varnothing$, then $P(A \cup B) = P(A) + P(B)$.



Complementation

Intersection

Union

Figure 2.1 Venn diagrams.

5. For countably infinite sets $A_1, A_2, \ldots$ such that $A_i \cap A_i = \emptyset$ for $i \neq j$, we have

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$$

All the properties of the probability can be derived from these axioms. A partial list of these properties is as follows:

1. For any event A, $P(A) \leq 1$.
2. $P(A^C) = 1 - P(A)$.
3. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$, for arbitrary A and B.
4. $P(A) \leq P(B)$ for $A \subset B$.

(2.1)

Example 2.1 The early studies of probability were connected with games of chance. These still make good examples.[2] A deck of ordinary playing cards consists of 52 cards divided into four suits: clubs, spades, diamonds, and hearts. The cards are also numbered from 2 to 10 and jack, queen, king, and ace. The probability of drawing any particular card, say, the ace of spades, is $1/52$. In the game of poker, one draws five cards in a hand. The probability of drawing any particular set of five cards is $1/52 \times 1/51 \times 1/50 \times 1/49 \times 1/48 = 3.20641 \times 10^{-9}$. Notice that cards are not replaced once they are drawn. A "royal flush" is the cards 10, jack, queen, king and ace, all of the same suit. Since there are four mutually exclusive suits, the probability of drawing a royal flush is found by an application of axiom 4 above to be

$$\begin{aligned} P(\text{royal flush}) &= 4 \times 1/52 \times 1/51 \times 1/50 \times 1/49 \times 1/48 = 4 \times 3.20641 \times 10^{-9} \\ &= 1.28256 \times 10^{-8} \end{aligned}$$

The event of any five cards of the same suit is simply a "flush". The probability of this event is

$$P(\text{flush}) = 4 \times 13/52 \times 12/51 \times 11/50 \times 10/49 \times 9/48 = 0.001980792$$

The event of drawing 10, jack, queen, king, and ace with no match of suits is called a "straight from the 10" and has the probability

$$P(\text{straight from 10}) = 20/52 \times 16/51 \times 12/50 \times 8/49 \times 4/48 = 3.94004 \times 10^{-4}$$

[2] The unsurpassed reference for these kinds of examples is Feller (1957).

From property 3, we have that the probability of either of these two preceding events is

$$\begin{aligned} P(\text{flush or straight from 10}) &= 0.001980792 + 3.94004 \times 10^{-4} - 1.28256 \times 10^{-8} \\ &= 0.0023747 \end{aligned}$$

The probability of neither a straight from the 10 nor a flush is

$$P(\text{not flush and not straight from 10}) = 1 - 0.0023747 = 0.997625$$

2.1.2 Conditional Probability and Independence

The idea of the probability of an event can be extended to establish other relationships between events. We define the *conditional* probability as

$$P(A/B) = \frac{P(A \cap B)}{P(B)} \tag{2.2}$$

We term $P(A/B)$ the probability of event A conditioned on the occurrence of event B. It can be shown that $P(A/B)$ is a valid probability inasmuch as it satisfies the axioms. The events A and B are said to be *independent* if $P(A \cap B) = P(A) \times P(B)$. From Equation (2.2) this implies that $P(A/B) = P(A)$, meaning that the occurrence of event B has no bearing on the probability of event A. Although they are opposite concepts, beginning students often confuse mutually exclusive events with independent events. Two events are mutually exclusive if they cannot occur at the same time. For mutually exclusive events, $P(A \cup B) = P(A) + P(B)$. (Note the contrast with independent events.)

Example 2.2 If the five cards that are drawn do not form a flush, what is the probability that the cards are a straight from the 10? Start with the following relation:

$$P(\text{not straight from 10/not flush}) = \frac{0.997625}{(1 - 0.001980792)} = 0.999605227$$

Relation 2 in (2.1) gives the final answer

$$P(\text{straight from 10/not flush}) = 1 - 0.999605227 = 0.000395$$

Notice that the additional information improves the odds slightly.

2.1.3 The Law of Total Probability and Bayes' Rule

A set of events A_i; $i = 1, 2, \ldots, n$ *partitions* the set of experimental outcomes if

$$\bigcup_{i=1}^{n} A_i = \Omega$$

and

$$A_i \cap A_j = \varnothing \quad \text{for } i \neq j$$

We can write any event B in terms of a set events $A_1, A_2, \ldots, A_n$ that partition the space of experimental outcomes:

$$B = \bigcup_{i=1}^{n} A_i \cap B$$

Since $A_i \cap B$; $i = 1, 2, \ldots, n$ are disjoint events, we have the following from axiom 4:

$$P(B) = \sum_{i=1}^{n} P(A_i \cap B) \tag{2.3}$$

This is the *law* of *total probability*, which is of considerable utility in the text. The application of this law to the definition of conditional probability leads to *Bayes*' rule, which states that

$$P(A_i/B) = \frac{P(A_i \cap B)}{\sum_{i=1}^{n} P(A_i \cap B)} = \frac{P(A_i)P(B/A_i)}{\sum_{i=1}^{n} P(A_i)P(B/A_i)} \tag{2.4}$$

In many problems, communication systems, for example, it is easy to compute conditional probabilities such as $P(B/A_i)$, but what is required are conditional probabilities of the form $P(A_i/B)$. Bayes' rule allows the computation of one from the other.

Example 2.3 The binary symmetric channel, which is an abstract encapsulation of the effects of coding and modulation in a communications channel provides a good example. The channel inputs and outputs are the binary symbols 0–1. From physical considerations, it is relatively easy to compute the probabilities $P(\text{output} = j/\text{input} = i)$; $i,\ j = 0,\ 1$; however, in order to decide what was transmitted, the probability that one wants to compute is $P(\text{input} = i/\text{output} = j)$. Let $P(\text{input} = 0) = 0.4$ and $P(\text{input} = 1) = 0.6$. The probabilities $P(\text{output} = j,$

input = i); $i, j = 0, 1$ are given in the following table:

Input\Output	0	1
0	$\frac{2}{3}$	$\frac{1}{3}$
1	$\frac{1}{4}$	$\frac{3}{4}$

By the use of (2.3), we find the probabilities $P(\text{input} = j/\text{output} = i)$; $i, j = 0, 1$

Output\Input	0	1
0	0.64	0.36
1	0.2286	0.7714

2.2 RANDOM VARIABLES—PROBABILITY DISTRIBUTIONS AND DENSITIES

2.2.1 The Cumulative Distribution Function

A *random variable* is a function on the set of experimental outcomes mapping the experimental outcomes onto the real line. *Discrete random variables* are mapped onto a countable set of points on the real line, possibly infinite. In contrast, *continuous random variables* are mapped into an interval on the real line. Discrete or continuous random variables are represented by capital letters (e.g., X, Y). They are characterized probabilistically by their *cumulative probability distribution functions*,[3] which are defined as

$$F_X(x) = P(X \leq x) \tag{2.5}$$

Thus the probability distribution function evaluated at x is the probability of a set of experimental outcomes that map into the interval $(-\infty, x)$

From the basic axioms of probability the following properties for the probability distribution function may be shown:

1. $F(-\infty) = 0$
2. $F(\infty) = 1$
3. $F_X(x_1) \leq F_X(x_2)$ for $x_1 \leq x_2$

From the probability distribution function one can calculate the probability of an event lying in an interval

$$P(x_1 < X \leq x_2) = F_X(x_2) - F_X(x_1) \tag{2.6}$$

[3] These are usually called the *probability distribution functions* or just the *distribution functions.*

2.2.2 Discrete Random Variables

For a *discrete* random variable the probability distribution function is a sequence of steps. The height of each step is the probability of the random variable assuming a particular value. We may write the probability distribution function in the discrete case as

$$F_X(x) = \sum_{i=0}^{\infty} P(X = x_i)U(x - x_i)$$

where $U(x)$ is the unit step function given by

$$U(x) = \begin{cases} 1; & \text{for } x \geq 0 \\ 0; & \text{for } x < 0 \end{cases}$$

The condition $F(\infty) = 1$ requires that the normalizing condition be observed:

$$\sum_{i=0}^{\infty} P(X = x_i) = 1$$

The probabilities $P(X = x_i)$; $i = 1, 2, \ldots$ are all of the possible values $x_1, x_2, \ldots$ that X may assume. For simplicity of presentation we shall assume that discrete random variables assume integer values, $0, 1, 2, \ldots$. We recognize that there may be a finite upper limit to the number of discrete values.

It is frequently the case that one is required to calculate the probability that a random variable falls within a specified range: $P(a < X \leq b)$. To find this quantity, one simply sums the probabilities of the values in the range:

$$P(a < X \leq b) = \sum_{i=a+1}^{b} P(X = i) = F_X(b) - F_X(a)$$

Expectation of Discrete Random Variables—Moments The distributions of random variables are characterized by a number, which is called the *expectation* or the *mean value* of a random variable. For discrete random variables, we have

$$E[N] = \sum_{i=0}^{\infty} iP(N = i) \tag{2.7}$$

$E[Y] = E[g(X)]$ can be found from the distribution function of X. In the discrete case we have

$$E[g(N)] = \sum_{i=0}^{\infty} g(i)P(N = i)$$

The most useful functions in terms of characterizing random variables are moments of the form $Y = X^k$. The kth moment of a random variable is $E[X^k]$, and the kth central moment is $E[(X - E[X])^k]$. The *variance* of a random variable is the second central moment:

$$\sigma_X^2 \triangleq \mathrm{Var}[X] = E[(X - E[X])^2] = E(X^2) - E(X)^2 \tag{2.8}$$

It is a standard notation in both the discrete and the continuous cases to call the square root of the variance the standard deviation:

$$\mathrm{SD}(X) = \sqrt{\mathrm{Var}(X)} = \sigma_X$$

The Probability-Generating Function Frequently, in queueing theoretic models, it is more convenient to express distributions in terms of transforms. The *probability-generating function* of the discrete random variable N is defined as the expectation of z^N where N is the discrete random variable and z is a complex variable. We have

$$N(z) = E[z^N] = \sum_{i=0}^{\infty} z^i P(N = i) \tag{2.9}$$

The probability-generating function has an obvious relationship to the z transform of the sequence $P[X = i]$, $i = 0, 1, 2, \ldots$. The probability-generating function contains all the information that the probability distribution does inasmuch as one can be obtained from the other. Notice that

$$N(z)|_{z=1} = \sum_{i=0}^{\infty} P(N = i) = 1 \tag{2.10}$$

Given the probability-generating function of a random variable, the moments of the random variable can be easily found by repeated differentiation. For the first moment we have

$$E[N] = \left.\frac{dN[z]}{dz}\right|_{z=1} = \left.\sum_{i=0}^{\infty} iz^{i-1} P(N = i)\right|_{z=1} = \sum_{i=0}^{\infty} iP(N = i) \tag{2.11}$$

Higher moments can be found from successive differentiation and manipulation. For example, it is easy to show that

$$E[N^2] = \left.\frac{d^2N[z]}{dz^2}\right|_{z=1} + \left.\frac{dN[z]}{dz}\right|_{z=1} \tag{2.12}$$

The moment-generating function can also yield probabilities. From (2.9), we have

$$P(N=i)=\frac{1}{i!}\frac{d^iN(z)}{dz^i}\bigg|_{z=0}=\frac{1}{i!}\frac{d^i\sum_{i=0}^{\infty}z^iP(N=i)}{dz^i}\bigg|_{z=0} \tag{2.13}$$

Four discrete distributions are important in study of the performance of telecommunications systems: the discrete uniform, the binomial, the Poisson distribution, and the geometric distribution.

Discrete Uniform Distribution The *discrete uniform distribution* serves as an illustration since it is the simplest. Consider the set of points $k+1, k+2, \ldots, k+m$. The discrete random variable N is uniformly distributed if $P(N=i)=1/m$; $i=k+1, k+2, \ldots, k=m$. The mean and the variance of the discrete uniform distribution are[4]

$$\frac{\sum_{i=k+1}^{k+m} i}{m}=k+\frac{m+1}{2} \tag{2.14}$$

and

$$\frac{\sum_{i=k+1}^{k+m} i^2}{m}=k(k+m+1)+\frac{(m+1)(m+2)}{6} \tag{2.15}$$

The probability-generating function of the discrete uniform distribution is

$$U(z)=\frac{\sum_{i=k+1}^{k+m} z^i}{m}=\frac{z^{k+1}(1-z^m)/(1-z)}{m} \tag{2.16}$$

Example 2.4 Suppose playing cards, that you are assigned number with the jack = 11, the queen = 12, the king = 13, and the ace = 14, without regard to suit. Suppose also that a single card is drawn. The resulting random variable is uniformly distributed between 2 and 14:

$$P(N=i)=\frac{1}{13}; \quad i=2,3,\ldots,14$$

The cumulative distribution and the probability distribution of the resulting random variable is shown in Figure 2.2. The distribution function contains all the information about the random variable. For example, the probability of it being greater than 5 and less than or equal to 11 is given by

$$P(5<N\leq 11)=\frac{10}{13}-\frac{4}{13}=\frac{6}{13}$$

The mean and the variance are 8 and 14, respectively.

[4] We use the identities $\sum_{i=1}^{M} i=M(M+1)/2$ and $\sum_{i=1}^{M} i^2=M(M+1)(2M+1)/6$.

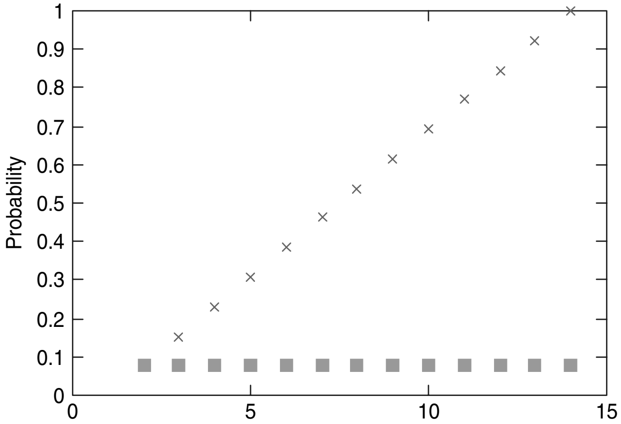


Figure 2.2 Discrete uniform.

Binomial Distribution The binomial distribution arises naturally from a sequence of independent experiments called *Bernoulli trials*. Suppose that an experiment can have two possible outcomes, which we call *success* and *failure*, respectively. These could be the outcomes of tosses of an unfair coin, for example. Let the probability of success be denoted as P. Clearly, the probability of failure is then $1 - P$. Now, suppose that this experiment is repeated n times with the stipulation that the trials are performed independently, where the outcome of a trial does not depend on the outcome of any other trial. We want to calculate the probability of k successes, $0 \leq k \leq n$. There are many ways that this can happen. For example, the first k trials successful and the last $n - k$ failures is such an elementary event; then, the probability of a particular sequence with k successes is $P^k(1 - P)^{n-k}$; $k = 0, 1, \ldots, n$. A more complex event, consisting of a number of elementary events, is the occurrence of k successes in any order. There are $\binom{n}{k}$ ways for there to be k successes in n trials, where each pattern is a mutually exclusive event. Since the events are mutually exclusive, we simply add their probabilities

$$P(k \text{ successes in } n \text{ trials}) = b(k; n, P) = \binom{n}{k} P^k(1 - P)^{n-k}; \quad k = 0, 1, 2, \ldots, n \tag{2.17}$$

This is the binomial distribution.

Finding the moments for the binomial distribution is straightforward. Substituting (2.17) into (2.7), we have

$$E(B) = \sum_{k=0}^{n} k \binom{n}{k} P^k(1 - P)^{n-k} = nP \sum_{k=0}^{n-1} \binom{n-1}{k} P^k(1 - P)^{n-k-1} = nP \tag{2.18}$$

From (2.8) and (2.17), we have

$$\begin{aligned}\mathrm{Var}(B) &= \sum_{k=0}^{n} k^2 \binom{n}{k} P^k (1-P)^{n-k} - (nP)^2 \\ &= nP \sum_{l=0}^{n-1} (l+1) \binom{n-1}{l} P^l (1-P)^{n-l-1} - (nP)^2 \\ &= nPP(n-1) + nP - (nP)^2 = nP(1-P) \qquad (2.19)\end{aligned}$$

The probability-generating function of the binomial distribution is

$$B(z) = \sum_{k=0}^{n} \binom{n}{k} P^k (1-P)^{n-k} z^k = (1-P+zP)^n \qquad (2.20)$$

A special case of the binomial distribution, which has it own niche in the literature, is that of $n = 1$, where the random variable can assume only two values, 0 and 1. We have simply that $P(B = 0) = 1 - P$ and $P(B = 1) = P$. The distinct name for this distribution is the *Bernoulli* distribution.

Example 2.5 In the Excel spreadsheet accompanying the text, the binomial distribution has to be worked out for the cases $n = 20$ and $P = 0.8$. The results are plotted in Figure 2.3, where (2.17) and the cumulative distribution are plotted. The mean and the variance are 16 and 3.2, respectively.

Poisson Distribution The Poisson distribution can be derived as a limiting case of the binomial distribution. Suppose that the number of trials n is increased without

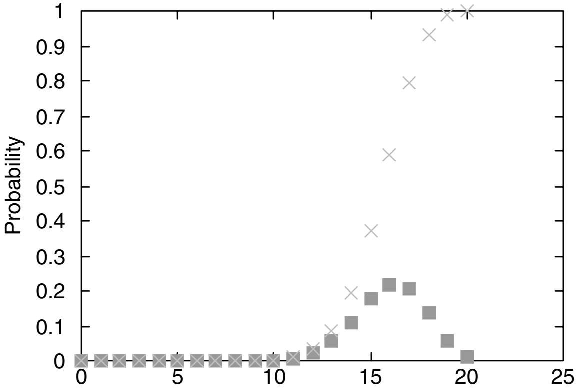


Figure 2.3 Binomial distribution.

limit and the probability of success P on any one trial is decreased toward zero in such a way that $\lambda = nP$ remains constant. We have

$$\lim_{\substack{n\to\infty \\ P\to 0}} b(0;\, n,\, P) = \lim_{\substack{n\to\infty \\ P\to 0}} (1-P)^n = \lim_{\substack{n\to\infty \\ P\to 0}} \left(1 - \frac{\lambda}{n}\right)^n = e^{-\lambda}$$

and

$$\lim_{\substack{n\to\infty \\ P\to 0}} \frac{b(k;\, n,\, P)}{b(k-1;\, n,\, P)} = \lim_{\substack{n\to\infty \\ P\to 0}} \frac{n-k+1}{k} \cdot \frac{P}{1-P} = \frac{\lambda}{k}$$

By induction it follows that the probability of k successes is

$$p(k,\, \lambda) = \frac{e^{-\lambda}\lambda^k}{k!}; \quad k = 0, 1, 2, \ldots \tag{2.21}$$

where λ is a parameter that characterizes the distribution. We shall see below that this is the average number of successes. Thus, this is the Poisson distribution with mean value λ.

Suppose that the n trials had been carried out in a one-second interval. Suppose also that the trails were carried out in exactly the same way for t seconds. The average rate of successes would be λt, and the probability of k successes is

$$P_k(t) = P(k \text{ successes in } t \text{ seconds}) = \frac{e^{-\lambda t}(\lambda t)^k}{k!}; \quad k = 0, 1, 2, \ldots \tag{2.22}$$

The mean and the variance of the Poisson distribution can be found from the same limiting argument that we have just used. Let $n \to \infty, P \to 0, nP \to \lambda t$. From (2.18) and (2.19), we have

$$E(P) = \lambda t \tag{2.23}$$

and

$$\mathrm{Var}(P) = \lambda t \tag{2.24}$$

The probability-generating function of the Poisson distribution is

$$P(z) = \sum_{k=0}^{\infty} \frac{e^{-\lambda t}}{k!} (\lambda t z)^k = e^{-\lambda t(1-z)} \tag{2.25}$$

Example 2.6 On the accompanying spreadsheet, the Poisson distribution for $\lambda t = 20$ has been plotted. The results are shown in Figure 2.4. We discontinue the calculation after $k = 50$, but it is simple enough to continue beyond this value.

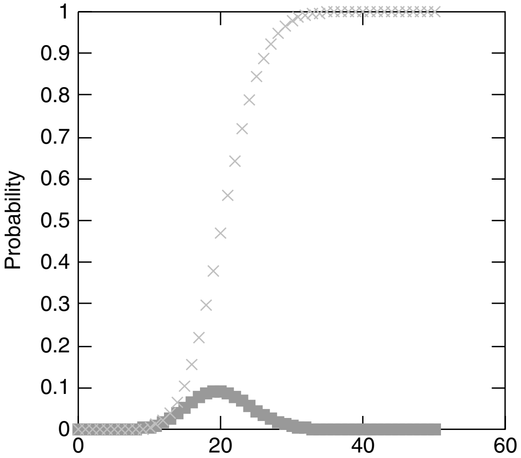


Figure 2.4 Poisson distribution.

Geometric Distribution Consider once again Bernoulli trials, where we are interested in the number of failures between successes. The probability of k failures until the first success is given by

$$\begin{aligned} P(\text{run of } k \text{ failures followed by a success}) &= g(k; P) \\ &= P(1-P)^k; \quad k = 0, 1, 2, \ldots \end{aligned} \tag{2.26}$$

Once again, the mean and the variance calculations are straightforward. We have

$$E(G_1) = \sum_{k=0}^{\infty} kP(1-P)^k = \frac{1-P}{P} \tag{2.27}$$

and

$$\operatorname{Var}(G_1) = \sum_{k=0}^{\infty} k^2 P(1-P)^k - \left(\frac{P}{1-P}\right)^2 = \frac{1-P}{P^2} \tag{2.28}$$

The probability-generating functions for the geometric distribution are as follows:

$$G_1(z) = \sum_{k=0}^{\infty} P(1-P)^i z^i = \frac{P}{1-z(1-P)} \tag{2.29}$$

In certain applications, the lowest value of the geometrically distributed random variable is 1 rather than zero. For example, the number of packets required to represent a message is modeled as geometrically distributed; however, a message having no packets makes no sense. In cases of this kind, we have

$$P(G_2 = k) = g(k; P) = P(1 - P)^{k-1}; \quad k = 1, 2, \ldots \tag{2.30}$$

In this case, it follows easily that

$$E(G_2) = \sum_{k=0}^{\infty} kP(1 - P)^k = \frac{1}{P} \tag{2.31}$$

and

$$\mathrm{Var}(G_2) = \sum_{k=0}^{\infty} k^2 P(1 - P)^k - \frac{1}{P^2} = \frac{1 - P}{P^2} \tag{2.32}$$

The probability-generating function for this form of the geometric distribution is

$$G_2(z) = \sum_{k=1}^{\infty} P(1 - P)^{i-1} z^i = \frac{zP}{1 - z(1 - P)} \tag{2.33}$$

Example 2.7 On the associated spreadsheet the probability distribution and the cumulative distribution function are shown of Figure 2.5 for the case $P = 0.1$. The mean and the variance for this case are 9 and 90, respectively. (See (2.27) and (2.2.8)).

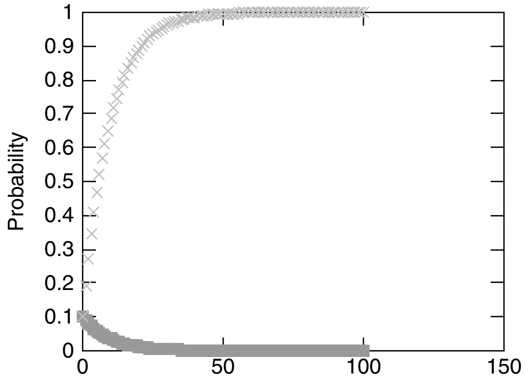


Figure 2.5 Geometric distribution.

2.2.3 Continuous Random Variables

For a *continuous* random variable, the probability distribution function is a continuous monotonically nondecreasing function. In this case the random variable may be characterized by the *probability density function*, which is defined as

$$f_X(x) = \frac{dF_X(x)}{dx} \tag{2.34}$$

or

$$F_X(x) = \int_{-\infty}^{x} f_X(\xi)d\xi \tag{2.35}$$

From (2.6) and (2.35), we may write $f_X(x)dx = P(x < X \leq x + dx)$. The properties of the probability distribution function imply the following properties for the density function:

$$f_X(x) \geq 0; \quad \forall x$$

and

$$\int_{-\infty}^{\infty} f_X(\xi)d\xi = 1$$

Notice that the density function is *not* a probability; accordingly, it is not necessarily less than one. In the continuous case, the probability that the random variable falls in a particular range can be expressed in terms of the density or the distribution function

$$P(a < T \leq b) = \int_{a}^{b} f_T(\tau)d\tau = F_T(b) - F_T(a) \tag{2.36}$$

Expectation of Continuous Random Variables—Moments For continuous random variables, expectation or the mean value is given by

$$E[X] = \int_{-\infty}^{\infty} x\, f_X(x)dx \tag{2.37}$$

In the continuous case, the expectation of a function of a random variable is given by

$$E[g(X)] = \int_{-\infty}^{\infty} g(x)\, f_X(x)dx \tag{2.38}$$

As in the discrete case, the variance is a very useful measure of a random variable. From (2.8), we have

$$\mathrm{Var}(X) = E(X^2) - E(X)^2 = \int_{-\infty}^{\infty} x^2 f_X(x)dx - \left[\int_{-\infty}^{\infty} x f_X(x)dx\right]^2 \tag{2.39}$$

The Laplace Transform of Continuous Random Variables In the case of continuous random variables, the Laplace transform of the probability density function plays a significant role in many applications. For a non-negative random variable X with probability density function $f_X(x)$, we have

$$L(f_X(x)) = E[e^{-sX}] = \mathcal{X}(s) = \int_0^{\infty} f_X(x)e^{-xs}dx \tag{2.40}$$

where s is a complex variable. Again there is a convenient check for the validity of a Laplace transform:

$$\mathcal{X}(s)|_{s=0} = \int_0^{\infty} f_X(x)dx = 1$$

Moments of the random variable can be calculated from the Laplace transform. From Equation (2.40) we have

$$\left.\frac{d^k\mathcal{X}(s)}{ds^k}\right|_{s=0} = \left.\int_0^{\infty} (-x)^k f_X(x)e^{-xs}dx\right|_{s=0} = (-1)^k E[X^k] \tag{2.41}$$

A related transform is the *characteristic function* of a random variable, which is defined as

$$\Phi_X(\omega) = \int_{-\infty}^{\infty} e^{j\omega x} f_X(x)dx$$

where $j = \sqrt{-1}$. The characteristic function is the Fourier transform of the probability density function. For the models that we shall be considering in the remainder of the text, the random variables are positive; accordingly, there is a simple relationship between the Laplace transform and the characteristic function. We have

$$\mathcal{X}(s) = \Phi_X(js)$$

There are three continuous distributions that are important for our work: the uniform distribution, the exponential distribution, and the Gaussian or normal distribution.

Uniform Distribution Perhaps the simplest example of a random variable is that of the *uniform distribution*. The density function of a uniformly distributed random variable is given by

$$f_T(t) = \begin{cases} \dfrac{1}{b-a}; & a \leq t \leq b \\ 0; & \text{otherwise} \end{cases} \tag{2.42}$$

The distribution function is given by

$$F_T(t) = \begin{cases} 0; & -\infty < t \leq a \\ \dfrac{t-a}{b-a}; & a < t \leq b \\ 1; & t > b \end{cases} \tag{2.43}$$

The mean value of a uniformly distributed random variable is found easily from the preceding:

$$E[X] = \frac{1}{b-a}\int_a^b x\,dx = \frac{a+b}{2} \tag{2.44}$$

We have for the variance

$$\text{Var}[X] = \frac{1}{b-a}\int_a^b x^2 dx - \left(\frac{a+b}{2}\right)^2 = \frac{(b-a)^2}{12} \tag{2.45}$$

The Laplace transform of the probability density for the uniform distribution is

$$U(s) = \frac{1}{b-a}\int_a^b e^{-sx} dx = \frac{e^{-as} - e^{-bs}}{s(b-a)} \tag{2.46}$$

Example 2.8 In the associated spreadsheet, the uniform distribution with $a = -4.5$ and $b = 3.2$ are worked out. The density and the distribution function are shown in Figure 2.6. The mean is -0.65 and the variance is 4.940833

Exponential Distribution The second example of a continuous random variable is that of the exponential distribution. The density function of an exponentially distributed random variable is given by

$$f_T(t) = \begin{cases} \mu e^{-\mu t}; & t \geq 0 \\ 0; & t < 0 \end{cases} \tag{2.47}$$

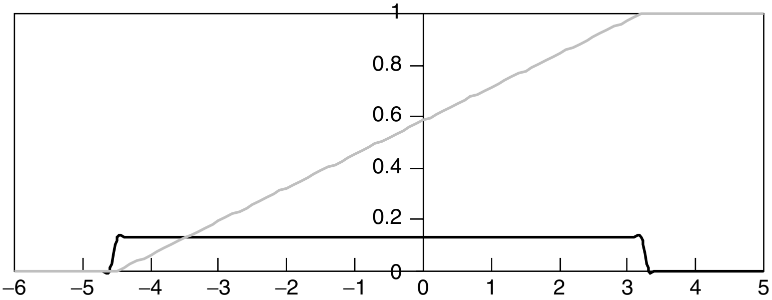


Figure 2.6 Uniform distribution.

The probability distribution function is

$$F_T(t) = \begin{cases} 1 - e^{-\mu t}; & t \geq 0 \\ 0; & t < 0 \end{cases} \tag{2.48}$$

The moments of the exponential distribution are

$$E[X] = \int_0^\infty x\mu e^{-\mu t} dx = \frac{1}{\mu} \tag{2.49}$$

$$\text{Var}[X] = \int_0^\infty x^2 \mu e^{-\mu t} dx - \frac{1}{\mu^2} = \frac{1}{\mu^2} \tag{2.50}$$

The Laplace transform of the probability density for the exponential distribution is

$$E(s) = \int_a^b \mu e^{-\mu x} e^{-sx} dx = \frac{\mu}{s + \mu} \tag{2.51}$$

Example 2.9 On the associated spreadsheet, the exponential distribution for the parameter $\mu = 0.5$ is worked out. The mean and the variance are 2 and 4, respectively. The density and the distribution are shown in Figure 2.7.

Gaussian (Normal) Distribution A probability distribution that finds wide application is the Gaussian distribution, also called the *normal distribution*. This distribution is most easily characterized by means of its probability density

$$f_X(x;\, \mu,\, \sigma) = \frac{e^{-(x-\mu)^2/2\sigma^2}}{\sqrt{2\pi\sigma^2}}; \quad -\infty < x < \infty \tag{2.52}$$

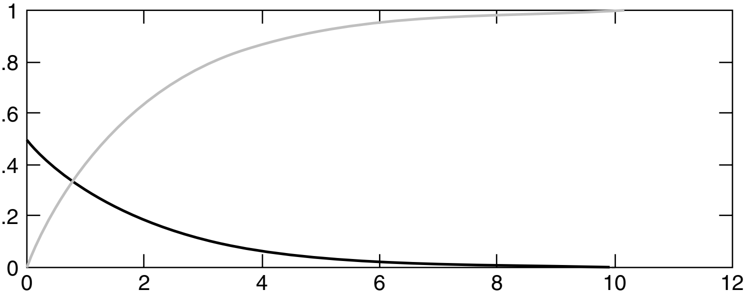


Figure 2.7 Exponential distribution.

As we shall see presently, the parameters μ and σ^2 are the mean and the variance, respectively. The probability distribution is given by

$$F_X(x;\, \mu,\, \sigma) = \int_{-\infty}^{x} \frac{e^{-(t-\mu)^2/2\sigma^2}}{\sqrt{2\pi\sigma^2}}\, dt \tag{2.53}$$

Unlike the cases of the uniform and the exponential distributions, there is no closed-form expression for the probability distribution function. A function that is frequently available is the error function, which is defined as $\text{erf}(x) = (1/\sqrt{2\pi}) \int_0^x e^{-t^2/2} dt$. It is a simple manipulation to show that $F_X(x;\, \mu,\, \sigma) = 0.5 + 0.5\text{erf}((x-\mu)/\sigma\sqrt{2})$.

Finding the moment for the Gaussian distribution is not straightforward integral calculus, as were the exponential and uniform cases, and we must use the results from a table of integrals: $\int_{-\infty}^{\infty} e^{-x^2/2} dx = \sqrt{2\pi}$ and $\int_{-\infty}^{\infty} x^2 e^{-x^2/2} dx = \sqrt{2\pi}$. With the appropriate changes of variables, we find that

$$E(X) = \int_{-\infty}^{\infty} x \frac{e^{-(x-\mu)^2/2\sigma^2}}{\sqrt{2\pi\sigma^2}} dx = \mu \tag{2.54}$$

and

$$\text{Var}(X) = \int_{-\infty}^{\infty} (x-\mu)^2 \frac{e^{-(x-\mu)^2/2\sigma^2}}{\sqrt{2\pi\sigma^2}} dx = \sigma^2 \tag{2.55}$$

The Laplace transform is not well defined for the Gaussian random variable since it can take on negative values; however, the characteristic function for a Gaussian random variable is

$$\Phi_X(\omega) = e^{j\omega\mu - \sigma^2\omega^2/2}$$

Analogous to the discrete case, all of these Laplace transforms equal one when $s = 0$ (or $\omega = 0$). Deriving the mean and the variance for each of these distributions from the Laplace transform is a useful exercise.

Much of the importance of the Gaussian distribution is due to the *central-limit theorem*, which roughly states that the sum of a number of independent random variables approach having a Gaussian distribution as the number increases. The random variables can be of any of the types that would be encountered in most physical situations. The interesting thing is that the number of random variables that are summed does not need to be very large. The sum of 10 independent uniformly distributed random variables is a fair approximation to the Gaussian distribution, provided one does not go far out on the tails in an effort to compute small probabilities.

Example 2.10 In Figure 2.8, the density functions, as calculated on the associated spreadsheet is plotted for the case $\mu = 1.5$ and $\sigma = 0.5$.

Approximation of Binomial and Poisson Distributions by the Gaussian Distribution An interesting application is the approximation of the binomial and the Poisson distributions by the Gaussian distribution. Computations of probabilities for the binomial distributions often involve large summations, too large for numerical stability. Once again, we recall that the central-limit theorem states that the distribution of a sum of a large number of independent random variables may be approximated as Gaussian. The binomial distribution can be viewed as the sum of independent Bernoulli random variables; hence the central-limit theorem can be applied. The Gaussian distribution is specified by its mean, nP, and variance, $nP(1-P)$:

$$\begin{aligned} P(n_1 \leq B \leq n_2) &= \sum_{i=n_1}^{n_2} \binom{n}{i} P^i (1-P)^{n-i} \\ &\cong \int_{n_1-1/2}^{n_2+1/2} \frac{\exp\left(-(x-nP)^2/2nP(1-P)\right)}{\sqrt{2\pi nP(1-P)}}\, dx \end{aligned} \tag{2.56}$$

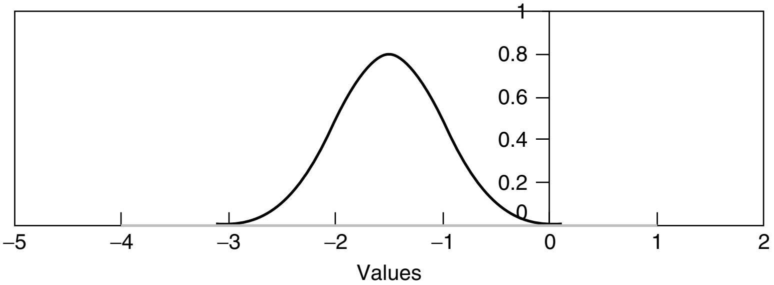


Figure 2.8 Gaussian distribution.

The factors $\frac{1}{2}$ in the integral limits are in recognition of the approximation of a discrete distribution by a continuous distribution. A good rule of thumb for the applicability of this approximation is $nP \geq 10$.

Recall that the Poisson distribution approximates the binomial when n is large and P is small. The Gaussian distribution may also be used to approximate the Poisson distribution if nP is large. The mean and the variance are both λt, and we have

$$P(P \geq j) = \sum_{i=j}^{\infty} \frac{(\lambda t)^i e^{-\lambda t}}{i!} \cong \int_{j-1/2}^{\infty} \frac{\exp(-(x-\lambda t)^2/2\lambda t)}{\sqrt{2\pi\lambda t}}\, dx \qquad (2.57)$$

The results discussed in this section are summarized in Table 2.1.

Table 2.1 Summary of Distributions, Densities and Moments

Discrete Random Variables	Distribution	Mean	Variance	Probability Generating Function
Discrete Uniform	$1/m;\ i = k+1,$ $k+2, \ldots, k+m$	$k + \frac{(m+1)}{2}$	$\frac{m^2-1}{12}$	$\frac{z^{k+1}(1-z^m)}{1-z}$
Binomial	$\binom{n}{k} P^k (1-P)^{n-k};$ $k = 0, 1, \ldots, n$	nP	$nP(1-P)$	$(1-P+zP)^n$
Poisson	$\frac{(\lambda t)^k}{k!} e^{-\lambda t};$ $k = 0, 1, \ldots$	λt	λt	$e^{-\lambda t(1-z)}$
Geometric-Type1	$P(1-P)^k;$ $k = 0, 1, \ldots$	$\frac{1-P}{P}$	$\frac{1-P}{P^2}$	$P/(1-z(1-P))$
Geometric-Type 2	$P(1-P)^{k-1};$ $k = 1, 2, \ldots$	$\frac{1}{P}$	$\frac{1-P}{P^2}$	$zP/(1-z(1-P))$

Continuous Random Variables	Density	Mean	Variance	Laplace transform
Uniform	$1/(b-a);\ a < x \leq b$ 0; Elsewhere	$(a+b)/2$	$(b-a)^2/12$	$\frac{e^{-as}-e^{-bs}}{b-a}$
Exponential	$\mu e^{-\mu t};\ t \geq 0$	$\frac{1}{\mu}$	$\frac{1}{\mu^2}$	$\frac{\mu}{s+\mu}$
Gaussian	$e^{-(x-\mu)^2/2\sigma^2}/\sigma\sqrt{2\pi};$ $-\infty < x < \infty$	μ	σ^2	$e^{j\mu\omega-\sigma^2\omega^2/2}$

2.3 JOINT DISTRIBUTIONS OF RANDOM VARIABLES

2.3.1 Probability Distributions

In this section we consider relations between two or more random variables. We define the *joint* probability distribution function of two random variables X and Y as

$$F_{XY}(x, y) = P(X \leq x, Y \leq y)$$

where x and y are real numbers. In words, this is the joint occurrence of the event $\{X \leq x\}$ *and* $\{Y \leq y\}$. The properties of the joint distribution function follow directly from the basic relations

1. $F_{XY}(-\infty, -\infty) = 0$
2. $F_{XY}(\infty, \infty) = 1$,
3. $F_{XY}(x_1, y) \leq F_{XY}(x_2, y)$ for $x_1 \leq x_2$
4. $F_{XY}(x, y_1) \leq F_{XY}(x, y_2)$ for $y_1 \leq y_2$
5. The *marginal distributions* are given by $F_X(x) = F_{XY}(x, \infty)$ and $F_Y(y) = F_{XY}(\infty, y)$

For discrete non-negative random variables, we have

$$F_{XY}(x, y) = \sum_{i=0}^{\infty}\sum_{j=0}^{\infty} P(X = x_i, Y = y_j)U(x - x_i)U(y - y_j)$$

where $P(X = x_i, Y = y_j)$ is the joint probability for the random variables X and Y. For continuous random variables, we define the joint probability density function

$$f_{XY}(x, y) = \frac{\partial^2 F_{XY}(x, y)}{\partial x \partial y}$$

or

$$F_{XY}(x, y) = \int_{-\infty}^{x}\int_{-\infty}^{y} f_{XY}(\xi, \varsigma)d\xi d\varsigma$$

In analogy with the one-dimensional case, we have the following properties for the joint density function: $f_{XY}(x, y) \geq 0$; $-\infty < x, y < \infty$ and $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{XY}(\xi, \varsigma)d\xi\, d\varsigma = 1$. The marginal distribution and density functions, when they exist, may be calculated from the joint density functions. We have

$$f_X(x) = \int_{-\infty}^{\infty} f_{XY}(x, \varsigma)d\varsigma$$

and

$$f_Y(y) = \int_{-\infty}^{\infty} f_{XY}(\xi, y)d\xi$$

Two random variables are *independent* if the joint distribution functions are products of the marginal distributions:

$$F_{XY}(x, y) = F_X(x)F_Y(y) \tag{2.58}$$

For independent discrete random variables, we have

$$P(M = i, N = j) = P(M = i)P(N = j) \tag{2.59}$$

For continuous random variables, the joint density function is the product of the marginal density functions

$$f_{XY}(x, y) = f_X(x)f_Y(y) \tag{2.60}$$

For random variables, which are not independent, the concepts of conditional distribution and joint density functions come into play. We can define the conditional distribution (see Section (2.1.2)):

$$F_{Y/X}(y/x) = P(Y \leq y/X \leq x) = \frac{F_{XY}(x, y)}{F_X(x)} \tag{2.61}$$

For discrete random variables, we have

$$P(M = i/N = j) = \frac{P(M = i, N = j)}{P(N = j)} \tag{2.62}$$

Similarly, for continuous random variables, we have

$$f_{Y/X}(y/x) = \lim_{\delta y \to 0} P(y < Y \leq y + \delta y/X = x) = \frac{f_{XY}(x, y)}{f_X(x)} \tag{2.63}$$

This is the probability density of Y conditioned on the event $X = x$.

In the text, we make particular use of the joint distribution of discrete and continuous random variables. Let N and X be discrete and continuous random variables, respectively. From the law of total probability, (2.2), we may write

$$F_x(x) = \sum_{j=0}^{\infty} P(N = j, X \leq x) = \sum_{j=0}^{\infty} P(N = j)P(X \leq x/N = j)$$

$$= \sum_{j=0}^{\infty} P_j F_{X/N}(x/j) \tag{2.64}$$

Although we have discussed joint density functions, joint distribution functions, and independence in connection with two random variables, these same concepts apply in a straightforward way to three or more random variables, $X_1, X_2, \ldots, X_N$. Particularly useful relations in this respect are

$$F_{X_1, X_2, \ldots, X_N}(x_1, x_2, \ldots, x_N)$$
$$= F_{X_1}(x_1)F_{X_1/X_2}(x_1/x_2) \cdots F_{X_N/X_1, \ldots, X_{N-1}}(x_N/x_1, \ldots, x_{N-1}) \tag{2.65}$$

and

$$f_{X_1, X_2, \ldots, X_N}(x_1, x_2, \ldots, x_N)$$
$$= f_{X_1}(x_1)f_{X_1/X_2}(x_1/x_2) \cdots f_{X_N/X_1, \ldots, X_{N-1}}(x_N/x_1, \ldots, x_{N-1}) \tag{2.66}$$

2.3.2 Joint Moments

The computation of expected values of functions of two random variable follows directly from the definition of expectation. Let $g(\cdot, \cdot)$ be an arbitrary function of two variables. For the discrete case, we have

$$E[g(X, Y)] = \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} g(i, j)P(X = i, Y = j) \tag{2.67}$$

and for the continuous case

$$E[g(X, Y)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y) f_{XY}(x, y) dx\, dy \tag{2.68}$$

An important property of independent random variables is that the expectations of products are equal to the products of expectations:

$$E[g(X)h(Y)] = E[g(X)]E[h(Y)] \tag{2.69}$$

This follows directly from (2.58) and (2.59) since the joint distribution separates into the product of the marginal distributions.

2.3.3 Autocorrelation and Autocovariance Functions

The expected value of the product of two random variables plays an important role. For discrete random variables, we have

$$E[XY] = \sum_{i=0}^{\infty}\sum_{j=0}^{\infty} ijP(X = i, Y = j) \tag{2.70}$$

and

$$E[XY] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xyf_{XY}(x, y)dx\,dy \tag{2.71}$$

This joint first moment is called the *correlation* or the *autocorrelation* of the random variables X and Y. The *covariance* or the *autocovariance* is defined as

$$\mathrm{Cov}[X, Y] = E[(X - E[X])(Y - E[Y])] \tag{2.72}$$

We say that two random variables are *uncorrelated* if $E[XY] = E[X]E[Y]$ or equivalently $\mathrm{Cov}(X, Y) = 0$. It is easy to show that independent random variables are uncorrelated. The reverse does not hold, except in the case of Gaussian random variables. (See Example 2.11.)

Example 2.11 The general form for the joint distribution of two Gaussian random variables is

$$f_{XY}(x, y)$$
$$= \frac{\exp\left\{\left(\frac{-1}{2(1-\rho_{XY}^2)}\right)\left[\left(\frac{x-\mu_X}{\sigma_X}\right)^2 - 2\rho_{XY}\left(\frac{x-\mu_X}{\sigma_X}\right)\left(\frac{y-\mu_Y}{\sigma_Y}\right) + \left(\frac{-1}{2(1-\rho_{XY}^2)}\right)\left(\frac{y-\mu_Y}{\sigma_Y}\right)^2\right]\right\}}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho_{XY}^2}}$$

where μ_X and μ_Y are the respective means and σ_X^2 and σ_Y^2 are the respective variances. The parameter is defined as $\rho_{XY} = \mathrm{Cov}(X, Y)/\sigma_X\sigma_Y$. Now, if X and Y are uncorrelated, this parameter is zero and the cross-term in the joint density disappears:

$$f_{XY}(x, y) = \frac{\exp\{-[(x-\mu_X)^2/2\sigma_X^2] - [(y-\mu_Y)^2/2\sigma_Y^2]\}}{2\pi\sigma_X\sigma_Y} = f_X(x)f_Y(y)$$

Thus, the joint distribution is the product of the marginal distributions and the random variables are independent.

2.4 LINEAR TRANSFORMATIONS

2.4.1 Single Variable

Random variables can be transformed to produce other random variables. Consider, for example, the linear transformation $Y = aX + b$, where X is a random variable with mean μ and variance σ^2 and a and b are constants. It is left as an exercise to show that the mean and the variance of Y are, respectively, $a\mu + b$ and $a^2\sigma^2$.

2.4.2 Sums of Random Variables

Perhaps the most important transformation is the sum of two random variables producing a third random variable. We begin with the sum of discrete random variables, $Y = X_1 + X_2$. We have

$$\begin{aligned} E[Y] &= \sum_{x_1=0}^{\infty}\sum_{x_2=0}^{\infty}(x_1 + x_2)P(X_1 = x_1, X_2 = x_2) \\ &= \sum_{x_1=0}^{\infty}\sum_{x_2=0}^{\infty}x_1P(X_1 = x_1, X_2 = x_2) + \sum_{x_1=0}^{\infty}\sum_{x_2=0}^{\infty}x_2P(X_1 = x_1, X_2 = x_2) \\ &= \sum_{x_1=0}^{\infty}x_1P(X_1 = x_1) + \sum_{x_2=0}^{\infty}x_2P(X_2 = x_2) = E(X_1) + E(X_2) \end{aligned}$$

The corresponding formula for the continuous case is

$$\begin{aligned} E(Y) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x_1 + x_2)f_{X_1X_2}(x_1, x_2)dx_1dx_2 \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}x_1f_{X_1X_2}(x_1, x_2)dx_1dx_2 + \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}x_2f_{X_1X_2}(x_1, x_2)dx_1dx_2 \\ &= \int_{-\infty}^{\infty}x_1f_{X_1}(x_1)dx_1 + \int_{-\infty}^{\infty}x_2f_{X_2}(x_2)dx_2 \\ &= E(X_1) + E(X_2) \end{aligned}$$

From these two results, we conclude that the expectation is a linear operation since the expectation of the sum of two random variables is the sum of the expected values of each individually. It is easy to show that this result obtains for the sum of more than two random variables:

$$E\left(\sum_{i=1}^{n}X_i\right) = \sum_{i=1}^{n}E(X_i) \tag{2.73}$$

It is important to point out that (2.73) holds regardless of whether the random variables are independent.

The calculation of the variance of the sum of random variables is also straightforward:

$$\begin{aligned}\mathrm{Var}[X+Y] &= E[(X+Y-E[X+Y])^2] \\ &= E[(X-E[X])^2]+E[(Y-E[Y])^2] \\ &\quad +2E[(X-E[X])(Y-E[Y])] \\ &= \mathrm{Var}[X]+\mathrm{Var}[Y]+2\mathrm{Cov}(X,\,Y) \qquad (2.74)\end{aligned}$$

Note that for uncorrelated random variables, the covariance is equal to zero and we have

$$\mathrm{Var}[X+Y]=\mathrm{Var}[X]+\mathrm{Var}[Y] \qquad (2.75)$$

Equation (2.75) says that the variance of the sum of independent random variables is equal to the sum of the variances.

Convolution The probability distribution of the sums of independent random variables can be found from the *convolution* operator, which looks different for discrete and continuous random variables. For the sum of independent discrete random variables, $N=K_1+K_2$, we have

$$P(N=n)=\sum_{i=0}^{n}P(K_1=i)P(K_2=n-i) \qquad (2.76)$$

(Recall that we assume that discrete random variables assume the values $0, 1, 2, \ldots$.) The application of the law of total probability is in evidence here.

Example 2.12 We consider the sum of two independent Poisson random variables with the respective mean values $\lambda_1 t$ and $\lambda_2 t$. From (2.76) we have

$$\begin{aligned}P(N=n) &= \sum_{i=0}^{n}\frac{e^{-\lambda_1 t}(\lambda_1 t)^i}{i!}\frac{e^{-\lambda_2 t}(\lambda_2 t)^{n-i}}{(n-i)!}=e^{-(\lambda_1+\lambda_2)t}\sum_{i=0}^{n}\frac{(\lambda_1 t)^i(\lambda_2 t)^{n-i}}{i!(n-i)!} \\ &= e^{-(\lambda_1+\lambda_2)t}\frac{((\lambda_1+\lambda_2)t)^n}{n!} \qquad (2.77)\end{aligned}$$

The last step here uses the binomial expansion.[5] Notice that the probability distribution of the sum of two independent Poisson random variables is itself

[5] The binomial theorem states that $(x+y)^n=\sum_{i=0}^{n}\binom{n}{i}x^i y^{n-i};\ n\geq 0$.

Poisson and the mean is the sum of the individual means, $(\lambda_1 + \lambda_2)t$. This property does not hold for the general case of discrete random variables.

The probability distribution of the sum of continuous independent random variables, $Y = X_1 + X_2$ is also given by a convolution operation

$$f_Y(y) = \int_{-\infty}^{\infty} f_{X_1}(z) f_{X_2}(z - y)dz \tag{2.78}$$

In the remainder of the text, we will be dealing with sums of independent, identically distributed (iid) exponential random variables. For the sum of 2, the probability density function is given by

$$f_Y(y) = \int_0^{\infty} \mu e^{-\mu x} \mu e^{-\mu(y-x)} dx = y\mu^2 e^{-\mu y}$$

For the sum of k, the probability density function can be found by repeated convolution to obtain

$$f_Y(y) = \frac{\mu(\mu y)^{k-1} e^{-\mu y}}{(k-1)!}; \quad y \geq 0, k \geq 1 \tag{2.79}$$

This distribution is called the *k-stage Erlang* or the simply the *Erlang k* distribution. The Erlang k distribution is equal to the *chi-square* (χ^2) distribution with $2k$ degrees of freedom.

Probability-Generating Function The probability-generating functions, for discrete random variables play an important role for the sums of independent discrete random variables since the generating functions of the sum of independent random variables is the product of the individual generating function. The definition of the generating function for the sum is a straightforward extension. If $M = N_1 + N_2$ then

$$M(z) = E[z^{N_1+N_2}] = \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} P(N_1 = i, N_2 = j) z^{i+j}$$

If N_1 and N_2 are independent, we can write

$$M(z) = E[z^{N_1+N_2}] = \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} P(N_1 = i) z^i P(N_2 = j) z^j = N_1(z) \cdot N_2(z) \tag{2.80}$$

Example 2.13 The probability-generating function can be used to provide another proof that the sum of independent Poisson random variables is itself Poisson. Let $Y = X_1 + X_2$, where X_1 and X_2 have mean values $\lambda_1 t$ and $\lambda_2 t$, respectively. Since the generating function of the sum is the product of the individual generating functions, we have

$$P_Y(z) = P_1(z)P_2(z) = e^{-\lambda_1 t(1-z)}e^{-\lambda_2 t(1-z)} = e^{-(\lambda_1+\lambda_2)t(1-z)} \tag{2.81}$$

The form in (2.81) is that of a Poisson random variable with mean value $(\lambda_1 + \lambda_2)t$. This is the only discrete distribution with this property.

Laplace Transforms and Characteristic Functions The Laplace transform for the sum of independent random variables is the product of the individual Laplace transforms, namely, for $Y = X_1 + X_2$. The derivation is a straightforward application starting with the definition:

$$\begin{aligned} E[e^{-sY}] = E[e^{-s(X_1+X_2)}] &= \int_0^\infty \int_0^\infty f_{X_1X_2}(x_1, x_2)e^{-s(x_1+x_2)}dx_1dx_2 \\ &= \int_0^\infty \int_0^\infty f_{X_1}(x_1)f_{X_2}(x_2)e^{-sx_1}e^{-sx_2}dx_1dx_2 \\ &= E[e^{-sX_1}] \cdot E[e^{-sX_2}] = X_1(s) \cdot X_2(s) \end{aligned} \tag{2.82}$$

It is also straightforward to show that the same applies to characteristic functions. The Laplace transform provides interesting results for the Erlang k distribution. Since this distribution is the distribution of k iid exponential random variables, the Laplace transform is the kth power of the individual Laplace transforms given by (2.82). For $Y = \sum_{i=1}^{k} X_i$, we have

$$Y(s) = \left[\frac{\mu}{\mu + s}\right]^k \tag{2.83}$$

The mean and the variance are, respectively

$$E[Y] = \frac{k}{\mu} \quad \text{and} \quad \mathrm{Var}[Y] = \frac{k}{\mu^2}$$

An interesting result, which ties together the distributions that we have considered, involves the sum of exponentially distributed random variables when the number of terms in the sum is itself geometrically distributed. Conditioned on the number of terms in the sum, the Laplace transform for the sum is given by (2.83). Averaging over the geometric distribution type 2 (see Table 2.1), we find

$$Y(z) = \sum_{n=1}^{\infty}(1 - P)^{n-1}P\left[\frac{\mu}{\mu + s}\right]^n = \frac{\mu P}{s + \mu P} \tag{2.84}$$

Thus, the sum is an exponentially distributed random variable with mean $1/\mu P$.

An interesting form of the Erlang k distribution occurs when the mean is normalized so as to be invariant with k. The individuals in the sum are multiplied by $1/k$ so that $E[Y] = 1/\mu$. The probability generating function is then

$$Y(z) = \left[\frac{k\mu}{k\mu + s}\right]^k$$

We take a limit as $k \to \infty$ to obtain

$$\lim_{k\to\infty} Y(z) = \lim_{k\to\infty}\left[1 + \frac{s}{k\mu}\right]^{-k} = e^{-s/\mu} \tag{2.85}$$

Equation (2.82) is the Laplace transform for a constant, that is, a degenerate random variable that assumes a single value, $f_X(x) = \delta(x - 1/\mu)$, where $\delta(x)$ is the Dirac delta function. We shall use this result in later chapters of the text when we model constant service times.

2.5 TRANSFORMED DISTRIBUTIONS

Consider the general transformation

$$Y = g(X)$$

where $g(\cdot)$ is some nonlinear function of its argument. If the function is monotonically nondecreasing, then the following simple relationship holds for the probability distribution functions

$$F_Y(y) = P(Y \le y) = P(g(X) \le y) = P(X \le g^{-1}(y)) = F_X(g^{-1}(y)) \tag{2.86}$$

where $F_Y(y)$ and $F_X(x)$ are the distribution functions for Y and X, respectively. $f^{-1}(y)$ is the inverse of $f(y)$. If X is a continuous random variable, then it follows that the probability density functions are related as $f^{-1}(y)$

$$f_Y(y) = f_X(g^{-1}(y))\left|\frac{dg^{-1}(y)}{dy}\right| \tag{2.87}$$

Thus, the simple transformation $Y = X + a$ where a is a constant yields

$$F_Y(y) = F_X(y - a)$$

and

$$f_Y(y) = f_X(y - a)$$

The random variable $Y = aX$ has the distribution function

$$F_Y(y) = F_X(y/a)$$

and density function

$$f_Y(y) = \frac{f_X(y/a)}{a}$$

Another useful example is when X is uniformly distributed between 0 and 1 and the transformation is

$$Y = -\ln(1 - X) \tag{2.88}$$

The distribution of X is the uniform distribution given by (2.42) with $a = 0$ and $b = 1$. The inverse of the transformation is

$$x = g^{-1}(y) = 1 - e^{-y}$$

The distribution of X is given by

$$F_X(x) = \begin{cases} 0; & -\infty < x \leq 0 \\ x; & 0 < t \leq 1 \\ 1; & t > 1 \end{cases}$$

Substituting into (2.83), we find

$$F_Y(y) = 1 - e^{-y}; \quad y \geq 0 \tag{2.89}$$

However, this is an exponentially distributed random variable with mean 1. This can be transformed into an exponentially distributed random variable with mean $1/\mu$ simply by multiplying by $1/\mu$. This result is applied in simulation programs. The standard random-number generator gives a number that is uniformly distributed between 0 and 1. If we use the transformation of (2.88) on the output of the random-number generator, we get an exponentially distributed random variable, Y, which has mean value 1. We will see more examples of this later, when we deal with simulation.

2.6 INEQUALITIES AND BOUNDS

Markov Inequality Several useful inequalities can be derived in a unified framework. Suppose that we have a positive nondecreasing function $h(x)$. Two examples that will be used later are $h(x) = x$ and $h(x) = e^{\alpha x}$. Next assume a random variable X with probability density function $f_X(x)$. From (2.38), and the assumed properties of $h(x)$, which makes the integrand positive, we have

$$E(h(X)) = \int_{-\infty}^{\infty} h(x)f_X(x)dx \geq \int_{t}^{\infty} h(x)f_X(x)dx \geq h(t)\int_{t}^{\infty} f_X(x)dx \geq h(t)P(X \geq t)$$

This can be written succinctly as the *Markov inequality*:

$$P(X \geq t) \leq \frac{E(h(X))}{h(t)} \tag{2.90}$$

This inequality can be used to find other useful inequalities. Suppose that X is a nonnegative random variable. Let $h(x) = xU(x)$. Substituting, we find the form of the *simple Markov inequality*:

$$P(X \geq t) \leq \frac{E(X)}{t}; \quad t \geq 0 \tag{2.91}$$

Example 2.14 We find the simple Markov inequality for an Erlang 4 distribution with $\mu = 2$. Since this distribution is the sum of four independent exponentially distributed random variables, each having mean 0.5, the bound is just $P(X \geq t) \leq 2.0/t$; $t \geq 0$. We can compare this with the true value in order to evaluate the value of the inequality. The Erlang k distribution is given by (2.83). From a table of integrals,[6] we have

$$P(X \geq t) = \int_{t}^{\infty} \frac{2(2x)^{k-1}}{(k-1)!}e^{-2x}dx = e^{-2t}\sum_{j=0}^{3}\frac{(2t)^{k-1-j}}{(k-1-j)!}$$

where $k = 4$.

The results are worked out on the accompanying spreadsheet and plotted in Figure 2.9. As we see, the simple Markov inequality is a rather loose bound. Its virtue is that it provides a way to get a quick estimate.

Chebyshev Inequality Now, given the moments of random variables, we can find statements about probabilities of random variables by application of the *Chebyshev inequality*, which can be derived from (2.90). Assume that the random variable X has a finite variance, σ_X^2, and define $Y = (X - E(X))^2$ and $h(y) = y$. In (2.90), let $t = \varepsilon^2$. We have

$$P(Y \geq \varepsilon^2) \leq \frac{E(Y)}{\varepsilon^2} \tag{2.92}$$

[6] See, for example, Gradshteyn and Ryzhik (1965), equation 2.321 #4.

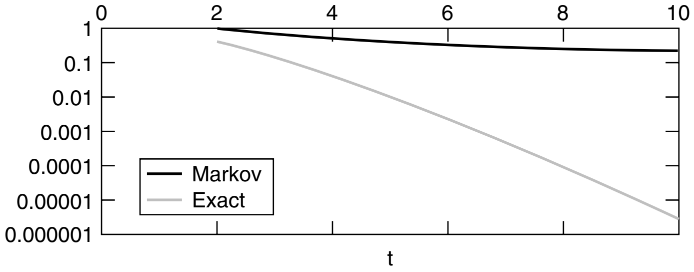


Figure 2.9 Markov inequality.

But $E(Y) = \text{Var}(X)$ and $Y \geq \varepsilon^2 \Leftrightarrow |X - E(X)| \geq \varepsilon$. Substituting into (2.92), yields

$$P(|X - E(X)| \geq \varepsilon) \leq \frac{\sigma_X^2}{\varepsilon^2} \tag{2.93}$$

An interesting observation concerns the normalized form of the Erlang k random variable considered in the previous section. Recall that this was the sum of k independent exponentially distributed random variables, each having mean $E[X] = 1/k\mu$. The mean and variance are then $1/\mu$ and $1/k\mu$, respectively. Application of the Chebyshev inequality shows that the normalized Erlang k distribution approaches that of a constant as k increases. This reinforces the result obtained at the end of the previous section.

Example 2.15 We find the Chebyshev inequality for the Erlang 4 distribution with $\mu = 2$. Since the terms in the sum are independent of one another, the variance of the sum is the sum of the variances, that is, 1.0. The inequality is then

$$P(|X - 2.0| \geq d) \leq \frac{1.0}{d^2}$$

We can compare this to an exact calculation. Since the random variable is positive, for $\varepsilon \geq 2.0$, we have $P(|X - 2.0| \geq d) = \int_{2.0+d}^{\infty} f_X(x)dx$. The results are plotted in Figure 2.10. Again we see that the bound is rather loose.

Chernoff Bound The final bound is the *Chernoff bound*, which follows when $h(t) = e^{-\alpha t}$; $\alpha \geq 0$. The result in (2.90) is

$$P(X \geq d) \leq e^{-\alpha d}E(e^{\alpha X}) = e^{-\alpha d}\mathsf{X}(-\alpha); \quad \alpha \geq 0 \tag{2.94}$$

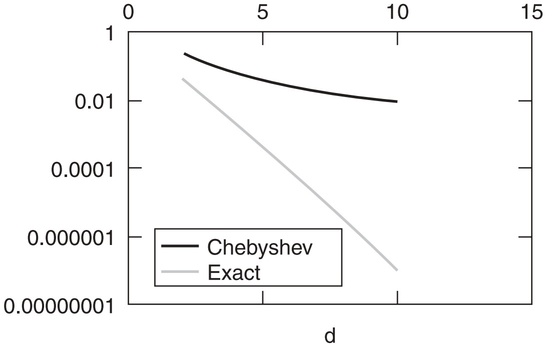


Figure 2.10 Chebyshev inequality.

where $X(-\alpha)$ is the Laplace transform of the probability density function evaluated at $-\alpha$. Notice that this bound applies for all $\alpha \geq 0$; thus, the bound can be made tighter by minimizing the right-hand side (RHS) of (2.94) (see Example 2.16).

Since the integral defining the Laplace transform does not converge for all values of the Laplace variable, one should be careful in the application of (2.94). A simple example for the exponential distribution illustrates this point. Substituting (2.51) into (2.94), we see that for $\alpha > \mu$ we have $P(X \geq d) \leq e^{-\alpha d}(\mu/(-\alpha + \mu)) < 0$, which is obviously wrong. The problem is that the $X(-\alpha)$ exists only for $\alpha < \mu$.

An alternative derivation of the Chernoff bound may furnish further insight. The probability can be written using a unit step function as follows

$$P(X > d) = \int_d^\infty f_X(x)dx = \int_0^\infty U(x-d)f_X(x)dx$$

where $U(x)$ is the unit step. Now for any real number α such that $\alpha \geq 0$, we have the inequality $e^{\alpha(t-d)} \geq U(t-d)$ (Fig. 2.11). Thus

$$P(X > d) \leq \int_0^\infty e^{\alpha(x-d)} f_X(x)dx = e^{-\alpha d} X(-\alpha)$$

Example 2.16 The Erlang k distribution has an interesting bound. Substituting (2.83) into the Chernoff bound (2.94), we have

$$P(X \geq t) \leq e^{-t\alpha}\left(\frac{\mu}{-\alpha + \mu}\right)^k$$

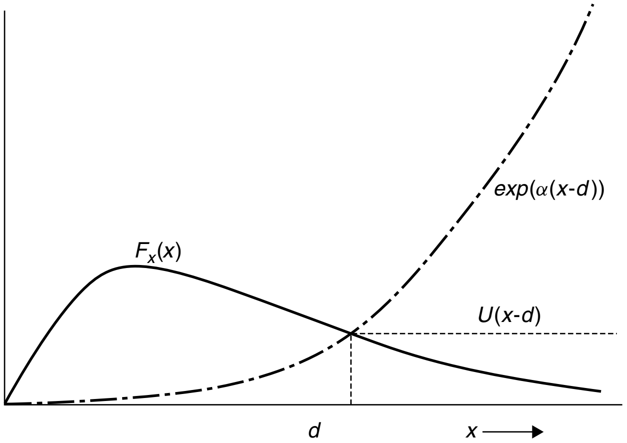


Figure 2.11 Chernoff bound.

Differentiating the RHS with respect to α and setting the result equal to zero, we find

$$-te^{-t\alpha}\left(\frac{k\mu}{-\alpha+k\mu}\right)^k+e^{-t\alpha}k\left(\frac{k\mu}{-\alpha+k\mu}\right)^{k-1}\frac{k\mu}{(-\alpha+k\mu)^2}=0$$

After canceling factors and solving for α, we find $\alpha=\mu-k/t$. (Notice that $\mu>\alpha$.)

Substituting into the upper bound, we find $P(X\geq t)\leq e^{-t\mu+k}(\mu t/k)^k$. This bound is plotted in Figure 2.12 along with the exact value. As is evident, the Chernoff bound is much better than the others that we have studied. It is significant that the Chernoff bound and the exact distribution have the same slope.

In our treatment of discrete random variables, we used the moment-generating function characterization of the distribution. The Chernoff bound can just as well be

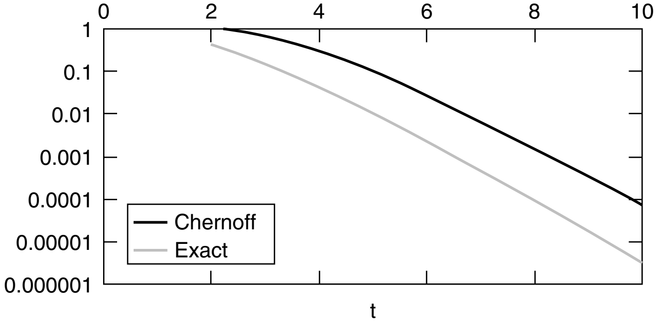


Figure 2.12 Chernoff bound versus exact distribution.

applied to the moment-generating function simply by setting $z = e^{\alpha}$ in the expression for the probability-generating function.

Weak Law of Large Numbers An application of the Chebyshev inequality results in the *weak law of large numbers*. Consider n independent, identically distributed (iid) random variables, $X_1, X_2, \ldots, X_n$, each with mean m and variance σ^2. The average $S = (1/n)\sum_{i=1}^{n} X_i$ has mean m and variance σ^2/n. Substituting into (2.93), we find

$$\lim_{n\to 0} P(|S - m| \geq \varepsilon) \leq \lim_{n\to 0} \frac{\sigma^2}{n\varepsilon^2} = 0; \quad \varepsilon \geq 0 \tag{2.95}$$

The implication of (2.95) is that the probability of S being different from the mean value, m, is zero in the limit; in other words, the arithmetic average of independent samples approaches the mean of the underlying distribution with probability one as the number of samples increases.

2.7 MARKOV CHAINS

2.7.1 The Memoryless Property

In this section we shall review the basic properties of Markov chains and Markov processes. Since this is only a review, no detailed proofs will be given.[7] Markov chains are members of the class of random processes, which assume a countable set of values and change state at regularly spaced intervals, which we write as T seconds. Markov chains are characterized by a certain memorylessness in the state transitions; the probability distribution of the state after the next transition depends only on the present state and not on the succession of states that led up to the present state. Let S_i indicate the state of the system at time iT. The memoryless property implies that

$$P(S_{N+1} = l_{N+1}/S_1 = l_1, S_2 = l_2, \ldots, S_N = l_N) = P(S_{N+1} = l_{N+1}/S_N = l_N) \tag{2.96}$$

where $l_1, l_2, \ldots, l_N, \ldots$ are the particular values, drawn from a countable set that the state assumes. *Bernoulli trials* provide an example of Markov chains. This consists of a sequence of equally spaced trials or tests. On each trial there is success with probability P, which is independent of the outcomes of past trials. We define the state on the Nth trial to be the accumulated number of successes. Let S_i denote the total number of successes after the ith trial. The probability distribution of the state after the $(N + 1)$th trial conditioned on the previous states is given by (2.96).

[7] A number of texts treat Markov chains. We cite four with which we are most familiar: Cox (1965), Karlin and Taylor (1975), Nelson (1995), and Papoulis and Pillai (2002).

In general, the probability of the transition from one state to another can change with time; however, for a *homogeneous* Markov chain there is no time dependence and the transition probability is a function only of the states. In this case we denote the probability of transition from state i to state j by $p_{ij} = P(S_{N+1} = j/S_N = i)$.

A convenient representation for the states in a homogeneous Markov chain is the *state transition probability diagram.* The possible states in the chain are connected by directed arcs, which are weighted by the probability of going from one state to another. In Figure 2.13, we show the state transition probability diagram for a Markov chain in which the state is the number of successes in a sequence of Bernoulli trials.

2.7.2 State Transition Matrix

If there are a finite number of states, state transitions may be represented by the state transition matrix whose elements are the state transition probabilities $\{p_{ij}\}$:

$$P = \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1M} \\ p_{21} & p_{22} & \cdots & p_{2M} \\ \vdots & \vdots & \ddots & \vdots \\ p_{M1} & p_{M2} & \cdots & p_{MM} \end{bmatrix} \tag{2.97}$$

Since the elements in the rows of P sum to one, $\sum_{j=1}^{M} p_{ij} = 1; i = 1, 2, \ldots, M$, this is a *stochastic matrix.*

Let $\pi_j^i; i = 0, 1, 2, \ldots, j = 1, 2, \ldots, M$ be the probability that the Markov chain is in state j after the ith transition $\pi_j^i = P(S_i = j)$. We define the vector of probabilities $\boldsymbol{\pi}^i = [\pi_1^i, \pi_2^i, \ldots, \pi_M^i]$. From the law of total probability, we can write the probability distribution at the $(N+1)$th in terms of the probability at the Nth step as

$$\begin{aligned} P(S_{N+1} = i) &= \sum_{j=1}^{M} P(S_{N+1} = i, S_N = j) \\ &= \sum_{j=1}^{M} P(S_N = j)P(S_{N+1} = i/S_N = j); \quad i = 1, 2, \ldots, M \end{aligned} \tag{2.98}$$

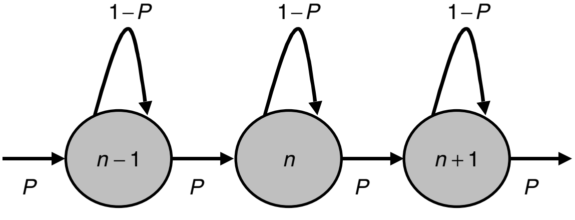


Figure 2.13 State transition diagram: Bernoulli trials.

Using the terms that we have defined, this can be written as

$$\pi_i^{n+1} = \sum_{j=1}^{M} \pi_j^n p_{ji}; \quad i = 1, 2, \ldots, M \tag{2.99}$$

In matrix notation, we have

$$\boldsymbol{\pi}^{n+1} = \boldsymbol{\pi}^n P \tag{2.100}$$

The probability distribution on successive steps can be found by simple iteration. Thus

$$\boldsymbol{\pi}^{n+k} = \boldsymbol{\pi}^n P^k \tag{2.101}$$

where the superscript is the *k*th power of *P*.

Example 2.17 Consider a Markov chain depicted by the state transition diagram in Figure 2.14. The transition matrix is

$$P = \begin{bmatrix} \frac{1}{6} & \frac{1}{2} & 0 & \frac{1}{3} \\ \frac{1}{4} & 0 & \frac{1}{2} & \frac{1}{4} \\ 0 & 0 & 1 & 0 \\ 0 & \frac{1}{4} & \frac{3}{4} & 0 \end{bmatrix}$$

Suppose that the system is in state 1 with probability 1 at the *N*th step. The successive probabilities have been worked out on the accompanying spreadsheet and are shown in the following table, where, as we see, probability accumulates in state 3. The reason is clear. Once the chain enters state 3, there is no escape; we have an *absorbing state*. Since all the other states have a nonzero probability path to state 3, the process we eventually end up in state 3, exclusively, $\boldsymbol{\pi}^{\infty} = [0010]$.

Step	State 1	State 2	State 3	State 4
0	0.166667	0.5	0	0.333333
1	0.152778	0.166667	0.5	0.180556
2	0.06713	0.121528	0.71875	0.092593
3	0.04157	0.056713	0.848958	0.052758
4	0.021107	0.033975	0.916884	0.028035
5	0.012011	0.017562	0.954897	0.015529
6	0.006392	0.009888	0.975325	0.008394
7	0.003537	0.005295	0.986565	0.004603

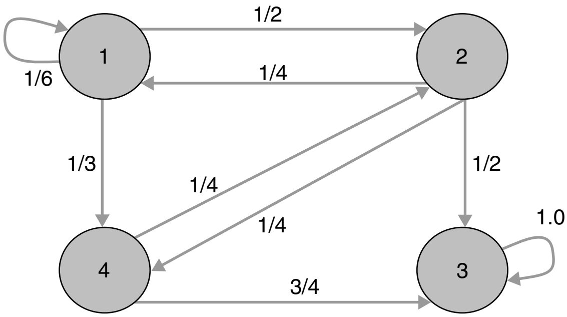


Figure 2.14 State transition diagram: Example 2.17.

Now, suppose that the transition matrix is changed so that there is a transition out of state 3:

$$P = \begin{bmatrix} \frac{1}{6} & \frac{1}{2} & 0 & \frac{1}{3} \\ \frac{1}{4} & 0 & \frac{1}{2} & \frac{1}{4} \\ 1 & 0 & 0 & 0 \\ 0 & \frac{1}{4} & \frac{3}{4} & 0 \end{bmatrix}$$

This is depicted in Figure 2.15.

The results are tabulated as follows:

Step	State 1	State 2	State 3	State 4
0	1	0	0	0
1	0.166667	0.5	0	0.333333
2	0.152778	0.166667	0.5	0.180556
3	0.56713	0.121528	0.21875	0.092593
4	0.343654	0.306713	0.130208	0.219425
5	0.264162	0.226683	0.317925	0.191229
6	0.418623	0.179888	0.256764	0.144725
7	0.371506	0.245493	0.198488	0.184513

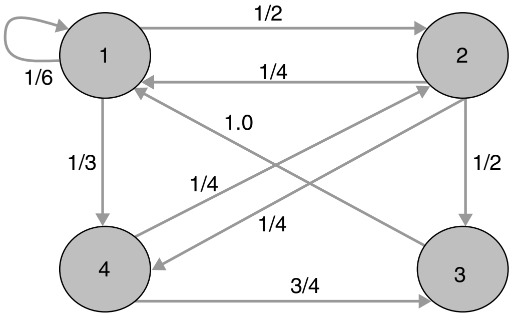


Figure 2.15 State transition diagram—Example 2.17 continued.

2.7.3 Steady-State Distribution

As we have seen in the previous subsection, the system can converge to a steady-state condition. Basically, the existence of steady-state solutions depends on the connectivity of states. We define p_{ij}^n as the (i, j) element of the transition matrix P^n, the probability of going from state i to state j in n trials. We set

$$p_{ij}^0 = \begin{cases} 1; & \text{if } i = j \\ 0; & \text{otherwise} \end{cases}$$

State j is defined to be *accessible* from state i if $p_{ij}^n > 0$ for some value of n. If two states i and j are accessible from each other, we use the notation $i \leftrightarrow j$. In the example of Figure 2.16, we have $2 \leftrightarrow 4$, for example. A set of states C is called *closed* if no state outside C can be reached from a state inside C. An absorbing state is a closed set with only one member. A Markov chain is called *irreducible* if there exist no closed sets other than the set of all states. General Markov chains can be decomposed into one or more irreducible chains. Since there is an absorbing state in the example in Figure 2.14, the Markov chain that the state transition diagram represents is not irreducible. The chain shown in Figure 2.16 does represent an irreducible Markov chain; there is a transition from three to one.

The long-term properties of the Markov chain depend on the recurrence of states. Because of the memorylessness of the Markov chain each time a particular state, say, j, recurs after a number of trials, we are in the same position in terms of future evaluation. This is called a *recurrent event*. The probability of returning to state j after l steps is p_{jj}^l. For each state j we define its *period* $d(j)$ as the greatest common divisor of all the integers $l \geq 1$ for which $p_{jj}^l > 0$. If $p_{jj}^l = 0$ for all $l \geq 1$, we define $d(j) = 0$. In Figure 2.15, for example, we have $d(j) = 1, j = 1, 2, 3, 4$. A Markov chain for which each state has period one is called *a periodic*.

The next concept that we consider is the *first return* to a state. Let f_{jj}^l be the probability that, given that state j occurred on the zeroth trial, it occurs again for the first time on the lth trial. We define $f_{jj}^0 = 0$ for consistency. In Figure 2.15, for example, $f_{11}^1 = \frac{1}{6}$, $f_{11}^2 = \frac{1}{8}$, $f_{11}^3 = \frac{1}{4} + \frac{1}{4} + \frac{1}{48} = \frac{25}{48}$, and so on. The events of the first return on trials $1, 2, \ldots$ are disjoint events, and by the basic axioms of probability

$$P(\text{eventual return to state } j) = \sum_{l=1}^{\infty} f_{ij}^l \leq 1 \tag{2.102}$$

If there is equality for some j, we say that the state j is *persistent*.

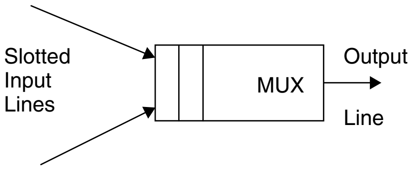


Figure 2.16 Simple multiplexer.

The probability of return to a state, not necessarily the first return, can be expressed in terms of first return probabilities. Consider the event that state i occurs at the lth trial given that i occurred on the zeroth trial. It may be that state i occurred a number of times in the intervening trials. Each pattern of occurrence constitutes a disjoint event. From the law of total probability, we may write

$$\begin{aligned} p_{ii}^{j} &= \sum_{j=1}^{l} P(\text{return on trial } l-j \text{ and first return on trial } k) \\ &= \sum_{j=1}^{l} P(\text{return on trial } l-j)P(\text{first return on trial } j) = \sum_{j=1}^{l} p_{ii}^{l-j} f_{ii}^{j} \end{aligned} \tag{2.103}$$

The key to writing (2.103) is the memorylessness property of Markov chains. Now let us define the following generating functions for $|z| < 1$

$$P_j(z) = \sum_{l=0}^{\infty} p_{jj}^{l} z^{l} \tag{2.104}$$

and

$$F_j(z) = \sum_{l=1}^{\infty} f_{jj}^{l} z^{l} \tag{2.105}$$

Multiplying both sides of (2.103) by z^l and summing from $l = 1$, we find

$$\sum_{l=1}^{\infty} p_{jj}^{l} z^{l} = P_j(z) - 1 = \sum_{l=1}^{\infty} \sum_{j=1}^{l} p_{ii}^{l-j} f_{ii}^{j} z^{l} = P_j(z) F_j(z)$$

or

$$P_j(z) = \frac{1}{1 - F_j(z)} \tag{2.106}$$

It is not difficult to see that if a state of a Markov chain is persistent, then $\lim_{z \to 1} F_i(z) = \sum_{j=0}^{\infty} f_{ii}^{j} = 1$. Furthermore, from (2.104), a necessary and sufficient condition for a state to be persistent is

$$\lim_{z \to 1} P_i(z) = \sum_{j=1}^{\infty} p_{ii}^{j} = \infty$$

A state is called *transient* $\sum_{j=1}^{\infty} f_{ii}^{j} < 1$. Again from (2.104), a necessary and sufficient condition for a transient state is $\sum_{j=1}^{\infty} p_{ii}^{j} < \infty$. The *mean recurrence* time for a state is

$$\mu_i = \sum_{j=1}^{\infty} j f_{ii}^{j} \tag{2.107}$$

A state j is a *persistent null state* if $\mu_j = \infty$. It can be shown that a necessary and sufficient condition for a persistent null state is $\sum_{l=1}^{\infty} p_{jj}^{l} = \infty$ but $\lim_{l\to\infty} p_{jj}^{l} = 0$. It follows directly that for a persistent null state

$$\lim_{l\to\infty} p_{ij}^{l} = 0$$

A state j has period $t > 1$ if $p_{jj}^{l} > 0$ only for $l = t,\ 2t,\ 3t, \ldots$. If j is persistent and periodic, then it can be shown[8] that

$$\lim_{l\to\infty} p_{ij}^{l} = \frac{\sum_{l=1}^{\infty} f_{ij}^{l}}{\mu_j}$$

where $\sum_{l=1}^{\infty} f_{ij}^{l}$ is the probability of ever reaching state j from state i. It can also be shown that

$$\lim_{l\to\infty} p_{ii}^{l} = \frac{1}{\mu_i}$$

For example, for the null state, $\mu_i = \infty$. Now if a state j is persistent and periodic with period t, then

$$\lim_{l\to\infty} p_{ij}^{lt} = \frac{t}{\mu_j}$$

Persistent states that are neither periodic states nor null states are called *ergodic* states. Our interest is in states of this sort.

Armed with these definitions, we can state two important theorems regarding Markov chains. These theorems give the key to finding the steady-state distribution of a Markov chain.

Theorem 2.1 In an irreducible Markov chain all states belong to the same class. They are all transient, all persistent null states, or all persistent nonnull states. In all cases, they have the same period. Moreover, every state can be reached from every other state.

In every chain the persistent states can be divided uniquely into closed sets. From any state of one of these sets, all other states in that set can be reached and no state outside

[8] The proofs of the following two equations are beyond the scope of this text. Proof can be found in the references listed at the end of the chapter.

the set can be reached. In addition to the closed sets, the chain may contain transient states in general. From this transient state, a closed set of states may be reached.

Examples. In the example of Figure 2.14 state 3 is a set of closed states. State 1 is a transient state. This can be seen from the fact that the series f_{ii}^{j}; $j = 1, 2, \ldots$ is dominated by 2^{-j} and $\sum_{i=1}^{\infty} 2^{-i} = 1$; consequently, $\sum_{j=1}^{\infty} f_{ii}^{j} < 1$. In the example of Figure 2.15, all the states are persistent.

Corollary. In a Markov chain with finitely many states there exist no null states, and not all states may be transient states.

Theorem 2.2 An irreducible aperiodic Markov chain belongs to one of the following two classes:

(a) The states are either all transient, or all null states; in this case $\lim_{l\to\infty} p_{ij}^{l} = 0$ for all i, j and a steady-state distribution does not exist.

(b) All states are ergodic:

$$\lim_{l\to\infty} p_{ij}^{l} = \frac{1}{\mu_j}$$

In this latter case of all ergodic states, $1/\mu_j$ is the unique steady-state (stationary) distribution.

A Markov chain has reached steady state when the probability distribution of the states does not change from step to step; accordingly, the steady-state distribution is independent of the initial condition. Let the steady-state distribution be, $\boldsymbol{\pi} = [\pi_1, \pi_2, \ldots, \pi_M]$, for stationarity, we must have [see (2.98)]

$$\pi_k = \sum_{j=1}^{M} \pi_j p_{jk} \tag{2.108}$$

or in vector terms

$$\boldsymbol{\pi} = \boldsymbol{\pi} P \tag{2.109}$$

In order to find the steady-state distribution given the state transition matrix, we solve Equation (2.108) together with the normalizing condition

$$\sum_{j=1}^{M} \pi_j = 1 \tag{2.110}$$

We recognize that $\boldsymbol{\pi}$ is the eigenvector of the matrix P with eigenvalue 1.

Example 2.18 The steady-state probability for the Markov chain depicted in Figure 2.16 is worked out on the accompanying spreadsheet. The result is $\boldsymbol{\pi} = [0.358566, 0.223108, 243028, 0.175299]$.

Example 2.19 Figure 2.16 shows two transmission lines, which feed into a multiplexer. Data in the form of fixed-length packets arrive in slots, which are synchronized among the input lines. The slots contain a packet with probability p, independently from one line to the other. Assume that the lines are synchronized so that there can be simultaneous packet arrival if both lines have messages in the current slot. The multiplexer transmits one packet in a slot time and has the capability of storing two packets. If there is no room for a packet in the buffer, it is simply lost.

In any slot time there are three possible arrival possibilities:

$$P(j \text{ cells arrive}) = \begin{cases} (1-p)^2; & j = 0 \\ 2p(1-p); & j = 1 \\ p^2; & j = 2 \end{cases}$$

We assume that the multiplexer is able to transmit one cell, before storing an incoming cell.

The number of packets in the buffer as a function of time is described by a Markov chain whose state transition diagram is given in Figure 2.17. The state transition matrix is given by

$$P = \begin{bmatrix} (1-p)^2 & 2p(1-p) & p^2 \\ (1-p)^2 & 2p(1-p) & p^2 \\ 0 & (1-p)^2 & 1-(1-p)^2 \end{bmatrix}$$

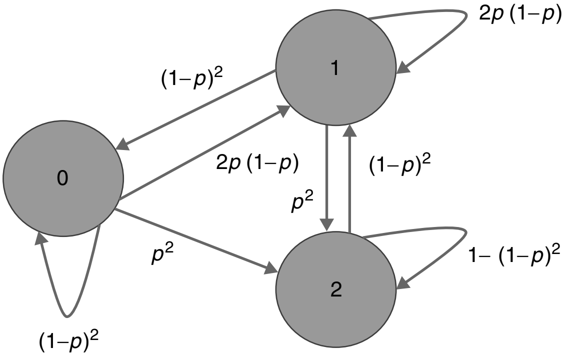


Figure 2.17 State transition diagram: multiplexer.

The average arrival rate of cells is simply *2p* per slot. The steady-state probabilities for this system are worked out on the linked accompanying spreadsheet. When $p = 0.3$, the steady-state probabilities for states 0, 1, and 2 are 0.413966, 0.430862, and 0.155172, respectively. A packet is lost when there are two cells arriving to a buffer with two cells. The probability of this event is $P(\text{two arrivals}) \times P(\text{state } 2) = 2p(1-p)P(\text{state } 2) = 0.013965517$ for $p = 0.3$. The remaining packets get through the multiplexer; accordingly, the throughput is 0.295810345.

For an application of Markov chain modeling, that evaluates the throughput of a wireless LAN with a single access point and number of terminals which takes into account of collisions at the MAC layer, transmission errors and the transport protocol mechanism, see H. Pan 2003.

2.8 RANDOM PROCESSES

2.8.1 Definition: Ensemble of Functions

Processes, such as arrivals and departures, that we study in the text are functions of time, which exhibit random properties. This kind of function can be modeled as a *random process*, which can be defined in a fashion similar to the definition of a random variable. Suppose that a function of time is associated with each outcome of a random experiment. The ensemble (or collection) of these functions is a random process. A simple example of a random process involves an experiment with a wheel, with marked angles, spun relative to a constant point. The outcome of the experiment is an angle, θ, which is uniformly distributed between $-\pi$ and π. This angle is mapped into the phase of a sinusoid with a specified carried frequency, f_0. The ensemble of functions

$$X_t(\omega) = \cos(2\pi f_0 t + \theta(\omega)); \quad -\infty < t < \theta, \quad -\pi \leq \theta(\omega) \leq \pi \qquad (2.111)$$

forms the random process. The dependence on the experimental outcome is indicated by the ω in (2.111). In the rest of this section, we suppress this variable to conform to the usual practice. It should be realized that the idea of an underlying experiment is a mathematical construct, which allows one to deploy the concepts of probability, random variables, probability distributions, and similar concepts. In many examples of random processes, the underlying experiment would be neither so simple nor so direct.

2.8.2 Stationarity and Ergodicity

Time samples of a random process are themselves random variables since they represent mapping from a space of an experimental outcome onto the real line. Moments of the samples can be calculated. For the random process defined in

(2.111), the first two moments are

$$E(X_t) = \frac{1}{2\pi}\int_{-\pi}^{\pi} \cos(2\pi f_0 t + \theta)d\theta = \frac{1}{2\pi}\sin(2\pi f_0 t + \theta)\big|_{-\pi}^{\pi} = 0$$

$$E(X_t^2) = \frac{1}{2\pi}\int_{-\pi}^{\pi} \cos^2(2\pi f_0 t + \theta)d\theta = \frac{1}{2\pi}\left[\frac{\theta}{2} - \frac{1}{4}\sin(4\pi f_0 t + 2\theta)\big|_{-\pi}^{\pi}\right] = \frac{1}{2}$$

Samples of the random process, which are taken at different times, are correlated random variables, in general. The *autocorrelation function* is a particularly important measure of this correlation. This function is defined as

$$R(t_1, t_2) = E(X_{t_1}X_{t_2}) \tag{2.112}$$

where $X(t_1)$ and $X(t_2)$ indicate samples taken at times t_1 and t_2, respectively. Similarly, the autocovariance function is defined as

$$V(t_1, t_2) = \frac{E((X_{t_1} - E(X_{t_1}))(X_{t_2} - E(X_{t_2})))}{E((X_t - E(X_t))^2)} = \frac{E(X_{t_1}X_{t_2}) - E(X_{t_1})E(X_{t_2})}{E((X_t - E(X_t))^2)} \tag{2.113}$$

For the sinusoidal process defined in (2.111), we have

$$\begin{aligned} R(t_1, t_2) &= \frac{1}{2\pi}\int_{-\pi}^{\pi} \cos(2\pi f_0 t_1 + \theta)\cos(2\pi f_0 t_2 + \theta)d\theta \\ &= \frac{1}{2\pi}\int_{-\pi}^{\pi}\left[\frac{1}{2}\cos(2\pi f_0(t_1 - t_2)) + \frac{1}{2}\cos(2\pi f_0(t_1 + t_2) + 2\theta)\right]d\theta \\ &= \frac{1}{2}\cos(2\pi f_0(t_1 - t_2)) \end{aligned} \tag{2.114}$$

Since the mean of the process is equal to 0, the covariance function is $V(t_1, t_2) = 2R(t_1, t_2)$.

We notice that the mean and the mean-square moments of the sinusoidal process are constant with time. Further, correlation between samples is a function of time difference and is not based on the absolute time. Processes with these properties are called *wide-sense stationary.* If this time invariance were true for the whole distribution of the process, and hence, all of its moments, we would have a *strict-sense stationary process.* For a wide-sense stationary process, we may write the correlation function and the autocovariance function in terms of a single variable.

$$R(t, t - \tau) = R(\tau)$$
$$V(t, t - \tau) = V(\tau)$$

Generally speaking, we may observe a single realization of a random process. A reasonable question would be "Is the realization typical of the other realizations in the ensemble?" We may, for example, compute time averages on the realization of the process that is available, $x(t)$. The mean and the mean-square time averages are, respectively given by

$$\begin{aligned}\overline{X_T} &= \lim_{T\to\infty}\frac{1}{T}\int_0^T x(t)dt \\ \overline{X_T^2} &= \lim_{T\to\infty}\frac{1}{T}\int_0^T x^2(t)dt\end{aligned} \tag{2.115}$$

Higher moments are computed in a similar fashion. Correlations between different sample points can be computed as

$$\overline{X(t_1)X(t_2)} = \lim_{T\to\infty}\frac{1}{T}\int_0^T x(t-t_1)x(t-t_2)dt \tag{2.116}$$

When a process is what is called *ergodic*, all the time-average moments are equal to the moments computed by the ensemble average. In an approximate way, an ergodic process is one where every realization of the process passes through the same set of states. The importance of the ergodic property lies in the fact that the particular realization of the process that is obtained from an observation is typical. We are assured that the time average will not vary from realization to realization.

2.8.3 Markov Processes

In the previous section, we specified Markov processes as having transition between states at points spaced a constant T seconds apart. The semi-Markov process is a generalization in which the interval between these transitions is a random variable. In general, the value of an interval is dependent, in a probabilistic sense, on the beginning and end states of the interval. A semi-Markov process can be described by the values of the successive intervals between state or level changes and the sequence of levels themselves. Now, if we focus on the levels alone, we are looking at a Markov chain *imbedded* in the semi-Markov process.[9]

If the interval between level transitions is exponentially distributed, the semi-Markov process becomes the *Markov process.* The term *Markov* is really synonymous with *memoryless*. The mean of the interval may depend on the current level; however, the future evolution of the process depends only on the current state or level. A particular class of Markov process is the *birth and death process*, where the state is the number of members in a population and the state changes by one unit at a time.

[9] As we shall see in the remainder of the text, the imbedded Markov chain is a powerful tool in the analysis of queuing systems.

REFERENCES

Cox, R., and H. D. Miller, *The Theory of Stochastic Processes*, Methuen, New York, 1965.

Feller, W., *An Introduction to Probability and Its Applications*, Vol. 1, Wiley, New York, 1957.

Gradshteyn, I. S., and I. M. Ryzhik, *Table of Integral, Series and Products*, Academic Press, New York, 1965.

Karlin, S., and H. M. Taylor, *A First Course in Stochastic Processes*, Academic Press, New York, 1975.

Kleinrock, L., *Queueing Systems*, Vol. 1: *Theory*, Wiley, New York, 1975.

Kobayashi, H., *Modeling and Analysis, An Introduction to System Performance Evaluation Methodology*, Addison-Wesley, Reading, MA, 1978.

Leon-Garcia, A., *Probability and Random Processes for Electrical Engineering*, Addison-Wesley, Reading, MA, 1994.

Nelson, R., *Probability, Stochastic Processes and Queueing Theory*, Springer-Verlag, New York, 1995.

Pan, H. et al., "On the throughput of an IEEE 802.11a Wireless LAN system with Terminals under Heterogeneous Radio conditions," 18th International Teletraffic Congress, Vol. 5b, Berlin, Germany, 31 August–5 September 2003, pp. 991–1000.

Papoulis, A., and Pillai, S.U., *Probability, Random Variables, and Stochastic Processes*, McGraw-Hill, New York, 2002

Seneta, E., *Non-Negative Matrices*, George Allen & Unwin Ltd., London, 1973.

EXERCISES

2.1 Express the event that contains all outcomes in A or B but not both.

2.2 Using the axioms of probability, prove the relations in (2.1).

2.3 Repeat the calculation in Example 2.1 under the assumption that one can replace cards after they are drawn from the deck.

2.4 Suppose that in a block of B bits, the individual bits are in error independently with probability P_e. The approximation, $P_B \cong BP_e$, is often made for the probability of a block being in error. Show the conditions under which this approximation is justified.

Several of the following exercises concern the *ARQ* protocol. In this method of error control, messages are accompanied by redundant bits, which allow transmission errors to be detected at the receiver. If an error is detected in a message, the transmitter is informed and the message is retransmitted.

2.5 In a block of 1024 bits using error-correcting codes, single-bit errors can be corrected; all other error patterns can be detected.

(a) If $P_e = 10^{-4}$ and bit errors are independent, what is the probability of retransmission?

(b) Plot the probability of block retransmission as a function of bit error probability.

(c) Now suppose that up to two errors can be corrected; repeat (b) on the same chart.

2.6 In many protocols, repeated failures result in remedial action, which may mean kicking the process to a higher level in the protocol hierarchy. Under the assumptions of Exercise 2.5(a), what is the probability that there are

(a) Exactly two retransmissions

(b) More than two retransmissions?

2.7 Suppose that a channel is supposed to have a probability of a message being in error $P_e = 10^{-3}$.

(a) If no errors are corrected, what is the probability that three retransmissions are necessary, under these assumptions?

(b) Now, supposing that there actually are three retransmissions, what would you say about the assumptions? (This sort of reasoning comes up again in Chapter 9.)

2.8 As you recall, in an additive white Gaussian noise channel, the probability of error for binary signals is given by

$$P_e = \int_{\gamma}^{\infty} \frac{\exp(-x^2/2)}{\sqrt{2\pi}} dx$$

where $\gamma = 2\text{SNR}$ for antipodal signals and $\gamma = \text{SNR}$ for orthogonal signals. Plot probability of error as a function of the signal-to-noise ratio SNR—both curves on the same sheet, please.

2.9 A wireless channel may be modeled as being in one of two states, good and bad, with probabilities 0.95 and 0.05, respectively. When in the bad state, the probability of a bit being in error is 0.01 and in the good, 0.00001. If an error is observed, what is the probability that the channel is in the bad state?

2.10 Make the same assumptions as Exercise 2.9.

(a) If there is one or more retransmissions under an ARQ protocol, what is the probability that the channel is in the bad state?

(b) As a test of the sensitivity of the result in (a), plot the a posteriori probability of being in the bad state as a function of the a priori probability of being in the bad state.

2.11 Suppose that, because of fading, a wireless channel can be one of two states, good and bad, with the respective probabilities of bit error $P_e = 10^{-4}$ and $P_e = 5 \times 10^{-4}$. The a priori probability of the channel being in the bad state is assumed to be 0.1. Assume that one bit error can be corrected. If errors are detected, what is the probability that the channel is in the bad state?

2.12 For *direct-sequence* optical transmission, bits are conveyed by the presence or absence of a pulse. Further, while the pulse is being transmitted, photons arrive at the receiver at a Poisson rate whose average is determined by the pulse power. Suppose that an optical pulse is detected if the receiver counts 100 photons in an interval.

(a) If the average rate of photon reception in 1 ns is 150, what is the input power into the detector? (Note that 1 MW $= 7.5 \times 10^{15}$ photons/s.)

(b) Under the assumptions of part (a) what is the probability that a positive pulse will not be detected?

(c) What power is required for a probability of error of $P_e = 10^{-8}$ at a Tbps rate? (*Suggestion*: Cut and try on a spreadsheet.)

2.13 Suppose that 10 terminals randomly access a synchronous TDMA line. In each slot a terminal tried to transmit its message with probability 0.1. Since there is no coordination among the terminals, the messages may collide, in which case the terminal retried in the next slot.

(a) What is the probability of collision?

(b) What is the probability that a message gets through in a slot?

(c) What is the average number of slots required for a message from a particular terminal to get transmitted successfully?

2.14 Messages arrive to a switch at a Poisson rate with an average of 10 messages in a 3-min interval. Calculate the probability of less than 175 messages in a busy hour. Do this two ways: the exact computation and the approximation.

2.15 A network node is fed by three input lines. The rates for lines 1, 2 and 3 are, respectively 15, 10, and 25 messages per minute.

(a) What is the probability that more than 5000 messages arrive in an hour?

(b) What is the probability that the traffic for line 3 exceeds the others? (This is a hard one.)

2.16 Find the Chernoff bound for a binomially distributed random variable.

2.17 Consider two independent binomially distributed random variables with parameters (M, P) and (N, P), respectively. Show that their sum is binomial by

(a) Convolution

(b) Probability-generating function

2.18 Using characteristic functions, show that the sum of independent Gaussian random variables is Gaussian. The Gaussian distribution is the only continuous distribution with this property.

2.19 Extend Example 2.19 to the case where three lines enter the multiplexer.

2.20 Continue Exercise 2.19 to the case where the buffer can store three packets.

2.21 Show that the process given in (2.111) is ergodic in the second order since computing (2.115) and (2.116) give the same result as the corresponding ensemble averages.

3

APPLICATION OF BIRTH AND DEATH PROCESSES TO QUEUEING THEORY

3.1 ELEMENTS OF THE QUEUEING MODEL

We now turn to consider the basic elements of queueing models. In order to describe these elements, a convenient notation has been developed. In its simplest form it is written $A/R/S$, where A designates the arrival process, R the service required by an arriving customer, and S the number of servers. For example, $M/M/1$ indicates a Poisson arrival of customers, an exponentially distributed service discipline, and a single server. The M may be taken to stand for memoryless (or Markovian) in both the arrival process and the service time. Since there is no indication, it may be assumed that there is no restriction on the available storage. In Chapter 6, we shall study the $M/G/1$ queue in which the service time follows a general distribution. A further embellishment on this notation indicating the size of the waiting room is denoted by an additional space. Thus, for example, $M/M/S/N$ indicates a system where there is Poisson arrival, exponential service, S servers, and room for N customers including those in service. We are particularly interested in the case where $S = N$ which is the subject of the famous *Erlang B* formula, which will be derived in the present chapter.

There are two equivalent ways to describe the arrival process in a queueing system: (1) the probability distribution for the number of arrivals in an interval of time and (2) the probability distribution for the interval between arrivals. As we shall see in the next section, the Poisson process can be described as having arrivals that follow the Poisson distribution in an interval or whose inter–arrival times are exponentially distributed.

Modeling and Analysis of Telecommunications Networks, by Jeremiah F. Hayes and
Thimma V. J. Ganesh Babu
ISBN 0-471-34845-7 © 2004 John Wiley & Sons, Inc.

In the present chapter, we will consider only the simple cases where the arrival process is Poisson with a constant rate. However, there are far more complex models, which can be used to approximate reality. For example, in Chapter 7 we study the *fluid flow model*, in which data arrive at a constant rate over a given time interval. This rate can fluctuate according to an underlying Markov chain. One of the virtues of the fluid flow model is that it captures the continuous flow of information on transmission lines in a more realistic fashion than does the Poisson model. A potential problem with the Poisson model is that messages arrive instantaneously. In Chapter 8 we study another widely used traffic source model, *Markov modulated Poisson process* (MMPP), in which an underlying Markov chain controls the average arrival rate.

Example 3.1 A simple arrival model has arrivals in 10 equal-length slots over a fixed length frame. Suppose that the probability of an arrival in a slot is 0.2, independently from slot to slot. The probability of three arrivals in a frame is governed by the binomial distribution and has a probability 0.201327.

In telecommunications systems there are two basically different forms for the service time. In the traditional circuit-switched network, the service time is simply the duration of a call. This same idea could be appropriate in a computer system, where a component is used exclusively by a user for a period of time. In packet-switched networks, the service time is the time required to transmit information. The generic term that we will use to describe the encapsulation of information is the *message*. The quantity of information in a message may be simply the number of bits. The term *packet* indicates the encapsulation of information in blocks to which are appended overhead bits for such functions as addressing and error control. In the asynchronous transfer mode (ATM) system, the packet is called a *cell* consisting of 384 information bits and 40 overhead bits. One of the important functions of the ATM system is to convert the variable-length messages into fixed-length cells for transmission over the network.

Example 3.2 Consider the ATM system in which messages are broken up into cells containing 384 information bits plus 40 overhead bits. For example, if a message has 6374 bits, it is represented by 17 cells. Bits are stuffed to the cell to round it out to a multiple of 384. Accompanying each cell are 40 overhead bits. The total number of bits that are needed to represent the message in the cell format is then 7208. Suppose that these messages are transmitted over a line at a rate of 1.344 Mbps, the time required to transmit this message is 5.363 ms.

Now suppose that the number of bits in a message may be assumed to be exponentially distributed with mean $1/\mu$. The probability that a message is represented by k cells is given by

$$P(k \text{ cells}) \cong \int_{384(k-1)}^{384k} \mu \exp(-\mu t)dt = e^{-384\mu(k-1)}(1 - e^{-384\mu}); \quad k = 1, 2, \ldots$$

Of course this is the geometric distribution (type 2) with parameter $P = 1 - e^{-384\mu}$ (see Table 2.1). Now suppose that $1/\mu = 10^4$. The mean number of cells in the message is 26.54487. At a rate of 1.344 Mbps, the average time required to transmit a message is 8.374 ms.

The final component of the queueing model refers to the number of servers. Beyond this there is the order of service, which constitutes the *service discipline*. This service discipline can be, for example, *first-come first-served* (FCFS), *last-come first-served* (LCFS), or even *service in random order*. Moreover, if the service time is exponentially distributed (memoryless), a server can interrupt service on one customer before completion to serve another without changing the dynamics. Such is the case in *round-robin* service, where the server cycles through all members of the queue, giving each at most a fixed amount of service. As this increment of service decreases, we approach a service discipline where the customer with the least required service is served first. The key in all of these cases is that the server is never idle while the queue is not empty.

3.2 LITTLE'S FORMULA

3.2.1 A Heuristic

We now turn to a relationship among average quantities that is widely applicable in queueing systems. Let $\bar{N}$ be the average number of messages in the system, $\bar{D}$ the average delay, and λ is the average arrival rate. *Little's formula* states that

$$\bar{N} = \lambda \times \bar{D} \tag{3.1}$$

This formula is eminently reasonable, as the following heuristic argument shows (see Fig. 3.1). Consider a system in which messages are served in order of arrival. Assume that departing messages have spent an average amount of time $\bar{D}$ seconds in the system. New customers arrive at an average rate of λ message per second; consequently, a departing message sees an average of $\bar{N}$ messages remaining in the

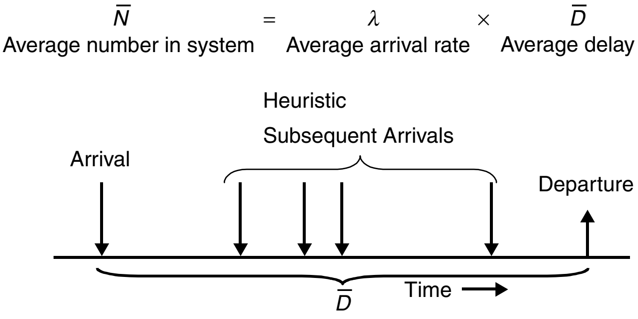


Figure 3.1 Little's formula.

system. Although this heuristic assumes service in order of arrival, Little's formula holds for any service discipline.

3.2.2 Graphical Proof

The following graphical proof shows that Little's formula holds for a wide range of conditions. Indeed, it would be a challenge to find conditions under which it does not hold. In Figure 3.2 the arrival of customers to and departures from an initially empty system in the interval (0, T) is shown. We assume, for the moment, a first-come first-served (FCFS) discipline. The arrival time of the ith customer is denoted as a_i; $i = 1, 2, \ldots, n$, and the departure time of the ith arriving customer is l_i; $i = 1, 2, \ldots, n$. The graph $A(t)$ represents the cumulative arrival process; with each arrival the graph $A(t)$ increments by one. Notice that $A(T)$ is the total number of arrivals in (0, T). (There are a total of $n = 8$ arrivals in Fig. 3.2.) Similarly, the graph $L(t)$ indicates the number of customers that have left the system up to time t. The quantity $N(t) = A(t) - L(t)$ in (0, t) is the number of customers in the system at time t.

Under the FCFS assumption, the time intervals shown as d_i in Figure 3.2 represent the intervals between the arrival of the ith message and its departure. The final time interval is simply the interval between the arrival of the last arriving message and the end of the interval. As we increase the value of T, the beginning and end intervals become insignificant. The quantity $\Delta(T)$ is the total area between $A(t)$ and $L(t)$. Since all of the steps are of height 1, we have

$$\Delta(T) = d_1 + d_2 + \cdots + d_n \tag{3.2}$$

Averaging over the whole interval, the average number of customers in the system during the interval (0, T), is given by

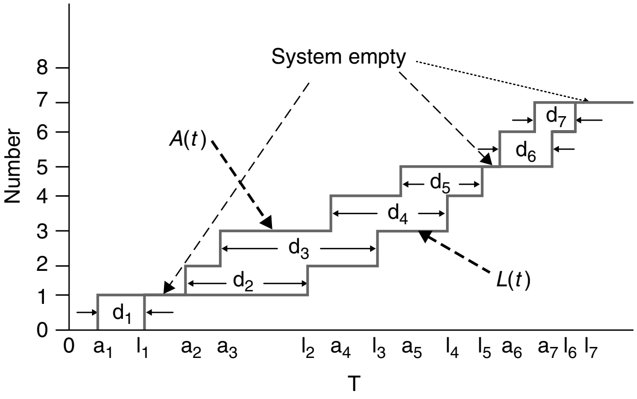


Figure 3.2 Arrivals and departures—FCFS.

$$\overline{N(T)} = \frac{\Delta(T)}{T} \tag{3.3}$$

Defining $\lambda(T)$ as the average arrival rate of customers in the interval (0, T), clearly

$$\lambda(T) = \frac{A(T)}{T} \tag{3.4}$$

From (3.2) we have, for large n, that the average waiting time of a message in the system is given by

$$\overline{D(T)} = \frac{\Delta(T)}{A(T)} \tag{3.5}$$

Equations (3.3)–(3.5) can be combined to form

$$\overline{N(T)} = \frac{\Delta(T)}{A(T)}\frac{A(T)}{T} = \overline{D(T)}\lambda(T) \tag{3.6}$$

Now, we let $T \to \infty$. Assuming that limits exist and are given by

$$\begin{aligned} \lim_{T\to\infty} \lambda(T) &= \lambda \\ \lim_{T\to\infty} \overline{D(T)} &= \bar{D} \end{aligned} \tag{3.7}$$

the limit for $\bar{N}(T)$ exists and is given by

$$\lim_{T\to\infty} N(T) = \bar{N} = \lambda\bar{D} \tag{3.8}$$

and we have a proof.

It is not difficult to show that the proof applies for any order of service. We begin by rewriting (3.2):

$$\Delta(T) = \sum_{i=1}^{n} d_i = \sum_{i=1}^{n} (l_i - a_i) = \sum_{i=1}^{n} l_i - \sum_{i=1}^{n} a_i \tag{3.9}$$

Now define $l_{(i)}$; $i = 1, 2, \ldots, n$, as the ith departure, not necessarily the departure of the ith arrival. To illustrate this, for the LCFS discipline, we have the sequence of departures $l_{(1)} = l_1$, $l_{(2)} = l_2$, $l_{(3)} = l_4$, $l_{(4)} = l_5$, $l_{(5)} = l_2$, $l_{(6)} = l_6$, $l_{(7)} = l_7$ (see Fig. 3.3). Notice that arrivals 3, 4, and 5 depart before arrival 2 since 2 was still in the system when 3, 4, and 5 arrived.

Returning to the general case, it is clear that the area between $A(t)$ and $L(t)$ is given by

$$\Delta(T) = \sum_{i=1}^{n} d_{(i)} = \sum_{i=1}^{n} (l_{(i)} - a_i) = \sum_{i=1}^{n} l_{(i)} - \sum_{i=1}^{n} a_i \tag{3.10}$$

However, each departure time $l_{(i)}$; $i = 1, 2, \ldots, n$ can be paired with a single departure from l_i; $i = 1, 2, \ldots, n$; consequently, we have

$$\sum_{i=1}^{n} l_{(i)} = \sum_{i=1}^{n} l_i \tag{3.11}$$

From (3.2), the area between $A(T)$ and $L(T)$ is the same as in the FCFS case and the proof goes through in the same fashion.

Example 3.3 Three classes of messages arrive at a service facility. Assume that the loads are such that no queueing is necessary. Class 1 messages arrive at an average rate of 120 per minute and depart after a quick examination lasting 200 ms. The second class is processed at an average rate of 6 messages per minute and arrive at a rate of 20 per minute. Each message in the third class requires 30 s of processing, and third-class messages arrive at a rate of 10 per minute. The average service times of each class in seconds are 0.2, 10, and 30, respectively. Since there is no storage and a server is always available, these are the average delays of each class in the system. Applying Little's formula to each class separately, we find $\bar{N}_1 = (0.2)(120/6) = 0.4$, $\bar{N}_2 = (10)(20/60) = 3.33$, $\bar{N}_3 = (30)(10/60) = 5.0$. The total number of messages is $\bar{N} = 8.73$ messages.

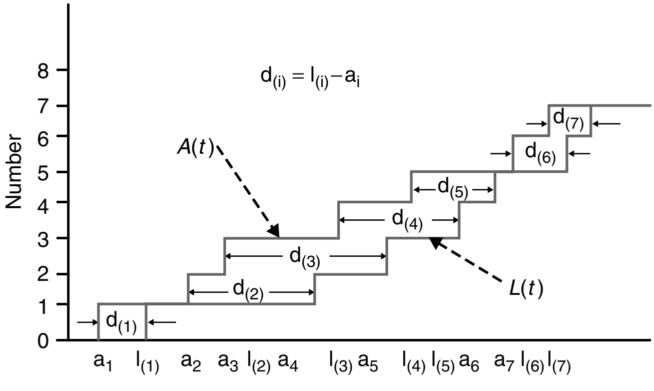


Figure 3.3 Arrivals and departures—arbitrary discipline.

Example 3.4 In a particular application, messages contain a geometrically distributed number of cells according to $P(k \text{ cells in a message}) = 0.4 \times (0.6)^{k-1}$; $k = 1, 2, \ldots$. From Table 2.1 the average number of cells in a message is $1/P = 2.5$. These messages arrive to a processor in which the time required to process a cell is 6.4 ms; consequently, the average processing time for a message is 16 ms. On the associated spreadsheet the relationship between the arrival rate and the average number of messages in the system is plotted. For example, if messages arrive at an average rate of 100/s, the number of messages in the system is then 1.6.

3.2.3 Basic Relationship for the Single-Server Queue

In order to illustrate the wide applicability of Little's formula, we use it to derive a basic relationship for the single-server queue. Consider the system shown in Figure 3.4. We consider the "system" to be the server alone; arrivals to the system cross the boundary A-A′ at rate λ. We disregard any queueing that may have taken place before the boundary. The server is either empty, with probability P_0, or full with one customer, with probability $1 - P_0$. The average number of customers in the system is then

$$\bar{N} = 1 - P_0 \tag{3.12}$$

The average time that a customer resides in the server is simply the average service time $\bar{M}$. From the application of Little's formula and (3.12), we have

$$\lambda\bar{M} = \bar{N} \tag{3.13}$$

Since $\bar{M}$ is the average service time, $\rho = \lambda\bar{M}$, and we have

$$\rho = 1 - P_0 \tag{3.14}$$

We will see this result later for specific systems; however in deriving (3.14) no assumptions were made regarding the probability distributions of the service time,

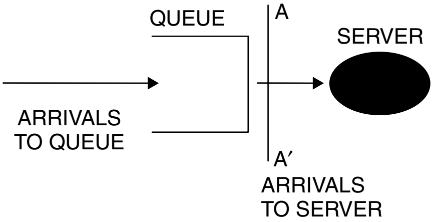


Figure 3.4 Arrivals to a single server.

the arrival rate, or the service discipline. All that we require are the averages of these quantities.

3.3 THE POISSON PROCESS

3.3.1 Basic Properties

Measurements of traffic in voice and in data systems have shown that, in a wide range of applications, call and message generation can be modeled as a Poisson process. In this instance nature is kind to the system analyst since the Poisson process is particularly tractable from a mathematical point of view. We shall examine the Poisson arrival process in some detail focusing on its unique properties, which are important in this chapter and throughout the text. As we shall see, combining and splitting processes results in Poisson processes. The connection between the Poisson process and the exponential distribution is shown. Finally, we show that the Poisson arrival process is a special case of the pure birth process. This leads directly to the consideration of birth and death processes, which model certain queueing systems in which customers having exponentially distributed service requirements arrive at a service facility at a Poisson rate.

We define the *Poisson process* to be a point process for which the number of events (successes) in a t-second interval is given by the Poisson distribution

$$P_k(t) = P(k \text{ successes in } t \text{ seconds}) = \frac{e^{-\lambda t}(\lambda t)^k}{k!}; \quad k = 0, 1, 2, \ldots \tag{3.15}$$

where the average rate of successes in the t-second interval is taken to be λ per second. In the queueing context the arrival of a message is a success.

As we have seen in Section 2.2.2, the Poisson process can be viewed as a limiting case of Bernoulli trials; accordingly, properties of the Poisson process may be inferred from the Bernoulli trials. The first of these is *memorylessness*. As we have seen in Section 2.7.1, the cumulative number of successes in Bernoulli trials form a Markov chain. It is entirely reasonable that the Markov property carries through. Letting X_t be the accumulated number of arrivals in $(0, t)$ and $t_1 \leq t_2 \leq \cdots \leq t_N \leq t_{N+1}$, we may write

$$\begin{aligned} P[X_{t_{N+1}} &= k_{t_{N+1}}/X_{t_1} = k_{t_1}, X_{t_2} = k_{t_2}, \ldots, X_{t_N} = k_{t_N}] \\ &= P[X_{t_{N+1}} = k_{t_{N+1}}/X_{t_N} = k_{t_N}] \end{aligned}$$

This simply states that the cumulative number of arrivals in an interval depends only on the number in the preceding interval.

What we call the *uniformity* property means that, given the number of arrivals in an interval, the arrivals are uniformly distributed throughout the interval. This property may be seen by an example based on Bernoulli trials. Consider the outcomes of five Bernoulli trials. Suppose that the number of successes is given as two. The set of outcomes with two successes are (SSFFF), (SFSFF), (SFFSF), (SFFFS), (FSSFF), (FSFSF), (FSFFS), (FFSSF), (FFSFS), (FFFSS). Observe that the number of successes is uniformly distributed throughout the five trials, so that the probability of a success on any trial is $2/5$. The same property can be inferred for the Poisson process.

Finally, we have *reversibility*; if we view the process in either the forward or the reverse direction, the same patterns are seen with the same probabilities. Again the property is obvious for the Bernoulli process and carries through to the Poissson process. This property will come into play when we deal with the reversibility of processes in Chapter 4.

3.3.2 Alternative Characterizations of the Poisson Process

Let us now consider an alternative characterization of the Poisson processes, in terms of the time interval between arrivals (refer to Fig. 3.5). Let T denote the random time of the first arrival after some initial time set arbitrarily at $t = 0$. The event that the time of first arrival in the interval $(0, t)$ means that there is at least one arrival in this interval.

We have from (3.15)

$$F_T(t) = P[T \leq t] = 1 - P[\text{no arrival in } (0, t)] = 1 - P_0(t) = 1 - e^{-\lambda t} \tag{3.16}$$

The probability density function for an arrival is

$$f_T(t)dt = P[t < T \leq t + dt] = \lambda e^{-\lambda t}dt; \quad t \geq 0 \tag{3.17}$$

In (3.16) and (3.17), we have an exponentially distributed random variable. Its mean value is

$$E(T) = \int_0^\infty t f_T(t)dt = \frac{1}{\lambda} \tag{3.18}$$

The Laplace transform of the density function is

$$L(\lambda e^{-\lambda t}) = \int_0^\infty e^{-st} f_T(t)dt = \frac{\lambda}{s + \lambda} \tag{3.19}$$

The relation between the Poisson and the exponential distributions is illustrated in Figure 3.6. The Poisson distribution is a discrete random variable, the number of

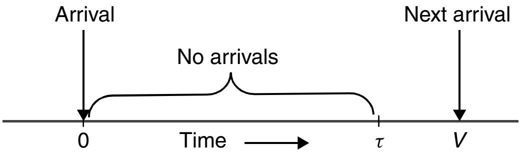


Figure 3.5 Interval between arrivals.

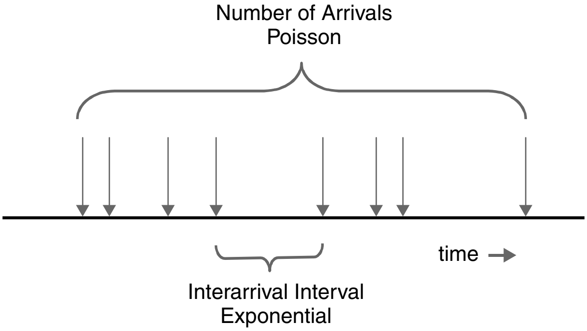


Figure 3.6 Poisson and exponential distributions.

arrivals in a time interval, while the exponentially distributed random variable is a continuous, random variable, the interval between arrivals.[1]

The exponential distribution also has a memoryless property. Suppose that we are given the event that there has been an arrival at time $t = 0$ and none in the interval $(0, t)$; what is the probability of an arrival in the interval $(0, t + \tau)$? This is the probability of an arrival in the interval $(t, t + \tau)$ conditioned on no arrival in the interval $(0, t)$. We designate the time of this first arrival as T. From (3.16) we have

$$\begin{aligned} P[T \leq t + \tau/T \geq t] &= P[t < T \leq t + \tau/T \geq t] = \frac{F_T(t + \tau) - F_T(t)}{1 - F_T(t)} \\ &= \frac{1 - e^{-\lambda(t+\tau)} - 1 + e^{-\lambda t}}{e^{-\lambda t}} = 1 - e^{-\lambda t} \qquad (3.20) \end{aligned}$$

Equation (3.20) reaffirms the memoryless character of the arrival process. The probability of an arrival in an interval depends only on the duration of the interval and not on previous arrivals. The equation also provides an alternate characterization of the Poisson process. The Poisson process is one in which the inter-arrival times are independent and exponentially distributed.

There is yet another fruitful alternative to the preceding definitions of the Poisson process. We consider an incremental time interval, δ, so small that $\lambda\delta \ll 1$ whatever the value of λ. From a Taylor series expansion of (3.15), we have

$$\begin{aligned} P_0(\delta) &= 1 - \lambda\delta + o(\delta) \\ P_1(\delta) &= \lambda\delta + o(\delta) \qquad (3.21) \\ P_i(\delta) &= o(\delta); \quad i \geq 2 \end{aligned}$$

[1] Poisson processes are a member of the class of independent increment processes in which the number of arrivals in different nonoverlapping intervals are independent random variables.

where $o(\delta)$ designate higher-order terms in δ such that $\lim_{\delta\to 0}[o(\delta)/\delta] = 0$. The significance of Equations (3.21) is that for an incremental interval the probability of more than one arrival is negligible and the probability of an arrival is proportional to the duration of the incremental interval, δ.

Although Equations (3.21) are a consequence of the previous definition of the Poisson process, it can be shown that an independent increment process that is stationary and that satisfies these equations is a Poisson process characterized by Equation (3.15). Let $P_n(t)$ denote the probability that there are n arrivals in a time interval t for this process. We examine the change in probability in the incremental interval $(t, t+\delta)$ beginning with $n = 0$. Since the process is independent increment, we may write

$$\begin{aligned} P_0(t+\delta) &= P[\text{no arrivals in } (0, t+\delta)] \\ &= P[\text{no arrivals in } (0, t)] \times P[\text{no arrivals in } (t, t+\delta)] \\ &= P_0(t)(1-\lambda\delta) \end{aligned} \tag{3.22}$$

For $n > 0$, we have two disjoint events (see Fig. 3.2):

$$\begin{aligned} P_n(t+\delta) &= P[n \text{ arrivals in } (0, t+\delta)] \\ &= P[n \text{ arrivals in } (0, t)] \times P[\text{no arrivals in } (t, t+\delta)] \\ &\quad + P[(n-1) \text{ arrivals in } (0, t)] \times P[\text{one arrival in } (t, t+\delta)] \\ &= P_n(t)(1-\lambda\delta) + P_{n-1}(t)\lambda\delta \end{aligned} \tag{3.23}$$

After simple algebra, we have

$$\frac{P_0(t+\delta) - P_0(t)}{\delta} = -\lambda P_0(t)$$

and

$$\frac{P_n(t+\delta) - P_n(t)}{\delta} = -\lambda P_n(t) + \lambda P_{n-1}(t); \quad n > 0$$

Letting $\delta \to 0$, we find the Kolmogorov forward differential equations:

$$\begin{aligned} \frac{dP_0(t)}{dt} &= -\lambda P_0(t) \\ \frac{dP_n(t)}{dt} &= -\lambda P_n(t) + \lambda P_{n-1}(t); \quad n > 0 \end{aligned} \tag{3.24}$$

It is an easy inductive proof to show that the solution to (3.24) is indeed given by (3.15).

Equation (3.24) may be expressed in vector form as

$$\frac{d\mathbf{P}(t)}{dt} = \Lambda \mathbf{P}(t) \tag{3.25}$$

where $\mathbf{P}(t) = (P_0(t), P_1(t), \ldots, P_n(t), \ldots)^T$ and Λ is the *infinitesimal generator matrix*

$$\Lambda = \begin{bmatrix} -\lambda & 0 & 0 & 0 & 0 & \cdots \\ \lambda & -\lambda & 0 & 0 & 0 & \cdots \\ 0 & \lambda & -\lambda & 0 & 0 & \cdots \\ 0 & 0 & \lambda & -\lambda & 0 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}$$

3.3.3 Adding and Splitting Poisson Processes

In the preceding section we have shown that if a process is such that the probability of an arrival over an incremental interval δ is proportional to the duration of the interval (i.e., $\lambda\delta$) and independent from interval to interval, then the process is Poisson. This result can be used to prove certain important properties of the Poisson process. For example, consider the sum of two arrival processes consisting of the total number of arrivals in an interval from either process (see Fig. 3.7). If the two processes are independent and Poisson with average arrival rates λ_1 and λ_2, respectively, then the sum process is Poisson with average arrival rate $\lambda_1 + \lambda_2$. To prove this, consider arrivals in an incremental interval δ. The probability of an arrival from either source is $\lambda_1\delta(1-\lambda_2\delta) + \lambda_2\delta(1-\lambda_1\delta) \cong (\lambda_1+\lambda_2)\delta$. The probability of arrivals from both is $\lambda_1\lambda_2\delta^2$, which is vanishingly small. It is easy to

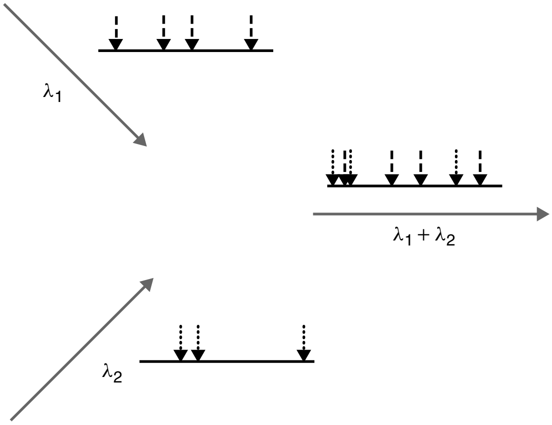


Figure 3.7 Sum of Poisson processes.

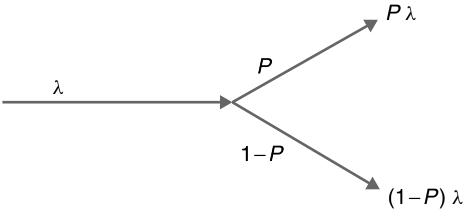


Figure 3.8 Bifurcation of Poisson arrival stream.

show that the conditions of stationary and independent increments hold for the sum process. Thus, the sum process has the same properties as each of the contributing processes, but with the average arrival rate $\lambda_1 + \lambda_2$. This same replication of Poisson processes holds for the sum of any finite number of independent Poisson processes.

This same approach can be applied to the random splitting or bifurcation of a Poisson arrival process. Consider the case depicted in Figure 3.8, where Poisson arrivals are placed in either one of two bins with probabilities P and $1 - P$, respectively. In an incremental interval the event of an arrival to bin 1 is the joint event of an arrival to the system and the selection for bin 1. Since the subevents are independent, the probability of an arrival to bin one is the product $P\lambda\delta$. Again the arrival process is a stationary independent increment process, consequently it is Poisson with average rate $P\lambda$ arrivals/s. Clearly the arrival process to queue 2 is also Poisson but with average arrival rate $(1 - P)\lambda$. The same basic results hold for any independent random splitting of a Poisson arrival process.

Example 3.5 ATM cells arrive to a communications buffer at a Poisson rate of 1000 cells per second. The transmitter activates when there are 10 cells in the buffer. Starting with an empty buffer, what is the probability of a delay of more than 15 ms in activating the transmitter? The probability of a delay of more than 15 ms accumulating 10 messages is just the probability of less than 10 messages arriving in this interval. This is given by the expression

$$P(\text{number of arrivals} < 10 \text{ in } 15\,\text{ms})$$

$$= \sum_{i=0}^{9} \frac{((15 \times 10^{-3})10^3)^i \exp((-15 \times 10^{-3})10^3)}{i!}$$

This has been found in the associated spreadsheet to be 0.069853661.

3.3.4 Pure Birth Processes

The Poisson process can be generalized by allowing the probability of an arrival in an incremental interval to be a function of the number already in the system. The probability of an arrival in an incremental interval is indicated as $\lambda_n\delta$ if there were

n previous arrivals. The general class of processes obtained in this fashion are called *pure birth processes*, indicating an application to population dynamics where the rate of increase is a function of the size of the population.

The set of differential equations governing pure birth processes can be derived in a fashion very similar to those for Poisson arrival processes. In analogy with (3.22) and (3.23), we write

$$\begin{aligned}
P_0(t+\delta) &= P[\text{no arrivals in } (0, t+\delta)] \\
&= P[\text{no arrivals in } (0, t)] \times P[\text{no arrivals in } (t, t+\delta)] = P_0(t)(1-\lambda_0\delta) \\
P_n(t+\delta) &= P[n \text{ arrivals in } (0, t+\delta)] = P[n \text{ arrivals in } (0, t)] \\
&\qquad \times P[\text{no arrivals in } (t, t+\delta)] + P[(n-1) \text{ arrivals in } (0, t)] \\
&\qquad \times P[\text{one arrival in } (t, t+\delta)] \\
&= P_n(t)(1-\lambda_n\delta) + P_{n-1}(t)\lambda_{n-1}\delta
\end{aligned}$$

Letting $\delta \to 0$ and rearranging terms, we find the Kolmogorov forward differential equations

$$\begin{aligned}
\frac{dP_0(t)}{dt} &= -\lambda_0 P_0(t) \\
\frac{dP_n(t)}{dt} &= -\lambda_n P_n(t) + \lambda_{n-1} P_{n-1}(t); \quad n > 0
\end{aligned} \tag{3.26}$$

We also have the normalizing condition

$$\sum_{n=0}^{\infty} P_n(t) = 1$$

These equations may be expressed in vector form as

$$\frac{d\mathbf{P}(t)}{dt} = \Lambda \mathbf{P}(t)$$

where $\mathbf{P}(t) = (P_0(t), P_1(t), \ldots, P_n(t), \ldots)^T$ and Λ is the *infinitesimal generator matrix*

$$\Lambda = \begin{bmatrix}
-\lambda_0 & 0 & 0 & 0 & 0 & \cdots \\
\lambda_0 & -\lambda_1 & 0 & 0 & 0 & \cdots \\
0 & \lambda_1 & -\lambda_2 & 0 & 0 & \cdots \\
0 & 0 & \lambda_2 & -\lambda_3 & 0 & \cdots \\
\vdots & \vdots & \vdots & \vdots & \vdots & \ddots
\end{bmatrix}$$

The diagonal terms of the infinitesimal generator matrix correspond to the flow out of a state in the global balance equation, while the off-diagonal terms in the same column correspond to the flow into a state. Notice that the columns of Λ sum to zero. In later chapters of the text, we shall devote more time to the consideration of the infinitesimal generator matrix.

Example 3.6 We now apply these results to a particular problem. We assume that each member of a population gives birth to a new member in an incremental interval with probability $\lambda\delta$. We then have $\lambda_n = n\lambda$. Equations (3.26) then become

$$\frac{dP_n(t)}{dt} = -n\lambda P_n(t) + (n-1)\lambda P_{n-1}(t); \quad n \geq 0$$

Now suppose that the initial condition is $P_k(0) = 1$ for some $k > 0$. An inductive proof can be used to verify that the solution to this differential equation with this initial condition is

$$P_n(t) = \binom{n-1}{n-k} e^{-n\lambda t}(1 - e^{-\lambda t})^{n-k}; \quad n \geq k,\ t \geq 0$$

This is the Yule–Furry process.

We note that although the probability of a birth increases with the size of the population, the process in Example 3.6 behaves reasonably inasmuch as growth remains stable. This reasonable behavior is not always the case. If the probability of a birth increases quickly enough with the population size, explosive growth takes place; an infinite number of births can take place in a finite time interval. It can be shown that a necessary and sufficient condition for stable growth is for the series $\sum_{n=0}^{\infty} 1/\lambda_n$ to diverge.[2] Clearly, in this case λ_n approaches zero at a rate fast enough to prevent explosive growth.

3.3.5 Poisson Arrivals See Time Averages (PASTA)

There is a property of the Poisson arrival process that plays an important role in the utility of queueing theoretic results in this and subsequent chapters. For the Poisson process, the distribution of customers in this system that is seen by an arrival is typical in the sense that it is a time average seen at randomly chosen points in time. This property is succinctly described by the acronym *PASTA* (Poisson arrivals see time averages).

We illustrate this property by means of a counter-example involving non-Poisson arrivals (see Fig. 3.9). Suppose that the inter–arrival times are 2 s. Each arrival is served for 1 s and departs the system. Clearly, each arrival sees an empty system; however, the system is occupied half the time and would be so observed by random observations.

[2] For a discussion of these processes, see Chapter XVII of Cooper (1972) and Chapter 4 of Feller (1957).

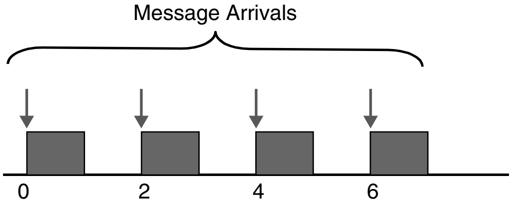


Figure 3.9 PASTA counterexample.

A rigorous proof of PASTA is beyond the scope of this text;[3] however, we shall present the relevant ideas that are involved. First, we recognize that the Poisson arrivals are at randomly chosen points in time. As we have seen, given the arrival of one or more customers in an interval, the arrivals are uniformly distributed throughout the interval. The second relevant concept is that the number of customers in the system has no effect on the arrival process.

3.4 BIRTH AND DEATH PROCESSES: APPLICATION TO QUEUEING

3.4.1 Steady-State Solution

In the previous section the pure birth process, generalized from the Poisson process, served as a model of an arrival process. In an incremental interval at most one arrival was possible. However, in queueing models we are interested in departures from the system as well as arrivals. This can be handled as an extension of the pure birth process. By allowing the number of members of the population to increase or decrease by at most one in an incremental interval, we have a birth and death process. As in the case of the pure birth process we look at the change of state in an incremental time interval $(t, t + \delta)$. Again it is assumed that δ is so small that at most one arrival with probability $\lambda_n \delta$ can occur when the population size is n. The probability of a departure in this same incremental interval is $\mu_n \delta$. The probability of both an arrival and a departure in an incremental interval is of the order of δ^2, and consequently it may be neglected. We follow the same approach as in the derivation of (3.22) and (3.23) beginning with the case of $n = 0$, where we recognize that no departures are possible and we have

$$\begin{aligned}P_0(t+\delta) &= P[0 \text{ in system at time } t] \times P[\text{no arrivals in } (t, t+\delta)] \\ &\quad + P[1 \text{ in system at time } t] \times P[\text{departure in } (t, t+\delta)] \\ &= (1 - \lambda_0 \delta) P_0(t) + \mu_1 \delta P_1(t)\end{aligned}$$

[3] See Section 7.1.5 of Nelson (1995) and Section 5.16 of Wolff (1989).

For the case $n > 0$ we have

$$\begin{aligned}
P_n(t+\delta) = {} & P[n \text{ in system at time } t] \\
& \times [\text{neither arrivals nor departures in}(t, t+\delta)/n \text{ in system}] \\
& + P[n-1 \text{ in system at time } t] \\
& \times P[\text{one arrival in } (t, t+\delta)/n-1 \text{ in system}] \\
& + P[n+1 \text{ in system at time } t] \\
& \times P[\text{one departure in } (t, t+\delta)/n+1 \text{ in system}] \\
= {} & (1 - \lambda_n\delta - \mu_n\delta)P_n(t) + \lambda_{n-1}\delta P_{n-1}(t) + \mu_{n-1}\delta P_{n+1}(t)
\end{aligned}$$

Here we recognize that $1 - \lambda_n\delta - \mu_n\delta \cong (1 - \lambda_n\delta)(1 - \mu_n\delta)$ is the probability of neither an arrival nor a departure. Letting $\delta \to 0$ we have the following set of differential equations:

$$\begin{aligned}
\frac{dP_0(t)}{dt} &= -\lambda_0 P_0(t) + \mu_1 P_1(t) \\
\frac{dP_n(t)}{dt} &= -(\lambda_n + \mu_n)P_n(t) + \lambda_{n-1}P_{n-1}(t) + \mu_{n+1}P_{n+1}(t); \quad n > 0
\end{aligned} \tag{3.27}$$

Equation (3.27) may be expressed in vector form as

$$\frac{d\mathbf{P}(t)}{dt} = M\mathbf{P}(t) \tag{3.28}$$

where $\mathbf{P}(t) = (P_0(t), P_1(t), \ldots, P_n(t), \ldots)^T$ and M is the *infinitesimal generator matrix*:

$$M = \begin{bmatrix}
-\lambda_0 & \mu_1 & 0 & 0 & 0 & \cdots \\
\lambda_0 & -\lambda_1 - \mu_1 & \mu_2 & 0 & 0 & \cdots \\
0 & \lambda_1 & -\lambda_2 - \mu_2 & \mu_3 & 0 & \cdots \\
0 & 0 & \lambda_2 & -\lambda_3 - \mu_3 & \mu_4 & \cdots \\
\vdots & \vdots & \vdots & \vdots & \vdots & \ddots
\end{bmatrix}$$

These differential equations are known as the *Kolmogorov differential equations*. We shall return to the vector form in subsequent chapters.

The solution to the set of Equation (3.28) is not as simple as the case of pure birth equations. Indeed the question of the existence and uniqueness of solutions to these equations is not trivial.[4] However, it can be shown that a positive solution such that $\sum_{n=0}^{\infty} P_n = 1$ always exists as long as $0 \leq \lambda_n < \mu_n$. The case $\sum_{n=0}^{\infty} P_n < 1$ is of theoretical interest; one can draw an analogy with the divergent birth processes

[4] For a discussion of this point, see Chapter XVII of Feller (1957).

discussed earlier. However, for the cases of practical interest solutions may be assumed to be unique and $\sum_{n=0}^{\infty} P_n = 1$. We will give conditions in specific cases.

It is possible to sidestep the difficulties inherent in solving (3.28) by shifting the focus of our attention. A pure birth process is monotonically increasing if $\lambda_n > 0$ for all n since there are no departures. If one waits long enough, the population size will surpass any bound. In this case it is natural to seek a transient solution focusing on such questions as population size at a particular time or the time required to reach a certain level. Now for birth and death processes as applied to queueing, we are interested in cases where the population does not increase without bound. Indeed, one of our concerns is to find conditions for such stability. Accordingly, our interest is not in transient but in steady-state solutions.[5]

In order to find the equilibrium distribution of the population, we let $dP_n/dt = 0$; $n = 0, 1, \ldots$ in (3.27). The implication is that there is no change with time in the probability distribution in (3.27). The resulting set of equations are

$$\begin{aligned} \lambda_0 P_0 &= \mu_1 P_1 \\ (\lambda_n + \mu_n) P_n(t) &= \lambda_{n-1} P_{n-1}(t) + \mu_{n+1} P_{n+1}(t) \end{aligned} \tag{3.29}$$

where P_n; $n = 0, 1, 2, \ldots$ indicates the probability that there are n customers in the system at equilibrium. In (3.29) the time dependence is suppressed. In order to obtain a solution, the normalizing condition is required:

$$\sum_{n=0}^{\infty} P_n = 1$$

Equations (3.29) are called the *global balance* or the *equilibrium equations*. The right-hand side (RHS) is the probability of entering state n, while the left-hand side (LHS) is the probability of leaving state n. Writing and solving the equilibrium equations is a useful approach to analyzing the behavior of systems. The equilibrium equations can be visualized by the state transition flow diagram shown in Figure 3.10, where the flows into and out of states are depicted.

By a simple rearrangement of the terms in Equations (3.29), an interesting alternative view of the process is obtained. We have

$$\begin{aligned} \mu_{n+1} P_{n+1} - \lambda_n P_n &= \mu_n P_n - \lambda_{n-1} P_{n-1}; \quad n = 1, 2, \ldots \\ \mu_1 P_1 - \lambda_0 P_0 &= 0 \end{aligned}$$

The solution to these equations are the *local balance or the detailed balance* equations:

$$\mu_n P_n = \lambda_{n-1} P_{n-1}; \quad n = 1, 2, \ldots \tag{3.30}$$

[5] For a transient solution to the birth–death equation when $\lambda_n = \lambda$ and $\mu_n = \mu$, see Kleinrock (1975), Chapter 2.

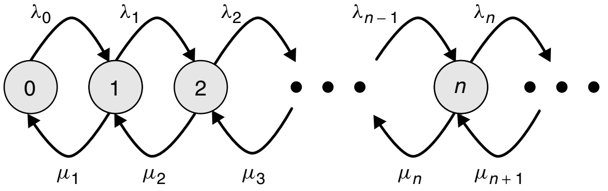


Figure 3.10 State transition flow diagram birth–death process.

Equation (3.30) states that the flow from state n to state $n - 1$, as represented by the RHS, must be balanced by the flow into state n from state $n - 1$. By a simple iteration on (3.30) we find

$$P_n = P_0 \prod_{i=1}^{n} \frac{\lambda_{i-1}}{\mu_i} \tag{3.31}$$

where, from the normalizing condition, we have

$$P_0 = \left[1 + \sum_{n=1}^{\infty} \prod_{i=1}^{n} \frac{\lambda_{i-1}}{\mu_i} \right]^{-1}$$

In the next section specific sets of values of λ_j and μ_j will model the behavior of different queueing models.

3.4.2 Queueing Models

In general the birth and death process models the situation where there is a Poisson arrival of customers at a service facility with each customer requiring an exponentially distributed amount of service. The facility gives each customer the amount of service that it requires. The completion of service to a customer results in the departure of the customer from the system. Customers awaiting service form a queue. In terms of the birth–death model, arrivals and departures correspond to births and deaths, respectively. By choosing different values of λ_j and μ_j, variations on this basic theme are achieved. Since service times are exponentially distributed, when there are j customers in the system, the probability of a departure in the next incremental interval is $\mu_j\delta$ independent of the amount of service previously given to a customer. As we shall see presently, the value of μ_j depends on the particular queueing model under consideration. Similarly, since arrival is Poisson, the probability of an arrival in the next incremental interval is $\lambda_j\delta$ when there are j customers in the system independent of the amount of time the system has been in this state.

If the service time is simply the duration of a voice call, the exponential assumption is a reasonable approximation to reality. For other applications, where the service is the time required to transmit a message, the exponential assumption is more difficult to justify. The implication is that messages have a continuous distribution and can be of any length greater than zero. In both aspects, we have an approximation since messages are in bits and messages should contain overhead information. Nevertheless, studies have shown that the exponential distribution is a reasonable approximation in many applications. As we shall see in Chapter 4, the assumption of exponential message length is useful in dealing with networks of queues. Moreover, as we shall see later under certain conditions, there is an *insensitivity* property whereby the distributions of buffer occupancy depend only on the mean of the service time, but not on its distribution.

The message arrival model in this application also merits comment. The assumption is that messages arrive at the buffer instantaneously. If the input lines are of the same speed or lower than the output lines, this is again an approximation. Processes for which arrivals do not occur instantaneously but over a period of time may be modeled as a fluid flow (see Chapter 7).

3.4.3 The M/M/1 Queue—Infinite Waiting Room

The first model that we shall consider is the M/M/1 queue with an infinite waiting room. In this case, customers arrive at a Poisson rate of λ per second on the average and each customer requires an exponentially distributed amount of service with mean $1/\mu$. From our previous discussion, we have the probability of an arrival in an incremental interval δ as $\lambda\delta$ irrespective of the number in the system. Since the service time is exponentially distributed with mean $1/\mu$, the probability of a completion of service in an incremental interval is $\mu\delta$. Arrivals correspond to births and departures to deaths; therefore the parameters of the process are $\lambda_n = \lambda$ and $\mu_n = \mu$. The state transition flow diagram is shown in Figure 3.11.

Substituting in (3.31), we find that the probability of n customers in the system, both in service and in queue is given by

$$P_n = (1 - \rho)\rho^n; \quad n = 0, 1, \ldots \tag{3.32}$$

where $\rho = \lambda/\mu$. We assume here that $\rho < 1$, which is reasonable since this means that the average number of arrivals during the average service time, $1/\mu$, is less than one. If it were otherwise, the queue would be unstable. We notice that the

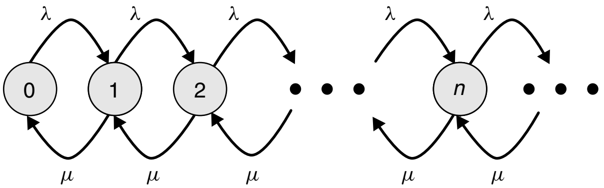


Figure 3.11 State transition flow diagram M/M/1 queue.

distribution of the number of customers in the system is geometric with the probability of an empty system, $P_0 = 1 - \rho$. Define N to be the random variables representing the number of customers in the system. The probability generating function of N is

$$P(z) = E(z^N) = \sum_{n=0}^{\infty} z^n (1-\rho)\rho^n = \frac{1-\rho}{1-z\rho} \tag{3.33}$$

with mean and variance

$$\bar{N} = \frac{\rho}{1-\rho} \tag{3.34a}$$

$$\text{Var}(N) = \frac{\rho}{(1-\rho)^2} \tag{3.34b}$$

The quantity ρ is called the *traffic intensity* or *load*. We shall follow the convention of measuring the load in *erlangs* with ρ defined as above. An alternative that is common in telephone traffic studies is the *call second* (CS) according to the portion of a busy hour that is used. The relation between the two units of measurement is CS $= 3600\rho$.

An application of Little's formula gives the average delay in the M/M/1 queue:

$$\bar{D} = \frac{\bar{N}}{\lambda} = \frac{1/\mu}{1-\rho} = \frac{1}{\mu - \lambda} \tag{3.35}$$

Note that, under very light loading $\rho \cong 0$ and $\bar{D} \cong 1/\mu$, the message transmission time. As congestion increases, $\rho \to 1$ we have $\bar{D} \to \infty$.

The average number of customers in the system and the average delay normalized to the message transmission time, $1/\mu$, are plotted as a function of load in Figure 3.12. (See the associated Excel spreadsheet.) The nonlinear form of these curves is characteristic of queueing results. For light loading, there is a linear increase with load. Beyond the knee of the curve the system may be characterized as unstable inasmuch as small changes in load bring large changes in the average number and in average delay.

The statistics that we have considered to this point lump together customers being served and those waiting in line. It is of interest to consider the number in the queue separately. From (3.32), the distribution of Q, the number of customers in the queue is given by

$$Q_n = P[n \text{ in queue}] = \begin{cases} (1-\rho)(1+\rho); & n = 0 \\ (1-\rho)\rho^{n+1}; & n = 1, 2, \ldots \end{cases} \tag{3.36}$$

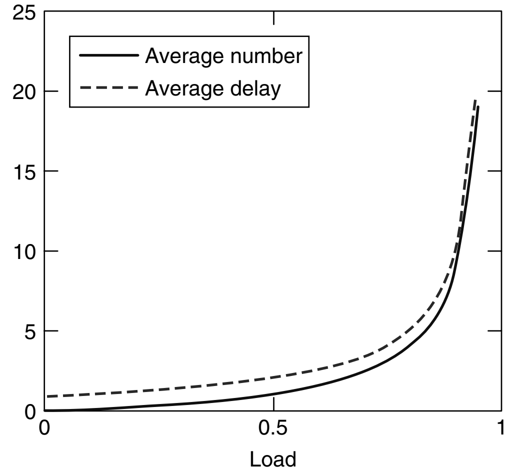


Figure 3.12 Traffic load versus average delay and number of customers.

It follows immediately that the average queue size is given by

$$\bar{Q} = \frac{\rho^2}{1-\rho} \tag{3.37}$$

Delay Distribution—M/M/1 Queue In this subsection, we derive the probability distribution of delay in the M/M/1 queue under the assumption that service is FCFS. In this case, all the customers who are left behind by a departing customer arrived while the departing customer was in the system. From PASTA, the probability of an arriving customer meeting n customers is given by (3.32). For a system in equilibrium, the distribution seen by a departing customer must be the same as that seen by an arriving customer; consequently

$$P(n \text{ customers in system/delay of departing customer} = t) = \frac{(\lambda t)^n e^{-\lambda t}}{n!}$$

Averaging over the delay, we find the probability of n customers in the system as discussed above. We have

$$P_n = \int_0^\infty \frac{(\lambda t)^n e^{-\lambda t}}{n!} d(t)dt; \quad n = 0, 1, 2, \ldots$$

Now, we calculate the probability-generating function for the number of customers in the system:

$$P(z) = \sum_{n=0}^{\infty} P_n z^n = \sum_{n=0}^{\infty} z^n \int_0^\infty \frac{(\lambda t)^n}{n!} e^{-\lambda t} d(t)dt$$

By exchanging integration and summation and the application of the Taylor series for the exponential, we find

$$P(z) = \int_0^\infty e^{-\lambda t(1-z)} d(t)dt = \mathcal{D}(\lambda(1-z)) \tag{3.38}$$

Recall that $\mathcal{D}(s)$ is the Laplace transform of the probability density of delay, $d(t)$. Here it is evaluated at $\lambda(1-z)$. We recognize that (3.38) is a general relationship between the probability-generating function of the number of Poisson arrivals during a random interval of time and the Laplace transform of the probability density of the time interval. We shall have occasion to use it again in the text.

The probability-generating function for the M/M/1 queue is given by (3.33). From (3.38), we have

$$\frac{1-\rho}{1-z\rho} = \mathcal{D}(\lambda(1-z))$$

If we substitute $z = 1 - s/\lambda$ into (3.31), we find

$$\mathcal{D}(s) = \frac{1-\rho}{1-(1-s/\lambda)\rho} = \frac{\mu-\lambda}{s+\mu-\lambda} \tag{3.39}$$

Thus, consulting Table 2.1, we see that the delay for the M/M/1 queue is exponentially distributed with mean

$$\bar{D} = \frac{1}{\mu-\lambda}$$

that is

$$d(t) = (\mu-\lambda)e^{-(\mu-\lambda)t}; \quad t \geq 0 \tag{3.40}$$

3.4.4 The M/M/1/L Queue—Finite Waiting Room

The case of the M/M/1/L queue represents the simplest case of a queue in which the arrival and the departure rates are not constant. Customers arrive at rate λ as long as there is less than L in the system, either in service or in the queue. When the waiting room becomes full, would be arriving customers are turned away.[6] This may be expressed as

$$\lambda_j = \begin{cases} \lambda; & j < L \\ 0; & j \geq L \end{cases}$$

[6] In the telephone network, for example, this overflow traffic might travel on alternative routes.

Since the service time of a customer in service is exponentially distributed, the departure rate remains at $\mu_n = \mu$ for all n. The state transition flow diagram for this case is shown in Figure 3.13.

In this case the steady-state distribution of the number in the system length is found by substitution into (3.31):

$$P_n = \begin{cases} \dfrac{(1-\rho)\rho^n}{1-\rho^{L+1}}; & n \leq L \\ 0; & n > L \end{cases} \tag{3.41}$$

Note that the normalizing condition here is

$$P_0 = \frac{1-\rho}{1-\rho^{L+1}} \tag{3.42}$$

The generating function of the number in the system is

$$P(z) = \frac{(1-\rho)[1-(z\rho)^{L+1}]}{(1-\rho^{L+1})(1-z\rho)} \tag{3.43}$$

A quantity that is of interest in systems with finite waiting rooms is the probability of a customer arriving to a full waiting room. From (3.34a) we find this to be given by

$$P_L = \frac{(1-\rho)\rho^L}{(1-\rho^{L+1})} \tag{3.44}$$

This blocking probability is shown as a function of the load ρ for several values of L in Figure 3.14 (see the associated Excel spreadsheet). As expected, the blocking probability increases with increases in load and decreases as the size of the waiting room increases. These M/M/1/L queues can be used to model the arrival of messages at a buffer whose output is a transmission line. The time required to "serve a customer" in this case is the length of the message in bits divided by the transmission rate. The messages may arrive at random from a number of different

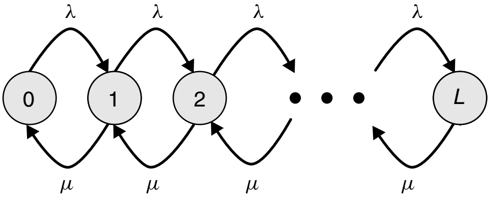


Figure 3.13 State transition flow diagram M/M/1/L queue.

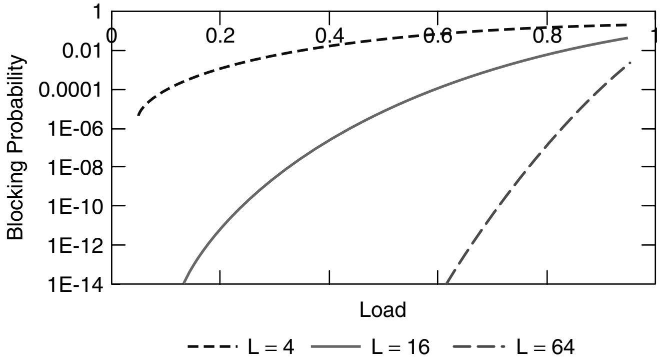


Figure 3.14 M/M/1/L queue.

sources. An infinite waiting room or buffer approximates the case where the probability of overflow because of too many messages in the buffer is negligible.

Example 3.7 We apply the M/M/1/L model to voice message multiplexer. A voice signal is digitized at a rate of 8000 bps. The average length of a voice message is 3 min. Messages are transmitted on a DS-1 line, which has a capacity of 1.344 Mbps. While waiting for transmission, the messages are stored in a buffer, which has a capacity of 10^6 bits. The blocking probability is calculated on the associated Excel spreadsheet and is shown on Figure 3.15 as a function of message arrival rate.

3.4.5 The M/M/S Queue—Infinite Waiting Room

As the next variation on the basic theme, we shall consider the M/M/S queue with an infinite waiting room in which there are $S \geq 1$ servers. If the number of customers is less than the number of servers, all customers receive service simultaneously.

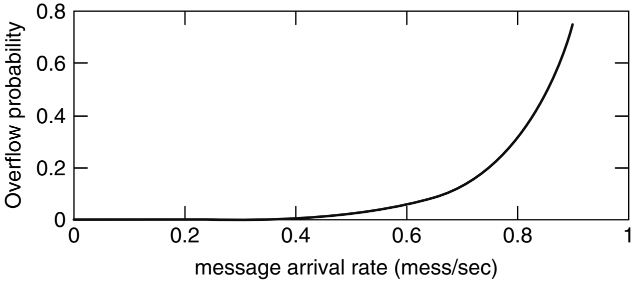


Figure 3.15 Voice messaging.

If the number of customers is greater than the number of servers, a queue forms. In any event, each of the customers receiving service may depart in an incremental interval with probability $\mu\delta$. If there are j customers simultaneously being served, the probability of one departure in an incremental interval is $j\mu\delta$. The probability of more than one departure in an incremental interval is negligible. If the waiting room is infinite, the arrival rate is independent of the number of customers in the system $\lambda_n = \lambda$. We have

$$\mu_n = \begin{cases} n\mu; & n \leq S \\ S\mu; & n > S \end{cases} \tag{3.45}$$

The state transition flow diagram for this case is shown in Figure 3.16.

From (3.31), we have

$$P_n = \begin{cases} \dfrac{P_0\rho^n}{n!}; & n \leq S \\ \dfrac{P_0\rho^n}{S!S^{n-S}}; & n > S \end{cases} \tag{3.46}$$

with

$$P_0 = \left[\sum_{n=0}^{S-1} \frac{\rho^n}{n!} + \frac{S\rho^S}{S!(S-\rho)}\right]^{-1}$$

In order for this solution to be valid, we must have $(\lambda/S\mu) = \rho/S < 1$. This is essentially the same condition for stability as for the M/M/1 queue.

In connection with this model, a quantity of interest is the probability that a customer arrives to find all the servers busy and must wait in queue. Recalling PASTA, the distribution seen by arrivals is given by (3.46); consequently, the probability of queueing is given by

$$P_c(S,\rho) = \sum_{n=S}^{\infty} P_n = \frac{S\rho^S/(S!(S-\rho))}{\sum_{n=0}^{S-1} \rho^n/n! + S\rho^S/(S!(S-\rho))} \tag{3.47}$$

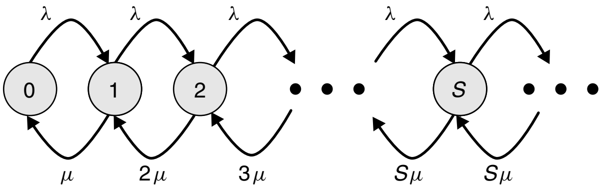


Figure 3.16 State transition flow diagram M/M/S queue.

This is the probability that all servers are busy and a newly arrived customer must wait. The formula is the historic *Erlang C* formula (after A. K. Erlang; see Chapter 1), which was applied in connection with voice traffic on the telephone network. The current implementation is call waiting, for which one is told "All of our operators are busy, etc." The service time in this case is the duration of or the *holding time* of a call, which may be reasonably modeled as exponentially distributed. The Poisson arrival model is also reasonable for telephone traffic. The servers are telephone trunks, which are occupied for the duration of a call. Since calls are queued, the case is called *blocked calls held*. The Erlang C formula is plotted in Figure 3.17, where P_c is shown as function of ρ with S as a parameter. These results have been worked out on the attached associated Excel spreadsheet.

Delay for the M/M/S Queue There is an interesting relationship between the Erlang C formula and the average number of messages in the queue. For the M/M/S queue, the average number of messages in the queue is given by

$$\bar{Q} = \sum_{j=0}^{\infty} jP_{j+S} = \frac{\sum_{j=0}^{\infty} jP_0\rho^{j+S}}{S!} = \frac{P_0\rho^S}{S!}\frac{S\rho}{(S-\rho)^2} \tag{3.48}$$

We take the ratio of this term to the blocking probability as expressed in (3.47) to find

$$\frac{\bar{Q}}{P_c} = \frac{\rho}{S-\rho} \tag{3.49}$$

We can use Little's formula to calculate delay in the M/M/S queue. From (3.49) and Little's formula, we have the average queue delay in terms of the blocking

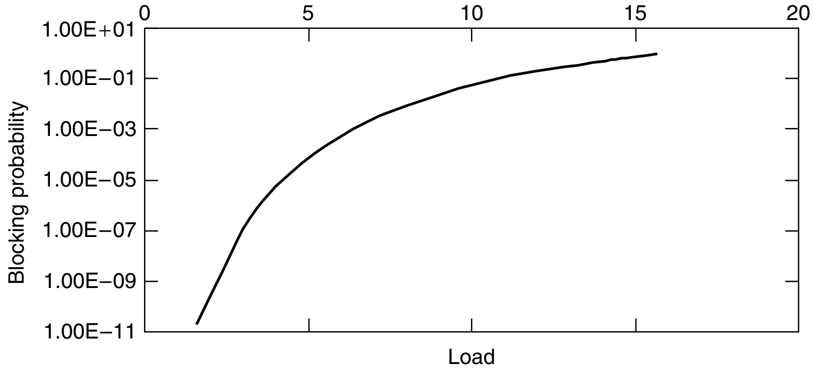


Figure 3.17 Blocking probability M/M/S queue.

probability:

$$\overline{D_Q} = \frac{\bar{Q}}{\lambda} = \frac{P_c \rho}{\lambda(S - \rho)}$$

By including the average service time, we find the average time that a message spends in the system

$$\bar{D} = \frac{1}{\mu} \frac{1 + P_c}{(S - \rho)} \tag{3.50}$$

Example 3.8 We consider in this section an application of some of our queueing theoretic results to a communications problem. Suppose that a 160-kbps line is available for data transmission. This line can be used in either of two ways. *Time-division multiplexing* (TDM) can be used to provide sixteen 10-kbps channels, or the full 160-kbps rate can be used to transmit messages. Let us assume that messages arrive at a Poisson rate and are exponentially distributed. As we shall see in Chapter 5, if average message lengths are long enough, the transmission line may be treated as a continuously available server in TDM systems. We may assume that this condition applies in the present case and average delay can be found from (3.40) and (3.50) for the M/M/S queue with $S = 16$. The results are plotted as a function of load in Figure 3.18 for an average message length of 2000 bits.

Also plotted in Figure 3.18 is the average delay when the full 160-kbps channel is used for all arriving messages. The model is the M/M/1 queue, where the single server is the 160-kbps line. The average delay is given by (3.28). Notice that $1/\mu$ in the numerator is just the time required to transmit a message. In this case there is a single server, but the average time required to transmit a message is one-fifth what it is for the TDM case.

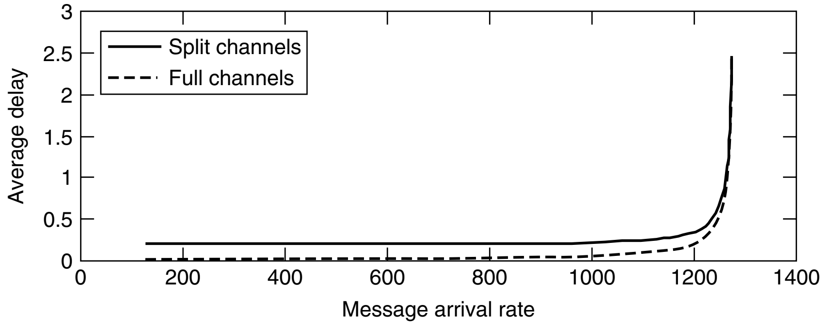


Figure 3.18 Multiplexing comparison: average delay versus load.

The plots show a result that is of particular interest in the computer communications where bursty traffic is a factor. For light loading the delay for TDM is 5 times what it is for the full channel. The reason is that, for light loading, there is little queueing delay and the time required to transmit a message dominates. As the load increases, queueing delay becomes more important and delays tend to be equal for both systems. For heavy loads, all the channels in TDM are in use and the departure rate of messages is the same as the full-channel case.

3.4.6 The M/M/S/L Queue—Finite Waiting Room

Another widely applicable formula is obtained by considering the model of the M/M/S queue with a finite waiting room. As in the earlier arrival case of a finite waiting room, there can be no new arrival when the waiting room is full, that is, $\lambda_n = 0$ for $j \geq L$, where L is the total number in the system. The following coefficients describe the system:

$$\mu_n = \begin{cases} n\mu; & n \leq S \\ S\mu; & n > S \end{cases}$$
$$\lambda_n = \begin{cases} \lambda; & n \leq L \\ 0; & n > L \end{cases} \tag{3.51}$$

The state transition flow diagram in this case is shown in Figure 3.19.

Again substituting these coefficients into (3.31) we can obtain the solution. The general solution is not difficult to find—just messy. We consider only the special case where $L = S$. In this case only the customers in the system are being served. The formula for the M/M/S/S queue is

$$P_n = \frac{P_0\rho^n}{n!}; \quad n = 1, 2, \ldots, S \tag{3.52}$$

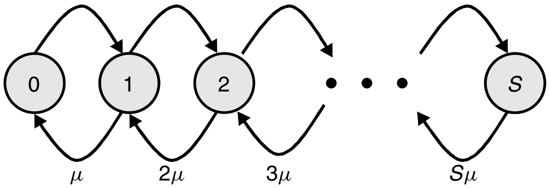


Figure 3.19 State transition flow diagram M/M/S/S queue.

with

$$P_0 = \left[\sum_{n=0}^{S} \frac{\rho^n}{n!} \right]^{-1}$$

When an arrival finds all S servers busy, it must leave the system since there is no waiting room. The probability of this event is the Erlang B formula, given by

$$P_B(S, \rho) = \frac{\rho^S}{S! \sum_{n=0}^{S} (\rho^n/n!)} \tag{3.53}$$

Equation (3.53) can be put in an iterative form. It is left to the reader to show that

$$P_B(S, \rho) = \frac{\rho P_B(S-1, \rho)}{S + \rho P_B(S-1, \rho)} \tag{3.54}$$

The iteration begins with

$$P_B(0, \rho) = 1$$

The historic application of the Erlang B formula is in sizing telephone trunks. As stated above, it is a reasonable assumption that the duration of telephone calls is exponentially distributed and the arrival rate is Poisson. The servers in this case are these trunks. The question of interest is to choose the number of trunks for a particular traffic volume so that an arriving call finds a trunk with sufficiently high probability. If all trunks are busy, blocked calls leave the system. This is called *blocked calls cleared.* This equation is plotted in Figure 3.20. It is of interest to note that there is a certain *insensitivity* property of the Erlang B formula. We shall see presently that the blocking probability does not depend on the probability distribution of the service time, only on its mean value.

A case of interest in connection with (3.52) is the infinite server queue, $S = \infty$. An arriving customer always has a server available. We have

$$\begin{aligned} P_0 &= \left[\sum_{n=0}^{\infty} \frac{\rho^n}{n!} \right]^{-1} = e^{-\rho} \\ P_n &= \frac{e^{-\rho}\rho^n}{n!}; \quad n = 1, 2, \ldots \end{aligned} \tag{3.55}$$

Of course, this is the Poisson distribution with mean ρ. We shall be dealing with this case in Chapter 4 in connection with networks of queues.

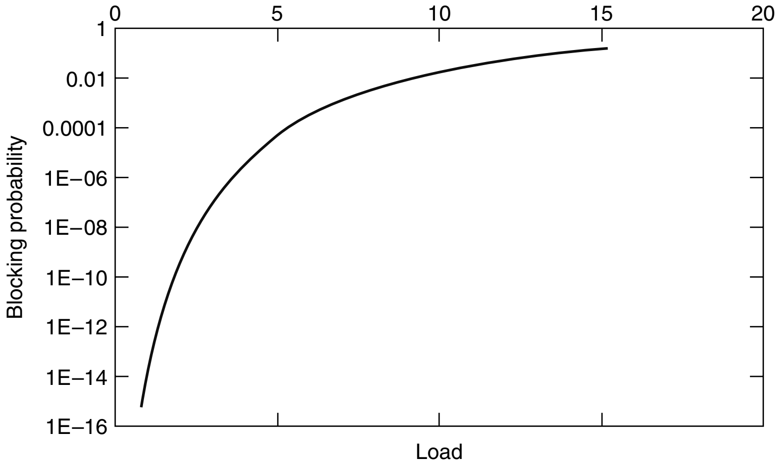


Figure 3.20 Call blocking probability versus call traffic load.

Example 3.9 Messages that are a constant 1000 bits long arrive at a multiplexer having 16 output lines, each operating at a 50 kbps rate. Suppose that the messages arrive at an average rate of 1,440,000 per busy hour. There is no storage; thus, if a message is not served immediately, it is lost. An arrival rate of 1,440,000 message per busy hour = 400 messages per second and the load $\rho = (400)(1000/50{,}000) = 8$, substituting into (3.53) gives Probability of blocking = 0.0045.

3.4.7 Finite Sources

Finally, in this section we consider the case of a finite number of sources (see Fig. 3.21). We assume N sources, each generating a message in an incremental interval, $(t, t+\delta)$, with probability $\delta\sigma$ when it is in the active state. The message length is exponentially distributed, and while it is being transmitted, the source enters the idle state and generates no new messages until the previously generated message has been transmitted. It is assumed that there are S servers. We also make the simplifying assumption that there is no storage; consequently,

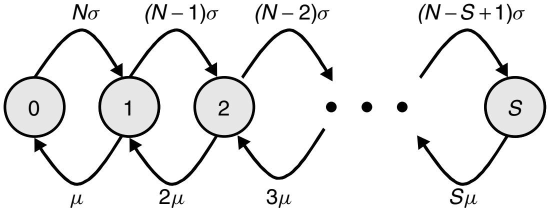


Figure 3.21 State transition flow diagram M/M/S/S queue.

if a source generates a message that finds no server free, the message is lost and the source returns to the idle state immediately. This is the finite-source analog to the M/M/S/S queue. The message arrival and departures rates are, respectively

$$\begin{aligned} \mu_n &= n\mu; \qquad n \leq S \\ \lambda_n &= (N-n)\sigma; \quad n \leq S \end{aligned} \tag{3.56}$$

From (3.31) we have the solution for the probability of n messages being transmitted

$$P_n = P_0 \binom{N}{n} \left(\frac{\sigma}{\mu}\right)^n; \quad n = 0, 1, \ldots, S \tag{3.57}$$

where

$$P_0 = \left[\sum_{k=0}^{S} \binom{N}{k} \left(\frac{\sigma}{\mu}\right)^k\right]^{-1}$$

From (3.57), we have

$$P_n = \frac{\binom{N}{n}\left(\frac{\sigma}{\mu}\right)^n}{\sum_{k=0}^{S}\binom{N}{k}\left(\frac{\sigma}{\mu}\right)^k}; \quad n = 0,\ 1, \ldots, S \tag{3.58}$$

This is known as the *Engset* distribution.

As in the case of the M/M/S/S queue, a call is blocked when all servers are occupied. The probability of this event is P_S. In the special case where $N \leq S$, there is no blocking. The denominator of (3.58) is just $[1 + \sigma/\mu]^N$, and we have a binomial distribution with N sources and the probability of a single source being active equal to $\sigma/(\sigma + \mu)$.

3.5 METHOD OF STAGES

3.5.1 Laplace Transform and Averages

In all our work to this point the service time distribution has been exponential. The method of stages, which was devised by Erlang himself, permits the analysis of systems with general distributions of service time by use of the memoryless property of the exponential distribution. Figure 3.22 shows the service received by a message in the form of a network with K exponential stages, which is called a *Cox network*

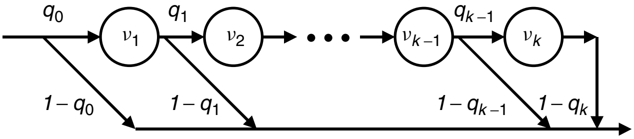


Figure 3.22 The method of stages.

[Cox, 1955]. A single message enters the network as shown and moves from stage to stage. After completing stage i, the message may leave the system with probability $1 - q_i$ or go on to the next stage with probability q_i. The probability that a message goes through exactly i stages is

$$\prod_{j=0}^{i-1} q_j(1 - q_i); \quad i = 1, 2, \ldots, K \tag{3.59}$$

where we take $q_0 = 1$ and $q_K = 0$. Clearly

$$\sum_{i=1}^{K} \prod_{j=0}^{i-1} q_j(1 - q_i) = 1 \tag{3.60}$$

since all possibilities are included. If the mean value for stage i is $1/\nu_i$, the Laplace transform for the service received at the ith stage is given by $\nu_i/(\nu_i + s)$. Since the service at each stage is independent, the Laplace transform for the service time received by a message that goes through exactly i stages is

$$\prod_{j=1}^{i} \frac{\nu_j}{s + \nu_j} \tag{3.61}$$

Averaging over the number of stages that a message goes through, we have for the Laplace transform of the time the message spends in the K-stage circuit

$$M(s) = \sum_{i=1}^{K} \prod_{j=0}^{i-1} q_j(1 - q_i) \prod_{l=1}^{i} \frac{\nu_l}{s - \nu_l} \tag{3.62}$$

It can be shown that any service time distribution, which may be represented by a rational function of s, can be put in the form of (3.62), which is called a *Coxian distribution*. The roots of the denominator polynomial of the rational function are given by ν_j in (3.62). The calculation of the various coefficients may be quite complicated. To solve such problems, we use the embedded Markov chain approach treated in Chapter 5. For us, the utility of the method of stages is as an analytic tool. For example, we later use it to demonstrate a certain insensitivity property of the

Erlang B formula. In Chapter 4, it is used to show this same insensitivity in networks of queues.

The average service time as found by differentiating with respect to s and setting $s = 0$:

$$\bar{M} = \sum_{i=1}^{K} \prod_{j=0}^{i-1} q_j(1 - q_i) \sum_{l=1}^{i} \frac{1}{\nu_l} \tag{3.63}$$

A little manipulation of (3.63) puts it into a much more useful form. Since $q_K = 0$, we have

$$\bar{M} = \frac{q_0}{\nu_1} + \sum_{i=2}^{K} \prod_{j=0}^{i-1} q_j \sum_{l=1}^{i} \frac{1}{\nu_l} - \sum_{i=1}^{K-1} \prod_{j=0}^{i} q_j \sum_{l=1}^{i} \frac{1}{\nu_l} \tag{3.64}$$

After a change of variables in the first summation on the RHS of (3.64), we have

$$\bar{M} = \frac{q_0}{\nu_1} + \sum_{m=1}^{K-1} \prod_{j=0}^{m} q_j \sum_{l=1}^{m+1} \frac{1}{\nu_l} - \sum_{i=1}^{K-1} \prod_{j=0}^{i} q_j \sum_{l=1}^{i} \frac{1}{\nu_l} = \sum_{m=1}^{K} \frac{\prod_{j=0}^{m-1} q_j}{\nu_m} \tag{3.65}$$

A commonly encountered case is $q_i = 1$ and $\nu_i = \nu$, for all i. In this case we have

$$M(s) = \left[\frac{\nu}{s + \nu}\right]^K \tag{3.66}$$

which has the probability density function, which we recognize to be the *Erlang K* distribution. The mean value of the service time is given by

$$\bar{M} = \frac{K}{\nu} \tag{3.67}$$

As we have seen in Section 2.5 in connection with the Chebyshev inequality, his distribution can be used to approximate a constant service time.

3.5.2 Insensitivity Property of Erlang B

In order to demonstrate the use of the method of stages and to illustrate the so-called *insensitivity* property of the Erlang B blocking formula, we carry out a couple of exercises.

Example 3.10 In the first of these, we consider a single server with K identical stages. We assume a pure blocking system in which there is no queueing. If a message arrives to find the server occupied, it is lost; otherwise, it enters the first stage of the server. We assume that the message arrives at a Poisson rate with an average of λ messages per second. The state of the system is the stage where the message in the system is being served. If the server is empty, we say that the system is in stage 0. Now we write the equilibrium equations (see Fig. 3.23) for the state transition flow diagram, where P_i is the probability of being in state i:

$$\begin{aligned} P_0\lambda &= \nu P_K \\ P_1\nu &= \lambda P_0 \\ P_2\nu &= \nu P_1 \\ &\vdots \\ P_K\nu &= \nu P_{K-1} \end{aligned} \tag{3.68}$$

with the normalizing condition

$$\sum_{n=0}^{K} P_n = 1$$

The solution to these equations is given by

$$\begin{aligned} P_0 &= \frac{1}{K\gamma+1} \\ P_n &= \frac{n}{K\gamma+1}; \quad n = 1, 2, \ldots, K \end{aligned} \tag{3.69}$$

where $\gamma = \lambda/\nu$. The probability that an arriving message is blocked is simply the probability of finding a message in the system

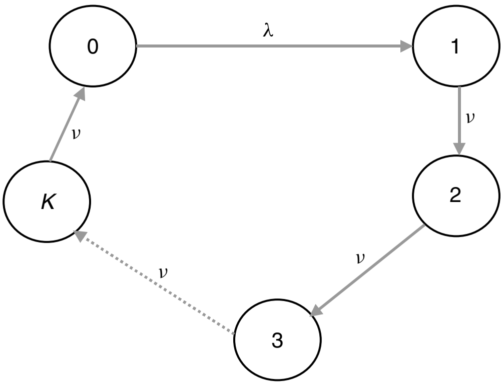


Figure 3.23 State transition flow diagram.

$$P_B = \sum_{i=1}^{K} P_i = \frac{K\gamma}{1+K\gamma} = \frac{K\lambda/\nu}{1+K\lambda/\nu} \tag{3.70}$$

From (3.67) we know that K/ν is the mean service time. We compare (3.70) to the Erlang B formula derived earlier for the M/M/S/S when $S = 1$ (3.53). We find that the same form when $S = 1$:

$$P_B = \frac{\text{(arrival rate)} \times \text{(average service time)}}{1 + \text{(arrival rate)} \times \text{(average service time)}} \tag{3.71}$$

In (3.53) the average service time is $1/\mu$, while in (3.70) it is K/ν. As long as these values are the same, the blocking probability will be the same function of the arrival rate even though the probability distributions of the service time are different.

Example 3.11 The same point is now made for an $M/E^{(2)}/2/2$, a system having two servers, each with two identical stages. Again, there is no storage and messages that arrive to systems serving two messages are lost. The state of the system is denoted (i,j), where i and j are, respectively, the number of messages in first and second stages. The state transition diagram for the stages (0,0), (0,1), (0,2), (1,0), (2,0), (1,1) is shown in Figure 3.24. The equilibrium equations are

$$\begin{aligned}
\lambda P_{00} &= \nu P_{01} \\
(\lambda+\nu)P_{01} &= 2\nu P_{02} + \nu P_{01} \\
2\nu P_{02} &= \nu P_{11} \\
(\lambda+\nu)P_{10} &= \lambda P_{00} + \nu P_{11} \\
2\nu P_{20} &= \lambda P_{10} \\
2\nu P_{11} &= \lambda P_{01} + 2\nu P_{20}
\end{aligned} \tag{3.72}$$

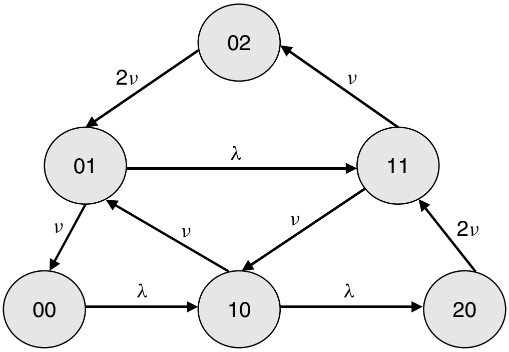


Figure 3.24 State transition flow diagram—two stages; each with two identical stages.

An arriving message is blocked if there are two messages in service:

$$P_B = P_{11} + P_{02} + P_{20} \tag{3.73}$$

Solving the equilibrium equations, we find that

$$P_B = \frac{2(\lambda/\nu)^2}{1 + 2(\lambda/\nu) + 2(\lambda/\nu)^2} \tag{3.74}$$

We see that this is the same form as the Erlang B formula in (3.53) with the mean service time $2/\nu$, and the insensitivity of the Erlang B formula to the probability distribution of the service time is demonstrated a second time. We shall encounter the insensitivity property again in Chapter 4 when we consider networks of queues. In this case as well, the method of stages is instrumental in finding a solution.

3.5.3 The Erlang B Blocking Formula: *N* Lines, Homogeneous Traffic

In Section 3.4.6, we derived the Erlang B formula, the blocking probability for the M/G/N/N queue under the assumption that the holding times are exponentially distributed. Just above, examples were given demonstrating that blocking probability does not depend on the distribution of the holding time, but only on the average holding time. In this subsection, we shall prove this result in the general case of arbitrarily distributed call duration.

We begin the derivation by assuming that the K-stage model shown in Figure 3.22 may describe the service given to each message. As we stated in connection with the method of stages, this model can approximate virtually any service distribution. We define the state of the system by the K-dimensional vector $(k_1, k_2, \ldots, k_K)$ where k_i; $i = 1, 2, \ldots, K$ is the number of messages in stage i. Notice that

$$\sum_{i=1}^{K} k_i \leq N$$

Messages arrive to the system with an average rate of λ messages per second. An arriving message goes into the first stage of an available server and is lost if none is available. This flow is into the first stage of the state space. We define the flow into stage i as ω_i; $i = 1, 2, \ldots, K$. Clearly

$$\begin{aligned} \omega_1 &= \lambda \\ \omega_i &= q_{i-1}\omega_{i-1}; \quad i = 2, 3, \ldots, K \end{aligned} \tag{3.75}$$

Next, we write the equilibrium equation for this system where the LHS indicates flow out of the state and the RHS flow into the state. If $\sum_{i=1}^{K} k_i < N$, we have

$$\begin{aligned}
&(\lambda + \sum_{i=1}^{K} k_i\nu_i)P(k_1, k_2 \ldots, k_K)\\
&\quad = \lambda P(k_1 - 1, k_2 \ldots, k_K)\\
&\qquad + \sum_{i=1}^{K-1}(k_i + 1)\nu_i(1 - q_i)P(k_1, k_2, \ldots, k_i + 1, \ldots, k_K)\\
&\qquad + \sum_{i=1}^{K-1}(k_i + 1)\nu_i q_i P(k_1, k_2 \ldots, k_i + 1, k_{i+1} - 1, \ldots, k_K)\\
&\qquad + \nu_K(k_K + 1)P(k_1, k_2, \ldots, k_K + 1) \qquad (3.76)
\end{aligned}$$

For the case $\sum_{i=1}^{K} k_i = N$, we have

$$\begin{aligned}
&\left(\sum_{i=1}^{K} k_i\nu_i\right)P(k_1, k_2, \ldots, k_K)\\
&\quad = \lambda P(k_1 - 1, k_2, \ldots, k_K)\\
&\qquad + \sum_{i=1}^{K-1}(k_i + 1)\nu_i q_i P(k_1, k_2, \ldots, k_i + 1, k_{i+1} - 1, \ldots, k_K) \qquad (3.77)
\end{aligned}$$

We now demonstrate that the solution to this equation is what is called the *detailed balance equation*, which is

$$\begin{aligned}
&k_i\nu_i P(k_1, k_2, \ldots, k_K) = \omega_i P(k_1, k_2, \ldots, k_{i-1} + 1, k_i - 1, \ldots, k_K);\\
&\quad i = 1, 2, \ldots, K \qquad (3.78)
\end{aligned}$$

We deal only with the case $\sum_{i=1}^{K} k_i < N$. The case $\sum_{i=1}^{K} k_i = N$ is a straightforward extension.

This equation expresses the flow between adjacent states. By successive iteration, we find an expression equivalent to (3.77)

$$P(k_1, k_2, \ldots, k_K) = P(0)\prod_{i=1}^{K}\frac{\rho_i^{k_i}}{k_i!} \qquad (3.79)$$

where $\rho_i = \omega_i/\nu_i$; $i = 1, 2, \ldots, K$ and $P(0)$ is a constant to be determined. We substitute (3.79) into (3.76) and cancel terms to obtain

$$\lambda + \sum_{i=1}^{K} k_i\nu_i = k_1\nu_1 + \sum_{i=1}^{K-1}(1 - q_i)\omega_i + \omega_K + \sum_{i=1}^{K-1} q_i\omega_i k_{i+1}\nu_{i+1}/\omega_{i+1} \qquad (3.80)$$

Substituting (3.75), and canceling terms again, we find

$$\lambda = \sum_{i=1}^{K-1} (1 - q_i)\omega_i + \omega_K \tag{3.81}$$

A repeated application of (3.75) shows that the RHS of (3.80) is λ and we have a solution. The constant term, $P(0)$, is found from the usual normalization equation.

$$P(0) = \left[\sum_{m=0}^{N} P\left(\sum_{i=1}^{K} k_i = m \right) \right] \tag{3.82}$$

Our interest is in the *total* number of messages in all stages, that is, the random variable

$$\sum_{i=1}^{K} k_i = M$$

We do a proof by induction. Assume that $K = 2$. We have

$$P(k_1 + k_2 = m) = P(0) \sum_{l=0}^{m} \rho_1^l \rho_2^{m-l}; \quad m \leq N$$

An application of the binomial theorem yields

$$P(k_1 + k_2 = m) = \frac{P(0)(\rho_1 + \rho_2)^m}{m!}; \quad m \leq N$$

Assuming that the formula holds for $K - 1$ stages, it is easy to show that

$$P\left(\sum_{i=1}^{K} k_i = m \right) = \frac{P(0)\left(\sum_{i=1}^{K} \rho_i\right)^m}{m!}; \quad m \leq N$$

but, from (3.75), we have

$$\sum_{i=1}^{K} \rho_i = \sum_{i=1}^{K} \frac{\omega_i}{\nu_i} = \lambda \sum_{i=1}^{K} \left(\prod_{j=0}^{i} q_j \right) \Big/ \nu_i \tag{3.83}$$

From (3.65), we see the RHS is the product of the mean service time and the mean arrival rate; thus; there is no dependence on the probability distribution of the service time. This demonstrates that the Erlang B formula

$$P_B = P\left(\sum_{i=1}^{K} k_i = N \right)$$

shows the same insensitivity with respect to the service time distribution; *all that matters is the mean of the service time.*

The insensitivity property has been shown for a more general case of blocking. Suppose that a number of sources, each with different service time requirements, feed into the same line. The blocking probabilities for each class can be shown to be sensitive only to the mean service times of the classes.[7]

REFERENCES

Cooper, R. B., *Introduction to Queueing Theory*, Macmillan, New York, 1972.

Cox, D. R., "A use of complex probabilities in theory of stochastic processes," *Proc. Cambridge Phil. Soc.* **51**: 313–319 (1955).

Cox, D. R., and D. Miller, *The Theory of Stochastic Processes*, Methuen, 1965.

Cox, D. R., and W. L. Smith, *Queues*, Methuen, New York, 1961.

Eilon, S., "A simpler proof of $L = \lambda W$," *Oper. Research* **17**: 915–916 (1969).

Feller, W., *An Introduction to Probability Theory and Its Applications*, Vol. 1, Wiley, New York, 1957.

Galliher, H. P., *Notes on Operations Research*, MIT Technology Press, Operations Research Center, Cambridge, MA, 1959, Chapter 4.

Hui, J. Y., "Resource allocation for broadband networks," *IEEE J. Select. Areas in Commun.* (9): 1598–1609 (Dec. 1988).

Jewell, W. S., "A simple proof of $L = \lambda W$," *Oper. Research* **15**(6): 109–116 (1967).

Kleinrock, L., *Queueing Systems*, Vol. 1: *Theory*, Wiley, New York, 1975.

Kobayashi, H., *Modeling and Analysis, An Introduction to System Performance Evaluation Methodology*, Addison-Wesley, Reading, MA, 1978.

Moran, P. A. P., *Theory of Storage*, Methuen, New York, 1959.

Mehdi, J., *Stochastic Models in Queueing Theory*, Academic Press, 2003.

Nelson, R., *Probability, Stochastic Processes and Queueing Theory*, Springer-Verlag, 1995.

Wolff, R. W., *Stochastic Modeling and the Theory of Queues*, Prentice-Hall, 1989.

EXERCISES

3.1 Show formally that given k successes in n Bernoulli trials, the probability of success for any single trial is k/n. (*Hint*: Count the number of ways that a success can occur on a particular trial.)

3.2 **(a)** What is the 10 probability of 10 straight successes in a sequence of Bernoulli trials?

(b) Given 10 straight successes, what is the probability of a success on the 11th trial?

[7] There is an elegant proof of this assertion in Hui (1988).

(c) The popular conception of the "law of averages" states that the portion of successes in n trials should be P. Show that this can be stated precisely using Chebyshev inequality and is the form of the strong law of large numbers. Reconcile (b) and (c).

3.3 Messages arrive at a telephone exchange at a Poisson rate of 6000 messages in a busy hour on the average.

(a) Write down an expression for the probability of more than 350 calls in 3 min.

(b) Find an approximate solution by assuming that the number of calls is a Gaussian random variable with the same mean and variance.

3.4 The duration D of a telephone conversation is distributed according to $P(D \leq t) = 1 - e^{-t/3}$; $t > 0$.

(a) What is the average duration of the call?

(b) If you are waiting outside a telephone booth (in the wind and rain), how long should you expect to wait, assuming that you arrived one minute after the call started?

(c) In general, how long, on the average, has a call encountered by an arriving customer been in progress?

(d) There is a paradox in (a)–(c). Explain.

3.5 Starting at time $t = 0$, messages arrive at a processor with a Poisson rate having an average of 100 ms between messages. In the first second, 10 messages were generated.

(a) What is the probability that 5 or more messages are generated in the interval (1.0, 1.5), where the time is in seconds?

(b) Repeat part (a) for the interval (0.75, 1.0).

3.6 Messages arrive to a service facility at a Poisson rate of four messages per second.

(a) Write down an expression for the probability of at least two messages in a 2-s interval. Messages arrive at times $T_1, T_2, \ldots$, where $t = 0$ is the beginning of the interval.

(b) Write down an expression for the probability that the first two arrivals occur before 2 s have elapsed. Suppose that four messages have arrived in the first 2 s.

(c) What are the mean and the variance of the number of messages that have arrived in the interval (1.5, 2.0)?

3.7 In the text we proved that the sum of independent Poisson processes is Poisson. Give an alternate derivation involving probability-generating functions.

3.8 We now work out an alternative proof of the fact that the sum of two independent Poisson processes is Poisson. Let T_1 and T_2 be the random variables indicating the interval between arrivals in the two streams. The arrival interval for the merged Poisson streams is $T = \min(T_1,\ T_2)$, accordingly. Show that T is exponentially distributed. What is the mean?

3.9 Referring to Exercise 3.8, show that the probability that the next arrival is from one or the other of the processes is $\lambda_i/(\lambda_1 + \lambda_2)$; $i = 1, 2$, respectively.

3.10 Suppose that a Poisson message arrival process is split deterministically, that is, alternate arrivals are placed in one of two different bins. Describe the arrival process at bin 2.

3.11 Suppose that a biological population starts with two members. Each member splits at a rate of two per second.

(a) Is the growth stable?

(b) What is the probability distribution as a function of time?

3.12 Suppose that a buffer holds exactly M fixed-length messages. Messages arrive at the buffer at a Poisson rate with an average of λ messages per second. As soon as the buffer is full, there are no new arrivals.

(a) Write down the set of differential equations for the probability of k messages in the buffer at time t.

(b) Solve these equations.

3.13 Consider a birth–death process with discouraged arrivals. We have

$$\lambda_n = 0 \qquad \text{for } n \geq N$$

Find P_n for all n.

3.14 Suppose that messages are generated from N sources. Once a message is in the queue or in service, there can be no more arrivals from that source until service completion. Assuming that there is a single server and as much storage as required, find the steady-state solution for the number of messages in the system P_n.

3.15 Consider an M/M/1 queue in which customers having completed service leave the system with probability q or immediately return to the queue with probability $1 - q$. Find the probability distribution for the number of customers in the system.

3.16 A person entering a banking facility finds that all three ATMs (automatic teller machines) are busy. Assume that the service times for all three are independent, identical exponential distributions.

(a) If there are no other customers waiting for service, what is the probability that the newly arrived customer will be the last to leave?

(b) If the average service time is 2 min, what is the average time the newly arrived customer will spend in the system?

(c) Repeat (a) under the assumption that there is one previously arrived customer waiting for service.

3.17 An M/M/1 queue has five messages in queue and one in service.

(a) Show that the interval until an arrival or a departure is exponentially distributed.

(b) What is the mean value in (a)?

(c) What is the probability of an arrival before a departure?

3.18 Repeat Exercise 3.16 for an M/M/6/6 with six messages in the system.

3.19 The window flow control technique limits the number of messages that may be in a network path to some maximum value. Show how this technique limits the average delay of a message in the path.

3.20 The telephone company wishes to provide trunks between two points. Assume that the rate of calling between the points is 360 calls per busy hour. The average duration of a call is 3 min. Assume that blocked calls are cleared. How many trunks are required so that the probability of blocking is less than 0.001?

3.21 In the text it was stated that the M/M/1 queue could be used to approximate the situation where randomly arriving messages of variable length are transmitted over a channel. Suppose that the average message length is 128 bytes (8 bits) and the line bit rate is 4800 bits per second. Messages arrive at an average rate of 7500 per busy hour.

(a) What is the probability of the buffer being empty? (Assume that the buffer holds messages being transmitted as well as messages waiting to be transmitted.)

(b) What is the average number of messages in the system?

(c) What is the average number of bits in the buffer?

(d) What is the average delay of a message?

3.22 As we have noted in the previous exercise, the number of bits in the buffer is a random sum of random variables. Show that the probability-generating function of the total number of bits in the buffer is given by $P(M(z))$, where $P(z)$ and $M(z)$ are the generating functions of the number of messages and the number of bits per message, respectively.

3.23 Assume a queue with the FCFS discipline and Poisson arrivals. What is the relationship between the variance of delay and the number of messages in the system?

3.24 Suppose that the Laplace transform for the duration of a telephone call is $C(s)$. During a call, other calls arrive at a Poisson rate. Find an expression for the probability-generating function of the number of these other arriving calls.

3.25 In an M/M/2 system the arrival rate is $\lambda = 3$ messages per second and the average transmission time is 0.5 s.

(a) What is the probability that arriving messages are blocked (i.e., what is the queue)?

(b) What is the average number of messages in the queue?

(c) What is the average time a message is in the system?

3.26 **(a)** For the arrival rate and the service time given in Exercise 3.20, find the number of servers necessary for the blocking probability to be less than 0.01.

(b) What is the average message delay for this system?

3.27 A communications link operates at a rate of 50 kbps. The link handles five calls each, generating packets at a rate of 5 packets/s. Assume that each packet is exponentially distributed with a mean of 1000 bits. Assume also that there is infinite storage.

(a) Assume that the line is split into five independent channels handling each of the calls. Find the average number of packets in the system and in the five channel queues.

(b) Find the average delay in each of the channel queues.

(c) Now suppose that all calls are transmitted over the full channel. Repeat parts (a) and (b).

3.28 Customers arrive at a Chinese restaurant at a rate of 160 per hour. A quarter of these are takeout customers. It takes 15 min on the average to prepare an order and 30 min to eat it. What is the average number of customers in the restaurant?

3.29 Three classes of messages arrive at a service facility. Assume that the loads are such that no queueing is necessary. Class 1 messages arrive at an average rate of 120 per minute and depart after a quick examination lasting 200 ms. The second class is processed at an average rate of 6 messages per minute and arrive at a rate of 20 per minute. Each message in the third class requires 30 s of processing and third-class messages arrive at a rate of 10 per minute. What is the average number of messages in the system?

3.30 People arrive at a taxi stand at a rate of 4 per minute. If the line has three passengers new arrivals depart without service. Taxis also arrive at a Poisson rate but at rate 6 per minute. Passengers depart immediately.

(a) Find the distribution for the line length.

(b) What is the average delay of a passenger?

3.31 A computational lab contains four terminals. Students arrive at a Poisson rate of four per hour and spend an average of 40 min at a terminal. If no terminals are available, students wait.

(a) What is the probability that a student will have to wait for a terminal?

Suppose that a student arrives to find all the terminals busy and two other students waiting.

(b) What is the expected time that the student needs to wait until a terminal becomes available?

3.32 The "leaky bucket" is a way of regulating the flow of traffic into a network. In order for a cell to be transmitted, it must have a token. We assume that tokens are generated at a Poisson rate and are stored in a pool, which is depleted by departing cells. Cells are also generated at a Poisson rate and are stored in a buffer if the token pool is empty. The operation of the system is such that either the cell buffer, the token pool, or both are empty at the same time.

(a) Show the state transition flow diagram.

(b) Find the probabilities of cell buffer and token pool occupancy under the following assumptions: cell arrival rate one per second, token arrival rate two per second, maximum of four cells stored, and maximum of two tokens in pool.

(c) Find the average cell delay.

3.33 A facility consists of two exponential servers, each having a different mean service time. A message entering an empty system is routed to the faster of the two servers, where it remains until its service completion. The system can hold one message in addition to those in service.

(a) Draw the state transition flow diagram for the system.

(b) Write down the equilibrium equations. Now suppose that a message can switch to a faster server if it is available.

(c) Repeat parts (a) and (b) above.

(d) Find the various probabilities in this case.

3.34 Consider an infinite server system in which each server consists of K identical exponential stages in series.

(a) Sketch the state transition flow diagram.

(b) Write down the equilibrium equations.

3.35 The servers in an M/G/3/3 queue is as depicted in Figure 3.25 as a two-stage hyperexponential distribution, H_2. A customer is routed with the probabilities shown to one of two parallel paths. The stages are exponential with the rates indicated.

(a) What is the average service time?

(b) Using the insensitivity property, find the probability of blocking when the arrival rate is 80 customers per second.

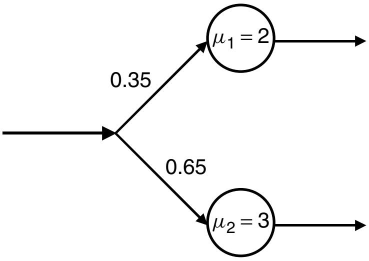


Figure 3.25

3.36 By purely probabilistic arguments, show that

$$\frac{\sum_{i=0}^{k-1} (\lambda t)^i e^{-\lambda t}}{i!} = \int_t^{\infty} \frac{\lambda(\lambda\tau)^{k-1}}{(k-1)!} e^{-\lambda\tau} d\tau$$

3.37 For the Erlang K distribution, substitute $\nu = K/m$ and show that the Laplace transform of the density function approaches e^{-ms} as $K \to \infty$. Recall that this is the Laplace transform for a constant service time.

4

NETWORKS OF QUEUES: PRODUCT FORM SOLUTION

4.1 INTRODUCTION: JACKSON NETWORKS

In many systems the appropriate model is a *network of queues.* Consider the model of a single node in the store-and-forward packet-switched network depicted in Figure 4.1. Messages enter the node from the external lines and are fed into an input buffer, where they are held for processing. The processor performs such operations as correcting and detecting errors, sending positive and negative acknowledgments

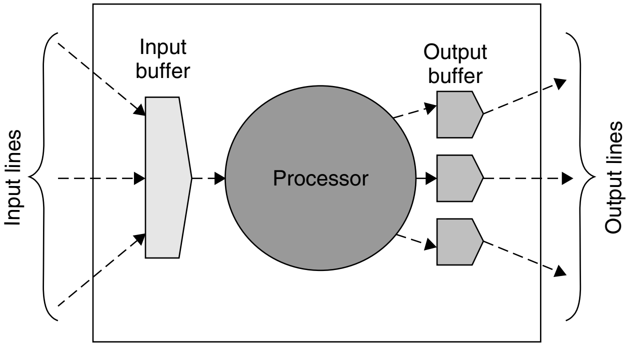


Figure 4.1 Model of message-switched node.

Modeling and Analysis of Telecommunications Networks, by Jeremiah F. Hayes and Thimma V. J. Ganesh Babu
ISBN 0-471-34845-7 © 2004 John Wiley & Sons, Inc.

and routing. After processing, the message is placed into an output queue for transmission over an output line. In this model the input buffer and processor form a single queue as does each output buffer and line. Questions of interest are the probability distribution of the number of messages in each queue and the distribution of the delay of a message through the node.

These networks may be modeled and analyzed by means of multidimensional birth and death processes, under a variety of assumptions. The salient result of this work is the *product form* solution, in which the joint probability distribution functions of queue occupancies resembles the product of marginal probability distribution functions of the number in the individual queues. We shall see that networks satisfying the proper set of assumptions are called *Jackson networks*, after J. R. Jackson (1963), who discovered the product form solution for networks under the following assumptions:

1. Poisson arrival to the network independent of the state of the network
2. Exponential service times at the nodes with the FCFS service discipline
3. Service time independent of the arrival process
4. Probabilistic routing whereby the next node, after service completion, is chosen independently from message to message

As a consequence of assumption 1, nodes with finite storage are not allowed.

Using the theory of Jackson networks, we shall find queue occupancy and delay in message-switched networks. Later in the chapter, we shall see that the same product form solution can be extended to different service time distributions and disciplines. Finally, we apply these results to several systems of practical interest.[1]

4.2 REVERSIBILITY: BURKE'S THEOREM

4.2.1 Reversibility Defined

A powerful tool in the analyses of networks of queues is *time reversibility.* A random process is time-reversible if the process, reversed in time, has the same probabilistic properties as the forward process. Consider the continuous-time process, $X(t)$. The *time-reversed process* derived from $X(t)$ is $X^r(t) = X(T - t)$ where T is an arbitrary point in time. For simplicity we set $T = 0$. Since the probabilistic properties are the same in both directions of time, the joint probabilities that the processes assume various values must be the same. A typical realization of a process and its reverse are shown in Figure 4.2. For the process to be reversible these segments must have the same probability.

[1] While writing this chapter we found several texts to be particularly useful: Gelenbe (1980), Gelenbe (1987), Kobayashi (1978), Nelson (1995), Robertazzi (1994), Walrand (1988), and VanDijk (1993).

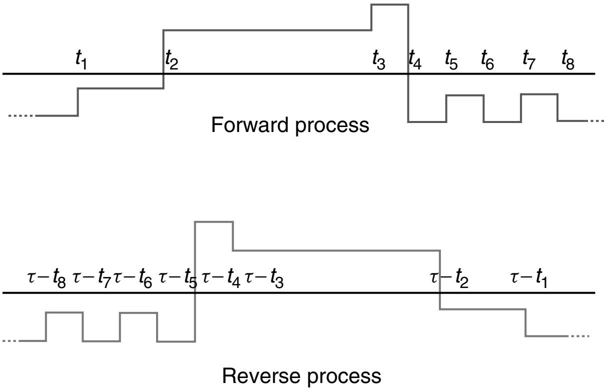


Figure 4.2 Forward and reverse processes.

Consider the respective sequences of times and points, $t_1, t_2, \ldots, t_m$ and $i_1, i_2, \ldots, i_m$. If the process is reversible, the joint probabilities are equal for all such sequences:

$$\begin{aligned} &P(X(t_1) = i_1, X(t_2) = i_2, \ldots, X(t_m) = i_m) \\ &\quad = P(X(\tau - t_m) = i_m, X(\tau - t_{m-1}) = i_{m-1}, \ldots, X(\tau - t_1) = i_1) \end{aligned} \qquad (4.1)$$

For a process to be reversible, it is not enough for the marginal probabilities for the forward and reverse processes to be equal. The circular process shown in Figure 4.3 illustrates this. The process $X(t)$ goes clockwise through the states, $i \rightarrow (i+1)\text{mod}12$, spending the same amount of time at each. The probability of being in any of the states is 1 in 12. The same is true of the reverse process, which travels counterclockwise. However, the reverse process is clearly not reversible since state 2 cannot follow state 1 in the reverse process, for example.

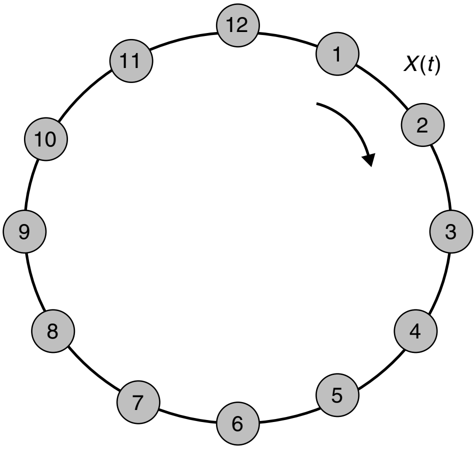


Figure 4.3 Circular process.

The Poisson process offers a simple example of a reversible process. Recall from Section 2.2.2 that we derived the Poisson process as the limit of Bernoulli trials. Were one to carry out the Bernoulli trials in the reverse order, the same succession of successes and failure would result since their respective probabilities are the same. The limiting process leading to the Poisson process would be the same as well.

4.2.2 Reversibility and Birth and Death Processes

A central role in our discussion is played by the birth–death process, which was treated in Chapter 3. Recall that the birth–death process was defined to be a continuous time, population process, which changed by at most one in an incremental interval. We now show that birth and death processes are reversible processes. We begin with the observation that the time the process spends at a certain population level is exponentially distributed, since it is terminated by an arrival or a departure. Assume that there are n members of the population. In the next incremental interval there will be an arrival or a departure with probabilities $\lambda_n\delta$ and $\mu_n\delta$, respectively; accordingly, the probability of either event is $(\lambda_n + \mu_n)\delta$ and the time spent in state n is exponentially distributed with mean value $1/(\lambda_n + \mu_n)$. The probability of eventually having a decrease is just the probability of a departure before an arrival. Let X and Y be the exponentially distributed intervals with means $1/\lambda_n$ and $1/\mu_n$, respectively.

$$P(\text{decrease}) = P(X \geq Y) = \int_0^\infty dx\ e^{-\lambda_n x} \int_0^x dy\ e^{-\mu_n y} = \frac{\mu_n}{\mu_n + \lambda_n}$$

Similarly, the probability of an increase is $\lambda_n/(\mu_n + \lambda_n)$.

We now prove that a necessary and sufficient condition for the birth–death process to be reversible is that the local balance equations[2] hold. We begin by proving necessity. Since the process $X(t)$ is reversible, we have from (4.1)

$$P(X(t) = j, X(t+\delta) = k) = P(X(t) = k, X(t+\delta) = j) \tag{4.2}$$

Inserting the probabilities of the population levels, P_j and P_k, we have

$$P_j P(X(t+\delta) = k/X(t) = j) = P_k P(X(t+\delta) = j/X(t) = k) \tag{4.3}$$

We now divide both sides by δ and let $\delta \rightarrow 0$. Since we are dealing with a birth–death process, let us assume, without loss of generality, that $k = j + 1$, that is, an arrival to population size j. We have

$$\begin{aligned} &\lim_{\delta \to 0} \frac{P(X(t+\delta) = j+1/X(t) = j)}{\delta} = \lambda_j \\ &\lim_{\delta \to 0} \frac{P(X(t+\delta) = j/X(t) = j+1)}{\delta} = \mu_{j+1} \end{aligned} \tag{4.4}$$

[2] This proof is adapted from a more general proof in Kelly (1979).

From (4.3) and (4.4), we have

$$P_j\lambda_j = P_{j+1}\mu_{j+1} \tag{4.5}$$

which is just the local balance equation. Clearly, we could just as well have a departure from population j, and the local balance equation would be derived.

Now, we consider sufficiency. As we have seen in Section 3.4, the local balance equations can be summed to derive the global balance equations, whose solution is the steady-state probabilities P_n; $n = 0, 1, 2, \ldots$. We consider a particular sequence of arrivals and departures in an interval of time $(0,T)$. Again, we do not lose generality by considering a specific example since it will be clear that the same results apply to any sequence. Recall that we showed at the beginning of this section that the sojourn time at any population level is exponentially distributed with a parameter that is the sum of the arrival and departure rates; furthermore, the probability of the next state is proportional to the rate into it. Let us assume that the population at time $t = 0$ is j and we have the sequence of arrivals and departures a, a, d, a, d, a with $t_1, t_2, t_3, t_4, t_5, t_6$, respectively as the time intervals between arrivals and departures as illustrated in Figure 4.4. The time interval t_7 rounds out the interval so that $\sum_{i=1}^{7} t_i = T$. At the end of the interval the population size is $j+2$. The probability of the sequence described above is

$$\begin{aligned}
&P_j \times \frac{\lambda_j}{\lambda_j+\mu_j} \times (\lambda_j+\mu_j)\exp(-(\lambda_j+\mu_j)t_1)dt_1 \times \frac{\lambda_{j+1}}{\lambda_{j+1}+\mu_{j+1}} \\
&\quad \times (\lambda_{j+1}+\mu_{j+1})\exp(-(\lambda_{j+1}+\mu_{j+1})t_2)dt_2 \times \frac{\mu_{j+2}}{\lambda_{j+2}+\mu_{j+2}} \times (\lambda_{j+2}+\mu_{j+2}) \\
&\quad \times \exp(-(\lambda_{j+2}+\mu_{j+2})t_3)dt_3 \times \frac{\lambda_{j+1}}{\lambda_{j+1}+\mu_{j+1}} \times (\lambda_{j+1}+\mu_{j+1})\exp(-(\lambda_{j+1}+\mu_{j+1})t_4)dt_4 \\
&\quad \times \frac{\mu_{j+2}}{\lambda_{j+2}+\mu_{j+2}} \times (\lambda_{j+2}+\mu_{j+2})\exp(-(\lambda_{j+2}+\mu_{j+2})t_5)dt_5 \times \frac{\lambda_{j+1}}{\lambda_{j+1}+\mu_{j+1}} \\
&\quad \times (\lambda_{j+1}+\mu_{j+1})\exp(-(\lambda_{j+1}+\mu_{j+1})t_6)dt_6 \times \exp(-(\lambda_{j+2}+\mu_{j+2})t_7)
\end{aligned}$$

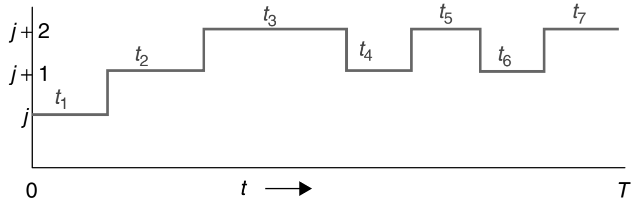


Figure 4.4 Birth–death process.

The term $(\lambda_j + \mu_j)\exp(-(\lambda_j + \mu_j)t_1)dt_1$ is the probability density of the first interval. The probability density of the next five intervals is similar. The probability that the process remains in the state $j + 2$ for at least t_7 seconds is given by $\exp(-(\lambda_{j+2} + \mu_{j+2})t_7)$. After the obvious cancellations, we have

$$
\begin{aligned}
&P_j\lambda_j \exp(-(\lambda_j + \mu_j)t_1)dt_1 \times \lambda_{j+1} \exp(-(\lambda_{j+1} + \mu_{j+1})t_2)dt_2\\
&\quad\times \mu_{j+2} \exp(-(\lambda_{j+2} + \mu_{j+2})t_3)dt_3 \times \lambda_{j+1} \exp(-(\lambda_{j+1} + \mu_{j+1})t_4)dt_4\\
&\quad\times \mu_{j+2} \exp(-(\lambda_{j+2} + \mu_{j+2})t_5)dt_5 \times \lambda_{j+1} \exp(-(\lambda_{j+1} + \mu_{j+1})t_6)dt_6\\
&\quad\times \exp(-(\lambda_{j+2} + \mu_{j+2})t_7)dt_7 \qquad (4.6)
\end{aligned}
$$

By successive substitutions of the local balance equation, starting with $P_j\lambda_j = P_{j+1}\mu_{j+1}$, and reversing order, we can show that

$$
P_j\lambda_j\lambda_{j+1}\mu_{j+2}\lambda_{j+1}\mu_{j+2}\lambda_{j+1} = P_{j+2}\mu_{j+2}\lambda_{j+1}\mu_{j+2}\lambda_{j+1}\mu_{j+2}\mu_{j+1} \qquad (4.7)
$$

By substituting (4.7) into (4.6) and rearranging terms, we have the expression

$$
\begin{aligned}
&P_{j+2}\mu_{j+2} \exp(-(\lambda_{j+2} + \mu_{j+2})t_7)dt_7 \times \lambda_{j+1} \exp(-(\lambda_{j+1} + \mu_{j+1})t_6)dt_6\\
&\quad\times \mu_{j+2} \exp(-(\lambda_{j+2} + \mu_{j+2})t_5)dt_5 \times \lambda_{j+1} \exp(-(\lambda_{j+1} + \mu_{j+1})t_4)dt_4\\
&\quad\times \mu_{j+2} \exp(-(\lambda_{j+2} + \mu_{j+2})t_3)dt_3 \times \mu_{j+1} \exp(-(\lambda_{j+1} + \mu_{j+1})t_2)dt_2\\
&\quad\times \exp(-(\lambda_j + \mu_j)t_1)dt_1 \qquad (4.8)
\end{aligned}
$$

which is the probability of the reverse process starting with population $j + 2$ and ending with j members. The term $e^{-(\lambda_i+\mu_i)t_1}$ is the probability that the process remains in state i for at least t_1 seconds. Again, it should be clear that this calculation goes through for any sequence of arrivals and departures.

4.2.3 Departure Process from the M/M/S Queue: Burke's Theorem

As we have seen in Chapter 3, the dynamics of the M/M/S queue can be described by a birth–death process, which is, as we have seen, reversible. Under the assumptions of the M/M/S queue, the arrivals to the forward process are Poisson. Since an increase in the forward process implies a decrease in the reverse process, the departures from the reverse process are also Poisson. Note that this would be true irrespective of the processes' reversibility. Reversibility requires that the rate of decrease or increase be the same in either direction. Thus, the departure process from the forward process must be the same as the departure process from the reverse process, that is, Poisson. This is the first part of *Burke's theorem* (Burke 1964, 1968).

Time reversibility also allows the proof of the second half of the theorem; specifically, at any time t, the number of messages in the system is independent of the sequence of departures prior to t. From reversibility, the departures prior to t in the forward process are the same as the arrivals after t in the reverse process. However, the arrival process in the reverse process is independent Poisson, which is independent of the current population.

We summarize: Burke's theorem states that the departure process from an M/M/S queue is Poisson and is independent of the contents of the queue.

Reversibility has also been used to prove the rest of the results of this chapter (Kelly 1979, Nelson 1995). Since our interest is not in theory per se, but in results and numerical computations, we adapt a different, algebraic approach, which we feel is clearer as an introduction to the subject.

4.3 FEEDFORWARD NETWORKS

4.3.1 A Two-Node Example

We start with what should be the simplest network example, the network formed by two queues in tandem, each having infinite storage and a single exponentially distributed server (see Fig. 4.5). Messages arrive at the first queue at a Poisson rate of λ per second. The service time has a mean value $1/\mu_1$. After completion of service, the message goes to a second queue where we assume that service is independent of the first and has mean $1/\mu_2$. We shall examine this independence assumption when we deal with store-and-forward message-switched networks in Section 4.4.

Using results from Chapter 3 in a straightforward fashion, we can find the joint distribution of the queue lengths in each queue. We shall do this under the assumption that each queue has only one server. Let the random variables Q_1 and Q_2 denote the number of messages in queues one and two in steady state, respectively. We define the joint probability

$$P(k_1, k_2, t) = P(Q_1(t) = k_1, Q_2(t) = k_2)$$

Certainly, the first of the tandem queues is unaffected by the second queue. We simply have Poisson arrivals to a single exponential server. The marginal distribution of the number in the system is that of the M/M/1 queue given by the geometric distribution

$$P(Q_1(t) = k_1) = P_1(k_1) = (1 - \rho_1)\rho_1^{k_1}; \quad k_1 = 0, 1, \ldots$$

where $\rho_1 = \lambda/\mu_1$.

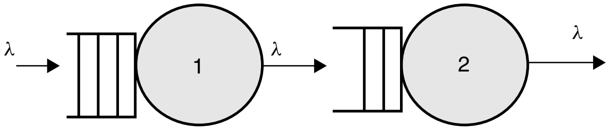


Figure 4.5 Two queues in tandem.

We now apply Burke's theorem. The first part states that the departure process from the M/M/1 queue is Poisson. Thus, the second of the tandem queues acts as an M/M/1 queue, and we have

$$P(Q_2(t) = k_2) = P_2(k_2) = (1 - \rho_2)\rho_2^{k_2}; \quad k_2 = 0, 1, \ldots$$

where $\rho_2 = \lambda/\mu_2$.

The second part of Burke's theorem states that the departure process from an M/M/1 queue is independent of the queue contents. This implies that the contents of the second queue are independent of the contents of the first queue; accordingly, the joint distribution in the steady state is simply the product of the marginal distributions:

$$P(k_1, k_2) = (1 - \rho_1)(1 - \rho_2)\rho_1^{k_1}\rho_2^{k_2}; \quad k_1, k_2 = 0, 1, 2, \ldots \tag{4.9}$$

Note that this solution is as though each of the queues existed in isolation. The delays through the queues are independent random variables. *In order to calculate the joint distribution, all that we need to know is the flow into the queue and the service distribution at the queue.* The wondrous quality of Jackson networks is that the same can be said of far more complex networks.

4.3.2 Feedforward Networks: Application of Burke's Theorem

An easy extension is that of more than two queues in tandem. Again, we assume that the single server in each queue is exponential and independent. If the arrival to the first queue is Poisson, the departure is Poisson and independent of the queue; consequently, the process replicates. The joint distribution of the number of messages in each queue is the product of geometric distributions, and we have

$$\begin{aligned} P(Q_1(t) = k_1, Q_2(t) = k_2, \ldots, Q_N(t) = k_N) &= P(k_1, k_2, \ldots, k_N) \\ &= \prod_{i=1}^{N} (1 - \rho_i)\rho_i^{k_i} \end{aligned} \tag{4.10}$$

where $\rho_i = \lambda/\mu_i$; $i = 1, 2, \ldots, N$ and $1/\mu_i$ is the average service time of the ith server. It is a straightforward exercise to show that these same results hold when there is more than one server at a facility.

The next extension is to *feedforward* or acyclic networks, that is, networks that do not contain feedback paths. An example of such a network is shown in Figure 4.6. Messages arrive from outside the network to each of the four nodes shown at independent Poisson rates with averages λ_1, λ_2, λ_3, λ_4, respectively. At each node the service time is independent and exponentially distributed. An essential feature of the network is *probabilistic routing*. For example, a message leaving queue 1 goes to queue 2 with probability 0.25 and to queue 3 with probability 0.75. Messages may leave the network in the same probabilistic fashion. For example, at node 3 messages

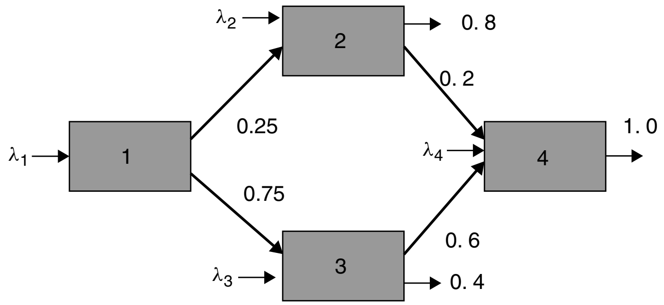


Figure 4.6 Open network with feedback.

depart with probability 0.4. Another important feature of flows is merging. At queues 2 and 3 traffic from queue 1 merges with external arrivals. The flow into queue 4 is the sum of flows from queues 2 and 3 and an external flow. Under these assumptions, all the flows within the network are Poisson. From Burke's theorem, the departure process from queue 1 is Poisson and independent of the queue. Further, as we have seen in Chapter 3, the random splitting of Poisson processes yields Poisson processes; consequently, the flows from 1 to 2 and from 1 to 3 are each Poisson. Since the sums of Poisson processes are Poisson, the total flows into 2 and 3, combining internal and external traffic, are each Poisson. This same line of reasoning shows that the total flow into node 4 is a Poisson process. Because of the independent Poisson flows into each of the nodes, we have a set of independent M/M/1 queues.

It is not difficult to generalize the results to any feedforward network. With Poisson arrivals to independent exponential servers with infinite buffers, the departure process is Poisson. Random splitting and combination of Poisson processes preserves the Poisson characteristic within the network and departure processes from the network are also Poisson. For an N-node network, each with single exponential servers, the joint probability distribution is also given by (4.10) with $\rho_i = \Lambda_i/\mu_i$, where Λ_i is the total flow into node i and $1/\mu_i$ is the average service time at each node.

4.3.3 The Traffic Equation

The traffic equation provides a systematic way of dealing with the flows in networks of all classes. We assume that the network is composed of N nodes, each containing an independent exponential server and an infinite buffer. The external arrival process to each node is Poisson with average rate, λ_i; $i = 1, 2, \ldots, N$. Message are routed probabilistically. The probability that a message is routed from node j to node i is given by q_{ji}; $i, j = 1, 2, \ldots, N$. The term q_{jN+1} indicates the probability of a message being routed outside the network. In general, it is assumed that $q_{jj} = 0$, $\forall j$ a node would not route a message to itself. Clearly, $\sum_{n=1}^{N+1} q_{jn} = 1$. Of course, messages are not routed in a probabilistic fashion in a telecommunications network (hopefully); nevertheless, the model applies under certain reasonable conditions. We assume a mixture of traffic with different destination so that the proportions in each direction are modeled by probabilities. The total flow into each node of the network, Λ_i; $i = 1, 2, \ldots, N$, is composed of external arrivals plus flows of each of the other nodes in the network. This relationship is represented as

$$\underbrace{\Lambda_i}_{\text{Total flow}} = \underbrace{\lambda_i}_{\text{external arrivals}} + \underbrace{\sum_{j=1}^{N} q_{ji}\Lambda_j}_{\text{Recycled flow}} \; ; \quad i = 1, 2, \ldots, N \tag{4.11}$$

The matrix version is

$$\mathbf{\Lambda} = \boldsymbol{\lambda} + \mathbf{\Lambda} Q \tag{4.12}$$

where Q is an $N \times N$ *routing matrix* with elements $\{q_{ij}\}$, $\mathbf{\Lambda} = [\Lambda_1, \Lambda_2, \ldots, \Lambda_N]$ and $\boldsymbol{\lambda} = [\lambda_1, \lambda_2, \ldots, \lambda_N]$. In either form, (4.11) or (4.12) is known as the *traffic equation* and also as the *conservation equation*. Normally, the inputs and the routing matrix are known, which we solve for the total input into each node

$$\mathbf{\Lambda} = \boldsymbol{\lambda}[I - Q]^{-1} \tag{4.13}$$

where I is the $N \times N$ identity matrix.

Example 4.1 It is a straightforward exercise to work out the joint distribution of the number of messages in each of the queues in Figure 4.6. Define Λ_i; $i = 1, 2, 3, 4$ as the total flow into node i. In terms of the external arrivals, we have the following set of equations for the total flow into each node:

$$\begin{aligned}
\Lambda_1 &= \lambda_1 \\
\Lambda_2 &= \lambda_2 + 0.25\Lambda_1 \\
\Lambda_3 &= \lambda_3 + 0.75\Lambda_1 \\
\Lambda_4 &= \lambda_4 + 0.2\Lambda_2 + 0.6\Lambda_3
\end{aligned}$$

The routing matrix is

$$Q = \begin{bmatrix} 0 & 0.25 & 0.75 & 0 \\ 0 & 0 & 0 & 0.2 \\ 0 & 0 & 0 & 0.6 \\ 0 & 0 & 0 & 0 \end{bmatrix}.$$

If the external flow is represented by the vector $\lambda = (\lambda_1, \lambda_2, \lambda_3, \lambda_4) = (2.0, 1.0, 0.5, 3.0)$, we can solve (4.13) to find $\Lambda = (\Lambda_1, \Lambda_2, \Lambda_3, \Lambda_4) =$

(2.0, 1.5, 2.0, 4.5). Now suppose that the respective average service rates are given by $\mu = (\mu_1, \mu_2, \mu_3, \mu_4) = (4.0, 6.0, 11.0, 9.9)$. The loads in each of the queues are $R = (R_1, R_2, R_3, R_4) = (0.5, 0.25, 0.18, 0.8, 0.4545)$, respectively and the joint distribution of the message in each of the queues is given by

$$P(Q_1 = k_1, Q_2 = k_2, Q_3 = k_3, Q_4 = k_4) = p(k_1, k_2, k_3, k_4)$$

$$= \prod_{i=1}^{4} (1 - R_i)R_i^{k_i} = (0.16741)(0.5)^{k_1}(0.25)^{k_2}(0.2091)^{k_3}(0.4424)^{k_4}$$

4.4 PRODUCT FORM SOLUTION FOR OPEN NETWORKS

4.4.1 Flows Within Feedback Paths

Feedforward networks fall within a broader class of *open networks*, which, in general, may contain *feedback paths*. In the class of open networks, at least one of the external flows into the network must be positive. Within the network, all the flow through a node is routed either to an internal node or to outside the network: i.e., $\sum_{i=1}^{N} q_{ji} \leq 1; j = 1, 2, \ldots, N$. Since there are arrivals to the network, there must be an exit path; thus $\sum_{i=1}^{N} q_{ki} < 1$ for at least one $k = 1, 2, \ldots, N$. Under these conditions on the routing, it can be shown that there is a valid solution to the traffic equation, (4.11), consisting of nonnegative flows through the nodes.[3] The solution of the traffic equation is the key to the calculation of probability distributions for occupancy and delay within the network.

We begin this section by demonstrating that the presence of a feedback path in a network destroys the Poisson character of the flow within the network. The network of two tandem queues shown in Figure 4.7a may illustrate this phenomenon. The arrivals are Poisson processes with an assumed average rate of 1 per hour. The output of the second queue is fed back to the first with probability $P = 0.999$. The service times in each of the queues are assumed to be independent and exponentially distributed with mean 1 ns! With this extreme set of parameters, the outputs of the first queue tend to be in bursts. A typical output sequence is shown in Figure 4.7b. This illustrates the general result, in that it can be shown that flow on links between queues within a feedback loop is not a Poisson process. Since the flows within the loop are not Poisson processes, we cannot use the preceeding analysis employing Burke's theorem to find the distribution of messages; nevertheless, as we shall see, the product form still holds. Moreover, the terms in the product are *the same as if the flows were Poisson.*

[3] The result is based on the Peron–Frobenius theorem for positive matrices. For details, see Chapter 2 of Seneta (1973).

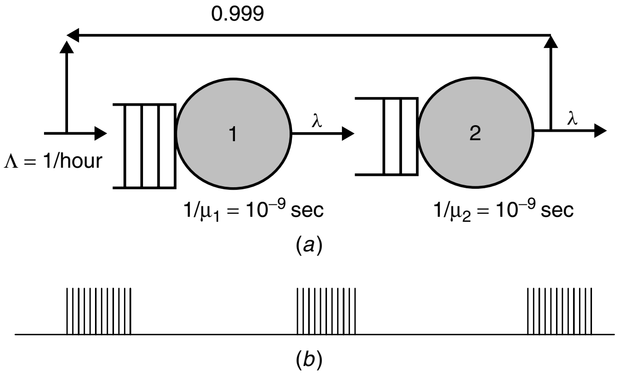


Figure 4.7 (a) Tandem queues with feedback; (b) typical sequence out of queue 1.

4.4.2 Detailed Derivation for a Two-Node Network

The theory of Jackson networks is based on multidimensional birth and death processes, which is a straightforward extension of the one-dimensional process studied in Chapter 3. The state of the system is described by a vector whose components are the number of messages in each of the queues. As in the one-dimensional case, at most one event can occur in an incremental interval. Thus, the component of a vector can change by at most one. The number of elements in the vectors that can change in an incremental interval is limited to at most two, corresponding to the single event of a transfer from one queue to another. A two-node network illustrating these concepts is shown in Figure 4.8. External traffic arrives at each of two nodes at rates λ_i; $i = 0, 1$, respectively. The routing between the queues is probabilistic with probabilities q_{01} and q_{10} as indicated. The traffic equation for this network can be written

$$\begin{aligned}\Lambda_0 &= \lambda_0 + q_{10}\Lambda_1 \\ \Lambda_1 &= \lambda_1 + q_{01}\Lambda_0\end{aligned}$$

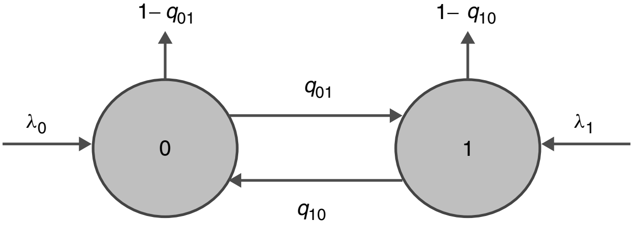


Figure 4.8 Two-node network.

The *state* of the system is the number of messages at each of the nodes. We define the probability $P(Q_0(t) = k_0, Q_1(t) = k_1) = P(k_0, k_1; t)$ as the probability of k_i; $i = 0, 1$ messages in the respective nodes at time t. We can derive the *Kolmogorov* differential equation for $P(k_0, k_1; t)$ using the same approach as in Section 3.4 for the one-dimensional case. For simplicity, we assume that there is only one exponential server in each node. We recognize that the probability of the state k_0, k_1 at time $t + \delta$ is the sum of the probabilities of seven disjoint events. For the case $k_0, k_1 > 0$, we write

$$\begin{aligned}P(k_0, k_1; t + \delta) = {} & P(k_0 - 1, k_1; t)\lambda_0\delta + P(k_0, k_1 - 1; t)\lambda_1\delta \\ & + P(k_0 + 1, k_1; t)\mu_0(1 - q_{01})\delta \\ & + P(k_0, k_1 + 1; t)\mu_1(1 - q_{10})\delta \\ & + P(k_0 + 1, k_1 - 1; t)\mu_0 q_{01}\delta \\ & + P(k_0 - 1, k_1 + 1; t)\mu_1 q_{10}\delta \\ & + P(k_0, k_1; t)(1 - \lambda_0\delta - \lambda_1\delta - \mu_0\delta - \mu_1\delta) \qquad (4.14)\end{aligned}$$

The terms on the RHS of (4.14) describe the state at time t and events in the interval $(t, t + \delta)$, which lead to the state k_0, k_1. The first two terms involve arrivals; the second two, departures from the system. The third pair of terms pertains to transfers between queues. The final term gives the probability of neither an arrival nor departure in the interval $(t, t + \delta)$. Equation (4.14) still holds for the cases $k_0 = 0$ and/or $k_1 = 0$ if we set probabilities with negative arguments equal to 0 and disallow departures from an empty queue. This shortcut dispenses with unnecessary detail. The next step in the derivation is to carry the term $P(k_0, k_1; t)$ to the LHS, divide both sides by δ, and let $\delta \to 0$. The result is the following differential equation:

$$\begin{aligned}\frac{dP(k_0, k_1; t)}{dt} = {} & P(k_0 - 1, k_1; t)\lambda_0 + P(k_0, k_1 - 1; t)\lambda_1 \\ & + P(k_0 + 1, k_1; t)\mu_0(1 - q_{01}) + P(k_0, k_1 + 1; t)\mu_1(1 - q_{10}) \\ & + P(k_0 + 1, k_1 - 1; t)\mu_0 q_{01} + P(k_0 - 1, k_1 + 1; t)\mu_1 q_{10} \\ & - P(k_0, k_1; t)(\lambda_0 + \lambda_1 + \mu_0 + \mu_1)\end{aligned}$$

Our interest is in steady-state probabilities; accordingly, we set $dP(k_0, k_1; t)/dt = 0$ and rearrange terms to obtain

$$\begin{aligned}P(k_0, k_1)(\lambda_0 + \lambda_1 + \mu_0 + \mu_1) = {} & P(k_0 - 1, k_1)\lambda_0 + P(k_0, k_1 - 1)\lambda_1 \\ & + P(k_0 + 1, k_1)\mu_0(1 - q_{01}) \\ & + P(k_0, k_1 + 1)\mu_1(1 - q_{10}) \\ & + P(k_0 + 1, k_1 - 1)\mu_0 q_{01} \\ & + P(k_0 - 1, k_1 + 1)\mu_1 q_{10}; \quad \forall k_0, k_1 \geq 0 \qquad (4.15)\end{aligned}$$

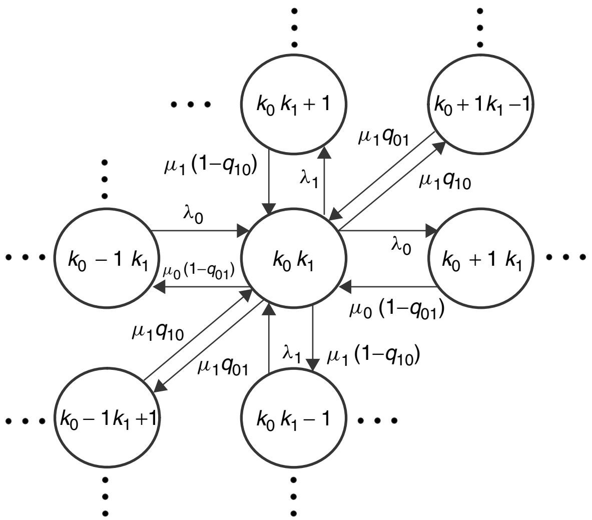


Figure 4.9 State transition flow diagram—two node network.

Note that we have removed the time dependence in this equation. The form here is one that we have seen in Section 3.4.1 (of Chapter 3). On the LHS we have flow out of the state k_0, k_1, while on the RHS we have flow into the state. The name for (4.15) with the widest currency is the *global balance equation*.[4] The term *equilibrium equation* is also used. The form is one that we have seen in Chapter 3. The state transition flow diagram is as shown in Figure 4.9.

Equation (4.15) represents a set of linear equations in the unknowns $P(k_1, k_2)$. These equations together with the normalizing condition $\sum_{k_1,k_2\geq 0} P(k_1, k_2) = 1$ can be solved, in principle, by standard matrix techniques. However, the form of (4.15) is such that the solution turns out to be quite simple. We substitute for λ_0 and λ_1 in (4.15) using the traffic equation (4.11) to obtain an equation in terms of Λ_0 and Λ_1, $\lambda_i = \Lambda_i - \sum_{j=0}^{1} \Lambda_j q_{ji}$; $i = 0, 1$. We switch all terms with a negative sign to the other side of the equation and rearrange the terms in a certain order:

$$\begin{aligned}
P(k_0, k_1)&(\Lambda_0 + \Lambda_1 + \mu_0 + \mu_1) + P(k_0 - 1, k_1)q_{10}\Lambda_1 \\
&+ P(k_0, k_1 - 1)q_{01}\Lambda_0 + P(k_0 + 1, k_1)\mu_0 q_{01} + P(k_0, k_1 + 1)\mu_1 q_{10} \\
&= P(k_0 + 1, k_1)\mu_0 + P(k_0, k_1 + 1)\mu_1 \\
&+ P(k_0 - 1, k_1)\Lambda_0 + P(k_0, k_1 - 1)\Lambda_1 \\
&+ P(k_0 + 1, k_1 - 1)\mu_0 q_{01} + P(k_0 - 1, k_1 + 1)\mu_1 q_{10} \\
&+ P(k_0, k_1)(q_{01}\Lambda_0 + q_{10}\Lambda_1);\ \forall k_0, k_1 \geq 0 \qquad (4.16)
\end{aligned}$$

[4] For further details on the global balance equation, see Gelenbe and Mitrani (1980), Gelenbe and Pujolle (1987) and Robertazzi (1994).

There are eight pairs of terms consisting of terms on each side of (4.16). If all of these pairs are equal, (4.16) is satisfied. We can break down the pairs to two basic types:

$$\begin{aligned} P(k_0, k_1)\Lambda_0 &= P(k_0 + 1, k_1)\mu_0 \\ P(k_0, k_1 + 1)\mu_1 &= P(k_0, k_1)\Lambda_1 \end{aligned}$$

which can be written

$$\begin{aligned} P(k_0 + 1, k_1) &= \rho_0 P(k_0, k_1) \\ P(k_0, k_1 + 1) &= \rho_1 P(k_0, k_1) \end{aligned} \tag{4.17}$$

where $\rho_i = \Lambda_i/\mu_i$; $i = 0, 1$. With the normalizing condition, (4.17) leads directly to the solution

$$P(k_0, k_1) = (1 - \rho_0)(1 - \rho_1)\rho_0^{k_0}\rho_1^{k_1};\ k_0, k_1 = 0, 1, \ldots \tag{4.18}$$

Recall that $\sum_{i=0}^{\infty} \rho_j^i = 1/(1 - \rho_j)$.

Equation (4.17) is reminiscent of the local balance or detailed balance equations that we have seen in Chapter 3.[5] We find it convenient to use the term *local balance* in referring to these equations later.

Example 4.2 By way of example, let us work out some numbers for this network of two nodes. Suppose that the externally arriving traffic is given by $\lambda_0 = 2.0$ messages/s and $\lambda_1 = 1.0$ messages/s, respectively. The routings are $q_{01} = 0.4$ and $q_{10} = 0.5$. The resulting traffic equations are $\Lambda_0 = 2.0 + 0.5\Lambda_1$ and $\Lambda_1 = 1.0+\ 0.4\Lambda_0$. The solution is $\Lambda_0 = 3.5$ and $\Lambda_1 = 2.5$. If the service rates in the nodes are $\mu_0 = 0.7$ and $\mu_1 = 1.5$, respectively, then, $\rho_0 = 0.2$ and $\rho_1 = 0.6$. The joint distribution is $P(k_0, k_1) = 0.32(0.2)^{k_0}(0.6)^{k_1}$; $k_0, k_1 = 0, 1, \ldots.$ The average number of messages in node 0 is 0.25 and in node 1 is 1.5.

4.4.3 *N*-Node Open Jackson Networks

We now consider the general case of open Jackson networks. We are given the probalistic routing matrix $Q = \{q_{ij}\}$ and the set of flows from external sources, λ_i; $i = 1, 2, \ldots, N$, which are assumed to be Poisson processes. The solution of (4.11) gives the set of total flows into each node of the network, Λ_i; $i = 1, 2, \ldots, N$. We assume that there is infinite storage at each of the nodes. Finally, we assume that

[5] For a good discussion of balance equations, see Nelson (1995), Chapter 10.

there are S_i servers at node i, each having an exponentially distributed service time for a message; accordingly, the departure rate is

$$\mu_i d(k_i) = \begin{cases} \mu_i k_i; & k_i \leq S_i \\ \mu_i S; & k_i > S_i \end{cases} \tag{4.19}$$

The queue at node i is stable if

$$\Lambda_i < \mu_i S_i; \quad i = 1, 2, \ldots, N \tag{4.20}$$

Two extreme cases are of particular interest: $S_i = 1$, the single-server queue, and $S_i = \infty$, the infinite server queue. A single server corresponds to a single line out of a node, while the infinite server case would correspond to a large number of trunks leaving a node—a number so large that it is rare that all are occupied.

The state of the system is the number of messages in each of the nodes, $k_1, k_2, \ldots, k_N$. The Kolmogorov differential equation for the state probability in this more general system can be derived in the same way as that for the two-node system in Figure 4.8. Again our interest is in the steady-state solution; accordingly, we set derivatives equal to zero. The resulting equations are a generalization of (4.15). The probability of this equilibrium state is denoted as $P(k_1, k_2, \ldots, k_N) = P(Q_1 = k_1, Q_2 = k_2, \ldots, Q_N = k_N)$. With this notation, the global balance equation is quite simply written in terms of the flow out of state $k_1, k_2, \ldots, k_N$ equated to the flow into a state $k_1, k_2, \ldots, k_N$. We have

$$\begin{aligned} P(k_1, k_2, \ldots, k_N)\left[\sum_{i=1}^{N} (\lambda_i + \mu_i d(k_i))\right] &= \sum_{i=1}^{N} P(k_1, k_2, \ldots, k_i - 1, \ldots, k_N)\lambda_i \\ &+ \sum_{i=1}^{N} P(k_1, k_2, \ldots, k_i + 1, \ldots, k_N)\mu_i d(k_i + 1)q_{i,N+1} \\ &+ \sum_{i=1}^{N} \sum_{j=1}^{N} P(k_1, k_2, \ldots, k_i + 1, \ldots, k_j - 1, \ldots, k_N) \\ &\times \mu_i d(k_i + 1)q_{ij}; \quad \forall k_1, k_2, \ldots, k_N \geq 0 \end{aligned} \tag{4.21}$$

In order to simplify notation, we agree that probabilities that show a negative argument are zero and that there can be no departure from an empty queue.

We solve this equation using the same approach as the two-node example above. We first solve the traffic equation, (4.11), for λ_i:

$$\lambda_i = \Lambda_i - \sum_{j=1}^{N} \Lambda_j q_{ji}; \quad i = 1,2, \ldots, N \tag{4.22}$$

We also write $q_{i,N+1} = 1 - \sum_{j=1}^{N} q_{ij}$. These expressions are substituted into (4.21), and terms with negative signs are switched from one side to the other. After rearranging the terms in a particular order we find

$$
\begin{aligned}
&\sum_{i=1}^{N} P(k_1, k_2, \ldots, k_N)\Lambda_i + \sum_{i=1}^{N} P(k_1, k_2, \ldots, k_N)\mu_i d(k_i) \\
&\quad + \sum_{i=1}^{N}\sum_{j=1}^{N} P(k_1, k_2, \ldots, k_j - 1, \ldots, k_N)\Lambda_i q_{ij} \\
&\quad + \sum_{i=1}^{N}\sum_{j=1}^{N} P(k_1, k_2, \ldots, k_i + 1, \ldots, k_N)\mu_i d(k_i + 1) q_{ij} \\
&= \sum_{i=1}^{N} P(k_1, k_2, \ldots, k_i + 1, \ldots, k_N)\mu_i d(k_i + 1) \\
&\quad + \sum_{i=1}^{N} P(k_1, k_2, \ldots, k_i - 1, \ldots, k_N)\Lambda_i \\
&\quad + \sum_{i=1}^{N}\sum_{j=1}^{N} P(k_1, k_2, \ldots, k_i + 1, \ldots, k_j - 1, \ldots, k_N)\mu_i d(k_i + 1) q_{ij} \\
&\quad + \sum_{i=1}^{N}\sum_{j=1}^{N} P(k_1, k_2, \ldots, k_N)\Lambda_i q_{ij} \qquad (4.23)
\end{aligned}
$$

We have arranged the terms in (4.23) in order to do a certain matching of terms on each side. The global balance equation holds if the following local balance equations hold:

$$
\begin{aligned}
P(k_1, k_2, \ldots, k_N)\Lambda_i &= P(k_1, k_2, \ldots, k_i + 1, \ldots, k_N)\mu_i d(k_i + 1) \\
P(k_1, k_2, \ldots, k_N)\mu_i d(k_i) &= P(k_1, k_2, \ldots, k_i - 1, \ldots, k_N)\Lambda_i \\
P(k_1, k_2, \ldots, k_j - 1, \ldots, k_N)\Lambda_i &= P(k_1, k_2, \ldots, k_i + 1, \ldots, \\
&\qquad k_j - 1, \ldots, k_N)\mu_i d(k_i + 1) \\
P(k_1, k_2, \ldots, k_i + 1, \ldots, k_N)\mu_i d(k_i + 1) &= P(k_1, k_2, \ldots, k_N)\Lambda_i
\end{aligned}
$$

These are all of the same form; consequently, the four equations can be distilled to one:

$$
P(k_1, k_2, \ldots, k_i + 1, \ldots, k_N) = \frac{R_i}{d(k_i + 1)} P(k_1, k_2, \ldots, k_N) \qquad (4.24)
$$

where $R_i = \Lambda_i/\mu_i$.

From (4.24), it is evident that the solution can also be written

$$
P(k_1, k_2, \ldots, k_N) = G^{-1} \prod_{i=1}^{N} \frac{R_i^{k_i}}{\prod_{j=1}^{k_i} d(j)}; \quad \forall k_1, k_2, \ldots, k_N \geq 0 \qquad (4.25)
$$

where G is the normalizing constant, which ensures that the probabilities sum to one. Clearly, G depends on the number of servers at each node. For the single-server queue, the solution is simply

$$P(k_1, k_2, \ldots, k_N) = \prod_{i=1}^{N} (1 - R_i) R_i^{k_i}; \quad \forall k_1, k_2, \ldots, k_N \geq 0 \tag{4.26}$$

while in the case of infinite server queues, we have

$$P(k_1, k_2, \ldots, k_N) = \prod_{i=1}^{N} \frac{e^{-R_i} R_i^{k_i}}{k_i!}; \quad \forall k_1, k_2, \ldots, k_N \geq 0 \tag{4.27}$$

It should be clear that both (4.26) and (4.27) sum to one. It should also be clear that the case of a network with a different number of servers at each node is a straightforward extension.

The solution here is really of the same form as that of feedforward networks. Given the flows from external sources and the routing matrix, one solves the traffic equation to find the total flows into each node. The probability distribution of the number of messages in the node is as if the total flow were Poisson and the contents of each of the nodes are statistically independent.

Performance Calculations A useful model of a communication network is that of M/M/1 *queues*. (We will discuss this model in detail below.) The marginal distribution for each node is, respectively,

$$P(\text{node } i \text{ has } k_i \text{ messages}) = (1 - R_i) R_i^{k_i}; \quad i = 1, 2, \ldots, N \tag{4.28}$$

Now, if the nodes in the system being modeled can hold no more than M messages, the probability of buffer overflow may be approximated as

$$\begin{aligned} P(\text{node } i \text{ has } M \text{ or more messages}) &= \sum_{k_i = M}^{\infty} (1 - R_i) R_i^{k_i} \\ &= R_i^M; \quad i = 1, 2, \ldots, N \end{aligned} \tag{4.29}$$

First- and second-order statistics are also of interest in measuring performance. The mean of a sum of random variables is *always* equal to the sum of the means. In this case we have

$$E(\text{total number of messages in } N \text{ nodes}) = \sum_{i=1}^{N} \frac{R_i}{1 - R_i} \tag{4.30}$$

The variance of a sum of *independent* random variables is equal to the sum of the individual variances, and we have

$$\text{Var(number of message in } N \text{ nodes)} = \sum_{i=1}^{N} \frac{R_i}{(1 - R_i)^2} \tag{4.31}$$

If there are enough nodes, the central-limit theorem would be applicable to estimate the probability distribution of the total number of messages in the system as being Gaussian with the mean and the variance given by (4.30) and (4.31), respectively.

Example 4.3 An example of an open network with feedback is shown in Figure 4.10. The routing matrix for this network is given by

$$Q = \begin{bmatrix} 0 & 1.0 & 0 & 0 \\ 0.2 & 0 & 0.5 & 0.3 \\ 0 & 0 & 0 & 0.6 \\ 0.4 & 0 & 0 & 0 \end{bmatrix}$$

If the input traffic is given by $\boldsymbol{\lambda} = (2.0, 1.0, 0.5, 0.3)$, the solution to (4.11) is worked out on the accompanying spreadsheet and is seen to be $\boldsymbol{\Lambda} = (5.91667, 6.91667, 9.895833, 6.33333)$.

Now, assume that the average service times in nodes 1 to 4 are 0.1, 0.07, 0.03, and 0.075, respectively. The resulting loads in each node are calculated on the accompanying spreadsheet as 0.5917, 0.4842, 0.2969, and 0.475, respectively. The probabilities of any combination of messages in the nodes can be found by simply substituting in (4.28). (See spreadsheet.) Further, by use of (4.29) the probability of buffer overflow can be approximated. A plot of overflow as a function of size is shown in Figure 4.11.

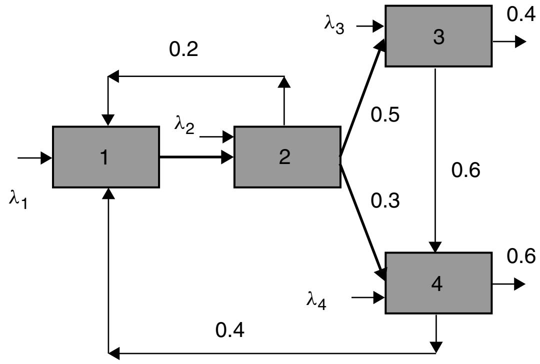


Figure 4.10 Open network with feedback.

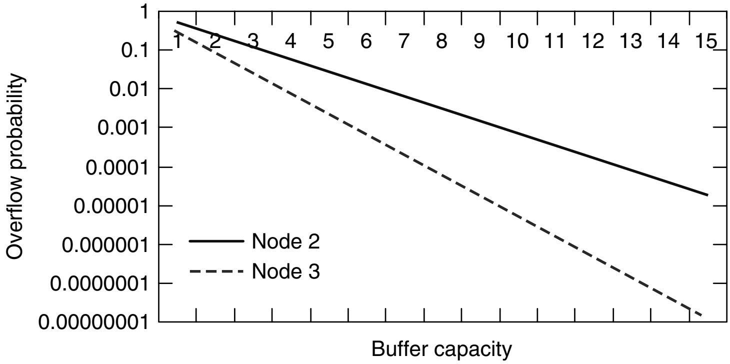


Figure 4.11 Overflow probability as a function of buffer size.

The mean and the variance of the number of messages in the network can be put to good use. Recall from Chapter 2 that the sum of independent random variables approached the Gaussian distribution as their number grew. Four nodes may be a little too few (we really want more than 10), but we go ahead anyhow in order to illustrate the approach. The mean and the standard deviation of the total number of messages in the network are 3.714574 and 2.77344, respectively. The probability of 10 or fewer messages in the network is 0.988283. The probability distribution is plotted on the spreadsheet. Note that the plot shows a continuous distribution, whereas the number of messages is a discrete random variable. Again we just are illustrating ideas. Later we show examples with many more nodes for which the approximation is more appropriate.

4.4.4 Average Message Delay in Open Networks

A widely used measure of performance in computer communication networks is *message delay*, which we define as the time interval between the arrival of a message to the network from an external source and its final departure from the network. We now consider the problem of calculating the average delay in a network.

The derivation of the average delay of a message is a nice application of Little's formula. The formula is used to find the average number of message in two ways: by summing over paths and by summing over nodes. Let $\bar{T}$ the average delay of a message in the network averaged over all starting nodes. The total rate of message flow into the network is just $\sum_{i=1}^{N} \lambda_i$. From Little's formula, the average number of messages in the network is

$$\bar{\rho} = \bar{T} \sum_{i=1}^{N} \lambda_i \tag{4.32}$$

We now compute this same quantity in two other ways. Let $\bar{T}_i$ indicate the average delay of a message that originates at node i; $i = 1, 2, \ldots, N$. Since messages enter the network at rate λ_i per second, the average number of messages in the network can be expressed as

$$\bar{\rho} = \sum_{i=1}^{N} \lambda_i \bar{T}_i \tag{4.33}$$

By equating the two expressions for the average number of messages, we find an expression for average delay

$$\bar{T} = \frac{1}{\alpha} \sum_{i=1}^{N} \lambda_i \bar{T}_i \tag{4.34}$$

where $\alpha = \sum_{i=1}^{N} \lambda_i$. Because of probabilistic routing, the path that messages follow in going through the network is complicated, and the calculation of $\bar{T}_i$ is not obvious. We find a tractable formula for average delay with another application of Little's formula. Assume for the moment that propagation delay is negligible. Let $\bar{D}_i$ denote the average delay of a message in node i. Recall that Λ_i is the total flow into node i, as found by solving the traffic equation. By Little's formula, the average number of messages in node i is $\Lambda_i \bar{D}_i$. Summing over all starting nodes, we find yet another expression for the total number of messages in the network:

$$\bar{\rho} = \sum_{i=1}^{N} \Lambda_i \bar{D}_i \tag{4.35}$$

From (4.32)–(4.35), we have the following equation for the average delay:

$$\bar{T} = \frac{1}{\alpha} \sum_{i=1}^{N} \Lambda_i \bar{D}_i \tag{4.36}$$

If each node has a single exponential server, the average delay of a message through such a node is

$$\bar{D}_i = \frac{\bar{M}_i}{1 - \rho_i}$$

where $\bar{M}_i$ is the average service time of a message and $\rho_i = \Lambda_i \bar{M}_i$. The average delay through the network is

$$\bar{T} = \frac{1}{\alpha} \sum_{i=1}^{N} \frac{\Lambda_i \bar{M}_i}{1 - \Lambda_i \bar{M}_i} \tag{4.37}$$

Example 4.4 Suppose that we have a five-node network with the routing matrix

$$Q = \begin{bmatrix} 0 & 0.25 & 0.1 & 0.3 & 0.15 \\ 0.4 & 0 & 0.125 & 0.25 & 0.05 \\ 0.35 & 0.1 & 0 & 0.2 & 0.1 \\ 0.3 & 0.05 & 0.1 & 0 & 0.05 \\ 0.1 & 0.4 & 0.2 & 0.2 & 0 \end{bmatrix}$$

and the input traffic $\lambda = [\,0.3 \quad 0.7 \quad 0.5 \quad 1.0 \quad 0.8\,]$. Suppose also that each server in the five nodes is exponential, having a uniform service rate of four messages per second. The example is worked out on the accompanying spreadsheet. The total flows into nodes 1, 2, 3, 4, and 5 are 3.076369, 2.502167, 1.791308, 3.252377, and 1.7328313, respectively. The average delay is 1.2063.

4.4.5 Store-and-Forward Message-Switched Networks

As we said at the beginning of this chapter, our interest in Jackson networks is partly motivated by their application to store-and-forward communication networks. In order to apply the results we have obtained, certain additional assumptions are necessary. In order to put these assumptions into the proper perspective, we need to consider a model of the store-and-forward node. Figure 4.12 depicts a network of three such nodes. Messages entering the node on an incoming link are fed directly to a central processor. The processor performs such tasks as checking for errors and determining the next node in the itinerary of a packet. Messages are routed from the central processor to a buffer feeding one of the output links to another node.

At both the central processor and the output links, messages are buffered; however, in many systems that are of interest, the speed of the central processor will be much faster than the transmission rates over the links. Thus, the processor queues can be ignored and the system can be modeled as consisting of *output buffer* queues, one for each link in the network. As indicated in Figure 4.13, there are six transmit buffers, one for each link in the network. The service rate is a function of the speed of the links connected to the buffer. The central processor routes messages to the appropriate output buffer.

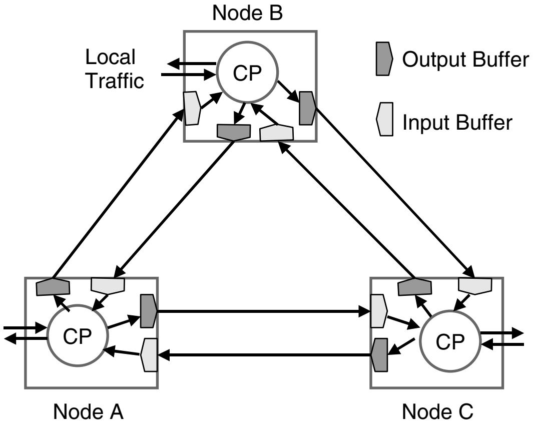


Figure 4.12 Three-node store-and-forward message-switched network.

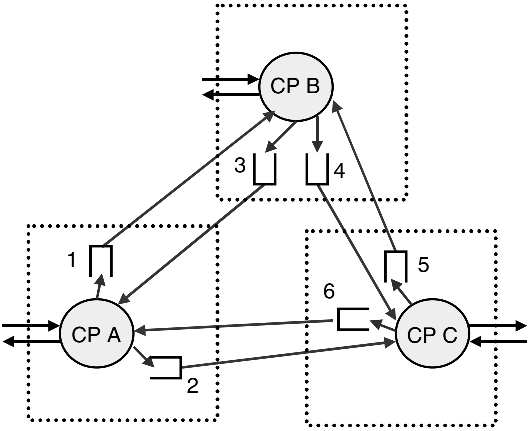


Figure 4.13 Three-node network.

Our focus is on the queueing delays in the output buffers; accordingly, we distill the network to that shown in Figure 4.14. The six nodes are mathematical constructs containing the output buffers served by the transmission lines. The routing function is also embodied in the node. Note that the nodes in Figure 4.14 are the mathematical entities that we have considered in the previous sections of this chapter. They carry out the functions of routing, buffering, and transmission, which actually reside in two of the physical nodes depicted in Figure 4.12. For example, the routing of messages from buffer 6 to buffers *i* and 2 is physically carried out in store-and-forward node *A* (Fig. 4.13), but in the mathematical model of Figure 4.14, routing is carried out in the same node as for buffer 2. The service time of a message is the time required to transmit it. We shall assume that this is an exponentially distributed random variable.

As stated above, messages are assembled and routed by central processors. Since the processor's operation is rapid, the arrival of messages to an output buffer is almost instantaneous and may be modeled as Poisson arrival. Clearly, the arrival

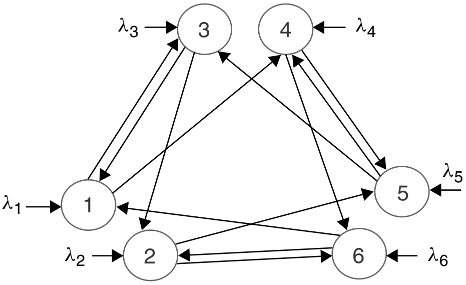


Figure 4.14 Six-node model of store-and-foward network.

of messages to a node is over a nonzero interval of time, and the arrival of several messages in a short time interval is impossible. By mixing together and assembling messages from several input lines, the Poisson arrival model is emulated. We also assume that the mixing of traffic is such that the routing may be modeled as being probabilistic.

The final assumption that we require is an *independence assumption*, which was made by Kleinrock (1964) in his classic work on store-and-forward networks. In the queueing networks we have dealt with up to this point, it was stated that the service times were associated with the server and that servers were independent. In communications networks this is not possible since the service time depends on the length of the message, which is carried from queue to queue. This introduces dependencies between the arrival process and the length of message or the service time. Under the independence assumption, it is assumed that the service time of a message is chosen independently at each node. The justification for this assumption is in networks where many different sources are feeding into the same node resulting in a mixing of message lengths. This assumption has been verified by means of Monte Carlo simulation on a digital computer.

We now derive a useful expression for average delay in the network. Suppose that the transmission rate in link i is C_i bps and the average number of bits in a message is $\bar{B}$. The average message transmission time on link i is $\bar{M}_i = \bar{B}/C_i$; $i = 1, 2, \ldots, N$. We substitute into (4.36) to get

$$\bar{T} = \frac{1}{\alpha}\sum_{i=1}^{N}\frac{\Lambda_i\bar{B}}{C_i - \Lambda_i\bar{B}} = \frac{1}{\alpha}\sum_{i=1}^{N}\frac{I_i}{C_i - I_i} \tag{4.38}$$

where $I_i = \Lambda_i\bar{B}$; $i = 1, 2, \ldots, N$ is the rate of transmitting information on link i. In deriving (4.38) we have neglected propagation delay between nodes. This is easily remedied. Let P_i be the delay in seconds on link i. The average number of messages in node i undergoing propagation delay is $\Lambda_i P_i$, consequently the average message delay taking into account both queueing and propagation is just

$$\bar{T} = \frac{1}{\alpha}\sum_{i=1}^{N}\left(\frac{I_i}{C_i - I_i} + \Lambda_i P_i\right) \tag{4.39}$$

In (4.39) we have expressions with quite tractable mathematical properties. It is the separable inasmuch as each term in the summation is a function only of quantities associated with a single link. Moreover, each term is a convex function in these quantities.

Another variation in this result is obtained by considering generalizations of the random component of the delay (Meister et al. 1971). We define the performance measure

$$\overline{T^k} = \left[\frac{1}{\alpha}\sum_{i=1}^{N}\left(\frac{I_i}{C_i - I_i}\right)^k\right]^{1/k} \tag{4.40}$$

This criterion reduces to average delay for $k = 1$. For $k = 2$, we have the standard deviation of flow in the network. This follows from the fact that the delay in an M/M/1 queue is exponentially distributed. For exponentially distributed random variables the variance is the square of the mean. Since the delays are assumed to be independent, from node to node, the variance of the sum is the sum of the variances. A final case of interest is for $k \to \infty$. In this case we have

$$\overline{T^k} = \lim_{k\to\infty}\left[\frac{1}{\alpha}\sum_{i=1}^{N}\left(\frac{I_i}{C_i - I_i}\right)^k\right]^{1/k} = \frac{I_{k^*}}{\alpha(C_{k^*} - I_{k^*})} \tag{4.41}$$

where k^* is the value of l for which $I_l/(C_l - I_l)$ is maximum over $l = 1, 2, \ldots, N$. The idea then is to allocate capacity so as to minimize the maximum link delay.

There are two basic applications for these results: capacity allocation and routing. Later, we will go through an example of capacity allocation where the I_i are fixed and the C_i are found so as to minimize delay. In the routing application, the capacities of the links, C_i, are fixed and the routing determining the values of I_i is found. The average delay given by one of the expressions in (4.38) and (4.41) is minimized over the set of I_i; $i = 1, 2, \ldots, N$ subject to a set of linear constraints ensuring conservation of flow in the nodes; that is, flow in equals flow out.

Example 4.5 Up to this point we have derived the conditions under which the product form solution is satisfied in a network of queues. We now illustrate these derivations by means of a numerical example involving the three-node network of Figure 4.12. The information that would be given is the network configuration, the amount of traffic generated at each store-and-forward node and the destination of this traffic. For a realistic network routing information would be either given or derived.

In our example it is assumed that there are no links between store-and-forward nodes B and C; accordingly, nodes 4 and 5, which represent output buffers carrying traffic between these nodes, are eliminated. The remaining links carry user information at a rate of 1.536 Mbps.[6] The traffic generated at each of the stations is $\Gamma_A = 1600$ messages/s, $\Gamma_B = 2000$ messages/s, and $\Gamma_C = 1800$ messages/s. The proportions of this traffic to each destination are tabulated below:

	A	B	C
A	—	0.25	0.75
B	0.6	—	0.4
C	0.2	0.8	—

In order to calculate any measure of performance, we must first find the flows in each of the four links out of nodes 1, 2, 3, and 6. From the table and inspection of the network, we see that these flows are $\Lambda_1 = 0.25\Gamma_A + 0.8\Gamma_C$, $\Lambda_2 = 0.75\Gamma_A + 0.4\Gamma_B$, $\Lambda_3 = \Gamma_B$, and $\Lambda_6 = \Gamma_C$. Substituting for Γ_A, Γ_B, and Γ_C and converting to

[6] The T1 line rate is 1.544 Mbps, of which 8000 bps is overhead.

bits, we find $I_1 = 0.78016$, $I_2 = 0.848$, and $I_4 = 0.7632$, where the units are Mbps. The capacity of each of the links is $C_i = 1.538$ Mbps; $i = 1, 2, 3, 6$. The total amount of traffic entering the network is $\alpha = 5400$ messages/s. Substituting into (4.38), we find the average message delay to be 216.5 μs.

4.4.6 Capacity Allocation

In Figure 4.15, an ATM switch with the attendant inputs and outputs are shown. The flow of traffic in an ATM network flow between source–destination pairs is organized into paths. The algorithms that control the flow of traffic assign capacity to these paths from a common channel. Of course, the sum of the capacities assigned to the paths must be less than or equal to the total capacity of the common channel. We have the constraint

$$C \geq \sum_{i=1}^{L} C_i \tag{4.42}$$

where C is the total capacity and C_i; $i = 1, 2, \ldots, L$ is the capacity in the L individual paths. We want to allocate the capacities C_i; $i = 1, 2, \ldots, L$ according to some criterion of performance. A reasonable criterion is average delay. We assume that the switch is the output queued variety and that most of the delay is in the output queue. All the other delays are common to all paths. We assume that, due to the mixing and splitting in the switching operation, message arrival to the output queue is a Poisson process. In order to carry the analysis forward, we assume that the messages have exponential length. Of course, this is inconsistent with the ATM context where cells are of constant length. As discussed above, the justification for the assumption is that it gives a formula that is mathematically easy to work with.

From the expression for average delay, (4.38)–(4.40), and the constraint on capacity, (4.42), the Lagrangian[7]

$$\frac{1}{\alpha}\sum_{i=1}^{L}\frac{I_i}{C_i - I_i} + \gamma \sum_{i=1}^{L} C_i$$

where γ is a Lagrange multiplier. We differentiate with respect to C_l, set the result equal to zero, and solve for the set C_l, $l = 1, 2, \ldots, L$. This gives the so-called square-root capacity assignment:

$$C_l = I_l + \left(\frac{I_l}{\gamma}\right)^{1/2}; \quad l = 1, 2, \ldots, L \tag{4.43}$$

The Lagrange multiplier γ is determined by the total cost constraint expressed by (4.42). We find

$$C_l = I_l + \frac{\left(C - \sum_{i=1}^{L} I_i\right) I_l^{1/2}}{\sum_{i=1}^{L} I_i^{1/2}}; \quad l = 1, 2, \ldots L \tag{4.44}$$

[7] See Fletcher (1987), p. 200.

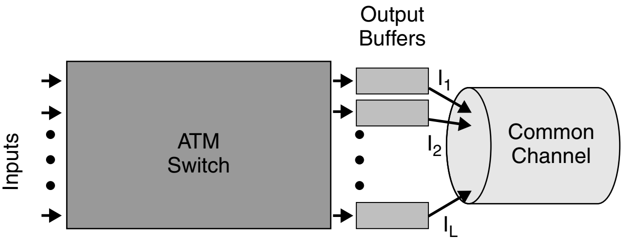


Figure 4.15 ATM multiplexing.

The solution presented by (4.44) stipulates that we allocate to the link its requirement I_l plus an excess in proportion to $(C - \sum_{i=1}^{L} I_i)$, which is the capacity remaining after the basic requirements are met in each link. Substituting into (4.38) we have for the average delay the expression

$$\bar{T} = \frac{1}{\alpha} \frac{(\sum_{i=1}^{L} I_i^{1/2})^2}{C - \sum_{i=1}^{L} I_i} \tag{4.45}$$

As one would expect, the delay is inversely proportional to the capacity margin. Recall that α is the total flow of traffic in messages per second. In this case, $\alpha = \sum_{i=1}^{L} I_i/424$, where 424 is the number of bits in an ATM cell.

Example 4.6 We now consider the capacity allocation problem in multiplexing ATM over the synchronous optical network (SONET) as illustrated in Figure 4.15. We have ten ATM streams feeding into an OC-12 line. On the accompanying spreadsheet, volumes of the individual streams are chosen at random and the optimum allocations are found for each set. We find the average delay for each allocation. A scatterplot shows the variation in delay. The mean and the variance are given in cells AC18 and AC19, respectively.

4.5 CLOSED JACKSON NETWORKS

4.5.1 Traffic Equation

We now consider *closed networks of queues* (Gordon and Newel, 1967), in which a fixed number of messages circulate within the network with neither arrivals to nor departures from the network. The classic application of the closed network model is the rudimentary model of the computer system depicted in Figure 4.16. A finite set of K jobs circulate between a file server and a set of workstations. Over a suitably short period of time, we can assume that jobs neither begin nor end so that there are neither departures nor arrivals. As in the case of open Jackson networks, we want to

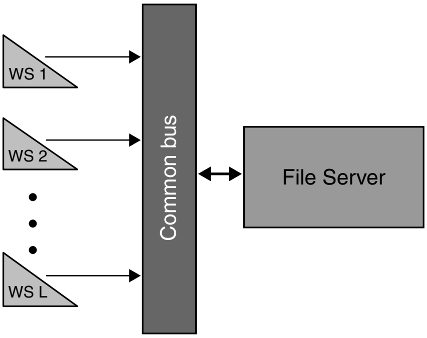


Figure 4.16 Model of file server and work stations.

find the probability distribution of the number of messages at each network node. We assume N nodes, each of which is modeled as an M/M/S queue. As in the case of open Jackson networks, we also assume probabilistic routing between nodes. Since there are no external arrivals, the traffic equation (4.12) becomes

$$\Lambda = \Lambda Q \tag{4.46}$$

where Q is a *stochastic matrix* in that all its rows sum to one, and $\sum_{k=1}^{N} q_{ik} = 1$; $i = 1, 2, \ldots, N$ since there are no departures from the network. With this condition on the routing matrix, there is a solution to the traffic equation for which at least one total flow is positive, $\Lambda_i > 0$. The solution is simply the eigenvector of Q corresponding to eigenvalue one. The solution is unique within a multiplicative constant. As we shall see, the multiplicative constant is absorbed in normalization. The solutions $\Lambda_1, \Lambda_2, \ldots, \Lambda_N$ represent the relative arrival rate of the message to the queues $1, 2, \ldots, N$, respectively. (Later, we will calculate the absolute flows into the nodes.)

Example 4.7 An example of a closed network is shown in Figure 4.17. The routing matrix for this network is given by

$$Q = \begin{bmatrix} 0 & 1.0 & 0 & 0 \\ 0.2 & 0 & 0.5 & 0.3 \\ 0.1 & 0.2 & 0 & 0.7 \\ 0.4 & 0 & 0.6 & 0 \end{bmatrix}$$

The normalized solution to (4.46) for this routing matrix is found on the associated spreadsheet as $\Lambda = (1.0, 1.306, 1.532, 1.464)$. Thus, for example, the K messages circulating in the network visit node 2 30.6% more often than node 1. We shall use this result to calculate the probability distribution of the messages among the queues for this example in the sequel. Later, we will also find the absolute values of flows.

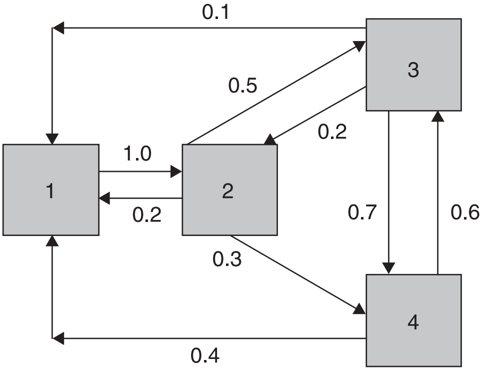


Figure 4.17 Closed network of queues.

4.5.2 Global Balance Equation—Solution

We now consider the global balance equations for closed networks of queues. This can be seen as an easy extension of the results for open networks. We assume that the servers in each of the network nodes are exponential. Once again, the global balance equation is written as the total flow into and out of the network states:

$$\sum_{i=1}^{N} \mu_i d(k_i) P(k_1, k_2, \ldots, k_N)$$

$$= \sum_{i=1}^{N} \sum_{j=1}^{N} \mu_i d(k_i + 1) q_{ij} P(k_1, k_2, \ldots, k_i + 1, \ldots, k_j - 1, \ldots, k_N) \qquad (4.47)$$

This is the same equation as (4.21) that would be obtained by letting $\lambda_i \to 0$ and $q_{i,N+1} \to 0$; $\forall i$ with the q_{ij}; $i, j = 1, 2, \ldots, N$ suitably renormalized. As this limit is approached, the solution for the joint probability is still the product form solution given in (4.26) and (4.27). After all, we are dealing with linear systems of equations in continuous variables; accordingly, there should be a continuity in the solution. We find

$$P(k_1, k_2, \ldots, k_N) = G(K, N)^{-1} \prod_{i=1}^{N} \left[R_i^{k_i} \Big/ \prod_{j=1}^{k_i} d(j) \right]$$

$$= \begin{cases} G(K, N)^{-1} \prod_{i=1}^{N} R_i^{k_i}; & \text{single servers} \\ G(K, N)^{-1} \dfrac{\prod_{i=1}^{N} R_i^{k_i}}{k_i!}; & \text{infinite servers} \end{cases} \qquad (4.48)$$

where $\Lambda_1, \Lambda_2, \ldots, \Lambda_N$ is the solution to the traffic equation. $R_i = \Lambda_i / \mu_i$; $i = 1, 2, \ldots, N$ and $G(K, N)$ is a constant to be determined. For a single server at each node, the denominator on the RHS of (4.48) is one, while for an infinite number of servers, it is just $k_i!$. The solution can be verified by substituting (4.48) into (4.47) and canceling like terms, we find

$$\sum_{i=1}^{N} \mu_i d(k_i) = \sum_{i=1}^{N} \sum_{j=1}^{N} \mu_i d(k_i + 1) q_{ij} \frac{R_i}{d(k_i + 1)} \frac{d(k_j)}{R_j}$$

$$= \sum_{i=1}^{N} \sum_{j=1}^{N} \mu_j d(k_j) q_{ij} \frac{\Lambda_i}{\Lambda_j} = \sum_{j=1}^{N} \frac{\mu_j d(k_j)}{\Lambda_j} \sum_{i=1}^{N} q_{ij} \Lambda_i$$

The application of (4.46) to the last term proves the identity.

4.5.3 Normalization Constant—Convolution Algorithm

In order for $P(k_1, k_2, \ldots, k_N)$ to be a valid probability, it must sum to one over all values of k_i. Conceptually this is simple; we sum (4.48) over all values of k_i such that $\sum_{i=1}^{N} k_i = K$ equate the sum to one and solve for $G(K, N)$. The problem is that summing over all possible combinations of k_i, turns out to be a formidable task since there are $\begin{pmatrix} N + K - 1 \\ N - 1 \end{pmatrix}$ ways for K messages to be arranged among the N nodes. This is a large number even for modest values of K and N. For example, if $N = 4$ and $K = 7$, there are 120 combinations. This difficulty is endemic to closed networks of queues. Fortunately, techniques have been found that considerably reduce the computational effort. These techniques are well suited to computation by spreadsheet. A number of these techniques are based on a convolution algorithm derived simultaneously (Buzen 1973, Reiser and Kobayashi, 1973). To begin, we shall consider this algorithm in its simplest form for the case of a single server at each network node. In this case, the service rate is independent of the number of messages in the node.

Define $S(k, n)$ as the set of $n = 1, 2, \ldots, N$ nonnegative integers that sum to k. $k = 1, 2, \ldots, K$:

$$S(k, n) = \left\{ k_1, k_2, \ldots, k_n \middle/ \sum_{i=1}^{n} k_i = k; \quad 0 \leq k \leq K; 1 \leq n \leq N \right\} \tag{4.49}$$

By summing over this set, $k_1, k_2, \ldots, k_n$, we define the quantity

$$G(k, n) = \sum_{S(k,n)} \prod_{i=1}^{n} R_i^{k_i} \tag{4.50}$$

This is the quantity that we want to evaluate as a function of k, n, and R_i. Consider what it represents. It is the sum over all possible ways of dispersing k messages among n nodes. We derive an algorithm for the computation of $G(k, n)$ by first splitting the summation in (4.50) into two parts based on the value of $k_n(= 0$ or $> 0)$, the number of message in the nth node:

$$G(k, n) = \sum_{\substack{S(k,n) \\ k_n=0}} \prod_{i=1}^{n} R_i^{k_i} + \sum_{\substack{S(k,n) \\ k_n>0}} \prod_{i=1}^{n} R_i^{k_i} \tag{4.51}$$

Consider each term on the RHS separately. Since the nth node is empty, the messages are spread only among $n - 1$ nodes. For the second term we can factor out R_n since we know that there is at least one message. Furthermore, we have the dispersal of $k - 1$ messages among n nodes. We have then the following iteration:

$$G(k, n) = G(k, n - 1) + R_n G(k - 1, n) \tag{4.52}$$

The initiating values for this replication are not difficult to derive. For the case $k \geq 1$, $n = 1$, all the messages reside at the one node, and we have

$$G(k, 1) = R_1^k; \quad k = 1, 2, \ldots, K \tag{4.53}$$

For $k = 0$, $n \geq 1$, there are no messages among the nodes; consequently, $\prod_{i=1}^{n} R_i^0 = 1$, and we have

$$G(0, n) = 1; \quad n \geq 1 \tag{4.54}$$

As a check, consider the case $k = 1$, $n \geq 1$, that is, one message among n nodes. Clearly there are only n ways to do this; consequently

$$G(1, n) = \sum_{i=1}^{n} R_i; \quad n = 1, 2, \ldots, N$$

This same equation could have been obtained by successive insertions of (4.54) into (4.52).

We can conceive of the iteration in (4.52) as filling an $K \times N$ array proceeding downward and to the right. It is never necessary to store more than N values. The iteration is represented in Figure 4.18. For the recent work on the recursive

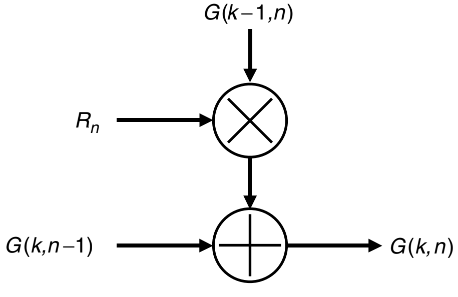


Figure 4.18 Convolutional algorithm, single-server case.

computation of normalisation constant with balanced fairness that ensures certain bandwidth sharing objective among competing flows in the network, see T. Bonald et al (2003).

Example 4.8 We can do an illustrative example of these computations. Suppose that we have a four-node network with the routing matrix

$$Q = \begin{bmatrix} 0 & 0.75 & 0.25 & 0 \\ 0.05 & 0 & 0.15 & 0.8 \\ 0.25 & 0.25 & 0 & 0.5 \\ 0.4 & 0.35 & 0.25 & 0 \end{bmatrix}$$

On the accompanying spreadsheet, the solution of (4.46) is shown to be $\Lambda = [1.0000 \quad 1.5844 \quad 0.9195 \quad 1.7273]$. We assume that there is a single server at each node with an average service time of 400 ms. Since we are calculating a normalization constant, all that matters is the relative values of the ρ_i. Now, let us assume that $K = 7$ messages are circulating among the four nodes of the network. Carrying out the iteration of Figure 4.17, we find the values for $G(k, n)$; $1 \leq k \leq K$; $1 \leq n \leq N$ shown on the associated Excel spreadsheet. We see that $G(7, 4) = 1.7036$, and the probability distribution is

$$P(k_1, k_2, k_3, k_4) = \frac{(0.4)^{k_1}(0.633756)^{k_2}(0.367812)^{k_3}(0.69098)^{k_4}}{1.7036}$$

The spreadsheet also has calculations for other quantities, which are now explained.

Mean Number of Messages in Each Queue It is serendipitous that intermediate values of $G(k, n)$, obtained in the course of calculating $G(K, N)$, are also useful. Consider first the probability that there are k or more messages in node i. If we "set aside" k messages at this node, the remaining $K - k$ messages are dispersed among the N nodes. We then have the following probability for this event:

$$\begin{aligned} P(Q_i \geq k) &= \frac{\sum_{\substack{S(K,N) \\ k_i \geq k}} \prod_{i=1}^{N} R_i^{k_i}}{G(K, N)} = \frac{R_i^k \sum_{S(K-k,N)} \prod_{i=1}^{N} R_i^{k_i}}{G(K, N)} \\ &= \frac{R_i^k G(K-k, N)}{G(K, N)}; \ k \geq 1 \end{aligned} \tag{4.55}$$

From (4.55) we have that the probability of *exactly* k messages at a node is given by

$$\begin{aligned} P(Q_i = k) &= P(Q_i \geq k) - P(Q_i \geq k+1) \\ &= \frac{R_i^k G(K-k, N) - R_i^{k+1} G(K-k-1, N)}{G(K, N)} \end{aligned} \tag{4.56}$$

The mean number of messages in the ith node is

$$E(K_i) = \sum_{k=0}^{K} kP(Q_i = k) = \sum_{k=0}^{K} \frac{k(R_i^k G(K-k, N) - R_i^{k+1} G(K-k-1, N))}{G(K, N)} \quad (4.57)$$

Now, we do a bit of manipulation in order to get a simpler form. We expand and change the variable, $j = k + 1$, in the second summation to find

$$E(K_i) = \sum_{k=0}^{K} \frac{kR_i^k G(K-k, N)}{G(K, N)} - \sum_{j=1}^{K} \frac{jR_i^j G(K-j, N)}{G(K, N)} + \sum_{j=1}^{K} \frac{R_i^j G(K-j, N)}{G(K, N)}$$

Notice that $G(i, N) = 0$ for $i < 0$. If we cancel terms, we have finally

$$E(K_i) = \frac{\sum_{k=1}^{K} R_i^k G(K-k, N)}{G(K, N)}; \quad i = 1, 2, \ldots, N \quad (4.58)$$

This calculation provides a good check. Since there are only K messages in the system, the sum of the averages in (4.58) over all nodes should be exactly K. We have found the averages for each node in Example 4.8 in the accompanying spreadsheet. We see that the results match since there are an average of seven messages in the network.

Absolute Flows As we have mentioned above, the solution of the traffic equation, (4.46), Λ_i; $i = 1, 2, \ldots, N$ gives only the relative value of flow through the nodes; however, we can use simple ideas that we developed previously to find the absolute values of these flows. Recall the simple relationship that we have derived in Chapter 2 for the relation between buffer occupancy and load, $P(\text{queue occupied}) = 1 - P_0 = \rho = \lambda\bar{M}$. From (4.55), we have that the probability of a node being occupied is given by

$$P(Q_i \geq 1) = P(Q_i > 0) = \frac{R_i G(K-1, N)}{G(K, N)}$$

Dividing by the average service time, we find for the absolute flow into the node

$$\Omega_i = \frac{R_i G(K-1, N)/G(K, N)}{\bar{M}_i} = \frac{\Lambda_i G(K-1, N)}{G(K, N)}; \quad i = 1, 2, \ldots, N \quad (4.59)$$

where Λ_i is the relative flow that is found as a solution to the traffic equation. The absolute flows are shown for Example 4.8 on the associated Excel spreadsheet.

Average Delays The average delay of a message through a node can be found from the application of Little's formula to (4.58) (average number) and (4.59) (average flow):

$$E(D_i) = \frac{E(K_i)}{\Omega_i} = \frac{\sum_{j=1}^{N} R_i^j (G(K-j, N)/G(K, N)}{\Lambda_i G(K-1, N)/G(K, N)} \tag{4.60}$$

Again, the example has been carried through on the associated spreadsheet.

4.5.4 Extension to the Infinite Server Case

The computation of the normalization constant for the case of infinite servers is simpler than that for the single server case. In this case the normalizing constant is

$$G(k, n) = \sum_{S(k,\,n)} \prod_{i=1}^{n} \frac{R_i^{k_i}}{k_i!} \tag{4.61}$$

In this case, finding the normalizing constant is a straightforward application of the multinomial expansion[8] with the result

$$G(k, n) = \frac{[\sum_{i=1}^{n} R_i]^k}{k!} \tag{4.62}$$

The joint distribution of the number of messages in each of the nodes is then

$$P(K_1 = k_1, K_2 = k_2, \ldots, K_N = k_N) = \frac{K! \prod_{i=1}^{N} R_i^{k_i}/k_i!}{[\sum_{i=1}^{N} R_i]^K} \tag{4.63}$$

The marginal distribution of the occupancy of the Nth node can also be found as an application of the multinomial expansion:

$$P(K_N = m) = R_N^m/m! \sum_{S(K-m,N-1)} \frac{\prod_{i=1}^{N-1} R_i^{k_i}/k_i!}{[\sum_{i=1}^{N} R_i]^K/K!} = \binom{K}{m} R_N^m \frac{[\sum_{i=1}^{N-1} R_i]^{K-m}}{[\sum_{i=1}^{N} R_i]^K} \tag{4.64}$$

This is easily applied to any other node. It is left as an exercise to show that the mean number of messages in the Nth node is given by

$$E(K_i) = \frac{KR_i}{\sum_{i=1}^{N} R_i}; \quad i = 1, 2, \ldots, N \tag{4.65}$$

Note that the sum of the averages over all nodes is equal to K. To find the absolute flows through a node, we apply Little's formula once again. Clearly,

[8] The multinomial expansion states that $(\sum_{i=1}^{n} x_i)^k = k! \sum_{S(k,n)} \prod_{i=1}^{n} (x^{k_i}/k_i!)$.

the average delay in a node for the infinite server case is just the average service time, $\bar{M}$; thus

$$\Omega_i = \frac{KR_i/\sum_{i=1}^{N} R_i}{\bar{M}_i} = \frac{K\Lambda_i}{\sum_{i=1}^{N} R_i}; \quad i = 1, 2, \ldots, N \tag{4.66}$$

Example 4.9 We change Example 4.8 by replacing the single servers by infinite servers. Of course, this change does not affect the solution to the traffic equation; consequently, the load in each node is the same. On the associated spreadsheet, we see that the average number of messages in each node is given by (1.3381, 2.1202, 1.2304, 2.3113). The average flow though the nodes is given by (3.3453, 5.3004, 3.076, 5.7783).

4.5.5 Mean Value Analysis of Closed Chains

Although it is not apparent from the preceding, there are numerical problems associated with the convolutional approach.[9] These difficulties may surface in the computation of large networks. An alternative approach to the convolutional algorithm is the mean value analysis. As its name implies, the mean value analysis yields averages rather than distributions; however, in many applications this is sufficient, particularly if computation is much simpler.

The technique is based on the arrival theorem[10] that states that, within a closed chain containing k messages, the distribution of the number of messages of its own class seen by a message arriving at a node is the steady-state distribution for the case of one less message in the chain, $k - 1$. In contrast, for Poisson arrivals in an open network, the steady-state distribution and the distribution seen by arriving messages are the same. We first apply this result to the single chain or ring network shown in Figure 4.19, in which each node has a single exponential server.

We start by assuming that the number of circulating messages is k. Let M_i denote the service times of messages in node i, $i = 1, 2, \ldots, N$. The delay of a message at a node is the sum of its own transmission time plus the transmission time of messages encountered on arrival. From the theorem we just quoted, an average of $\bar{n}_i(k-1)$ messages are encountered. The mean delay of an internal message in node i is given by

$$\bar{d}_i(k) = \bar{M}_i[1 + \bar{n}_i(k-1)]; \; k = 1, 2, \ldots, K; \quad i = 1, 2, \ldots, N \tag{4.67}$$

where $\bar{M}_i$ is the mean service time and $\bar{n}_i(k-1)$ is the mean number of messages in queue i in the steady state when there are $k - 1$ messages in circulation. From Little's formula we have for the throughput

$$\lambda(k) = \frac{k}{\sum_{i=1}^{N} \bar{d}_i(k)}; \quad k = 1, 2, \ldots, K \tag{4.68}$$

[9] For a general treatment on the subject, see Conway and Georganas (1989).

[10] See Reiser and Laverberg (1980).

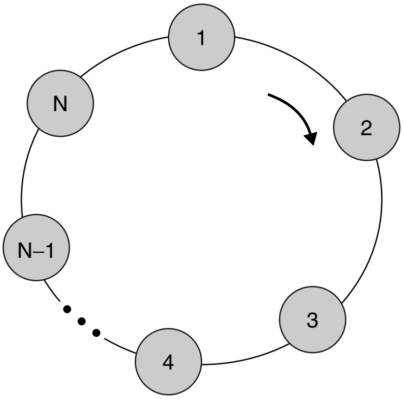


Figure 4.19 Single chain of exponential servers.

where $\sum_{i=1}^{N} \bar{d}_i(k)$ is the total delay around the chain. We also have from Little's formula for each node

$$\bar{n}_i(k) = \lambda(k)\bar{d}_i(k); \quad k = 1, 2, \ldots, K; \; i = 1, 2, \ldots, N \tag{4.69}$$

From (4.67)–(4.69) a recursive solution can be found:

$$\begin{aligned}
n_i(0) &= 0; \quad i = 1, 2, \ldots, N \\
\bar{d}_i(1) &= \bar{M}_i[1 + \bar{n}_i(0)] = \bar{M}_i; \quad i = 1, 2, \ldots, N \\
\lambda(1) &= \frac{1}{\sum_{i=1}^{N} \bar{d}_i(1)} \\
\bar{n}_i(1) &= \lambda(1)\bar{d}_i(1); \quad i = 1, 2, \ldots, N \\
&\vdots \\
\bar{d}_i(K) &= \bar{M}_i[1 + \bar{n}_i(K-1)]; \quad i = 1, 2, \ldots, N \\
\lambda(K) &= \frac{K}{\sum_{i=1}^{N} \bar{d}_i(K)} \\
\bar{n}_i(K) &= \lambda(K)\bar{d}_i(K); \quad i = 1, 2, \ldots, N
\end{aligned} \tag{4.70}$$

Example 4.10 We consider an example for a simple six-node ring network with 14 messages circulating, $N = 6$, $K = 14$. We assume that the mean service times for each of the three nodes are, respectively, 2.5, 0.75, 0.03, 0.2, 0.5, and

1.2. The results on the associated Excel spreadsheet show that the average number of messages in each of the six nodes is 12.29992, 0.428513, 0.012145, 0.086951, 0.249978, and 0.92249, respectively. Note that the sum over all nodes is 14.

4.5.6 Application to General Networks

In general, a network can be viewed as consisting of a number of closed chains, one for each source–destination pair. For example, what is called *credit-based* flow control can be viewed as having a fixed number of messages circulating between source–destination pairs. Let us represent the number of such chains as J. If in an N-node network there is a single class of traffic between each pair of nodes, then $J = N(N-1)$. Notice that in the general case, there may be more than one class of traffic circulating between a pair of nodes. Each chain may contain a different number of messages, which we denote as $K^j; j = 1, 2, \ldots, J$. Let $J(i); i = 1, 2, \ldots, N$ denote the set of chains having queue i in common, and let $Q(j); j = 1, 2, \ldots, J$ denote the set of queues in chain J. The average number of messages of each class in a node is a function of the total number of messages circulating in the network, in all R chains. We denote this average as

$$\bar{n}_i^j(K^1, K^2, \ldots, K^J); \quad i = 1, 2, \ldots, N; \quad j = 1, 2, \ldots, J$$

From the basic theorem, we have that the average number encountered at node i by a message in chain j is

$$\bar{n}_i^j(K^1, K^2, \ldots, K^j - 1, \ldots, K^J); \quad i = 1, 2, \ldots, N; \quad j = 1, 2, \ldots, J$$

Again the delay experienced by a message at a node is its own transmission time plus the transmission times of the messages already there. These latter would include messages from other chains. The total delay may then be expressed as

$$\bar{d}_i^j = \bar{M}_i^j(1 + \bar{n}_i^j(K^1, K^2, \ldots, K^j - 1, \ldots, K^J)); \\ i = 1, 2, \ldots, N; \quad j = 1, 2, \ldots, J \tag{4.71}$$

Again we have, from Little's formula, the two equations

$$\lambda(K^j) = \frac{K^j}{\sum_{i \in Q(j)} \bar{d}_i^j}; \quad j = 1, 2, \ldots, J \tag{4.72}$$

and

$$\bar{n}_i^j(K^1, K^2, \ldots, K^j) = \lambda(K^j) d_i^j; \quad i = 1, 2, \ldots, N; \quad j = 1, 2, \ldots, J \tag{4.73}$$

As in the one-dimensional case, a recursive solution can be found. We begin with the initial values

$$n_i^j(\mathbf{0}) = 0; \quad \mathrm{i} = 1, 2, \ldots, \mathrm{N}; \ \mathrm{j} = 1, 2, \ldots, \mathrm{J}$$

4.6 BCMP NETWORKS

4.6.1 Overview of BCMP Networks

In this section, we consider BCMP networks. The acronym stands for the initials of the authors, Baskette, Chandy, Muntz, and Palacios, of a well-known paper (Baskette et al. 1975), which consolidated and summarized results in the area. Jackson networks are a special case of BCMP networks. As noted in the previous chapter, routing among the nodes of the network is probabilistic. However, in addition to the first-come first-served (FCFS) exponentially distributed service studied in the previous chapter, there are three other kinds of service for which the product form of the joint distribution holds. The first of these is the *infinite-server model*, in which a message is immediately assigned a server as it enters the system. Thus, if there are n messages each with average duration $1/\mu$, they are all transmitted at a rate μn. The second service type is *processor sharing*, in which each message in the queue receives equal simultaneous service. Notice that for both processor sharing and the infinite-server case all messages are simultaneously in service. The final service type for which the product form holds is *preemptive resume last-come first-served* (LCFS), in which newly arrived messages are served immediately. Displaced messages are queued and resume where they were left off when the server is available again. When the service times of messages vary widely, these alternatives to the FCFS discipline may offer more satisfactory service. For example, when most messages are short, with only the occasional long message, processor sharing has the effect of giving priority to short messages.

In all three of these disciplines the product form holds for an arbitrary service distribution. All that is required is that the Laplace transform of the probability density be a rational function. A rational Laplace transform can approximate any service time distribution. Recall from Section 3.5 (of Chapter 3) that a service distribution with a rational Laplace transform can be realized by a Cox network. In fact, the probability distributions have an *insensitivity property* inasmuch as the occupancy distributions depend only on the mean service time and not on the service time distribution. A second major feature of BCMP networks is that more than one class of message is allowed. Furthermore, for each of these three service disciplines—processor sharing, infinite server, and LCFS—different classes of messages may have different service time distributions. Moreover, messages may also change class probabilistically; that is, in an incremental interval a message may go from one class to another with a given probability. Networks may be mixed with respect to class in that they may be closed for one class, no external arrivals, and no departures, and open for another. Finally, arrivals may be dependent on the state of the network, under certain conditions. We have seen a good example of such dependence models: limited storage.

The concept of probabilistic routing can be generalized to Markov routing. Under this discipline, customers are allowed to switch classes, in a probabilistic fashion as they are routed between nodes. A customer of class k leaving node i is switched to class l and routed to node j with probability q_{ij}^{kl}. The traffic equation can still be written. It is as though a customer that changes class at a node i leaves the network at that node and its replacement class enters at node j. There is a coupling among the classes of customers. If there are N nodes and C customer classes, the traffic equation is written $\Lambda_i^k = \lambda_i^k + \sum_{j=1}^{N} \sum_{l=1}^{C} \Lambda_j^l p_{ji}^{lk}$.

4.6.2 Single Node—Exponential Server

The subject of BCMP networks covers a lot of ground; accordingly, we deal with it in a piecewise fashion. In this and the next three subsections we shall consider the probability distribution of messages in a single node with Poisson arrivals with a single message class and with multiple message classes beginning in this subsection with the exponential server. When there is a single-server class, the probability distribution for the number of messages in a node is derived in Chapter 3 and is given by (3.26), which we repeat here, for convenience:

$$P(Q = n) = P(n) = (1 - \rho)\rho^n; \quad n = 0, 1, 2, \ldots \tag{4.74}$$

where $\rho = \lambda/\mu$, and μ and λ are the arrival and service rates, respectively.

Now, suppose that there are C message classes each with the same average service time $\bar{M} = 1/\mu$. Each class arrives at a Poisson rate with averages λ_i; $i = 1, 2, \ldots, C$. Since the sum of Poisson streams is Poisson, the aggregate arrival rate is Poisson with rate $\lambda = \sum_{i=1}^{C} \lambda_i$. Thus, probability distribution for the total number of messages of all classes is given by (4.74) with $\rho = \sum_{i=1}^{C} \lambda_i/\mu$. Now the probability that any one of these is from a particular class is proportional to the arrival rate for the class $p_i = \lambda_i / \sum_{i=1}^{C} \lambda_i = \lambda_i/\lambda$; $i = 1, 2, \ldots, C$. Conditioned on the total number of messages being m, the joint probability distribution for the number of messages of each class follows the multinomial distribution

$$P\left(N_1 = n_1, N_2 = n_2, \ldots, N_C = n_C / \sum_{i=1}^{C} N_i = m\right) = m! \prod_{i=1}^{C} \frac{p_i^{n_i}}{n_i!} \tag{4.75}$$

Multiplying by the distribution for the total number of messages given by (4.74), we find the joint distribution of the individual classes and the total number

$$P\left(N_1 = n_1, N_2 = n_2, \ldots, N_C = n_C, \ \sum_{i=1}^{C} N_i = m\right) = (1 - \rho)m! \prod_{i=1}^{C} \frac{\rho_i^{n_i}}{n_i!} \tag{4.76}$$

where $\rho_i = \lambda_i/\mu$; $i = 1, 2, \ldots, C$.

Implicit in (4.76) is the condition that $m = \sum_{i=1}^{C} n_i$. As a check, we can sum over all n_i satisfying this condition to find the distribution of (4.74).

4.6.3 Single Node—Infinite Server

We now turn to the case of an infinite number of servers at a node. The first step is to prove the *insensitivity* property for the infinite-server case for a single-message class. We show that for a Poisson arrival process with rate λ, the probability distribution for the number of messages in the system is given by

$$P(Q = m) = \frac{e^{-\rho}\rho^m}{m!}; \quad m = 0, 1, 2, \ldots \tag{4.77}$$

where $\rho = \lambda \bar{M}$ and $\bar{M}$ is the average message service time.

In this and subsequent cases, we assume that each server is modeled by the K-stage Cox network shown in Figure 4.20 where $q_0 = 1$ and $q_K = 0$. Recall that in Section 3.4 we derived the mean value for service time represented by this network as

$$\bar{M} = \sum_{m=1}^{K} \frac{\prod_{l=0}^{m-1} q_l}{\nu_m} \tag{4.78}$$

The server system consists of an unlimited number of servers of the form shown in Figure 4.20. We define the state of this system by the K-dimensional vector $(k_1, k_2, \ldots, k_K)$, where k_i; $i = 1, 2, \ldots, K$ is the number of messages in stage i. Messages arrive to the system with an average rate λ messages per second. An arriving message goes into the first stage of a server, one of which is always available. This flow is into the first stage of the state space. We define the flow into stage i as ω_i; $i = 1, 2, \ldots, K$. Clearly

$$\begin{aligned} \omega_1 &= \lambda \\ \omega_{i+1} &= \omega_i q_i; \quad i = 1, 2, \ldots, K-1 \end{aligned} \tag{4.79}$$

Next, we write the global balance equation for this system. The system departs from the state with either an arrival or a departure of a message from one of the stages. The state $(k_1, k_2, \ldots, k_K)$ is entered by an arrival, necessarily to the first stage or by a departure

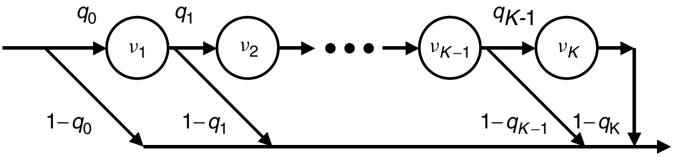


Figure 4.20 The method of stages.

from a stage. This departure can be from the system entirely or a shift to the next stage. We then have

$$\begin{aligned}&\left(\lambda + \sum_{i=1}^{K} \nu_i k_i\right) P(k_1, k_2, \ldots, k_K)\\ &\quad = \lambda P(k_1 - 1, k_2, \ldots, k_K) + \sum_{i=1}^{K} \nu_i (1 - q_i)(k_i + 1) P(k_1, k_2, \ldots, k_i + 1, \ldots, k_K)\\ &\quad + \sum_{i=1}^{K-1} \nu_i q_i (k_i + 1) P(k_1, k_2, \ldots, k_i + 1, k_{i+1} - 1, \ldots, k_K) \qquad (4.80)\end{aligned}$$

Now, we postulate a solution of the form

$$\nu_i k_i P(k_1, k_2, \ldots, k_K) = \omega_i P(k_1, k_2, \ldots, k_i - 1, \ldots, k_K); \quad i = 1, 2, \ldots, K \qquad (4.81)$$

Note that this is the local balance condition that we have seen previously. A form equivalent to (4.25) is

$$P(k_1, k_2, \ldots, k_K) = G^{-1} \prod_{i=1}^{K} \frac{\gamma_i^{k_i}}{k_i!} \qquad (4.82)$$

where $\gamma_i = \omega_i / \nu_i$; $i = 1, 2, \ldots, K$ and G is a constant to be determined. In order to test our solution, we substitute (4.82) into (4.80) and cancel terms to obtain

$$\lambda + \sum_{i=1}^{K} \nu_i k_i = k_1 \nu_1 + \sum_{i=1}^{K} (1 - q_i)\omega_i + \sum_{i=1}^{K-1} q_i \omega_i k_{i+1} \nu_{i+1} / \omega_{i+1} \qquad (4.83)$$

Employing (4.79) and canceling terms again, we find

$$\lambda = \sum_{i=1}^{K-1} (1 - q_i)\omega_i + \omega_K \qquad (4.84)$$

A repeated application of (4.79) shows that the RHS of (4.84) is λ, and we have verified our solution.

Since the stages are simply a mathematical device, and the number of messages in each step is not important, our interest is in the *total* number of messages in all stages, namely, the random variable

$$\sum_{i=1}^{K} k_i = m \qquad (4.85)$$

For the case $K = 2$, we have

$$P_r(k_1 + k_2 = m) = G^{-1} \sum_{i=1}^{m} \frac{\gamma_1^i \gamma_2^{m-i}}{i!(m-i)!}; \quad \forall m$$

An application of the binomial theorem yields

$$P_r(k_1 + k_2 = m) = G^{-1} \frac{(\gamma_1 + \gamma_2)^m}{m!}; \quad \forall m$$

It is an easy inductive proof to show that

$$P\left(\sum_{i=1}^{K} k_i = m\right) = G^{-1} \frac{\left(\sum_{i=1}^{K} \gamma_i\right)^m}{m!}; \quad \forall m$$

but

$$\rho = \sum_{i=1}^{K} \gamma_i = \sum_{i=1}^{K} \frac{\omega_i}{\nu_i} = \lambda \sum_{i=1}^{K} \frac{\left(\prod_{j=0}^{i} q_j\right)}{\nu_i} \tag{4.86}$$

We see the RHS of (4.86) is the product of the mean service time and the mean arrival rate; thus; there is no dependence on the probability distribution of the service time. The constant term, G is found from the usual normalization equation:

$$G^{-1} = \left[\sum_{m=0}^{\infty} P_r\left(\sum_{i=1}^{K} k_i = m\right)\right]^{-1} = e^{-\rho}$$

and we have

$$P(m) = P_r\left(\sum_{i=1}^{K} k_i = m\right) = \frac{e^{-\rho}\rho^m}{m!}; \quad \forall m \tag{4.87}$$

These results can be extended in an analogous fashion to nodes serving several different classes of messages, each with its own arrival rate and service time distribution. We carry out the derivation for the case of two classes. The extension to any number of classes will be obvious. The arrival rates for each of the classes are λ_i; $i = 1, 2$, respectively. The service times of the classes are represented by Cox networks with, respectively, K_i; $i = 1, 2$ stages, transition probabilities q_{ij}; $i = 1, 2; j = 0, 1, \ldots, K_i$, and average transition rates ν_{ij}; $i = 1, 2; j = 0, 1, \ldots, K_i$. The state of the system is $k_{11}, k_{21}; k_{12}, k_{22}; \ldots;$

k_{1K_1}, k_{2K_2}, where $k_{i,j}$; $i = 1, 2; j = 0, 1, \ldots, K_i$ is the number of class i messages that are in stage j. We write the global balance equation as

$$\begin{aligned}
&\left(\sum_{i=1}^{2}\left(\lambda_i + \sum_{j=1}^{K_i} \nu_{ij}k_{ij}\right)\right)P(k_{11}, k_{21}; k_{12}, k_{22}; \ldots; k_{1K_1}, k_{2K_2})\\
&\quad = \lambda_1 P(k_{11} - 1, k_{21}; k_{12}, k_{22}; \ldots; k_{1K_1}, k_{2K_2})\\
&\qquad + \lambda_2 P(k_{11}, k_{21} - 1; k_{12}, k_{22}; \ldots; k_{1K_1}, k_{2K_2})\\
&\qquad + \sum_{i=1}^{2}\sum_{j=1}^{K_i} \nu_{ij}(k_{ij} + 1)(1 - q_{ij})P(k_{11}, k_{21}; k_{12}, k_{22}; \ldots; k_{ij} + 1; \ldots; k_{1K_1}, k_{2K_2})\\
&\qquad + \sum_{i=1}^{2}\sum_{j=1}^{K_i - 1} \nu_{ij}(k_{ij} + 1)q_{ij}P(k_{11}, k_{21}; k_{12}, k_{22}; \ldots; k_{ij} + 1; k_{ij+1} - 1, \ldots; k_{1K_1}, k_{2K_2})
\end{aligned} \tag{4.88}$$

(In order to save some space, we have varied notation a bit on the RHS of (4.88).

On the basis of the same local balance argument that led to (4.82), we postulate a solution of the form

$$P(k_{1,1}, k_{2,1}; k_{1,2}, k_{2,2}; \ldots; k_{1,K_1}, k_{2,K_2}) = G^{-1}\prod_{i=1}^{2}\prod_{j=1}^{K_i}\frac{\gamma_i^{k_{i,j}}}{k_i!} \tag{4.89}$$

where $\gamma_{ij} = \omega_{ij}/\nu_{ij}$; $i = 1, 2; j = 1, 2, \ldots, K_i$, and ω_{ij} is the flow of class i customers into stage j. Substituting (4.89) into (4.88) and canceling like terms, we get an equation analogous to (4.84) with each term replaced by a pair of terms corresponding to each message class. The remainder of the derivation proceeds in the same way as in the single-class case. Again, as in the single-class case, we aggregate message in each of the stages to find the probability distribution for the total number of messages in each class. If there are C classes in total, the joint distribution is

$$P(n_1, n_2, \ldots, n_C) = e^{-\rho}\prod_{i=1}^{C}\frac{\rho_i^{n_i}}{n_i!}; \quad \forall n_i \geq 0 \tag{4.90}$$

where $\rho = \sum_{i=1}^{C}\rho_i$, $\rho_i = \lambda_i\bar{M}_i$; $i = 1, 2, \ldots, C$ and $\bar{M}_i$; $i = 1, 2, \ldots, C$ is the mean service time for the ith class. The probability distribution for the total number of message of all classes in the node is

$$P\left(\sum_{i=1}^{C} n_i = m\right) = \frac{e^{-\rho}\rho^m}{m!}; \quad m \geq 0 \tag{4.91}$$

4.6.4 Single Node—Processor Sharing

In this section, we consider the case of a single node in which the processor is shared between two classes of messages each with different service times. As will be evident presently, the extension to an arbitrary number of message classes is straightforward. For both classes, messages arrive at a Poisson rate with means, λ_i; $i = 1, 2$, respectively. Again the service times of the classes are represented by Cox networks with, respectively, K_i; $i = 1, 2$ stages, transition probabilities q_{ij}; $i = 1, 2$; $j = 0, 1, \ldots, K_i$, and average transition rates ν_{ij}; $i = 1, 2$; $j = 0, 1, \ldots, K_i$. As in the infinite server case, the state of the system is k_{11}, k_{21}; k_{12}, k_{22}; $\ldots$; k_{1,K_1}, k_{2,K_2}, where $k_{i,j}$; $i = 1, 2$; $j = 0, 1, \ldots, K_i$ is the number of class i messages which are in stage j. Because of the sharing of the processor, the departure rate of a message from a stage is $\nu_{ij}k_{ij}/m$, where $m = \sum_{i=1}^{2}\sum_{j=1}^{K_i} k_{ij}$ is the total number of messages in the system.

As in all the previous cases, we begin by writing the global balance equation:

$$\begin{aligned}
&\left(\sum_{i=1}^{2}\left(\lambda_i + \frac{\sum_{j=1}^{K_i}\nu_{ij}k_{ij}}{m}\right)\right)P(k_{11}, k_{12}, \ldots, k_{1K_1}; k_{21}, k_{22}; \ldots; k_{2K_2})\\
&\quad = \lambda_1 P(k_{11} - 1, k_{12}, \ldots, k_{1K_1}; k_{21}, k_{22}; \ldots; k_{2K_2})\\
&\quad + \lambda_2 P(k_{11}, k_{12}, \ldots, k_{1K_1}; k_{21} - 1, k_{22}; \ldots; k_{2K_2})\\
&\quad + \sum_{i=1}^{K_1}\nu_{1i}\frac{k_{1i}+1}{(m+1)}(1 - q_{1i})P(k_{11}, k_{12}, \ldots, k_{1i} + 1, \ldots, k_{1K_1}; k_{21}, k_{22}; \ldots; k_{2K_2})\\
&\quad + \sum_{i=1}^{K_2}\nu_{2i}\frac{k_{2i}+1}{(m+1)}(1 - q_{2i})P(k_{11}, k_{12}, \ldots, k_{1K_1}; k_{21}, k_{22}; \ldots, k_{2i} + 1, \ldots; k_{2K_2})\\
&\quad + \sum_{i=1}^{K_1-1}\nu_{1i}\frac{k_{1i}+1}{m}q_{1i}P(k_{11}, k_{12}, \ldots, k_{1i} + 1, k_{1i+1} - 1, \ldots, k_{1K_1}; k_{21}, k_{22}; \ldots; k_{2K_2})\\
&\quad + \sum_{i=1}^{K_2-1}\nu_{2i}\frac{k_{2i}+1}{m}q_{2i}P(k_{11}, k_{12}, \ldots, k_{1K_1}; k_{21}, k_{22}; \ldots, k_{2i} + 1, k_{2i+1} - 1, \ldots; k_{2K_2})
\end{aligned} \tag{4.92}$$

The term on the LHS of (4.92) indicates the departure from the state. The first two terms on the RHS are the arrival to the state due to the arrival of a message to the system. The second two terms represent a departure of a message from the system. Finally, the last two terms are for the shift of a message from one stage to another. We will now show that the solution to (4.92) is

$$\begin{aligned}
&P(k_{11}, k_{12}, \ldots, k_{1K_1}, k_{21}, k_{22}, \ldots, k_{2K_2})\\
&\quad = (1-\rho)m!\lambda_1^{k_1}\lambda_2^{k_2} \times \left[\frac{\prod_{i=1}^{K_1}\left(\prod_{j=0}^{i-1} q_{1j}/\nu_{1i}\right)^{k_{1i}}}{k_{1i}!}\right]\left[\frac{\prod_{i=1}^{K_2}\left(\prod_{j=0}^{i-1} q_{2j}/\nu_{2i}\right)^{k_{2i}}}{k_{2i}!}\right]
\end{aligned} \tag{4.93}$$

where $k_1=\sum_{i=1}^{K_1}k_{1i}, k_2=\sum_{i=1}^{K_2}k_{2i}$ and ρ is a constant to be determined. We substitute (4.93) into (4.92) and cancel common terms to get

$$\sum_{n=1}^{2}\left(\lambda_n+\sum_{j=1}^{K_n}\frac{\nu_{nj}k_{nj}}{m}\right)=\frac{\lambda_1k_{11}}{(\lambda_1q_{10}/\nu_{11})m}+\frac{\lambda_2k_{21}}{(\lambda_2q_{20}/\nu_{21})m}$$

$$+\sum_{i=1}^{K_1}\nu_{1i}(1-q_{1i})\frac{k_{1i}+1}{m+1}\frac{(m+1)\lambda_1}{k_{1i}+1}\prod_{j=0}^{i-1}q_{1j}/\nu_{1i}$$

$$+\sum_{i=1}^{K_2}\nu_{2i}(1-q_{2i})\frac{k_{2i}+1}{m+1}\frac{(m+1)\lambda_2}{k_{2i}+1}\prod_{j=0}^{i-1}q_{2j}/\nu_{2i}$$

$$+\sum_{i=1}^{K_1-1}\nu_{1i}q_{1i}\frac{k_{1i}+1}{m}\frac{\lambda_1\prod_{j=0}^{i-1}q_{1j}/\nu_{1i}}{k_{1i}+1}\frac{k_{1i+1}}{\lambda_1\prod_{j=0}^{i}q_{1j}/\nu_{1i+1}}$$

$$+\sum_{i=1}^{K_2-1}\nu_{2i}q_{2i}\frac{k_{2i}+1}{m}\frac{\lambda_2\prod_{j=0}^{i-1}q_{2j}/\nu_{2i}}{k_{2i}+1}\frac{k_{2i+1}}{\lambda_2\prod_{j=0}^{i}q_{2j}/\nu_{2i+1}}$$

After a number of obvious cancellations, we find

$$\sum_{n=1}^{2}\left(\lambda_n+\sum_{i=1}^{K_n}\frac{\nu_{ni}k_{ni}}{m}\right)=\frac{\nu_{11}k_{11}}{q_{10}m}+\frac{\nu_{21}k_{21}}{q_{20}m}+\sum_{i=1}^{K_1}\lambda_1\prod_{j=0}^{i-1}q_{1j}(1-q_{1i})$$
$$+\sum_{i=1}^{K_2}\lambda_2\prod_{j=0}^{i-1}q_{2j}(1-q_{2i})+\sum_{i=1}^{K_1-1}\frac{\nu_{1i+1}k_{1i+1}}{m}+\sum_{i=1}^{K_2-1}\frac{\nu_{2i+1}k_{2i+1}}{m} \tag{4.94}$$

We recognize that $q_{10}=q_{20}=1$ and $\sum_{i=1}^{K_1}\prod_{j=0}^{i-1}q_{1j}(1-q_{1i})=\sum_{i=1}^{K_2}\prod_{j=0}^{i-1}q_{2j}\times(1-q_{1i})=1$; hence, equality holds in (4.94) and the solution is given by (4.93).

Our interest is not in the number of messages in each stage but in the total number of messages of each class that are in the system; accordingly, we sum over all $k_{l,i}$ such that $\sum_{i=1}^{K_l}k_{li}=n_l$; $l=1, 2$. Continued application of the binomial theorem attains the result

$$P(n_1,n_2)=P\left(\sum_{i=1}^{K_l}n_{1i}=n_1,\ \sum_{i=1}^{K_2}n_{2i}=n_2\right)$$
$$=(1-\rho)(n_1+n_2)!\frac{\lambda_1^{n_1}}{n_1!}\left(\sum_{i=1}^{K_1}\prod_{j=0}^{i-1}\frac{q_{1j}}{\nu_{1i}}\right)^{n_1}\frac{\lambda_2^{n_2}}{n_2!}\left(\sum_{i=1}^{K_2}\prod_{j=0}^{i-1}\frac{q_{2j}}{\nu_{2i}}\right)^{n_2} \tag{4.95}$$

As in the previous cases, we substitute (4.78) to obtain

$$P(n_1, n_2) = (1 - \rho)(n_1 + n_2)! \frac{\lambda_1^{n_1}}{n_1!} \bar{M}_1^{n_1} \frac{\lambda_2^{n_2}}{n_2!} \bar{M}_2^{n_2} \tag{4.96}$$

where $\bar{M}_1$ and $\bar{M}_2$ and the mean service times for class 1 and 2 messages, respectively. Defining $\rho_i = \lambda_i \bar{M}_i$; $i = 1, 2$, (4.96) becomes

$$P(n_1, n_2) = \frac{(1 - \rho)\rho_1^{n_1} \rho_2^{n_2} (n_1 + n_2)!}{n_1! n_2!}$$

Another application of the binomial theorem shows that the normalization constant is given by $\rho = \rho_1 + \rho_2$. As the following manipulation shows, ρ is the product of the average arrival rate and the average message length, $\bar{M}$, assuming that $\bar{M}_1 = \bar{M}_2$.

If there were C classes of messages rather than two, it is not difficult to see that the joint probability density would be given by

$$P(n_1, n_2, \ldots, n_C) = (1 - \rho) m! \prod_{i=1}^{C} \frac{\rho_i^{n_i}}{n_i!} \tag{4.97}$$

where $\rho = \sum_{i=1}^{C} \rho_i = \sum_{i=1}^{C} \lambda_i \bar{M}_i$ and $\sum_{i=1}^{C} n_i = m$ is the total number of messages in the system. Note that this is the same form as the joint distribution for the queue with exponential service given by (4.76). If we were to sum (4.97) over all possible values of $(n_1, n_2, \ldots, n_C)$ such that $\sum_{i=1}^{C} n_i = m$, we would find that the total number in the system obeys (4.77).

4.6.5 Single Node—Last Come First Served (LCFS)

For the last case, we turn to the preemptive resume last come first served discipline. Consider now the state of the LCFS queue. For the LCFS discipline a message at the head of the line having reached a certain stage in the service process may be displaced by a new arrival. This message may be displaced in turn by a new message, and so on it goes. If there are m messages in the system, its state may be written $(l_1, l_2, \ldots, l_m)$. The $l_1, l_2, \ldots, l_m$ indicate the stages in the network that messages have attained, where l_1 is the stage of the message currently in service, l_2 is the next in line, and so on until l_m the stage of the last in line. Presumably, the second place message had reached stage l_2 when the message that is currently in service arrived. In writing the global balance equations, we distinguish two cases, $l_1 = 1$ and $l_1 > 1$. In either case the system departs the state with either an arrival to the system or a departure of the message in service from the first stage. The system enters the state $(1, l_2, l_3, \ldots, l_m)$ when there is an arrival to the system or when a message departs entirely and the second in line is in stage 1:

$$\begin{aligned} &P(1, l_2, \ldots, l_m)(\lambda + \nu_1) \\ &\quad = P(l_2, \ldots, l_m)\lambda + \sum_{n=1}^{K} P(n, l_2, \ldots, l_m)\nu_n(1 - q_n) \end{aligned} \tag{4.98}$$

We hypothesize that the solution to this equation is given by the joint density function

$$P(l_1, l_2, \ldots, l_m) = G^{-1} \prod_{i=1}^{m} \lambda \prod_{j=0}^{l_i-1} \frac{q_j}{\nu_{l_i}} \tag{4.99}$$

where G is a constant to be determined. Substituting (4.99) into (4.98) and canceling like terms, we find

$$\frac{\lambda}{\nu_1}(\lambda + \nu_1) = \lambda + \frac{\lambda}{\nu_1}\sum_{n=1}^{K} \lambda \prod_{j=0}^{n-1} \frac{q_j}{\nu_n}(\nu_n(1-q_n)) = \lambda + \frac{\lambda^2}{\nu_1}\sum_{n=1}^{K}\prod_{j=0}^{n-1} q_j(1-q_n) \tag{4.100}$$

Since we assume that $q_0 = 1$ and $q_K = 0$, it is a matter of simple manipulation to show that the summation on the RHS of (4.100) is equal to one and the equality follows.

Consider now the case $l_1 > 1$. As mentioned above, the departure from a state is the same. The arrival to the state $l_1, l_2, \ldots, l_m$ involves a shift from one stage to another or a departure from the system. The global balance equation is

$$\begin{aligned} P(l_1, l_2, \ldots, l_m)(\lambda + \nu_{l_1}) &= P(l_1 - 1, l_2, \ldots, l_m)\nu_{l_1-1} q_{l_1-1} \\ &\quad + \sum_{n=1}^{K} P(n, l_1, l_2, \ldots, l_m)\,\nu_n(1-q_n) \end{aligned} \tag{4.101}$$

The solution is the same as that given in (4.99). Substituting into (4.100) and canceling terms yields

$$\lambda + \nu_{l_1} = \frac{\nu_{l_1}}{q_{l-1}\nu_{l_i-1}} \nu_{l_i-1} q_{l-1} + \sum_{n=1}^{K} \lambda \prod_{j=0}^{n-1} \frac{q_i}{\nu_n} \cdot \nu_n(1-q_n) = \nu_{l_1} + \lambda \tag{4.102}$$

Thus, we have found the solution to the joint distribution of the number of messages in each stage.

We sum (4.99) over the stage values to find the distribution for the total number of messages in the node:

$$\begin{aligned} P(m) &= \sum_{l_1=1}^{K}\sum_{l_2=1}^{K}\cdots\sum_{l_m=1}^{K} P(l_1, l_2, \ldots, l_m) \\ &= \sum_{l_1=1}^{K}\sum_{l_2=1}^{K}\cdots\sum_{l_m=1}^{K} G^{-1}\prod_{i=1}^{m}\lambda\prod_{j=0}^{l_i-1}\frac{q_j}{\nu_{l_i}} \\ &= G^{-1}\prod_{i=1}^{m}\left(\sum_{l_i=1}^{K}\lambda\prod_{j=0}^{l_i-1}\frac{q_j}{\nu_{l_i}}\right) \end{aligned} \tag{4.103}$$

From (3.65), we see that the form of (4.103) is

$$P(Q = m) = G^{-1}(\lambda\bar{M})^m; \quad m = 0, 1, 2, \ldots \tag{4.104}$$

After normalization, we find the solution to be the one given in (4.74).

Finally, in our treatment of individual nodes we consider multiple classes of messages in a LCFS node. As in the previous cases, we assume two classes with Poisson arrivals having rate with averages λ_i; $i = 1, 2$, respectively. As previously, the service times of the classes are represented by Cox networks with, respectively, K_i; $i = 1, 2$ stages, transition probabilities q_{ij}; $i = 1, 2$; $j = 0, 1, \ldots, K_i$, and average transition rates ν_{ij}; $i = 1, 2; j = 0, 1, \ldots, K_i$. We begin with the assumption that the system contains m messages. The state of the system is $((c_1, l_1), (c_2, l_2), \ldots, (c_m, l_m)$, where c_i and l_i are, respectively, the class of the ith message and the stage attained in the service of the ith message. Again, we begin with the global balance equation for the case $l_1 = 1$. We have

$$\begin{aligned}
&P((c_1, 1), (c_2, l_2), \ldots, (c_m, l_m))(\lambda_1 + \lambda_2 + \nu_{c_{11}}) \\
&\quad = P((c_2, l_2), \ldots, (c_m, l_m))\lambda_{c_1} \\
&\qquad + \sum_{i=1}^{2}\sum_{n_i=1}^{K_i} P((i, n_i), (c_1, 1), (c_2, l_2), \ldots, (c_m, l_m))\nu_{in_i}(1 - q_{in_i})
\end{aligned} \tag{4.105}$$

The term (i, n_i) in the summation on the RHS of (4.105) indicates the class and the stage of the message in service. For the case $l_1 > 1$, we have a similar expression:

$$\begin{aligned}
&P((c_1, l_1), (c_2, l_2), \ldots, (c_m, l_m))(\lambda_1 + \lambda_2 + \nu_{c_{1l_1}}) \\
&\quad = P((c_1, l_1 - 1), (c_2, l_2), \ldots, (c_m, l_m))\nu_{c_1 l_1 - 1} q_{c_1 l_1 - 1} \\
&\qquad + \sum_{i=1}^{2}\sum_{n_i=1}^{K_i} P((i, n_i), (c_1, 1), (c_2, l_2), \ldots, (c_m, l_m))\nu_{in_i}(1 - q_{in_i})
\end{aligned} \tag{4.106}$$

The solution to (4.105) and (4.106) is

$$P((c_1, l_1), (c_2, l_2), \ldots, (c_m, l_m)) = G^{-1}\prod_{i=1}^{m}\lambda_{c_i}\prod_{j=0}^{l_i - 1}\frac{q_{c_i j}}{\nu_{c_i l_i}} \tag{4.107}$$

as can be seen by substitution. By following a now familiar process, we can aggregate members of the same class in different stages of the Cox network to find the joint distribution among the classes.

For a particular combination of m_1 number of class 1 messages and m_2 number of class 2 messages from the vector, $(c_1, l_1), (c_2, l_2), \ldots, (c_m, l_m)$, we have,

$$\begin{aligned}
&P\{m_1 \text{ of class 1 and } m_2 \text{ of class 2 from } (c_1, l_1), (c_2, l_2), \ldots, (c_m, l_m)\} \\
&\quad = \sum_{l_1=1}^{K_1}\sum_{l_2=1}^{K_1}\cdots\sum_{l_{m_1}=1}^{K_1}\sum_{l_{m_1+1}=1}^{K_2}\cdots\sum_{l_{m-1}=1}^{K_2}\sum_{l_m=1}^{K_2} P(l_1, l_2, \ldots, l_m) \\
&\quad = G^{-1}\prod_{i=1}^{m_1}\lambda_1\sum_{l_1=1}^{K_1}\prod_{j=0}^{l_i-1}\frac{q_{1j}}{\nu_{1l_i}}\prod_{r=1}^{m_2}\lambda_2\sum_{l_r=1}^{K_2}\prod_{j=0}^{l_r-1}\frac{q_{2j}}{\nu_{2,l_r}} \\
&\quad = G^{-1}(\lambda_1\bar{M}_1)^{m_1}(\lambda_2\bar{M}_2)^{m_2}
\end{aligned}$$

Since there are $\binom{m}{m_1}$ ways the same outcome can happen, we can write,

$$P\{Q_1 = m_1, Q_2 = m_2\} = (1-\rho)m! \frac{\rho_1^{m_1}}{m_1!} \frac{\rho_2^{m_2}}{m_2!}$$

where $m = m_1 + m_2$, $\rho = \rho_1 + \rho_2$. In the general case, where there are C classes, we have,

$$P(Q_1 = n_1, Q_2 = n_2, \ldots Q_C = n_C) = (1-\rho)m! \prod_{i=1}^{C} \frac{\rho_i^{n_i}}{n_i!} \tag{4.108}$$

where $\rho = \sum_{i=1}^{C} \rho_i = \sum_{i=1}^{C} \lambda_i \bar{M}_i$ and $m = \sum_{i=1}^{C} n_i$. Comparisons of (4.108) with (4.74) show *exactly* the same form. Adding together all of the different classes shows that the probability distribution for the total number of messages is simply given by (4.74).

We now summarize our results on single nodes. The listing is in the order, which is customary in the literature. From (4.77), (4.91), (4.97), and (4.108), we have

$$P(n_1, n_2, \ldots, n_C) = \begin{cases} \text{(1)} \quad \text{FCFS—exponential service:} \\ \qquad (1-\rho)m! \prod_{i=1}^{C} \rho_i^{n_i}/n_i! \\ \text{(2)} \quad \text{Processor sharing, arbitrary service:} \\ \qquad (1-\rho)m! \prod_{i=1}^{C} \rho_i^{n_i}/n_i!; \quad \rho_i = \dfrac{\lambda_i}{\mu_i} \\ \text{(3)} \quad \text{Infinite number of servers, arbitrary service:} \\ \qquad e^{-\rho} \prod_{i=1}^{C} \rho_i^{n_i}/n_i!; \quad \rho_i = \dfrac{\lambda_i}{\mu_i} \\ \text{(4)} \quad \text{Preemptive LCFS, arbitrary service:} \\ \qquad (1-\rho)m! \prod_{i=1}^{C} \rho_i^{n_i}/n_i!; \quad \rho_i = \dfrac{\lambda_i}{\mu_i} \end{cases} \tag{4.109}$$

4.7 NETWORKS OF BCMP QUEUES

We now move on to consider networks each of whose nodes are one of the four types considered above. As we shall demonstrate, the analysis of the performance of these networks is a marriage of the results that we have derived in the immediately preceding section with the results on Jackson networks that we obtained in Sections 4.4 and 4.5. As in the case of Jackson networks, it is assumed that routing is probabilistic and there is infinite storage at each of the queues. In the case of open networks, it is assumed that external arrivals are Poisson processes. In all cases, the arrival rate of messages to the node is the total arrival rate to the node found by solving the traffic equation, (4.11). As we shall show, the distribution of the total number of messages at each node is given by the product of expressions given in (4.109). To be specific, suppose that the disciplines at nodes 1 to i are drawn from (1), (2), or (4) of (4.109) and the disciplines in nodes $i+1$ to N from discipline (3) of (4.109). Let n_i be the number of messages in node i. If there is a single class of messages, the joint distribution of the number of messages at all of the nodes is of the product form given by

$$P(n_1, n_2, \ldots, n_N) = \prod_{j=1}^{i} (1 - R_j) R_j^{n_j} \prod_{k=i+1}^{N} \frac{e^{-R_k} R_k^{n_k}}{n_k!} \tag{4.110}$$

where $R_j = \Lambda_j \bar{M}_j$, where Λ_i is the solution to the traffic equation and $\bar{M}_j$ is the average service time at node j. The average delay of a message passing through the network is found in the same way as for Jackson networks. The only difference is that the average number of messages in the infinite server nodes is given by R_k; $k = i+1, i+2, \ldots, N$. For the others, it is $R_j/(1 - R_j)$; $j = 1, 2, \ldots, i$.

If the network is *closed*, the same considerations as in the Jackson network apply. If there are a fixed number K of messages circulating among the N nodes of the network, then it is necessary to find the normalization constant $G(K, N)$. If all of the nodes have service disciplines (1), (2) or (4) with single servers, the joint distribution is given by

$$P(n_1, n_2, \ldots, n_N) = G(K, N)^{-1} \prod_{i=1}^{N} R_i^{n_i} \tag{4.111}$$

The constant $G(K, N)$ can be found from the algorithm given in Section 4.5.3.

Now suppose that there are K or more servers at each node. The network would behave as though there were an infinite number of servers at each node. The joint distribution of messages is given by

$$P(n_1, n_2, \ldots, n_N) = G(K, N)^{-1} \prod_{i=1}^{N} \frac{R_i^{n_i}}{n_i!} \tag{4.112}$$

The normalizing constant is found by applying the multinomial expansion as in section 4.5.4. Since it is concerned only with mean values, the mean-value analysis proceeds in the same way as in Section 4.6.5.

The approach to proving the product form solution for BCMP networks should be quite familiar by now. The global balance equations for the network are written. It is postulated that the solutions to the global balance equations are the product of the solutions for individual nodes given in (4.109). Clearly, this procedure would entail a great deal of equation manipulation without a great deal of insight; accordingly, we content ourselves with examples drawn from telecommunications networks, which demonstrate the principles that are involved.

4.7.1 Store-and-Forward Message-Switched Nodes

In order to show the product form solution for BCMP networks, we consider a model for the store-and-forward node of a message-switched network, which is shown in some detail in Figure 4.21.

We assume that messages arrive at a Poisson rate. This assumption would not be tenable if there were a single input line since any number of messages could arrive in a small interval without allowance for the time of arrival of a message. The assumption relies on the effect of several input lines giving the appearance of Poisson arrival. Further, we assume that there is no limit on the storage in the node so that all arriving messages are accepted.

As mentioned above, messages enter the node from a number of input lines. The messages are first treated by the central processor, which checks for errors and determines whether there is space for the message in the node. Depending on the outcome of these tests, an ACK or a NACK is sent to the transmitting node. We can assume that both of these are piggybacked on information packets so that they have a negligible effect on traffic. The processor then determines the destination of an accepted message and places it one of O output channel buffers from which they are transmitted over the appropriate output line. In accordance with our model, a message is routed to output line i with probability q_i.

The central processor is modeled as a processor sharing discipline with a constant service time. At the output channel buffer the service time is the time required to transmit a message. This is assumed to be exponentially distributed. This is perhaps the most dubious of the assumptions that are required to analyze the node. If messages are of durations on the order of hundreds of bits or more, the effect of representing a discrete quantity by a continuous random variable should be negligible. The exponential assumption is simply necessary for mathematical tractability.

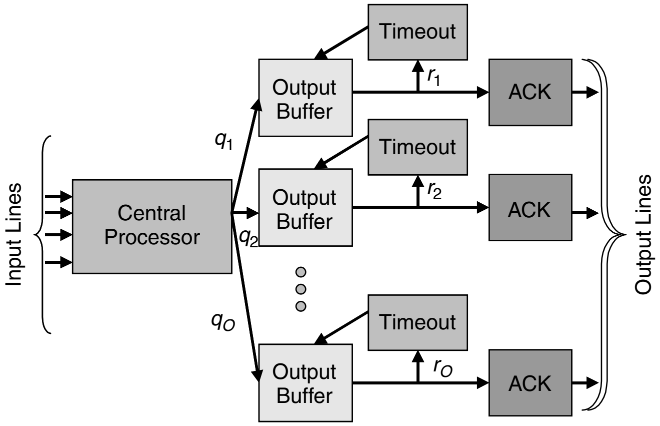


Figure 4.21 Store-and-forward packet-switching node.

The automatic repeat request (ARQ) process is modeled by the timeout and the ACK boxes shown in Figure 4.21. It is assumed that with probability r_j; $j = 1, 2, \ldots, O$ the attempted transmission over output channel j fails either because there is a channel error or there is not enough storage at the next node. We model this event as having the message enter the timeout box, where it resides for a random interval of time and then returns to the output buffer for retransmission. A successful transmission is modeled as having the message enter the ACK box for a random time interval after which it leaves the system. Of course, this event occurs with probability $1 - r_j; j = 1, 2, \ldots, O$. The residence times in both the timeout and the ACK boxes, respectively, represent the interval after transmission until a NACK or an ACK is received. This interval is composed of round trip propagation over the line and processing times at the receiver. Until a NACK or an ACK is received, messages are stored. Both the timeout and the ACK boxes are modeled as infinite server queues. The service times represent the duration of the protocol. Each may have different distributions. We assume that these residence times are independent of the service times in the central processor and the output channel buffer. These models for the timeout and the ACK boxes fit in with a selective reject policy where messages are held in the transmitter until a positive acknowledgment is received.

We now consider the traffic equation for the flows in the node. Let the flow from each of the input lines, in messages per second, be denoted as γ_i; $i = 1, 2, \ldots, I$, where I indicates the total number of input lines. We assume that a common buffer stores all of these input flows while they await processing. The total input into the central processor is $\Gamma = \sum_{i=1}^{I} \gamma_i$ messages per second. The total flow into output buffer i is

$$\Lambda_i^O = q_i \Gamma + \Lambda_i^T \tag{4.113}$$

where Λ_i^T is the total flow into timeout box i. The flows into the timeout box and the ACK box are, respectively

$$\begin{aligned} \Lambda_i^T &= r_i \Lambda_i^O; \quad i = 1, 2 \ldots, O \\ \Lambda_i^A &= (1 - r_i)\Lambda_i^O; \quad i = 1, 2 \ldots, O \end{aligned} \tag{4.114}$$

The solution to these equations is straightforward. Substituting into (4.113), we have

$$\Lambda_i^O = \frac{q_i \Gamma}{1 - r_i}; \quad i = 1, 2, \ldots O \tag{4.115}$$

Product Form Solution We shall now show that the preceding assumptions allow us to represent the node as a Jackson network for which the product form solution holds. The full network has a total of $3N + 1$ queues. However, the basic principles can be demonstrated by means of a much simpler network. We consider only an isolated branch of the network beginning with a channel output buffer. As a further simplification, we exclude the ACK box for the moment.

The network that results from the foregoing simplifications is the open network consisting of the output channel buffer and the timeout box, shown in Figure 4.22.

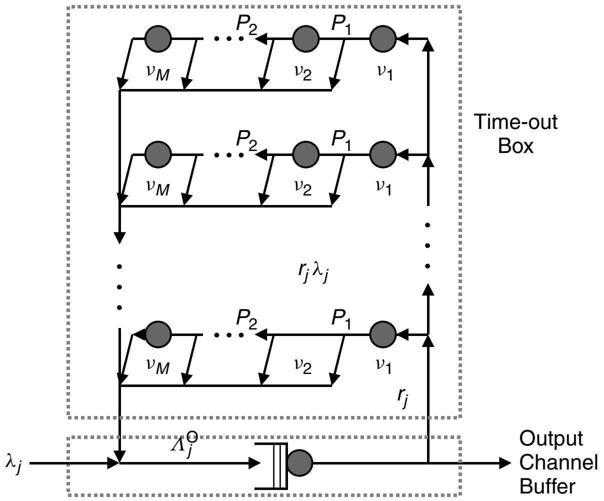


Figure 4.22 Portion of store-and-forward node.

The distribution can be represented by a Cox network. This representation is used in Figure 4.22 for each of the servers in the infinite-server queue representing the timeout box. There are a total of M stages for each server; after each stage, a message may depart. The average service time in the output channel buffer is l/ν_j seconds. The arrival rate from the central processor is Poisson with average $\lambda_j = q_j\Gamma; j = 1, 2, \ldots, O$ messages per second. We assume that messages that are not admitted are lost. After residence in the output channel buffer, messages are routed to the timeout box with probability r_j. The traffic equations giving the flow into portions of the network are easily found from observation. The flow into the output channel buffer is denoted as Λ_j^O, and the flow into the ith stage of the timeout box is denoted as ω_j. We have the following traffic equations for the branch:

$$\begin{aligned} \Lambda_j^O &= \sum_{i=1}^{M-1} \omega_i(1 - P_i) + \omega_M + \lambda_j \\ \omega_1 &= r_j\Lambda_j^O \\ \omega_{i+1} &= \omega_i P_i; \quad i = 1, 2, \ldots, M-1 \end{aligned} \tag{4.116}$$

The state of the network in Figure 4.22 is the $(M+1)$-dimensional vector $(n, k_1, k_2, \ldots, k_M)$, where n is the number of messages in the output channel buffer and k_i is the number of messages in the ith stage of the timeout box, $i = 1, 2, \ldots, M$. As in our previous study of networks of queues, we consider changes in an incremental interval. Because of the limited scope of the model, there are a limited number of events that can take place. A message can depart any stage of the timeout box and join the output channel buffer. Similarly, a message can depart the output channel buffer and join the timeout box or depart from the system entirely. Finally, an external message may arrive to the output channel buffer. We can write the global balance equation as

$$\begin{aligned}
&\left[\mu_j + \lambda_j + \sum_{i=1}^{M} k_i \nu_i\right] P(n, k_1, k_2, \ldots, k_M) \\
&\quad = r_j \mu_j P(n+1, k_1 - 1, k_2, \ldots, k_M) \\
&\qquad + (1 - r_j)\mu_j P(n+1, k_1, k_2, \ \ldots, k_M) \\
&\qquad + \lambda_j P(n-1, k_1, k_2, \ldots, k_M) + \sum_{i=1}^{M-1} \nu_i (1 - P_i)(k_i + 1) \\
&\qquad \times P(n-1, k_1, k_2, \ldots, k_i + 1, \ldots, k_M) \\
&\qquad + \nu_M (k_M + 1) P(n-1, k_1, k_2, \ldots, k_M + 1) \\
&\qquad + \sum_{i=1}^{M-1} \nu_i P_i (k_i + 1) P(n, k_1, k_2, \ldots, k_i + 1, k_{i+1} - 1, \ldots, k_M)
\end{aligned} \tag{4.117}$$

Note that k_i messages in the ith stage implies that the departure rate is $k_i \nu_i$. We understand, of course, that there can be no departures from an empty queue. The next step is to follow the same procedure as in Section 4.4.3. By substituting for λ_j using (4.116) and rearranging terms, we can show that the equilibrium equation is satisfied by the following local balance equations:

$$\begin{aligned}
\mu_j P(n, k_1, k_2, \ldots, k_M) &= \Lambda_j^O P(n-1, k_1, k_2, \ldots, k_M) \\
k_i \nu_i P(n, k_1, k_2, \ldots, k_M) &= \omega_i P(n-1, k_1, k_2, \ldots, k_i - 1, \ldots, k_M); \\
&\qquad\qquad\qquad i = 1, 2, \ldots, M
\end{aligned} \tag{4.118}$$

The solution to (4.118) is

$$P(n, k_1, k_2, \ldots, k_M) = G^{-1} \left(\frac{\Lambda_j^O}{\mu_j}\right)^n \prod_{i=1}^{M} \frac{\rho_i^{k_i}}{k_i!} \tag{4.119}$$

where G^{-1} is a normalizing constant that ensures that the probabilities sum to 1 and where $\rho_i = \omega_i / \nu_i$, $i = 1, 2, \ldots, M$. We shall consider the calculation of G^{-1} in due course. That (4.119) is a solution to the equilibrium equation (4.117) can be shown by substitution into (4.117) and application of the traffic equations (4.116). This is the same procedure as used in Section 4.3.

Our interest is not so much in the number of messages in individual stages as the total number in the timeout box. The form of the distribution allows a simple answer. Consider, for example, the joint distribution when there are only two stages in the timeout box:

$$P(n, k_1, k_2) = G^{-1} \left(\frac{\Lambda_j^O}{\mu_j}\right)^n \frac{\rho_1^{k_1}}{k_1!} \frac{\rho_2^{k_2}}{k_2!}$$

Summing over k_1 and k_2 such that $k_1 + k_2 = k$, we find that

$$P(n, k) = G^{-1}\left(\frac{\Lambda_j^O}{\mu_j}\right)^n \sum_{l=0}^{k} \frac{\rho_1^l}{l!}\frac{\rho_2^{k-l}}{(k-l)!} = P(0)\left(\frac{\Lambda_j^O}{\mu_j}\right)^n \frac{(\rho_1+\rho_2)^k}{k!}$$

Now let $k = \sum_{i=1}^{M} k_i$. By induction, it can be shown that the joint distribution of n and k is given by

$$P\left(n, \sum_{i=1}^{M} k_i = k\right) = G^{-1}\left(\frac{\Lambda_j^O}{\mu_j}\right)^n \frac{\rho_T^k}{k!} \tag{4.120}$$

where $\rho_T = \sum_{i=1}^{M} \rho_i$. Now consider the term ρ_T in (4.120). From (4.116), we have

$$\rho_T = \sum_{i=1}^{M} \frac{\omega_i}{\nu_i} = \omega_1 \sum_{i=1}^{M} \prod_{j=1}^{i-1} \frac{P_j}{\nu_i} = \omega_1 \bar{M}_T \tag{4.121}$$

Note that ω_1 is the input rate to the timeout box and $\bar{M}_T$ is the mean processing time of a message in the timeout box. Thus (4.120) and (4.121) show that the distribution of the total number in the timeout box is a function only of the mean and not of the distribution of the processing time.

These observations demonstrate that the product form solution holds for this simple network of BCMP queues in the same way that it did for Jackson networks (see Section 4.4.3). The form of the results obtained so far allows us to derive the probability distribution of message occupancy for the whole node. We begin with a single branch. The message arrival rates can be expressed in terms of the input rate. From (4.116), we have

$$\begin{aligned} \Lambda_j^O &= \frac{\lambda_j}{(1-r_j)} \\ \omega_1 &= r_j\Lambda_j^O = \frac{\lambda_j r_j}{(1-r_j)} \\ \Lambda_j^A &= (1-r_j)\Lambda_j^O = \lambda_j \end{aligned} \tag{4.122}$$

The form of the results obtained so far allows us to derive the probability distribution of message occupancy for the whole node. We begin with a single branch. By going through the same steps as above, we can show that the joint probability of n messages in the output channel buffer and l_j messages in the timeout and ACK boxes of the jth branch of the node is

$$P(n_j, l_j) = G^{-1}\left(\frac{\Lambda_j^O}{\mu_j}\right)^{n_j} \frac{[r_j\Lambda_j^O\bar{M}_T + \bar{M}_A\lambda_j]^{l_j}}{l_j!} \tag{4.123}$$

where $\bar{M}_A$ is the average time spent in the ACK box and λ_j is the flow in messages per second through the ACK box. This is a straightforward application of the familiar convolution algorithm that has most recently been used to obtain (4.120). Equation (4.123) makes the point that was made at the beginning of this section. The solution is in the form of the product of individual marginal distributions. Assume that the average duration of a message, in bits, is denoted as $\bar{B}$. The transmission rate in bits per second on each output line is given by C_i; $i = 1, 2, \ldots, O$. The average transmission time of a message is $1/\mu_j = \bar{B}/C_j; j = 1, 2, \ldots, O$.

The constant G^{-1} can be found simply by summing (4.123) over all values. We find

$$P(n_j, l_j) = \frac{(1-\rho_{O_j})\rho_{O_j}^{n_j} e^{-R_j} R_j^{l_j}}{l_j!} \;;\quad j = 1, 2, \ldots, O \tag{4.124}$$

where $R_j = r_j \Lambda_j^O \bar{M}_T + \bar{M}_A \lambda_j$ and $\rho_{O_j} = \Lambda_j^O \bar{B}/C_j; j = 1, 2, \ldots, O$.

It should be evident that the joint distribution of messages over all nodes simply has the product form. We indicate the number of messages in the central processor as n_0. We have

$$P(n_0, n_1, \ldots, n_M, l_1, \ldots, l_M) = (1-\rho_P)\rho_P^{n_0} \prod_{j=0}^{M} (1-\rho_{O_j})\rho_{O_j}^{n_j} e^{-R_j} R_j^{l_j}/l_j! \tag{4.125}$$

where ρ_P is the load in the central processor, $\rho_P = \Gamma\bar{p}$, the product of the total input rate to the node and the processing time of a message. We have used the fact that the central processor is modeled as a shared processor in writing (4.125). The form of (4.125) is that of the product of marginal distributions. The central processor and the output buffers have a geometric distribution, and the timeout and the ACK boxes have Poisson distributions of messages.

Total Number of Messages in the Node Up to this point we have assumed that there is no limitation on the number of messages that can be stored in the node. Now, we try to estimate the probability of message overflow. We assume that all the components use a common shared memory. The probability of overflow for the real, finite system is approximated as the probability of the number of messages in the infinite system exceeding the value of the available storage. When the probability is small, as it should be in a real system, the approximation is a good one.
Since the number of messages in the timeout and the ACK boxes are independent and Poisson, their sum is Poisson and we can write

$$P(n_0, n_1, \ldots, n_M, l) = \frac{(1-\rho_P)\rho_P^{n_0}\left[\prod_{j=0}^{M}(1-\rho_{O_j})\rho_{O_j}^{n_j}\right] e^{-R} R^l}{l!} \tag{4.126}$$

where l is the total number of messages in the timeout and the ACK boxes and

$$R = \sum_{j=1}^{M} (r_j \Lambda_j^O \bar{M}_T + \bar{M}_A \lambda_j)$$

The distribution for the total number of messages in the timeout and the ACK boxes, which is a convolution of geometric distributions, does not have a simple form. Of course, a numeric solution is possible, but tedious. A simpler approach is to call on the central-limit theorem, since we have the sum of a number of independent random variables. All that is required is to find the mean and the variances of the number of messages in the various components. Let the random number of messages residing in each component be denoted as N^P, N_i^O, N_i^T and N_i^A; $i = 1, 2, \ldots, O$, for the central processor, the output buffers, and the timeout and the ACK boxes, respectively. The means and variances of each of these quantities are, respectively

$$\begin{aligned} E(N^P) &= \frac{\rho_P}{1-\rho_P} \\ E(N_i^O) &= \frac{\rho_{O_i}}{1-\rho_{O_i}} \;; \quad i = 1, 2, \ldots, O \\ E(N_i^A + N_i^T) &= \rho_{AT_i} = \rho_{A_i} + \rho_{T_i}; \quad i = 1, 2, \ldots, O \end{aligned} \tag{4.127}$$

and

$$\begin{aligned} \mathrm{Var}(N^P) &= \frac{\rho_P}{(1-\rho_P)^2} \\ \mathrm{Var}(N_i^O) &= \frac{\rho_{O_i}}{(1-\rho_{O_i})^2}; \quad i = 1, 2, \ldots, O \\ \mathrm{Var}(N_i^A + N_i^T) &= \rho_{AT_i}; \quad i = 1, 2, \ldots, O \end{aligned} \tag{4.128}$$

The mean and the variance of the total number of messages in the node are the sum of these quantities. If the total is a Gaussian random variable, the calculation of the probability of overflow is straightforward (see Example 4.11).

Average Message Delay through a Node We now find the average delay of a message passing through a node. This is just the application of (4.36). We consider each component in turn. The central processor is modeled as a shared processor, which has the average delay

$$\overline{D_P} = \frac{\bar{P}}{1-\rho_P}$$

the total flow into the central processor is Γ. The average delay of a message in the output buffer is

$$\overline{D_j} = \frac{\bar{B}/C_j}{1-\rho_{Oj}} = \frac{\bar{B}}{C_j - \Lambda_j^O \bar{B}} = \frac{\bar{B}}{C_j - I_j}$$

where $I_j = \Lambda_j^O \bar{B}; j = 1, 2, \ldots, M$ is the average information flow on output line j. Since the ACK and the timeout boxes are modeled as infinite server queues, the average delays of messages is just $\overline{M_A^j}$ and $\overline{M_T^j}; j = 1, 2, \ldots, M$, respectively.

Putting all this together, we find for the average delay

$$\bar{D} = \frac{\bar{P}}{1 - \Gamma\bar{P}} + \frac{1}{\Gamma}\sum_{j=1}^{M}\left[\frac{I_j}{C_j - I_j} + \lambda_j\overline{M_A^j} + \frac{\lambda_j r_j}{1 - r_j}\overline{M_T^j}\right] \qquad (4.129)$$

In (4.129) we recognize that different conditions on the output line may require different timing in the timeout and the ACK boxes.

Example 4.11 We have worked out an example on the associated spreadsheet for the case of four input and four output lines to a node. The exogenous variables are the input line flows, the routings to the output lines, the retransmission probabilities, processing times, message length, line rates, and propagation times. We compute several quantities: including the average delay.

4.7.2 Example: Window Flow Control—A Closed Network Model

We consider an end-to-end window flow control scheme, which limits the number of messages flowing between a particular source–destination pair at any given point in time (Pennotti and Schwartz 1975). The path between source and destination is depicted in Figure 4.23. There are N nodes, which we shall model as having independent exponentially distributed service time with mean value $1/\mu_i$, $i = 1, 2, \ldots, N$. Several messages are in transit at any given point in time. The maximum number of such messages is W. When the limit has been reached, no new messages from the originating station may enter the path until one of these messages has been acknowledged. We model the forward direction and the return path as a chain as shown in Figure 4.24. The possible difference between message and

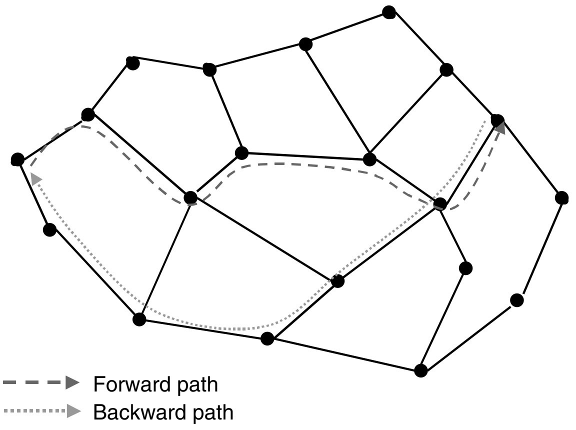


Figure 4.23 Forward and feedback paths through a network.

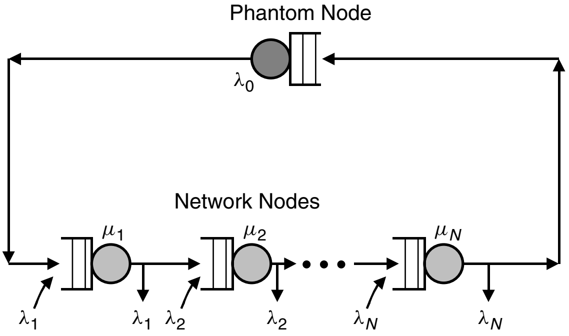


Figure 4.24 Logical network chain.

acknowledgment lengths may be taken into account by the different average service times in each node.

It is difficult to model the process of holding the messages until there is room in the chain and we seek a simplifying approximation. We assume that messages arrive at the source node at an average rate of λ_0 messages per second and that message arriving to a full chain are lost. This will allow us to calculate the throughput of the path between the source and the destination. This last model can be represented as shown in Figure 4.24. The original N nodes are augmented by a phantom node. The service rate of this node is λ_0 messages per second. Messages from the source–destination pair that we are interested in, circulate in this closed chain. If W internal messages are outstanding, the phantom node is empty and there can be no new arrivals to node 1. When there are less than W messages in nodes 1 to N, there is at least one message in the phantom node. In this case messages arrive to node 1 at a rate of λ_0 messages per second.

The flow between any source-destination pair in the network can be modeled by means of a closed chain. Clearly, the same node may be part of more than one chain. Traffic flowing in chains other than the one we are interested in is modeled as entering and departing as shown in Figure 4.24. The effect of this external traffic is to impede the flow within the chain. For simplicity, we shall assume that the same volume of traffic, λ_i in messages per second, enters the chain at the input to node i and departs the loop at the output of node i, $i = 1, 2, \ldots, N$. As we shall see, there is no loss of generality in this assumption. In keeping with its function in the model, there can be no external traffic in the phantom node. In this model the internal traffic and the external traffic form separate classes of messages. Because of the mixing of traffic within the network we assume that the independence assumption discussed above holds and the service times are independent from node to node.

With the internal traffic and the external traffic, we have a mixed network of queues. The state of this network is given by the number of messages of each class in each of the queues. We distinguish between externally arriving messages and messages that are circulating internally. The phantom node is designated as node

0. We define the state as $(k_0, k_1, \ldots, k_N; l_1, l_2, \ldots, l_N)$, where k_i, $i = 0, 1, \ldots, N$ and $l_j, j = 1, 2, \ldots, N$ are, respectively, the number of internal and external messages in node i. As in the case of an open network, there can be only a limited number of changes in the state in an incremental interval. An internal message may go from one queue to another within the chain. External messages arrive or depart. The equilibrium equation is then

$$\begin{aligned}
&\left(\sum_{i=0}^{N} \lambda_i + \sum_{i=1}^{N} \mu_i\right) P(k_0, k_1, k_2, \ldots, k_N; l_1, l_2, \ldots, l_N) \\
&\quad = \lambda_0 P(k_0 + 1, k_1 - 1, k_2, \ \ldots, k_N; l_1, l_2, \ldots, l_N) \\
&\qquad + \sum_{i=1}^{N} \lambda_i P(k_0, k_1, k_2, \ldots, k_N; l_1, \ldots, l_i - 1, \ldots, l_N) \\
&\qquad + \sum_{i=1}^{N-1} \mu_i \frac{k_i + 1}{k_i + l_i + 1} P(k_0, \ldots, k_i + 1, k_{i+1} - 1, \ldots, k_N; l_1, \ldots, l_N) \\
&\qquad + \mu_N \frac{k_N + 1}{k_N + l_N + 1} P(k_0 - 1, k_1, \ldots, k_N + 1; l_1, \ldots, l_N) \\
&\qquad \sum_{i=1}^{N} \mu_i \frac{l_i + 1}{k_i + l_i + 1} P(k_0, k_1, k_2, \ldots, k_N; l_1, \ldots l_i + 1, \ldots, l_N)
\end{aligned} \tag{4.130}$$

In Equation (4.130) terms such as $[(k_i + 1)/(k_i + l_i + 1)]$ reflect the portion of the contents of node i, which is internal traffic. Clearly, this equation cannot be true for all values of k_i and l_i. First, there can be no departures from an empty queue. If, for example, $k_i = 0$, then the term μ_i is omitted on the LHS. Furthermore, (4.130) holds only for states such that $\sum_{i=0}^{N} k_i = W$.

The flows of internal and external traffic into node i are λ_0 and λ_i messages per second, respectively. A sufficient condition for (4.130) to hold are the local balance equations, which involve a single node at a time:

$$\begin{aligned}
&\mu_i \frac{k_i}{k_i + l_i} P(k_0, k_1, \ \ldots, k_N; l_1, l_2, \ldots, l_N) \\
&\quad = \mu_{i-1} \frac{k_{i-1} + 1}{k_{i-1} + 1 + l_{i-1}} P(k_0, k_1, k_{i-1} + 1, k_i - 1, \ \ldots, k_N; l_1, l_2, \ldots, l_N); \\
&\quad i = 1, 2, \ldots, N
\end{aligned} \tag{4.131}$$

and

$$\mu_i \frac{l_i}{k_i + l_i} P(k_0, k_1, \ldots, k_N; l_1, l_2, \ldots, l_N)$$
$$= \lambda_i P(k_0, k_1, \ldots, k_N; l_1, l_2, \ldots, l_i - 1, \ldots, l_N); \quad i = 1, 2, \ldots, N$$

Simple substitution shows that the solution to these equations is of the product form given by

$$P(k_0, k_1, \ldots, k_N; l_1, l_2, \ldots, l_N) = G^{-1} \prod_{i=1}^{N} \left(\frac{\lambda_0}{\mu_i}\right)^{k_i} \frac{\rho_i^{l_i}(k_i + l_i)!}{k_i! l_i!} \tag{4.133}$$

Again, we see that the solution is a product of marginal distributions for individual nodes where G^{-1} is a normalizing constant such that the probabilities sum to one over the allowable states and where $\rho_i = \lambda_i/\mu_i$, $i = 1, 2, \ldots, N$. The value of G is determined by summing (4.133) over all allowable values of k_i, $i = 0, 1, \ldots, N$ and l_i, $i = 1, 2, \ldots, N$. For the latter variable this is simple. Since there is no restriction on storage in the node, and since external traffic flows in an open chain, there is no constraint on the numbers of external messages in the nodes 1 through N. We sum over l_i to obtain the following, after a rearrangement of terms:

$$P(k_0, k_1, \ldots, k_N) = G^{-1} \prod_{i=1}^{N} \left(\frac{\lambda_0}{\mu_i}\right)^{k_i} \sum_{l_i=0}^{\infty} \frac{\rho_i^{l_i}(k_i + l_i)!}{k_i! l_i!} \tag{4.134}$$

Consider one of the summations in (4.134) as a separate entity for the moment. As a further temporary simplification, we suppress the dependence on i:

$$\frac{1}{k!} \sum_{l=0}^{\infty} \frac{\rho^l (k + l)!}{l!} = \frac{1}{k!} \sum_{l=0}^{\infty} \frac{d^k(\rho^{k+l})}{d\rho^k} = \frac{1}{k!} \frac{d^k[\rho^k/(1-\rho)]}{d\rho^k} = \frac{1}{(1-\rho)^{k+1}} \tag{4.135}$$

where $d^k/d\rho^k$ is the kth derivative with respect to ρ. Substituting (4.135) into (4.134) in the same fashion, we find

$$\begin{aligned} P(k_0, k_1, \ldots, k_N) &= G^{-1} \prod_{i=0}^{N} \left(\frac{\lambda_0}{\mu_i}\right)^{k_i} \left[\frac{1}{(1-\rho_i)^{k_i+1}}\right] \\ &= \frac{G^{-1}}{\prod_{i=1}^{N}(1-\rho_i)} \prod_{i=0}^{N} \left[\frac{\lambda_0}{(\mu_i - \lambda_i)}\right]^{k_i} \\ &= G^{-1}(W, N) \prod_{i=0}^{N} R_i^{k_i} \end{aligned} \tag{4.136}$$

where $R_0 = 1$, by definition and $R_i = \lambda_0/(\mu_i - \lambda_i)$; $i = 1, 2, \ldots, N$.

In (4.136) the term $\prod_{i=0}^{N}(1-\rho_i)$ has been absorbed into the constant since it does not depend on k_i. The resulting equation shows that the only effect of the external traffic on the internal traffic is to reduce the rate at which internal messages are served from μ_i to $\mu_i - \lambda_i$. Clearly, stability requires that $\lambda_i < \mu_i$. The form of the probability distribution for internal messages is independent of the external traffic. The constant $G(W, N)$ can be found by means of the algorithm of (4.52) and Figure 4.18. The mean number of internal messages in each node can be found from (4.58).

Results Applied to Window Flow Control These results can be used to give insight to the window flow control technique. We focus on two measures of performance. As stated earlier, new internal messages are blocked when the phantom node is empty. The blocking probability is found from equation (4.56) with $k_0 = 0$ as

$$P_B = P(k_0 = 0) = 1 - \frac{G(W-1, N)}{G(W, N)} \tag{4.137}$$

The second measure of performance that we use is the increase of the delay of external messages due to internal traffic. We begin by calculating the average number of external messages in a node. Let L_m and K_m denote the number of external and internal messages, respectively, at node m; $m = 1, 2, \ldots, N$. From (4.134), we have

$$\begin{aligned} E[L_m] &= \sum_{l_1=0}^{\infty}\sum_{l_2=0}^{\infty}\cdots\sum_{l_N=0}^{\infty} l_m \sum_{S(W,N)} G^{-1}\prod_{i=1}^{N}\left(\frac{\lambda_0}{\mu_i}\right)^{k_i}\frac{\rho_i^{l_i}(k_i+l_i)!}{(k_i!l_i!)} \\ &= \sum_{l_1=0}^{\infty}\sum_{l_2=0}^{\infty}\cdots\sum_{l_N=0}^{\infty}\sum_{S(W,N)} G^{-1} l_m \frac{\rho_m^{l_m}(k_m+l_m)!}{(k_m!l_m!)}\prod_{\substack{i=1\\ i\neq m}}^{N}\left(\frac{\lambda_0}{\mu_i}\right)^{k_i}\frac{\rho_i^{l_i}(k_i+l_i)!}{(k_i!l_i!)}; \\ &\quad m = 1, 2, \ldots, N \end{aligned}$$

After a change of variable $l'_m = l_m - 1$, we find

$$E[L_m] = \sum_{l_1=0}^{\infty}\sum_{l_2=0}^{\infty}\cdots\sum_{l_N=0}^{\infty}\sum_{S(W,N)} G^{-1}(k_m + 1 + l_m)\rho_m \prod_{i=1}^{N}\left(\frac{\lambda_0}{\mu_i}\right)^{k_i}\frac{\rho_i^{l_i}(k_i+l_i)!}{(k_i!l_i!)}$$

which we recognize as

$$E[L_m] = \rho_m[E[K_m] + E[L_m] + 1]$$

Solving for $E[L_m]$, we find that the average number of external messages in a node is

$$E[L_m] = \frac{\rho_m}{1-\rho_m}[E[K_m] + 1] \tag{4.138}$$

When there is no internal traffic, $E_0[L_m] = \rho_m/(1 - \rho_m)$. The average number of internal messages, $E[K_m]$, at node m is a straightforward application of (4.58). Applying Little's formula, we see that the normalized difference in average delay due to internal traffic in a node is

$$\frac{\Delta D_m}{D_{m_0}} = \frac{\dfrac{\rho_m}{\lambda_m(1 - \rho_m)} E[K_m]}{\dfrac{\rho_m}{\lambda_m(1 - \rho_m)}} = E[K_m]$$

Averaging over the message arrival rate for all nodes, we find

$$\frac{\Delta D}{D_0} = \frac{\sum_{m=1}^{N} \lambda_m E[K_m]}{\sum_{i=1}^{N} \lambda_i} \tag{4.139}$$

where $E[K_m]$ is given by (4.58) with $R_m = \lambda_0/(\mu_m - \lambda_m)$. Note that λ_0 is an arbitrary constant, which will wash out in the normalization.

Example 4.12 These results have been applied to the case of a chain consisting of five nodes with a window of ten messages. It is assumed that service rates in the nodes are, respectively $\mu_1 = 3$, $\mu_2 = 4$, $\mu_3 = 2$, $\mu_4 = 6$, $\mu_5 = 3$. The service rate in the phantom node is $\lambda_0 = 0.6$, and the arrival rates of external traffic are, respectively, $\lambda_1 = 2$, $\lambda_2 = 0.5$, $\lambda_3 = 0.8$, $\lambda_4 = 4$, $\lambda_5 = 1$. On the accompanying spreadsheet the normalization constant was found to be $G(10, 5) = 12.05682$. Once again, a good check is to calculate the average number of internal messages in each node and sum over all nodes. Clearly, the answer must be 10. The probability of blocking is calculated as 0.296093017. The change in delay of external messages due to internal traffic is 0.688370317.

4.7.3 Cellular Radio

We can model a cellular radio system as a network of queues. The geographic area covered by the network is segmented into hexagonal cells, each of which is served by a base station that broadcasts messages to the mobile users in the cell. Frequencies are assigned to the cell in a repeating pattern that is designed to minimize interference. As mobile users move between cells, they switch frequency bands. A portion of the mobile radio system is shown in Figure 4.25, where 14 cells can be seen.

Since each cell is modeled as having an infinite number of servers, the probability distribution of the number of users in each cell is

$$P(K_1 = k_1, K_2 = k_2, \ldots, K_N = k_N) = \prod_{i=1}^{N} \frac{e^{-\rho_i} \rho_i^{k_i}}{k_i!} \tag{4.140}$$

where N is the number of cells in the system under study. The marginal distribution of the number of users in any cell is just

$$P(K = k) = e^{-\rho} \rho^k / k!; \quad k = 0, 1, 2, \ldots \tag{4.141}$$

where we have suppressed the dependence on a cell for simplicity.

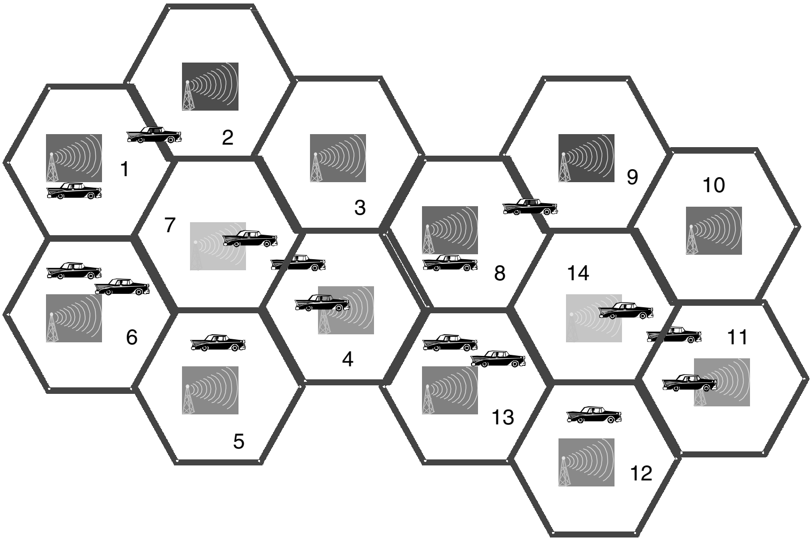


Figure 4.25 Mobile radio system.

Now, each of the users contends for L channels. The model is the same as the finite-source model studied in Chapter 3. A user is blocked if all channels are occupied. The probability of this event is

$$Q_L = \frac{\binom{K}{L}\gamma^L}{\sum_{i=0}^{L}\binom{K}{i}\gamma^i} \tag{4.142}$$

where $\gamma = \sigma/\mu$, where σ is the probability that a single source goes from off to on in an incremental interval and μ is the probability of change in the opposite direction. We find the overall probability of blocking in a particular cell by averaging Q_L over the probability distribution of the number of users in a cell given by (4.141). We find

$$P_B = \sum_{k=L+1}^{\infty} \underbrace{\frac{e^{-\rho}\rho^k}{k!}}_{\text{number of users}} \underbrace{\frac{\binom{k}{L}\gamma^L}{\sum_{i=0}^{L}\binom{k}{i}\gamma^i}}_{\text{blocking probability}} = \frac{e^{-\rho}\gamma^L}{L!}\sum_{k=L+1}^{\infty}\frac{\rho^k}{k!(k-L)!\sum_{i=0}^{L}\gamma^i/(i!(k-i)!)}$$

Example 4.13 In terms of the network of queues model, the users correspond to messages and the cells, to queues. We assume that the queue in a cell is modeled as an infinite number of servers. As we have seen for an infinite number of servers, the product form solution holds for an arbitrarily distributed service time; however, it would seem that the residence time is most appropriately modeled as by the exponential distribution. We assume that the 14-node system depicted in Figure 4.25, is embedded in a larger system. We assume that the users are equally likely to move through each of the six cell boundaries. Thus, for cell 1, half of its traffic leaves the subsystem; while for cell 7, all its traffic remains in the subsystem. The routing matrix for the subsystem is

$$Q=\begin{bmatrix}
0 & \frac{1}{6} & 0 & 0 & 0 & \frac{1}{6} & \frac{1}{6} & 0 & 0 & 0 & 0 & 0 & 0 & 0\\
\frac{1}{6} & 0 & \frac{1}{6} & 0 & 0 & 0 & \frac{1}{6} & 0 & 0 & 0 & 0 & 0 & 0 & 0\\
0 & \frac{1}{6} & 0 & \frac{1}{6} & 0 & 0 & \frac{1}{6} & \frac{1}{6} & 0 & 0 & 0 & 0 & 0 & 0\\
0 & 0 & \frac{1}{6} & 0 & \frac{1}{6} & 0 & \frac{1}{6} & \frac{1}{6} & 0 & 0 & 0 & 0 & \frac{1}{6} & 0\\
0 & 0 & 0 & \frac{1}{6} & 0 & \frac{1}{6} & \frac{1}{6} & 0 & 0 & 0 & 0 & 0 & 0 & 0\\
\frac{1}{6} & 0 & 0 & 0 & \frac{1}{6} & 0 & \frac{1}{6} & 0 & 0 & 0 & 0 & 0 & 0 & 0\\
\frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0\\
0 & 0 & \frac{1}{6} & \frac{1}{6} & 0 & 0 & 0 & 0 & \frac{1}{6} & 0 & 0 & 0 & \frac{1}{6} & \frac{1}{6}\\
0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{6} & 0 & \frac{1}{6} & 0 & 0 & 0 & \frac{1}{6}\\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{6} & 0 & \frac{1}{6} & 0 & 0 & \frac{1}{6}\\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{6} & 0 & \frac{1}{6} & 0 & \frac{1}{6}\\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{6} & 0 & \frac{1}{6} & \frac{1}{6}\\
0 & 0 & 0 & \frac{1}{6} & 0 & 0 & 0 & \frac{1}{6} & 0 & 0 & 0 & \frac{1}{6} & 0 & \frac{1}{6}\\
0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & 0
\end{bmatrix}$$

The arrival rate of users from adjacent cells is proportional to the number of sides of a cell, which interface the larger system. The input vector users per minute entering a cell is then

$$\boldsymbol{\lambda}=\begin{bmatrix}3 & 3 & 2 & 1 & 3 & 3 & 0 & 1 & 3 & 3 & 3 & 3 & 2 & 0\end{bmatrix}$$

On the associated Excel spreadsheet, we find that the resulting total flows into each cell as

$$\Lambda=\begin{bmatrix}6 & 6 & 6 & 6 & 6 & 6 & 6 & 6 & 6 & 6 & 6 & 6 & 6 & 6\end{bmatrix}$$

Of course, this is what is expected for the uniform case. If the residence time of users in the cells is an average of 3 min, the load in each cell is 18. On the associated Excel spreadsheet, we have calculated the blocking probability for the specific case of $L = 5$ (available lines) in a cell and $\gamma = 0.2$ (source activity). The result in cell M75 shows a blocking probability of 0.04949103. The reader can try other cases being careful to include enough terms in the sums in L53–L66.

It is of quite considerable interest in telecommunication networks and production systems to analyse stochastic networks with unreliable servers. Starting with BCMP networks and introducing degradable network models (which incorporates breakdowns of servers and their repair), with different rules for customer's routing connected with nodes in repair status, steady state distribution of product form have been obtained in Sauer and Daduna (2003).

REFERENCES

Baskette, F., M. Chandy, R. Munty and F. G. Palacios, "Open, closed and mixed networks of queues with different classes of customers," *J. Assoc. Comput. Mach.* **22**: 248–260 (1975).

Bonald, T., A. Proutiére, J. Roberts and J. Virtamo, "Computational aspects of balanced fairness," *18th International Teletraffic Congress*, Vol. 5b, Berlin, Germany, 31 August–5 September 2003, 801–810.

Burke, P. J., "The dependence of delays in tandem queues," *Ann. Math. Stat.* **35**: 874–875 (1964).

Burke, P. J., "The output process of a stationary M/M/S queueing system," *Ann. Math. Stat.* **37**(4): 18.44–18.52 (1968).

Buzen, J. P., "Computational algorithms for closed networks with exponential servers," *Commun. ACM* **16**: 527–531 (Sept. 1973).

Camarda, P., G. Schiraldi and F. Talucci, (Politecnico de Bari) "Mobility modeling in cellular communication networks," *Proc LAN Conf.*, Minneapolis, Fall, 1996.

Conway, A. E., and N. D. Georganas, *Queueing Networks—Exact Computational Algorithms: A Unified Theory Based on Decomposition and Aggregation*, MIT Press, 1989.

Fletcher, R., *Practical Methods of Optimization*, second edition, Wiley, 1987.

Gelenbe, E., and I. Mitrani, *Analysis and Synthesis of Computer Systems*, Academic Press, 1980.

Gelenbe, E., and G. Pujolle, *Introduction to Queueing Networks*, Wiley, 1987.

Gordon, W. J., and G. F. Newel, "Closed queueing system with exponential servers", *Oper. Research*, **15**: 254–265 (1967).

Jackson, J. R., "Networks of waiting lines," *Oper. Research* **5**: 518–521 (1957).

Jackson, J. R., "Jobshop-like queueing systems," *Manage. Sci.*, **10**: 131–142, (1963). *Reversibility and Stochastic Networks*, Wiley, 1979.

Kleinrock, L., *Queueing Systems*, Vol. 1: *Theory*, Wiley, New York, 1975.

Kleinrock, L., *Communication Nets: Stochastic Message Flow and Delay*, McGraw-Hill, New York, 1964 (out of print; reprinted Dover, New York, 1972).

Kobayashi, H., *Modeling and Analysis, An Introduction to System Performance Evaluation Methodology*, Addison-Wesley, Reading, MA 1978.

Meister, B., H. R. Mueller, and H. Rudin, "New optimization criteria for message-switched networks," *IEEE Trans. Commun. Technol.*, **COM-19**(3): 256–260 (June 1971).

Nelson, R., *Probability, Stochastic Processes and Queueing Theory*, Springer-Verlag, 1995.

Pennotti, M. C., and M. Schwartz, "Congestion control in store and forward tandem links", *IEEE Trans. Commun.* **COM-23**(2): 1434–1443 (Dec. 1975).

Reich, E., "Waiting time when queues are in tandem," *Ann. Math. Stat.*, **28**: 768–773 (1957).

Reiser, M., and H. Kobayashi, "Numerical solution of semiclosed exponential server queueing networks", *Proc 7th Asilomar Conf. Circuits, Systems and Computers*, Nov. 1973.

Reiser, M., and S. S. Lavenberg, "Mean value analysis of closed multichain queueing networks," *J. ACM* **22**: 313–322, (April 1980).

Robertazzi, T. G., *Computer Networks and Systems*, Springer-Verlag, 1994.

Ross, S. M., *Stochastic Processes*, Wiley, New York, 1983.

Sauer, C., and H. Daduna, "Separable networks with unreliable servers," 18th International Teletraffic Congress, Vol. 5b, Berlin, Germany, 31 August–5 September 2003.

Seneta, E., *Non-Negative Matrices*, George Allen & Unwin Ltd., 1973.

VanDijk, N. M., *Queueing Networks and Product Forms: A Systems Approach*, Wiley, 1993.

Walrand, J., *Queueing Networks*, Prentice-Hall, 1988.

EXERCISES

4.1 Demonstrate the sufficiency of local balance in birth and death processes using the arrival–departure process ddad at time intervals t_1, t_2, t_3, t_4, respectively, starting at time $t = 0_o$. Assume that the initial number in the system is 2.

4.2 Suppose that each of the nodes in the network shown in Figure 4.6 contains two independent exponential servers.

- **(a)** Find the joint distribution of the number of messages in each node.
- **(b)** What is the average number of messages in each queue?
- **(c)** Repeat (a) and (b) for the case of an infinite number of servers at each node.

4.3 Suppose that two classes of jobs arrive at a service facility each at an independent Poisson rate. The facility is equipped with K servers. The class 1 jobs require a single server for an exponentially distributed time interval. In contrast, the class 2 jobs require all K servers for an exponentially distributed time period. Assume that there is no room to store jobs so that a job that cannot be served immediately departs. The two service rates may be different.

- **(a)** Sketch the state transition flow diagram for the number of jobs of both types at the facility.
- **(b)** Write down the equilibrium equations.
- **(c)** Find the joint probability distribution for the number of jobs of each class in the facility.

4.4 In deriving the global balance equations for the N-node Jackson network [see (4.21)], we assumed that $q_{ii} = 0; \forall i$. What would be the change in this equation if the assumption were not true?

4.5 Consider the simple network of two queues in tandem shown below. Messages arrive at the first queue at a Poisson rate. The service times in each queue are independent and exponentially distributed with different mean values. Assume that the output of the second queue is fed back to the input of the first queue with probability P. With probability $1 - P$, a customer leaves the system after queue 2. Assume infinite waiting rooms in each queue.

(a) What are the conditions on the arrival rate and the service times for a product form solution?

(b) Find an expression for joint distribution of the number of messages in each queue.

(c) Given $\lambda_1 = 10$ messages/s, $P = 0.8$, $1/\mu_1 = 0.01$ s, and $1/\mu_2 = 0.005$ s, find the average delay.

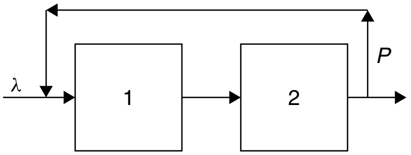


Figure 4.26

4.6 Suppose that we have a four-node network with the routing matrix

$$\begin{bmatrix} 0 & 0.6 & 0.15 & 0 \\ 0 & 0 & 0.1 & 0.75 \\ 0.2 & 0.25 & 0 & 0.3 \\ 0.4 & 0.35 & 0.25 & 0 \end{bmatrix}$$

and the input traffic $\lambda = [4.0, 1.5, 3.0, 1.0]$. Find the total flow into each node.

4.7 Now suppose that for the network of Exercise 4.2, each server in the four nodes has a uniform service rate of 30 messages per second.

(a) Find the joint distribution of the number of messages in each node.

(b) What is the average delay?

4.8 Consider the three-node network shown below. The arrival rate to the network is 120 messages per second with 1000 bits per message. All lines operate at a rate of 0.5 Mbps.

(a) If all nodes are served by one line, what is the joint distribution of the number of messages in each node?

(b) What is the average delay of a message through the network?

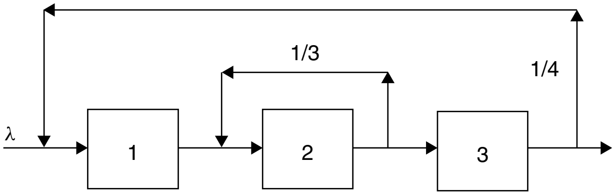


Figure 4.27

4.9 Repeat Exercise 4.8 for the case of two lines out of each node, both of which operate at a rate of 0.25 Mbps.

4.10 Consider the store-and-forward network of Figure 4.12. Suppose that nodes *A* and *C* are close together and *B* is far from both. In order to reduce costs, we remove the links from *B* to *A* and from *C* to *B*. (It is reasonable to assume that the cost of a communications link increases with line length.) Assuming the same flows as in Example 4.5, find the average message delay.

4.11 A fortuneteller and a stockbroker share the same waiting room in an office building. Clients for each of these arrive at a Poisson rate and require an exponentially distributed period of consultation. Assume that the waiting room can hold no more than two clients of either type. Potential clients arriving at a full waiting room take their business elsewhere.

(a) Sketch the state transition flow diagram for the number of clients of either type in consultation or in the waiting room.

(b) Write down the global balance equations.

(c) Find steady-state joint probability distribution for the number of clients of either type.

4.12 Suppose that in Figure 4.10 we have the following values for the various quantities depicted:

(i) $\lambda_1 = 8.0$, $\lambda_2 = 32.0$, $\lambda_3 = 16.0$, $\lambda_4 = 8.0$, all in messages per second,

(ii) The average service times in seconds in each node are $1/\mu_1 = 0.025$, $1/\mu_2 = 0.0167$, $1/\mu_3 = 0.02$, $1/\mu_4 = 0.04$.

(iii) $q_{12} = \frac{1}{2}$, $q_{21} = \frac{1}{4}$, $q_{23} = \frac{1}{2}$, $q_{41} = \frac{1}{2}$.

(iv) Each node has a single server.

(a) Write down the steady-state distribution of the joint queue distributions.

(b) What is the average delay?

4.13 Suppose that we have a five-node network with the routing matrix

$$Q = \begin{bmatrix} 0 & 0.15 & 0.3 & 0.2 & 0.10 \\ 0.5 & 0 & 0 & 0.4 & 0.05 \\ 0.35 & 0.1 & 0 & 0.05 & 0.1 \\ 0.2 & 0.25 & 0.1 & 0 & 0.15 \\ 0.2 & 0.2 & 0.1 & 0.15 & 0 \end{bmatrix}$$

Suppose also that the message rates, in thousands of messages per second, into the nodes are $\boldsymbol{\lambda} = \begin{bmatrix} 1.7 & 2.25 & 3.15 & 4.0 & 1.8 \end{bmatrix}$. Also, assume that each of the five nodes has a single exponentially distributed server. What is the minimum allowable service rate for each of these servers if the system is to remain stable?

4.14 Suppose that we have five ATM streams feeding into an OC-3 line. In Mbps the volumes of traffic for each line are, respectively, 20, 14, 18.5, 30 and 5. What is the optimum allocation of capacity with the cost function given in (4.38)?

4.15 Assume the traffic parameters given in Exercise 4.13. Assume also that messages have an average length of 1 kbyte. Suppose also that there is a constant delay of 10 μs between nodes. Finally, assume that 5 Gbps of capacity is to be allocated among the links coming out of each node. What is an optimum allocation of capacity?

4.16 Consider the following routing matrix in a closed network of queues:

$$Q = \begin{bmatrix} 0 & 0.15 & 0.3 & 0.2 & 0.35 \\ 0.5 & 0 & 0 & 0.4 & 0.1 \\ 0.35 & 0.3 & 0 & 0.05 & 0.3 \\ 0.2 & 0.25 & 0.4 & 0 & 0.15 \\ 0.2 & 0.45 & 0.2 & 0.15 & 0 \end{bmatrix}$$

Each node has an exponentially distributed server, having an average service length of 100 ms. Suppose that three jobs are circulating among the nodes.

(a) What is the probability distribution of the jobs in each node?

(b) What are the average delays through each node?

4.17 Repeat Exercise 4.16 under the assumption that each node has an infinite number of servers with the same service rate.

4.18 Consider a closed path consisting of five nodes. Assume that seven messages are circulating in among the nodes. We assume that the service rates in messages per second for each of the five nodes are, respectively, 6, 2, 4, 5, and 1. Find the average number of messages in each node.

4.19 Find the optimum capacity allocation for arbitrary k in Equation (4.40). What is the general solution when $k \to \infty$?

4.20 The symmetric network shown below has the externally arriving traffic $\lambda_1 = \lambda_4 = 1$ messages per second and $\lambda_2 = \lambda_3 = 2$ messages per second. All the links between stations are two-way, but capacity in each direction may be different.

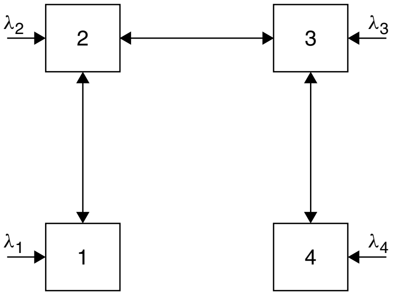


Figure 4.28

The routing matrix between stations is as indicated below:

$$Q = \begin{bmatrix} 0 & 0.5 & 0 & 0 \\ 0.5 & 0 & 0.25 & 0 \\ 0 & 0.25 & 0 & 0.5 \\ 0 & 0 & 0.5 & 0 \end{bmatrix}$$

(a) Find the total traffic in messages per second into each node. (*Hint:* Use symmetry.)

(b) Find the flows in messages per second on each link.

(c) Find the average message delay. State the assumptions that you need to do this. Assume that the only delays in the network are due to multiplexing on the links. Assume that the average message lengths are 1000 bits long and that the capacities of the links are 4000 bps.

4.21 The closed communication network shown below has four jobs circulating among the terminals and the central processing unit (CPU). The CPU is modeled as a processor, which is shared among the jobs. Each terminal generates a separate class. The processing time is a constant 50 ms. The terminals may be modeled as consisting of single exponential servers with average service times 50, 20, 20, and 10 ms, respectively, for terminals 1, 2, 3, and 4. Assume infinite storage at all the nodes. The probability of a message being routed from the CPU to terminals 1, 2, 3, and 4 is $\frac{1}{2}$, $\frac{1}{4}$, $\frac{1}{8}$, and $\frac{1}{8}$, respectively.

(a) What is the probability distribution of the number of messages in the CPU and the four terminals?

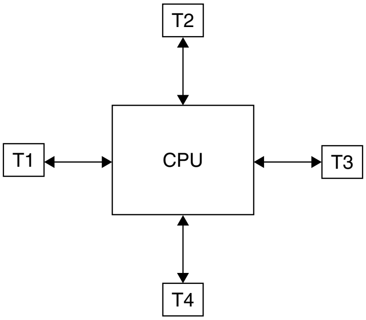


Figure 4.29

(b) What is the average number of messages in the CPU?

4.22 Repeat Exercise 4.21 under the assumption that the processor is modeled as an infinite number of servers and the terminals are single exponential servers.

4.23 Consider once again the network of Exercise 4.21. Assume that the CPU is a processor that is shared by the jobs in its queue. The terminals are LCFS with constant service time. Up to four messages may circulate in this system. External arrivals are to the terminals at a rate of only 2.5 messages per second to each. Assume that two-thirds of the jobs are completed in the CPU and do not need to be sent back to the terminals.

(a) Find the probability distribution for the number of jobs in the CPU.

(b) What is the average number of jobs in terminal 4?

4.24 Consider the double-ring network shown below. There are two classes of messages, those circulating among nodes 1, 2, and 3 and those circulating between nodes 3 and 4. Assume that both classes have two messages. Each of these nodes has a single exponential server. The mean service times are as follows: nodes 1, 2 and 4—2 s, node 3—1 s.

(a) Using the mean-value analysis, find the average number of messages in each node.

(b) Find the average delay around the left ring network.

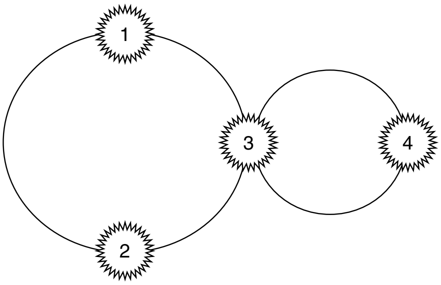


Figure 4.30

4.25 This exercise is based on Example 4.11. On a spreadsheet go through a similar example, but for three input lines and five output lines.

4.26 Rework Example 4.12 for a seven-node chain with the following set of parameters: $\mu_1 = 1.4$, $\mu_2 = 3$, $\mu_3 = 2$, $\mu_4 = 5$, $\mu_5 = 2.1$, $\mu_6 = 2$, $\mu_7 = 2.3$. The service rate in the phantom node is $\lambda_0 = 0.4$, and the arrival rates of external traffic are, respectively, $\lambda_1 = 0.9$, $\lambda_2 = 2.1$, $\lambda_3 = 0.5$, $\lambda_4 = 2.3$, $\lambda_5 = 1$, $\lambda_6 = 0.55$, $\lambda_7 = 0.8$.

4.27 Rework Example 4.13 for the following input vector:

$$\lambda = \begin{bmatrix} 2 & 1 & 2 & 1 & 3 & 2 & 3 & 1 & 3 & 1 & 3 & 2 & 2 & 2 \end{bmatrix}$$

5

MARKOV CHAINS: APPLICATION TO MULTIPLEXING AND ACCESS

5.1 TIME-DIVISION MULTIPLEXING

In Section 2.7, Markov chains were introduced and their basic characteristics were outlined. In this chapter, the Markov chain is used to model techniques for multiplexing and access in telecommunications networks.

We begin with the treatment of time-division multiplexing (TDM). The first widely deployed example of TDM was the T1 carrier system depicted in Figure 5.1. The line flow is 1.544 Mbps segmented into frames 125 μs in duration containing

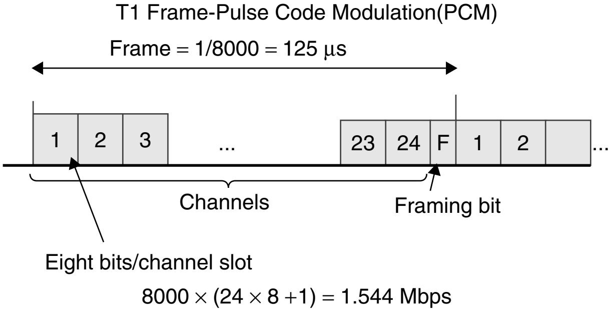


Figure 5.1 T1 frame structure.

Modeling and Analysis of Telecommunications Networks, by Jeremiah F. Hayes and Thimma V. J. Ganesh Babu
ISBN 0-471-34845-7 © 2004 John Wiley & Sons, Inc.

193 bits. The system provides 24 channels, where each channel is allocated a one-byte slot in a frame. There is an additional single bit devoted to synchronization in each frame. Throughout the chapter, we continue the concept of segmenting the flow into fixed-length slots. The Markov chains model the sequences formed by the number of information units, bytes, for example, in the system at the slot boundaries. From the point of view of analysis of performance, time-division multiplexing and time-division multiple access (TDMA) are equivalent.

5.2 THE ARRIVAL PROCESS

5.2.1 Packetization

In order for a source to be multiplexed onto a slotted line, its output must be segmented into fixed-size units, generally called *packets* and in ATM, called *cells*. This is the process of packetization. In the asynchronous transfer mode (ATM), for example, all information—voice, video, and data—is packaged into cells containing 48 octets (8-bit bytes).

In terms of the analysis of performance, packetization is simply a transformation of random variables. Suppose that the probability distribution of the number of bits in a message is denoted by $B(i) = P(\text{message} = i \text{ bits})$. If the number of information bits in packet is denoted as I, the probability distribution of the number of packets in a message is given by

$$M(k) \overset{\Delta}{=} P(k \text{ packets in a message}) = \sum_{i=(k-1)I+1}^{kI} B(i) \tag{5.1}$$

"Stuff" bits are used to round out messages to an integral number of packets. In each packet, overhead bits supplement the information bits. The total number of bits in a packetized message has the distribution $P(k) \overset{\Delta}{=} B(k(I + O))$, where O is the number of overhead bits in a packet.

Example 5.1 Suppose that the number of bits in a message is the sum of a constant number of overhead bits, O, plus a geometrically distributed component with distribution $B(i) = B^{i-1}(1 - B)$; $i = 1, 2, \ldots$. Let the number of information bits in a packet be denoted by I. Assuming that $I > O$, the probability distribution of the number of packets that are required to convey a message is given by

$$P(\text{one packet in a message}) = P(M \leq I - O) = \sum_{j=1}^{I-O} B^{j-1}(1 - B)$$

$$= 1 - B^{I-O}$$

$$
\begin{aligned}
P(k \text{ packets in a message; } k > 1) &= \sum_{i=I-O+1+(k-2)I}^{I-O+(k-1)I} B^{j-1}(1-B) \\
&= \sum_{j=0}^{I-1} B^{j+(k-1)I-O}(1-B) \\
&= B^{(k-1)I-O} \sum_{j=0}^{I-1} B^{j}(1-B) \\
&= B^{(k-1)I-O}(1-B^{I}); \quad k = 2, \ldots,
\end{aligned}
$$

For this distribution, $M(z) = z(1 - B^{I-O}) + zB^{I-O}(1 - B^{I})/(1 - zB^{I})$. Note that, when $O = 0$ here, we have $M(z) = z(1 - B^{I})/(1 - zB^{I})$; thus, the number of packets in a message has a geometric distribution with mean $1/(1 - B^{I})$. On the associated spreadsheet, these formulas are used to work out means and variances for the case of an ATM cell.

5.2.2 Compound Arrivals

As well as containing a random number of packets, messages may arrive in a slot according to a random arrival pattern. For example, suppose that n sources are connected to a multiplexer and that each source generates a message in a slot independently with probability P. In this case, the total number of messages generated follows the binomial distribution with parameters, n and P. As we have seen in Section 2.2, for a large number of sources, the number of message arrivals is governed by the Poisson distribution with average arrival rate $\lambda = nP$.

Suppose that the number of message arrivals in a slot has an arbitrary distribution with probabilities $a_1, a_2, \ldots$. The generating function is then $A_S(z) = \sum_{k=1}^{\infty} a_k z^k$. Now, conditioned on j message arrivals in an interval, the probability-generating function of the number of arrived packets is $M^{j}(z)$. Averaging over the arrival distribution, we have for the generating function of the arrival process

$$
A(z) = \sum_{k=0}^{\infty} a_k M^{k}(z) = A_S(M(z)) \tag{5.2}
$$

The average packet arrival rate in a slot, which is designated as ρ, the traffic intensity or load, a quantity of considerable interest. Differentiating (5.2), we find the average number of packets arriving in a slot

$$
\rho = \bar{A} = \left.\frac{dA_S(M(z))}{dz}\right|_{z=1} = A'_S(M(1))M'(1) = \bar{A}_S\bar{M} \tag{5.3}
$$

where $\bar{A}_S$ is the average number of messages arriving in a slot and $\bar{M}$ is the average number of packets per message.

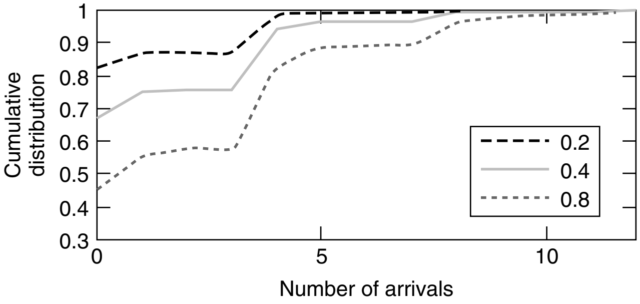


Figure 5.2 Arrival distribution.

Example 5.2 Suppose that the arrival process is Poisson with an average rate of λ messages per second, and that messages have two lengths, $P(\text{one packet}) = 0.3$ and $P(\text{four packets}) = 0.7$. The probability-generating function for the total number of packets arriving in a slot of duration T second is

$$A(z) = \sum_{j=0}^{\infty} \left(0.3z + 0.7z^4\right)^j \frac{(\lambda T)^j e^{-\lambda T}}{j!} = e^{-\lambda T(1-0.3z-0.7z^4)}$$

The moments and the probability distribution can be found by successive differentiation as discussed in Section 2.2. These operations have been done in Maple and evaluated at $\lambda T = 0.2, 0.4, 0.8$. The results are shown on the associated spreadsheet and some are plotted on Figure 5.2, where the cumulative distribution of arrivals is shown. Note the relatively large jumps at multiples of four.

5.3 ASYNCHRONOUS TIME-DIVISION MULTIPLEXING

Since it is the simplest of the techniques, we begin with an analysis of asynchronous time-division multiplexing (ATDM). In many applications data sources may be characterized as "bursty," meaning that there are long idle periods interspersed with periods of activity. Because of this characteristic of data sources, dedicating transmission facilities to individual sources may not be efficient; the channels dedicated to some sources are often empty while there is congestion on others. An appropriate technique for multiplexing a number of bursty sources is called *asynchronous time-division multiplexing* (ATDM) (see Fig. 5.3). Packets from each of the sources connected to the multiplexer are transmitted in order of arrival using the full capacity of the transmission line. In order to distinguish among packets from different sources sharing the line, addressing information must accompany the packets. While awaiting transmission, packets are held in a common buffer. The contents of this buffer are the focus of analysis. In the next section, we deal with

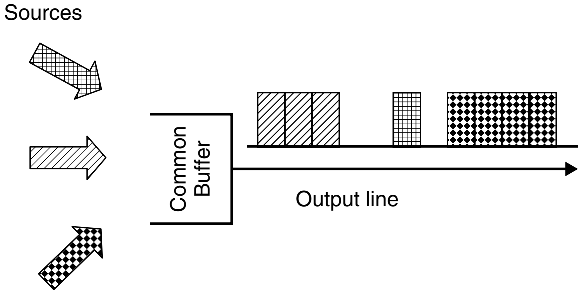


Figure 5.3 Asynchronous time-division multiplexing.

synchronous time-division multiplexing (STDM) to indicate the technique where capacity is dedicated to sources. The T1 carrier system depicted in Figure 5.1 is a synchronous system.

We assume that the aggregate arrival process to the multiplexer from all sources is λ messages per second. We also assume that the transmission rate on the synchronous line out of the buffer is $1/T$ slots per second, where a slot carries exactly one packet. Before transmission, newly arrived packets are stored in a buffer. The first question that we shall address is the probability distribution of the number of packets in the buffer.

We imbed a Markov chain at the points in time between slots. Let N_i be the number of packets in the buffer at the end of the ith slot. We let $A_i(N_i)$ denote the number of packets that arrive during the ith slot, assuming that there are N_i packets in the system. In the case of a finite buffer, it is necessary to account for this dependence since the buffer can be full and no new arrivals to the multiplexer are possible. Now, suppose that the ith departing message leaves behind a nonempty system: $N_i > 0$. The state of the system at the end of the next slot is given by

$$N_{i+1} = N_i - 1 + A_{i+1}(N_i); \quad N_i \geq 1$$

If the system is empty at the beginning of a slot, a slightly different equation describes the process. At the end of the next slot, the system, contains only messages that arrived during the slot:

$$N_{i+1} = A_{i+1}(N_i); \quad N_i = 0$$

The state dynamics can be summarized in the following equation:

$$N_{i+1} = N_i - U(N_i) + A_{i+1}(N_i) \tag{5.4}$$

where $U(\cdot)$ is the unit step $U(x) = \begin{cases} 0, & x \leq 0 \\ 1, & x > 0 \end{cases}$. Note that packets arriving in a slot cannot be transmitted until the beginning of the next slot.

5.3.1 Finite Buffer

We consider first the case where the buffer is finite. We begin, as in Section 2.7.2, by deriving the state transition matrix. If the buffer can hold a maximum of B packets, then the state transition matrix has $B+1$ rows and $B+1$ columns. Let $a_n = P(n$ arrivals in a slot). We start by calculating the elements of the first row in the matrix. These are the transition probabilities from an empty buffer at the beginning of a slot. At the end of the next slot, the content of the buffer is just what arrived during the slot. The buffer will be full if B or more packets arrive. The transition probabilities are as shown in (5.5). The result would be the same if there were just one packet in the buffer at the beginning of the slot since the one packet will be removed. If there are two packets in the buffer at the beginning of a slot, the number at the end would be the arrivals plus one residual. Note that the buffer cannot be empty at the end of the slot in this case. The result is in the third row in (5.5). The pattern replicates. There can be a decrease of at most one, and the buffer content is residuals plus new arrivals. The state transition matrix is then

$$R = \begin{bmatrix} a_0 & a_1 & a_2 & \cdots & a_{B-1} & \sum_{j=B}^{\infty} a_j \\ a_0 & a_1 & a_2 & \cdots & a_{B-1} & \sum_{j=B}^{\infty} a_j \\ 0 & a_0 & a_1 & \cdots & a_{B-2} & \sum_{j=B-1}^{\infty} a_j \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & a_1 & \sum_{j=2}^{\infty} a_j \\ 0 & 0 & 0 & \cdots & a_0 & \sum_{j=1}^{\infty} a_j \end{bmatrix} \tag{5.5}$$

Note that the rows all sum to one.

At the outset of our analysis we shall assume that equilibrium has been attained so that the probability distribution of buffer occupancy does not change from slot to slot. Since we have a finite buffer, stability does not require that the average number of packets arriving in a slot be less than one. If the arrival rate were very large, the buffer would simply be full all the time. Let P_i; $i = 0, 1, \ldots, B$ denote the steady-state probability that there are i packets in the buffer at the beginning of a slot. The probabilities are encapsulated in the row vector $\mathbf{P} = (P_0, P_1, \ldots, P_B)$. As we have seen in Section 2.7.3, the steady-state probability is the solution to the equation

$$\mathbf{P} = \mathbf{P}R \tag{5.6}$$

In the present case, the state transition matrix is such that solution is obtained fairly easily. This is because the number of packets can decrease by at most one between adjacent slots. The individual equations in (5.6) are

$$P_i = \sum_{j=0}^{B} P_j r_{ji} = \begin{cases} P_0 a_i + \sum_{j=0}^{i} P_{j+1} a_{i-j}; & 0 \leq i < B \\ P_0 \sum_{j=B}^{\infty} a_j + \sum_{j=0}^{B-1} P_{j+1} \sum_{k=B-j}^{\infty} a_k; & i = B \end{cases} \tag{5.7}$$

In each of these equations, we can solve for P_{i+1}:

$$P_{i+1} = \frac{P_i - \sum_{k=1}^{i} P_k a_{i-k+1} - P_0 a_i}{a_0}; \quad 0 \leq i < B \tag{5.8}$$

The reason for this simplicity lies in the fact that the decrease over a slot time can be at most one packet. In order for this solution to be valid, we must have $a_0 > 0$; otherwise, the average number of arrivals in a slot would be greater than one and the system would be full all the time.

The values obtained from (5.8) depend on the initial value P_0, which can be found from the normalizing relation

$$\sum_{i=0}^{B} P_i = 1 \tag{5.9}$$

There is a simple procedure to find $P_0, P_1, \ldots, P_B$ based on (5.8) and (5.9). We initialize $P'_0 = 1$ and find $P'_1, P'_2, \ldots, P'_B$ from (5.8). The final values are found from normalization

$$P_i = \frac{P'_i}{\sum_{j=0}^{B} P'_i}, \quad i = 0, 1, \ldots, B \tag{5.10}$$

Several quantities are of interest in connection with a finite buffer. For example, one can calculate the average number of packets in the buffer from

$$\overline{N_P} = \sum_{i=0}^{B} iP_i \tag{5.11}$$

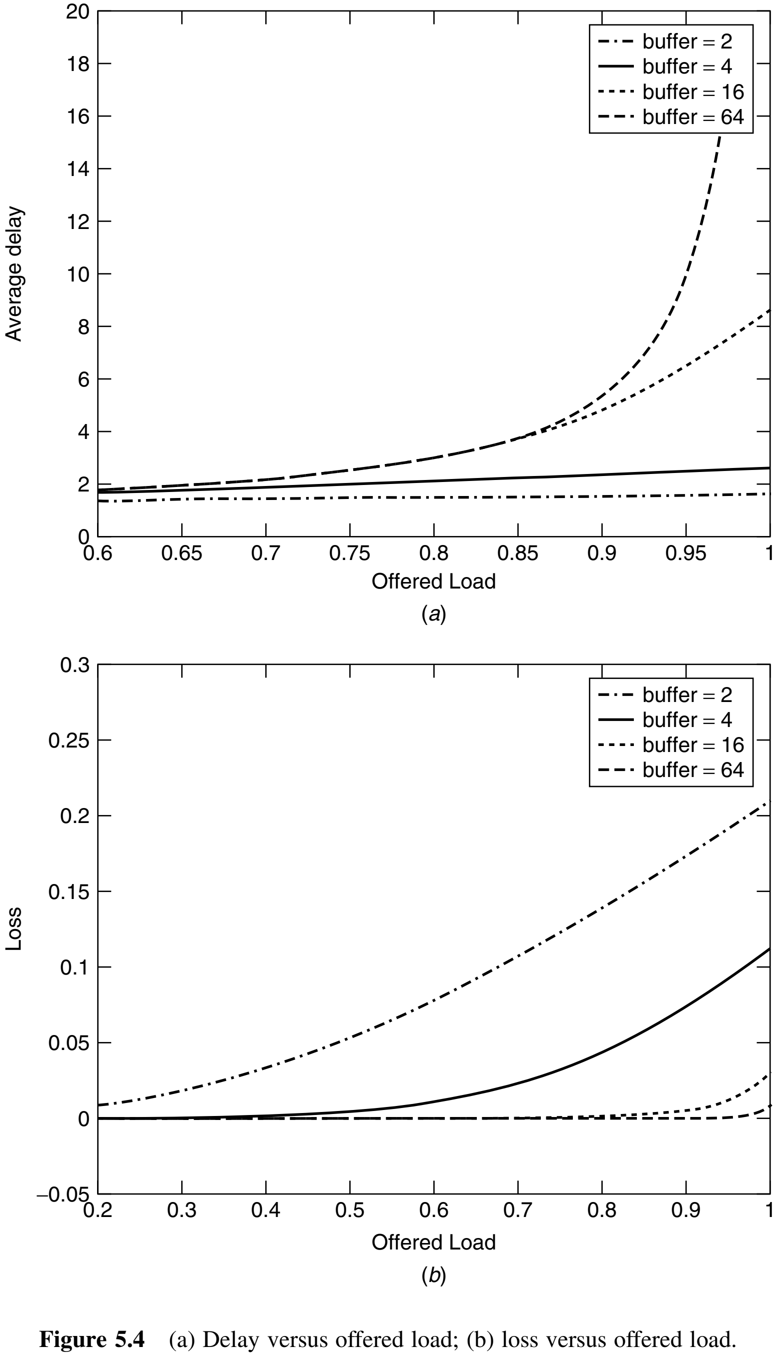
Another interesting quantity is the loss rate. An outgoing slot is empty only if the buffer is empty; accordingly, the rate on the output line is $(1 - P_0)/T$ packets/s. Since the rate at which packets arrive is $\sum_{i=1}^{\infty} ia_i/T$ packets/s, the rate at which packets are lost, normalized to the input rate, is

$$\bar{L} = 1 - \frac{1 - P_0}{\sum_{i=1}^{\infty} ia_i} \tag{5.12}$$

From Little's formula, the average delay of a packet in seconds is

$$\overline{D_P} = \frac{T\overline{N_P}}{1 - P_0} = \frac{T\sum_{i=0}^{B} iP_i}{1 - P_0} \tag{5.13}$$

Example 5.3 We assume that the arrival process is composed of the output of N independent sources with each source generating a packet with probability q in each slot. The calculation is carried out on the associated Matlab program. The results are plotted on Figures 5.4a and 5.4b, respectively. The former shows average delay and the latter, loss as a function of offered load, Nq.



(*a*)

(*b*)

Figure 5.4 (a) Delay versus offered load; (b) loss versus offered load.

5.3.2 Infinite Buffer

When the buffer is so large that the probability of overflow is negligibly small, certain simplifications in the analysis can be effected. We take expectations on both sides of (5.4). Under the assumption that a steady-state solution exists

$$\lim_{i\to\infty} E[N_{i+1}] = \lim_{i\to\infty} E[N_i] = \bar{N}$$

and we have

$$\lim_{i\to\infty} E[U(N_i)] = E[A_{i+1}] = \bar{A} \tag{5.14}$$

where $\bar{A}$ is the expected number of arrivals in a slot. Note that we have dropped the dependence on the number of packets in the buffer. Since the buffer is infinite, no packets are lost. As we have seen, this is designated as ρ. The term $U(N_i)$ is the indicator function of the event that the number of messages in the system is greater than 0. Accordingly, we have

$$E[U(N_i)] = P(N_i > 0) = 1 - P_0$$

where P_0 is the probability that the system is empty. From (5.14), we have

$$P_0 = 1 - \rho \tag{5.15}$$

There is an analysis, that leads directly to the probability-generating function for the number of packets in the buffer. We define the probability-generating function for the number of packets in the buffer as $P_i(z) = E(z^{N_i}) = \sum_{i=0}^{\infty} P_i z^i$. From (5.4), we have

$$P_{i+1}(z) \overset{\Delta}{=} E[z^{N_{i+1}}] = E[z^{N_i - U(N_i) + A_{i+1}}] = E[z^{N_i - U(N_i)}]E[z^{A_{i+1}}] \tag{5.16}$$

The term on the right here follows from the independence of the arrival process and the number of packets in the system. This is the key step following from the fact that the buffer is assumed to be infinite and has no effect on the arrival process.

Now consider the term $E[z^{N_i - U(N_i)}]$. We do a couple of simple manipulations:

$$E[z^{N_i - U(N_i)}] = \sum_{k=0}^{\infty} z^{k-U(k)} P(N_i = k) = P_0 + \sum_{k=1}^{\infty} z^{k-1} P(N_i = k)$$

$$= P_0 + z^{-1}\left[\sum_{k=0}^{\infty} z^k P(N_i = k) - P_0\right] = P_0 + z^{-1}[P_i(z) - P_0] \tag{5.17}$$

From (5.16) and (5.17), we have

$$P_{i+1}(z) = \{P_0 + z^{-1}(P_i(z) - P_0)\}A_{i+1}(z) \tag{5.18}$$

We assume stationarity of the process so that we can suppress the dependence on i; consequently $\lim_{i\to\infty} P_{i+1}(z) = \lim_{i\to\infty} P_i(z) = P(z)$, and we have

$$P(z) = \frac{P_0(1-z)A(z,T)}{A(z,T)-z} \tag{5.19}$$

This generating function allows us to find the moments of the distribution of the number of packets in the buffer. Simply setting $z = 1$ here leads to an indeterminate form. It is easier to clear fractions in (5.19) and differentiate. We have

$$\begin{aligned} &P'(z)[A(z,T)-z] + P(z)[A'(z,T)-1] \\ &\quad = P_0(-1)A(z,T) + P_0(1-z)A'(z,T) \end{aligned} \tag{5.20}$$

We let $z = 1$; since $A(1, T) = P(1) = 1$, and $A'(1,T) = \lambda T\bar{M} = \rho$, we have $P_0 = 1 - \lambda T\bar{M} = 1 - \rho$, which corresponds to the result we obtained in Section 3.2.3. Note that $\rho = \lambda T\bar{M}$ is the average number of packet arrivals in a slot. For stability, we must have $\rho < 1$.

Differentiating (5.20), we find

$$\begin{aligned} &P''(z)[A(z,T)-z] + 2P'(z)[A'(z,T)-1] + P(z)A''(z,T) \\ &\quad = 2P_0(-1)A'(z,T) + P_0(1-z)A''(z,T) \end{aligned}$$

Setting $z = 1$ and solving for $P'(1)$ gives

$$\overline{N_P} = P'(1) = \frac{(1-\rho)A'(1,T)}{1-A'(1,T)} + \frac{A''(1,T)}{2[1-A'(1,T)]} = \rho + \frac{A''(1,T)}{2(1-\rho)} \tag{5.21}$$

Since the buffer is infinite, there is no packet loss and the arrival rate to the buffer in packets per second is $\bar{A}/T = A'(1,T)/T$. The average delay of a packet is

$$\overline{D_P} = T + \frac{TA''(1,T)}{2\rho(1-\rho)} \tag{5.22}$$

By repeating the differentiation, higher-order moments can be found. Also, as we have seen in Section 2.2.2, the probability distribution of the number of packets in the buffer can be found.

We have used the probability generating function to obtain the results of this section. This approach will be continued and expanded in the next chapter when we treat the M/G/1 queue. The probability generating function approach is widely used in the analysis of queues [see Mehemet Ali et al. (2003) and Bruneel (1993).]

Example 5.4 We consider an example for which traffic is generated by N sources. In a slot, each source, acting independently, generates a message with probability P. The messages have two possible lengths: one packet and four packets. The probability of one packet is designated as Q. It is not difficult to show that the probability-generating function (PGF) for the number of packet arrivals in a slot is $A(z, T) = [P(Qz + (1 - Q)z^4) + 1 - P]^N$. The PGF for the number of packets in the buffer is found by substituting this expression into (5.19). On the associated *Maple program*, a calculation of the first two moments is carried out. For the parameters $P = 0.05$, $Q = 0.8$, and $N = 10$, the load is 0.8 and we find that the mean and the mean-square numbers of packets are 5.24 and 62.4816, respectively.

5.4 SYNCHRONOUS TIME-DIVISION MULTIPLEXING

If the flow from individual sources is high enough, dedicating transmission capacity to an individual source is warranted. This is just the situation depicted in Figure 5.1. Transmission capacity in the form of recurring time slots is dedicated to each of the data sources sharing the line. Flow on the line is blocked into fixed-length frames. The frame is allocated so that periodically recurring slots can carry only the output of a particular source. As in the case of ATDM, messages are stored in buffers until the portion of a frame dedicated to a source is available. Since the capacities allocated to sources are independent of one another, queueing for each source may be analyzed independently.

Our analysis focuses on a particular source. We take the beginning of a TDM multiplexing frame to be the slot for the particular source under study (see Fig. 5.5). We assume that the duration of the frame is T_F seconds. This would include slots for the multiplexed sources as well as guard time and a frame synchronization slot. Figure 5.5 shows the case where only one slot in a frame is dedicated to a source; however, in general, $b \geq 1$ slots may be allocated to a source. In each frame, at most b packets are removed from the buffer allocated to the source.

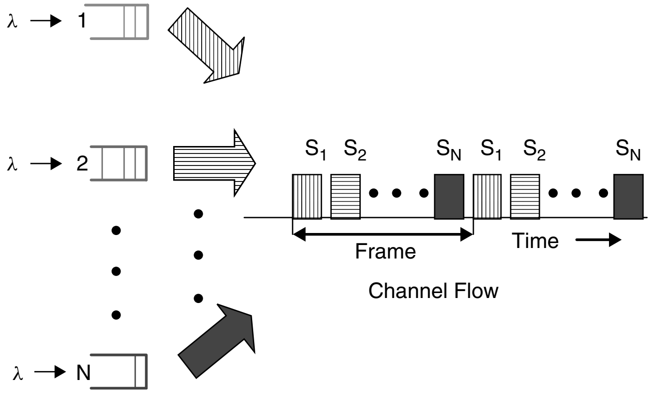


Figure 5.5 Synchronous time-division multiplexing.

We assume a gated or “please wait” strategy whereby packets arriving during a frame must wait until the next frame even though the slots may be empty. The alternative to “please wait” is “come right in,” in which newly arrived messages may be transmitted in the same frame. We write the equation for the imbedded Markov chain for the number of packets in the buffer as

$$N_{i+1} = \max(0, N_i - b) + A_{i+1} \tag{5.23}$$

where, as previously, N_i is the number of packets in the system at the beginning of the ith frame and A_i is the number of packets arriving during the ith frame. (Again, we assume an infinite buffer, so no packets are lost.) The imbedded points are the beginnings of frames. Any messages arriving during a frame are held over to be transmitted at the beginning of the next frame. In (5.23), the complicating factor is whether $N_i - b$ is negative. (In earlier analyses we had $b = 1$.) Again, we calculate the PGF of the number of packets in the buffer:

$$P_{i+1}(z) \overset{\Delta}{=} E[z^{N_{i+1}}] = E[z^{\max(N_i - b, 0) + A_{i+1}}] = \sum_{j=0}^{\infty} P_j^i z^{\max(j-b,0)} E[z^{A_{i+1}}]$$

As in Section 5.3.4, the step on the right depends on the independence of the arrival process and the number of packets in the buffer. After simple manipulation, we have

$$P_{i+1}(z) = \left[\sum_{j=0}^{b-1} P_j^i + \sum_{j=b}^{\infty} P_j^i z^{j-b}\right] E[z^{A_{i+1}}] \tag{5.24}$$

We assume that there is a steady-state solution so that $\lim_{i\to\infty} P_i(z) = P(z)$ and $\lim_{i\to\infty} P_j^i = P_j$. Continuing on, we find that

$$\begin{aligned} P(z) &= \left[\sum_{j=0}^{b-1} P_j + z^{-b}\left(\sum_{j=0}^{\infty} P_j z^j - \sum_{j=0}^{b-1} P_j z^j\right)\right] A(z, T_F) \\ &= A(z, T_F) \sum_{j=0}^{b-1} P_j (1 - z^{-b+j}) + z^{-b} A(z, T_F) P(z) \end{aligned}$$

where $A(z, T_F)$ is the PGF of the number of packets arriving in a frame. Solving for $P(z)$ yields

$$P(z) = \frac{A(z, T_F) \sum_{j=0}^{b-1} P_j (z^b - z^j)}{[z^b - A(z, T_F)]} \tag{5.25}$$

What is noteworthy in (5.25) is the presence of b unknowns $P_0, P_1, \ldots, P_{b-1}$. When $b = 1$, (5.25) reduces to the same form as (5.19). There is a single unknown P_0 that can be found from the normalizing condition $P(1) = 1$. When $b > 1$, an additional $b - 1$ equations must be obtained in order to solve for the b unknowns. These equations can be obtained by an application of Rouche's theorem.[1] Although more general solutions are possible, we restrict ourselves to the case of Poisson arrival.

5.4.1 Application of Rouche's Theorem

In order to obtain $b - 1$ additional equations we consider the properties of the functions $P(z)$, z^b, and $A(z, T_F)$. We begin by noting that the function $P(z)$ is analytic within the unit disk $|z| \leq 1$:

$$|P(z)| = \left| \sum_{i=0}^{\infty} z^i P_i \right| \leq \sum_{i=0}^{\infty} |z|^i P_i \leq \sum_{i=0}^{\infty} P_i = 1 \tag{5.26}$$

We now consider the denominator of (5.25), in particular zeros of the expression $z^b - A(z, T_F)$. Recall that $A(z, T_F)$ is the generating function of the arrival process. From Equation (2.24), we have for the PGF of the Poisson arrival process as $P(z) = e^{-\lambda T(1-z)}$. Substituting into (5.2), we find for a compound Poisson arrival process the generating function

$$A(z, T_F) = \exp(-\lambda T_F(1 - M(z))) \tag{5.27}$$

where $M(z)$ is the PGF of the number of packets in a message, λ is the number of message arrivals per second, and T_F is the duration of the frame. The assumption of Poisson arrivals simplifies the analysis; however, it can be shown that the same results can be obtained for any process for which arrivals are independent from frame to frame. Clearly, for the system to be stable, we must have

$$b > \lambda T_F M'(1) = \lambda T_F \bar{M} \tag{5.28}$$

where $\bar{M}$ is the mean number of packets in a message. We now consider points z_i, within the unit disk for which the denominator in (5.25) is equal to zero. We have

$$z_i^b = A(z_i, T_F) = e^{-\lambda T_F[1-M(z_i)]} \tag{5.29}$$

For any such point z_i, such that $|z_i| \leq 1$; $i = 1, 2, \ldots, b$, we must have a simple root. If there were a multiple root, that is, a factor $(z - z_i)^j$; $j > 1$, then the derivatives of both sides of (5.29) would be equal, and we have

$$b z_i^{b-1} = \lambda T_F M'(z_i) e^{-\lambda T_F[1-M(z_i)]} \tag{5.30}$$

Substituting (5.29) into (5.30), we have

$$b = \lambda T_F M'(z_i) z_i \tag{5.31}$$

[1] The application of Rouche's theorem to bulk departures is due to Bailey (1954). See also Boudreau et al. (1962).

But

$$\left|M'(z_i)z_i\right| = \left|\sum_{j=0}^{\infty} jz_i^{j-1}P(M=j)z_i\right| \leq \sum_{i=0}^{\infty} iP(M=i) = \bar{M} \tag{5.32}$$

implying that

$$b \leq \lambda T_F \bar{M}$$

but this contradicts the stability assumption in (5.28). Thus all roots within the unit circle must be simple if the system is stable. The next issue is how many of these simple roots there are within the unit circle. The number of zeros within the unit circle can be found by an application of Rouche's theorem.

Rouche's Theorem Given functions $f(z)$ and $g(z)$ analytic in a region R, consider a closed contour C in R; if on C, we have $f(z) \neq 0$ and $|f(z)| > |g(z)|$ then $f(z)$ and $f(z) + g(z)$ have the same number of zeros within C. We emphasize that under the conditions of the theorem, the functions $f(z)$ and $f(z) + g(z)$ have the same *number* of roots, not necessarily the same roots.

In applying this theorem, we make the following identifications: $z^b = f(z)$ and $-e^{-\lambda T_F(1-M(z))} = g(z)$. The region R consists of all z such that $|z| \leq 1 + \delta$, where $\delta > 0$. If δ is small enough, z^b and $-e^{-\lambda T_F(1-M(z))}$ are analytic in R since they are analytic in $|z| \leq 1$. We define the contour C to be $|z| = 1 + \delta'$, where $0 < \delta' < \delta$. From Taylor series expansions we have

$$\left|z^b\right| = \left|1 + \delta'\right|^b \cong 1 + b\delta' \tag{5.33}$$

$$\left|-e^{-\lambda T_F(1-M(z))}\right| \cong e^{-\lambda T_F(1-M(1))}\left|1 + \lambda T_F M'(1)\delta'\right| = 1 + \lambda T_F \bar{M}\delta' \tag{5.34}$$

Thus, from (5.33) and (5.34), we see that the conditions of Rouche's theorem are satisfied inasmuch as $|z^b| \geq |\exp(-\lambda T_F(1 - M(z)))|$ inside the contour C; accordingly, z^b and $z^b - e^{-\lambda T_F(1-M(z))}$ have the same number of roots within $|z| = 1 + \delta'$. But z^b has a root of multiplicity b at the origin; therefore $z^b - e^{-\lambda T_F(1-M(z))}$ has b roots, which are, as we have seen, distinct. One of these roots is at $z = 1$, which we designate as z_0. We designate the roots other than 1 as $z_1, z_2, \ldots, z_{b-1}$.

Since $P(z)$ is bounded on the unit disk, both the numerator and the denominator in (5.25) must be zero for the same values of z. This condition yields the set of b simultaneous equations in b unknowns $P_0, P_1, \ldots, P_{b-1}$. We have

$$\sum_{j=0}^{b-1} P_j(z_i^b - z_i^j) = 0; \quad i = 1, 2, \ldots, b-1 \tag{5.35}$$

A final equation comes from the normalizing condition $P(z)|_{z=1} = 1$. We apply l'Hôpital's rule to (5.25), obtaining

$$\sum_{j=0}^{b-1} P_j(b-j) = b - A'(1, T_F) \tag{5.36}$$

Equations (5.35) and (5.36) can be put in matrix form:

$$\mathbf{P}\Delta = \mathbf{A} \tag{5.37}$$

where $\mathbf{P} = [P_0, P_1, \ldots, P_{b-1}]$, $\mathbf{A} = [b - A'(1, T_F), 0, \ldots, 0]$ and

$$\Delta = \begin{bmatrix} b & z_1^b - 1 & \cdots & z_{b-1}^b - 1 \\ b-1 & z_1^b - z_1 & \cdots & z_{b-1}^b - z_{b-1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & z_1^b - z_1^{b-1} & \cdots & z_{b-1}^b - z_{b-1}^{b-1} \end{bmatrix}$$

By examining the determinant, we can determine whether these equations have a solution. By subtracting adjacent rows from one another, we find

$$\det \Delta = \begin{vmatrix} 1 & z_1 - 1 & \cdots & z_{b-1} - 1 \\ 1 & z_1^2 - z_1 & \cdots & z_{b-1}^2 - z_{b-1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & z_1^b - z_1^{b-1} & \cdots & z_{b-1}^b - z_{b-1}^{b-1} \end{vmatrix}$$

Each column $i = 1, 2, \ldots, b-1$ has a common factor $(z_{i-1} - 1)$, and we have

$$\det \Delta = \begin{vmatrix} 1 & 1 & \cdots & 1 \\ 1 & z_1 & \cdots & z_{b-1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & z_1^{b-1} & \cdots & z_{b-1}^{b-1} \end{vmatrix} \sum_{j=1}^{b-1} (z_j - 1)$$

Clearly this determinant cannot vanish since all of the roots, z_i, are different and none is equal to one. Thus the matrix Δ is nonsingular, and the set of linear equations has a solution.

5.4.2 Calculations Involving Rouche's Theorem

Now, we summarize the results of the preceding subsection. Finding the generating function of the number of messages is accomplished in two steps:

1. Find $(b - 1)$ roots $(z_1, z_2, \ldots, z_{b-1})$ of the equation $z^b - e^{-\lambda T_F[1-M(z)]}$.
2. Solve the linear equations (5.36) and (5.35); $i = 1, 2, \ldots, b - 1$.

Because of the form of $A(z,\ T_F)$, there is some simplification in finding the roots $z_1, z_2, \ldots, z_{b-1}$. For any integer k, we have $\exp(j2\pi k) = 1$; $k = 0, 1, \ldots, b-1$; therefore, for any root z_k the equation $z_k^b - \exp(j2\pi k - \lambda T_F(1 - M(z))) = 0$; $k = 0, 1, \ldots, b-1$ holds. Thus the multiple roots can be found by finding the roots of the $(b-1)$ equations

$$z_k = \exp\left(j\frac{2\pi k}{b} - \frac{\lambda T_F}{b}(1 - M(z))\right); \quad k = 0,1,\ldots, b-1 \tag{5.38}$$

The mean value of $\bar{N}$ can be found by differentiating with respect to z and setting $z = 1$. We have from Equation (5.25)

$$[z^b - A(z,\ T_F)]P(z) = A(z,\ T_F)\sum_{j=0}^{b-1} P_j(z^b - z^j)$$

Differentiating twice gives the two equations

$$\begin{aligned}&[bz^{b-1} - A'(z,\ T_F)]P(z) + [z^b - A(z,\ T_F)]P'(z)\\ &\qquad = A(z,\ T_F)\sum_{j=0}^{b-1} P_j(bz^{b-1} - jz^{j-1}) + A'(z,\ T_F)\sum_{j=0}^{b-1} P_j(z^b - z^j)\end{aligned}$$

and

$$\begin{aligned}&[b(b-1)z^{b-2} - A''(z,\ T_F)]P(z) + 2[bz^{b-1} - A'(z,\ T_F)]P'(z) + [z^b - A(z,\ T_F)]P''(z)\\ &\quad = A(z,\ T_F)\sum_{j=0}^{b-1} P_j[b(b-1)z^{b-2} - j(j-1)z^{j-2}]\\ &\qquad + 2A'(z,\ T_F)\sum_{j=0}^{b-1} P_j(bz^b - jz^{j-1}) + A''(z,\ T_F)\sum_{j=0}^{b-1} P_j(z^b - z^j)\end{aligned}$$

Substituting $z = 1$, we find

$$\begin{aligned}&[b(b-1) - A''(1,\ T_F)] + 2[b - A'(z,\ T_F)]P'(1)\\ &\qquad = \sum_{j=0}^{b-1} P_j[b(b-1) - j(j-1)] + 2A'(z,\ T_F)\sum_{j=0}^{b-1} P_j(b-j)\end{aligned}$$

solving for $P'(1) = \bar{N}$, we have

$$\bar{N} = \frac{\begin{bmatrix} \sum_{j=0}^{b-1} P_j[b(b-1) - j(j-1)] \\ +2A'(1,\ T_F)\sum_{j=0}^{b-1} P_j(b-j) - [b(b-1) - A''(1,\ T_F)]\end{bmatrix}}{2[b - A'(1,\ T_F)]}$$

Next we substitute $A'(1, T) = \lambda T_F \bar{M} = \rho$ and $A''(1, T) = \lambda T_F M''(1) + (\lambda T_F \bar{M})^2$ to obtain, finally

$$\bar{N} = \frac{\{\sum_{j=0}^{b-1} P_j[b(b-1) - j(j-1) + 2\rho(b-j)]\} - b(b-1) + \lambda T_F M''(1) + \rho^2}{2[b-\rho]} \tag{5.39}$$

Note that when $b = 1$, (5.39) reduces to (5.21). Rouche's theorem is being applied widely to obtain performance measures of queueing models. For a typical application of Rouche's theorem to the case of obtaining bounds on the mean delay and mean queue size for the discrete-time multi-server queue as applied to cable networks, see Denteneer et al. (2003).

Example 5.5 In this case, we assume the Poisson arrival of messages, which are geometrically distributed. The PGF for the message length is $M(z) = Pz/(1-(1-P)z)$ in (5.2). For this particular message distribution, we have $M'(1) = 1/P$ and $M''(1) = 2(1-P)/P^2$. The associated Matlab program covers an example with the parameters $P = 0.8$, $b = 4$. For an average rate of message arrival in a slot of 2, the roots are

$$z_0 = 1.0000, \quad z_1 = -0.1176 + 0.5540i,$$
$$z_2 = -0.5049, \quad z_3 = -0.1176 - 0.5540i$$

Having obtained the roots, we move on to solve for the probabilities. In another associated Matlab program, the matrix Δ is calculated and (5.37) is solved. The result is

$$P_0 = 0.1037, \quad P_1 = 0.1778, \quad P_2 = 0.1926, \quad P_3 = 0.1666$$

The average number of packets in the buffer given by (5.39) is also calculated as $\bar{N} = 1.3249$.

5.4.3 Message Delay[2]

We now consider the delay of a message as measured by the time interval between its arrival in the system and its transmission under the assumption of service in order of arrival. We shall carry through an analysis of delay under the assumption that only one packet is removed during a frame, $b = 1$. The analysis for $b > 1$ uses the same ideas, but is more complicated. As in the previous section, we assume that messages arriving after the beginning of a frame are held over until the next frame. Again we imbed a Markov chain at the beginning of the frame. There is a T-second slot dedicated to the source that we are focusing on at the beginning of each frame.

The components of message delay are depicted in Figure 5.6. We assume that a tagged message consisting of M_{L+1} packets arrives τ seconds after the beginning of a

[2] The analysis of the subsection is contained in Hayes (1974). For an alternative approch, see Lam (1977).

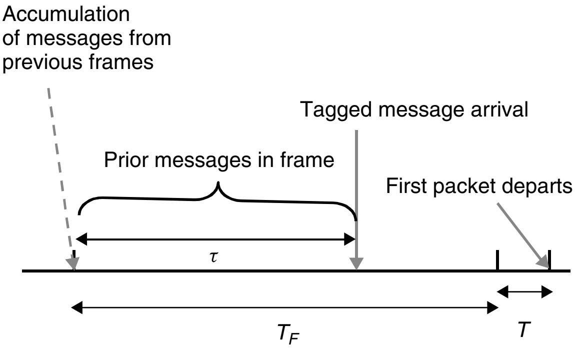


Figure 5.6 Components of message delay.

frame. The delay suffered by this message is the time required to transmit these packets plus packets already in the system when the tagged message arrived. We categorize the packets of previously arrived messages into two classes: packets held over from previous cycles and packets that have arrived during the time interval τ. Let N denote the number of packets held over from previous frames. Since we imbed the Markov chain at the beginning of the frame, we must take into account the fact that a packet is removed in each frame leaving a total of $N - U(N)$ to be transmitted in future frames. The number of messages arriving in the interval $(0, \tau)$ is denoted by the random variable L. The packets of these messages are designated $M_1, M_2, \ldots, M_L$. Thus the total number of packets to be transmitted before the tagged message is $N - U(N) + \sum_{l=1}^{L} M_l + M_{L+1}$. The first of these depart after a delay of $T_F - t + T$ seconds. Each of the remaining packets is transmitted at intervals of T_F seconds. The total delay is then

$$D_T = [N - U(N)]T_F + T_F \sum_{l=1}^{L} M_l + (T_F - \tau) + T + (M_{L+1} - 1)T_F \tag{5.40}$$

We calculate the Laplace transform for the delay by calculating the Laplace transform for independent components on the right-hand side of (5.40). We begin by calculating the probability generating function for the term $N - U(N)$. It is a straightforward manipulation to show that

$$\tilde{P}(z) = E[z^{N-U(N)}] = 1 - \rho + z^{-1}[P(z) - (1 - \rho)] \tag{5.41}$$

where $P(z)$ is the probability-generating function for n given by (5.25) for $b = 1$. The Laplace transform is found easily from

$$\tilde{P}(e^{-sT_F}) = E[e^{-s(N-U(N))T_F}] = 1 - \rho + e^{sT_F}[P(e^{-sT_F}) - (1 - \rho)] \tag{5.42}$$

Further, from basic considerations

$$E[N - U(n)]T_F = (\bar{N} - \rho)T_F \tag{5.43}$$

where $\bar{N}$ is as in (5.39) for $b = 1$. The Laplace transforms for the fourth and fifth terms in (5.40) can be found in a straightforward fashion. Let $r(t)$ denote the density function of the random variable $T_F M_l$. We have the Laplace transform

$$\mathcal{R}(s) \overset{\Delta}{=} L[r(t)] = E[e^{-sM_l T_F}] = M(e^{-sT_F}) \tag{5.44}$$

where $M(z)$ is the probability-generating function of M_i.

The second and third terms are independent of the other terms but not each other. We assume that the message in question, arrives τ seconds after the start of a frame. The probability of L previous messages arriving in the interval $(0, \tau)$ is

$$P(L \text{ messages in } \tau) = \frac{(\lambda\tau)^L e^{-\lambda\tau}}{L!} \tag{5.45}$$

As we have seen in Chapter 3 when we discussed the properties of the Poisson process, given that there is at least one arrival in an interval, the arrival time of a particular message is uniformly distributed in the interval. This is true for all messages. Consider the random variable $F \overset{\Delta}{=} T_F \sum_{l=1}^{L} M_l + (T_F - \tau)$ and the probability that it lies in the interval $(t, t + dt)$. There are three random variables involved here: L, which is distributed as a Poisson random variable as indicated in (5.45); $M_i T_F$, which has density function $r(t)$; and T, which is uniformly distributed in the interval $(0, T_F)$. We condition first on L and τ

$$P[t < F \leq t + dt | L, \tau] = r^{(L)}(t - (T_F - \tau))dt$$

where $r^{(L)}(t)$ is the L-fold convolution of $r(t)$. Averaging over L first, we have

$$P[t < F \leq t + dt | \tau] = \sum_{L=0}^{\infty} e^{-\lambda\tau} \frac{(\lambda\tau)^L}{L!} r^{(L)}(t - (T_F - \tau))dt$$

Finally, averaging over τ, which is uniformly distributed in $(0, T_F)$, gives

$$f(t)dt = P[t < F \leq t + dt] = \frac{1}{T_F} \int_0^{T_F} d\tau \sum_{L=0}^{\infty} e^{-\lambda\tau} \frac{(\lambda\tau)^L}{L!} r^{(L)}(t - (T_F - \tau))dt \tag{5.46}$$

The Laplace transform of this probability density function can be shown to be

$$\mathcal{F}(s) = \frac{e^{-\lambda T_F}[1 - \mathcal{R}(s)] - e^{-sT_F}}{T_F\{s - \lambda[1 - \mathcal{R}(s)]\}} \tag{5.47}$$

where $\mathcal{R}(s)$ is the Laplace transform of $r(t)$ [see (5.44)]. Differentiating (5.47) with respect to s and setting $s = 0$, we have for the mean value

$$\bar{F} = \left.\frac{d\mathcal{F}(s)}{ds}\right|_{s=0} = \frac{T_F}{2} + \frac{\lambda T_F^2 \bar{M}}{2} \tag{5.48}$$

Recall that the random variable F is the sum of $T_F - \tau$ and the time required to transmit the messages that arrive in $(0, \tau)$. This is certainly an intuitively appealing result. Since τ is uniformly distributed in $(0, T_F)$, its mean is $T_F/2$. The average number of packets that arrive in τ seconds is $\bar{M}\lambda\tau$. T_F seconds are required to transmit each packet. Since the four terms $(N - U(N))T_F$, $F = T_F \sum_{l=1}^{L} M_l +$ $(T_F - \tau)$, $(M_{L+1} - 1)T_F$ and T are independent of one another, the Laplace transform of the total delay can be found by multiplying the Laplace transforms of the density functions of individual terms. We have

$$D(s) = \tilde{P}(e^{-sT_F})F(s)M(e^{-sT_F})e^{s(T_F - T)} \tag{5.49}$$

By differentiating with respect to s and setting $s = 0$ we find

$$\bar{D} = T_F[N - \rho] + \bar{F} + T + (\bar{M} - 1)T_F$$

From (5.39) and (5.48), respectively, we find expressions for $\bar{N}$ and $\bar{F}$. We have, finally

$$\bar{D} = \bar{M}T_F + T - \frac{T_F}{2} + \frac{\lambda T_F^2 \overline{M^2}}{2(1 - \rho)} \tag{5.50}$$

where $\rho = \lambda T_F \bar{M}$ and T_F is the frame duration. In (5.50), $\bar{M}T_F$ is the average transmission time of a message that has queued for service. The factor $T - T_F/2$ accounts for the random arrival of a message during a frame. Finally, the last term is the queueing delay.

Example 5.6 We continue with the message length distribution given in Example 5.5. (*Note*: $b = 1$ for this example.) For the geometric distribution appropriate to message length distributions, we have $\bar{M} = 1/P$ and $\overline{M^2} = (2 - P)/P^2$. We apply

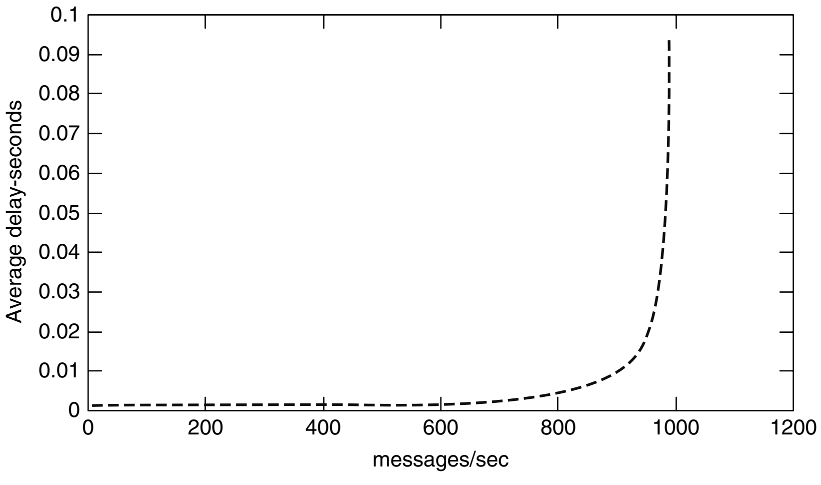


Figure 5.7 Average delay versus message arrival rate.

this to transmission over a T1 line. The frame duration is $T_F = 125\ \mu s$. This contains twenty-four 8-bit user slots and a single framing bit; accordingly, the duration of a slot is $T = 5.18\ \mu s$. The results are shown on the associated spreadsheet, where average delay is shown as a function of load with P as a parameter. The particular example shown in Fig. 5.7 is for $P = 0.125$.

5.5 RANDOM ACCESS TECHNIQUES

5.5.1 Introduction to ALOHA[3]

The salient result that emerges from the analyses of TDM systems is the deterioration of performance as the number of terminals increases irrespective of the total traffic in the system. This is readily seen from (5.50). Increasing the number of users requires a linear increase in T_F. The reason is that capacity is allocated to sources, even when the source has nothing to send. This observation motivates the consideration of random access techniques, which allocate bandwidth purely in response to traffic. As we shall see, random access systems are relatively insensitive to the number of active terminals in the systems so that performance is best when there are a large number of lightly loaded sources. The difficulty is that these systems are very sensitive to increases in the level of aggregate traffic. In this section, we use the Markov chain technique to analyze the performance of random access systems.

The genesis of the random access technique was in the ALOHA system at the University of Hawaii. The ALOHA system was a packet-switched network in which a radio channel was shared among a number of users. When a terminal generates a message, it is transmitted immediately regardless of the activity at the other terminals in the system (see Fig. 5.8). Of course, there is the obvious difficulty that messages from two or more terminals transmitting simultaneously collide and are thereby rendered unintelligible. By various mechanisms the terminal can be apprised of a collision. For example, in systems where every terminal can hear every other, collisions will be detected simply by monitoring the line. Another possibility is for the receiving terminal to return positive acknowledgments when a message is received correctly. If a transmitting terminal has not received a positive acknowledgment after an appropriate interval, it is presumed that a collision has taken place. When a terminal discovers, by whatever means, that a collision has occurred, the message is retransmitted. In order to avoid repeated collisions involving the same message, each terminal should have a different timeout interval. A simple and fair way to do this is for the timeout interval for each terminal to be a random variable. If the variance of this random variable is large enough, the probability of repeated collisions is small. However, choosing a variance that is too large increases the delay of a message.

A significant feature of the ALOHA system is instability—a property that we shall illustrate by means of a simple analysis[4] that illustrates the basic instability in

[3] For further details, see Abramson (1970, 1973) and Rom and Sidi (1990).

[4] For a rigorous, comprehensive treatment of instability in random access systems, the reader is referred to the work in Meditch and Lea (1983); see also Tsybakov (1985).

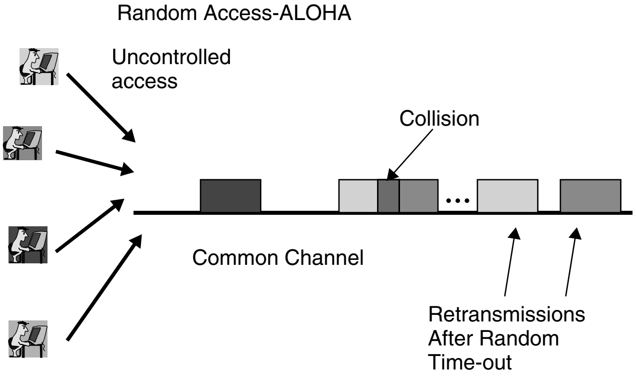


Figure 5.8 Line flow on the pure ALOHA channel.

the ALOHA system. Assume that all the terminals sharing the same channel generate fixed-length messages at a Poisson rate. Assume also that there are many lightly loaded terminals so that each holds no more than a single message at a time. Let the time in seconds required to transmit a message over the channel be denoted as m, and let λ denote the average rate, in messages per second, at which messages are generated by all sources. The dimensionless quantity $\rho = m\lambda$ is the, by now familiar, load offered to the communications channel. In order for the system to be stable, we must have $m\lambda < 1$ in the absence of overhead traffic. The flow on the line will consist of retransmitted as well as newly generated messages. Let the total average rate of flow of new and retransmitted messages be denoted as Λ messages per second. In order to carry the analysis forward it is assumed that the total traffic on the line has a Poisson distribution. In making the Poisson assumption, the reliance is on the mixing of traffic from a large number of lightly loaded terminals. A message will be involved in a collision if another terminal transmits in a window $2m$ seconds in duration about the message (see Fig. 5.9).

From the Poisson assumption, we have

$$P(\text{collision}) = 1 - e^{-2\Lambda m}$$

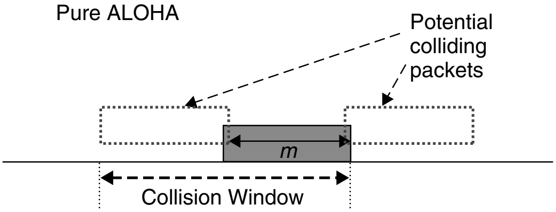


Figure 5.9 Pure ALOHA collision mechanism.

The average rate at which messages are retransmitted is $\Lambda(1 - e^{-2\Lambda m})$. Adding the newly generated and the retransmitted message traffic, we have

$$\Lambda = \lambda + \Lambda P(\text{collision}) = \lambda + \Lambda(1 - e^{-2\Lambda m}) \qquad (5.51)$$

The quantities $\rho = \lambda m$ and $R = \Lambda m$ are, respectively, the newly generated load and the total carried load on the channel. From (5.51), we have

$$\rho = Re^{-2R} \qquad (5.52)$$

We point out in connection with (5.52) that ρ is the independent variable and R the dependent one. Equation (5.52) is plotted in Figure 5.10. (See the associated spreadsheet.) We see that for ρ small enough, the relation is linear; however, as ρ increases, the proportion of retransmission increases and there is a point of saturation. By solving the equation $d\rho/dR = 0$, the point of saturation can be shown to be $R = \frac{1}{2}$ and $\rho = \frac{1}{2}e \cong 0.18$. Thus, although the delay at very light loading is minimal, less than one-fifth of the channel capacity is available to carry newly generated traffic. It is true that the analysis here is based on a Poisson assumption; however, simulation studies show that (5.52) holds for more realistic traffic flow distributions. We point out that there is potential instability in that for a wide range of newly offered traffic there are two possible loads on the line, values of R. The larger of the two implies a larger volume of retransmitted messages, a phenomenon that implies larger message delay.

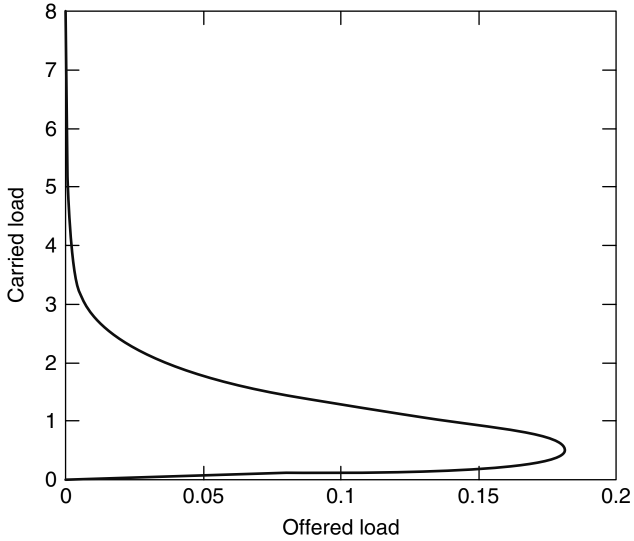


Figure 5.10 ALOHA throughput.

The throughput performance of the ALOHA system can be improved dramatically by adding a small amount of structure in the form of timing pulses broadcast to all terminals so that flow in the line is slotted, as in TDMA. The improved system is called *slotted ALOHA* (Roberts 1975). We assume that the messages fit into a slot m seconds in duration. Newly arrived messages are transmitted in the next available slot. Since only messages generated in the same slot in between can interfere with one another, the probability of a message being involved in a collision is considerably reduced (see Fig. 5.11). For very light loading this leads to a delay between message generation and message transmission averaging $m/2$ seconds—a small price to pay.

Again we assume that the total traffic flow newly generated and retransmitted is Poisson with average rate Λ messages per second. The probability of a message being involved in a collision is

$$P(\text{collision}) = 1 - e^{-\Lambda m}$$

Following the same line of reasoning that led to (5.52) gives in the case of slotted ALOHA

$$\rho = Re^{-R} \tag{5.53}$$

The point of saturation is doubled, residing at $\rho = 1/e \cong 0.36$. Thus with a small amount of overhead, which causes a minimal deterioration of performance, the channel throughput has doubled. (See the associated spreadsheet for the calculation.) We notice once again that there is potential instability since there are two possible loads on the line for the same offered load. We will see this phenomenon again, in the next subsection when we analyze delay.

5.5.2 Analysis of Delay

For both ALOHA and slotted ALOHA we have seen evidence for unstable operation in that there are two operating points for the same input rate to the

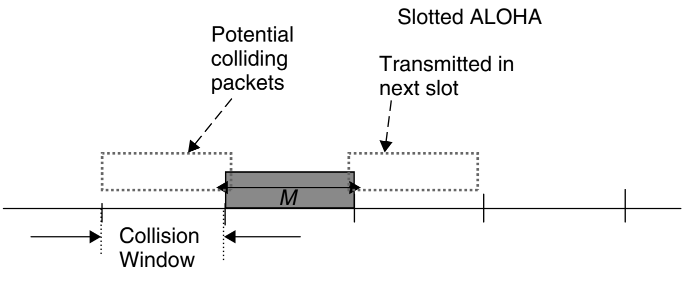


Figure 5.11 Slotted ALOHA.

system. In this section, we analyze the performance of slotted ALOHA by means of a Markov chain embedded at the slot boundaries (Kleinrock and Lam (1975)). We shall assume that there are N terminals sharing the common line. As earlier, it is assumed that each message has a constant transmission time equal to the duration of a slot. The general description of the system would require the number of messages at each terminal and the collision-retransmission history for each terminal. Clearly, this is far too complex for our purposes; consequently simplifying assumptions are in order. First, we assume that each terminal can hold only one message. This assumption fits the application, namely, systems with many lightly loaded terminals. It would model, for example, a single user at a terminal who generates messages in an interactive process. The second assumption is memorylessness of the message arrival process. We assume that the generation of messages is independent from slot to slot and that the probability of a message being generated in a slot is σ. This arrival process would hold if the interval to a message generation at an empty terminal is exponentially distributed with mean value $1/\lambda$. We have

$$\sigma = 1 - e^{-\lambda T} \tag{5.54}$$

where T is the duration of a slot.

A final simplifying assumption is a certain memorylessness in connection with the random retransmission interval after a collision. We assume that after an initial transmission that suffers a collision, the message is retransmitted in every succeeding interval with probability α until a successful transmission has taken place. The average retransmission interval is $1/\alpha$. However, simulation studies have shown that performance is not sensitive to the form of the probability distribution of the retransmission interval as long as the mean value is the same. For example, suppose that instead of a geometrically distributed retransmission interval, the interval is uniformly distributed in slots $D+1, D+2, \ldots, D+K$, where D is a detection interval after initial transmission. In the latter case, the mean retransmission interval is $D + K/2$. Equating mean retransmission times for the two cases, we find that the performances of the two techniques will be the same as long as $\alpha \cong 1/(D + K/2)$.

Now, we use the Markov chain that we have just defined to calculate the average delay. We derive the state transition matrix for the chain. The solution to the steady-state equilibrium distribution of the chain allows one to find the average number of messages in the system. The application of Little's formula yields average delay. This technique used in the first analysis of delay for slotted ALOHA has provided a basic methodology, which has been used in a number of subsequent analyses of more sophisticated random access systems (see Tobagi (1982), for example).

Recall that the state of the system is the number of backlogged terminals or, equivalently, the number of messages in the system, since a terminal may hold no more than one message and the probability of a message arriving in a slot is σ and that the probability of retransmitting a message in a slot is α. We begin the analysis with the following probabilities from the binomial distribution.

$$P(j \text{ transmissions in a slot}/i \text{ backlogged terminals}) = \binom{i}{j}\alpha^j(1-\alpha)^{i-j};\quad j = 0, 1, \ldots, i \tag{5.55}$$

and

$$P(j \text{ arrivals in a slot}/i \text{ backlogged terminals}) = \binom{N-i}{j}\sigma^j(1-\sigma)^{N-i-j};\quad j = 0, 1, \ldots, N-i \tag{5.56}$$

As in the case of TDMA, we derive the state transition matrix with the elements

$$t_{ij} = P(j \text{ backlogged terminal slot } n/i \text{ backlogged terminal slot } n-1) \tag{5.57}$$

We can find expressions for this transition probability. Consider the first row of the matrix, which shows the transition probabilities from a system with no backlogged terminals. A system that is empty will remain empty if there are one or no arrivals. (Recall that newly arrived messages must be immediately transmitted.) We then have from (5.56)

$$t_{00} = (1-\sigma)^N + N\sigma(1-\sigma)^{N-1} \tag{5.58}$$

If two or more terminals generate messages, then there will be a collision and all messages remain in the system

$$t_{0j} = \binom{N}{j}\sigma^j(1-\sigma)^{N-j};\quad j = 2, 3, \ldots, N \tag{5.59}$$

If there are two or more arrivals, there is a conflict, and these packets remain in the system. Notice that there is no transition from 0 to 1, $t_{01} = 0$.

Now suppose that there are $i > 0$ backlogged terminals, each of which transmits in the next slot with probability α. Also, there are arrivals to each of the empty terminals with probability σ. Since only one message can be transmitted in a slot, we have

$$t_{ii-k} = 0;\quad k > 1 \tag{5.60}$$

The number of backlogged terminals will be reduced by one, if there are no arrivals to empty stations and only one of the backlogged packet transmits:

$$t_{ii-1} = i\alpha(1-\alpha)^{i-1}(1-\sigma)^{N-i};\quad 0 < i \le N \tag{5.61}$$

There are three mutually exclusive ways for the number of backlogged packets to remain constant: (1) no more than one new arrival and no transmission from backlogged terminals, (2) no new arrivals and only one retransmission from the backlogged terminals, and (3) no new arrivals and transmission from two or more backlogged terminals:

$$t_{ii} = (N-i)\sigma(1-\sigma)^{N-i-1}(1-\alpha)^i + (1-\alpha)^i(1-\sigma)^{N-i}$$
$$+ (1-\sigma)^{N-i}(1-(1-\alpha)^i - i\alpha(1-\alpha)^{i-1}); \quad 0 < i \le N-1 \tag{5.62}$$

There will be an increase by one if there is only one new arrival and if one or more backlogged terminals transmit:

$$t_{ii+1} = (N-i)\sigma(1-\sigma)^{N-i-1}[1-(1-\alpha)^i]; \quad 0 < i \le N-1 \tag{5.63}$$

Finally, if there are two or more arrivals in a slot, the number of backlogged packets increases by this amount no matter what the backlogged terminals do. We have

$$t_{ii+j} = \binom{N-i}{j}\sigma^j(1-\sigma)^{N-i-j}; \quad 0 < i \le N-1,\ 1 < j \le N-i \tag{5.64}$$

Since $t_{ij} = 0$ for $j < i-1$, the state transition matrix has the same upper triangular form as (5.5); consequently, the same solution technique can be applied. If we assume a steady state probability for the number of backlogged terminals, $P_0, P_1, \ldots, P_N$ respectively, we can write, similar to (5.7)

$$P_i = \begin{cases} \sum_{j=0}^{i+1} P_j t_{ij}; & 0 \le i \le N-1 \\ \sum_{j=0}^{N} P_j t_{iN}; & i = N \end{cases}$$

As in subsection 5.3.2, the solution is straightforward. We solve for P_{i+1}

$$P_{i+1} = \frac{P_i - \sum_{j=0}^{i} P_j t_{ji}}{t_{i+1i}} \tag{5.65}$$

In order for this solution to be valid, we require $t_{i+1i} > 0$; otherwise, the system is unstable. The final, normalized solution is found by following the steps of (5.10). The steady-state probabilities are found from the normalizing condition analogous to (5.9).

The average number of backlogged terminals in the system is given by

$$\bar{K} = \sum_{i=0}^{N} iP_i \tag{5.66}$$

Since messages cannot arrive to a backlogged terminal, the average arrival rate per slot to the system is

$$S_{\text{in}} = \sigma(N - \bar{K}) \tag{5.67}$$

Now we use Little's formula to find the average delay of a message. Since each backlogged terminal holds one message, there are $\bar{K}$ of these. To these must be added newly arrived messages in nonbacklogged terminals, which are awaiting transmission in the next slot. The average number of these messages is just S_{in} giving a total of $\bar{K} + S_{\text{in}}$ message in the system on average. From (5.67), average delay of a message in slots is

$$\bar{D} = 1 + \frac{\bar{K}}{\sigma(N - \bar{K})} \tag{5.68}$$

The average delay here is the interval between message arrival and its final transmission when it departs the system. In a realistic system allowance must be made for time to determine whether a collision has taken place. Thus, this detection time must be added to the average delay determined in (5.68) if one were to compute the average time that a message must be held at the transmitter until it can be said to have been received correctly.

Example 5.7 The average delay $\bar{D}$ of the system as a function of S_{in} the average message arrival rate in a slot, with α as a parameter is shown in Figure 5.12. As we see, there is clear evidence of instability. Instability is indicated by two values for delay with the same input. Increasing α causes instability since collisions increase.

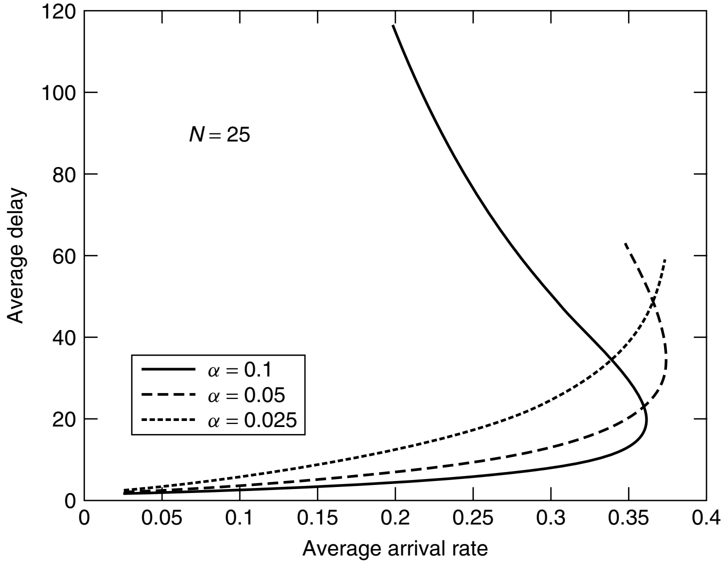


Figure 5.12 Average delay versus average message arrival rate.

REFERENCES

Abramson, N., “The ALOHA system—another alternative for computer communications,” *1970 Fall Joint Computer Conf., AFIPS Conf. Proc.*, Vol. 37, 1970, pp. 281–285.

Abramson, N., “The ALOHA system,” in *Computer Communications Networks*, N. Abramson and Kuo, F., eds., Prentice-Hall, Englewood Cliffs, NJ, 1973.

Bailey, N. T. J., “On queueing processes with bulk service,” *J. Roy. Stat. Soc.* **16**(2), pp. 80–87, (Feb. 1954).

Boudreau, P. E., J. S. Griffin, and M. Kac, “An elementary queueing problem,” *Am. Math. Monthly* **69**: 713–724 (1962).

Bruneel, H., and B. J. Kim, *Discrete-Time Models for Communication Systems Including ATM*, Kluwer Academic Publishers, 1993.

Carleial, A. B., and M. E. Helman, “Bistable behavior of ALOHA-type systems,” *IEEE Trans. Commun.* **COM-23**(4): 401–410 (April 1975).

Chu, W. W., “A study of asynchronous TDM for time sharing computer systems,” *AFIPS Conf. Proc., Fall Joint Computer Conf.*, Vol. 35, 1969, pp. 669–675.

Chu, W. W., and A. G. Konheim, “On the analysis and modeling of a class of computer communication systems,” *IEEE Trans. Commun.* **20**(3): 11, 645–660 (June 1972).

Denteneer, D., et al., “Bands for a discrete-time multi-server queue with an application to cable networks,” 18th International Teletraffic Congress, Berlin, Germany, 31 August–5 September 2003.

Hayes, J. F., “Performance models of an experimental computer communications network,” *Bell Syst. Tech. J.* **53**(2): 225–259 (Feb. 1974).

Hayes, J. F., and D. N. Sherman, “A study of data multiplexing techniques and delay performance,” *Bell Syst. Tech. J.* **51**: 1985–2011 (Nov. 1972).

Kleinrock, L., and S. S. Lam, “Packet switching in a multiaccess broadcast channel: performance evaluation,” *IEEE Trans. Commun.* **COM-23**(4): 410–422 (April 1975).

Lam, S. S., “Delay analysis of a time-division multiple access channel,” *IEEE Trans. Commun.* **COM-25**(12): 1489–1494 (Dec. 1977).

Lam, S. S., and L. Kleinrock, “Packet switching in a multiaccess broadcast channel: Dynamic control procedures,” *IEEE Trans. Commun.* **COM-23**(9): 891–904 (Sept. 1975).

Meditch, J. S., and C.-T. A. Lee, “Stability and optimization of the CSMA and CSMA/CD channels,” *IEEE Trans. Commun.* **COM-31**(6): 763–774 (June 1983).

Mehemet Ali, M., X. Zhang, and J. F. Hayes, “A performance analysis of a discrete-time queueing system with server interuptions for modeling wireless ATM multiplexers,” *Performance Evaluation*, **51**, pp 1–31, 2003.

Roberts, L. G., “ALOHA packet system with and without slots and capture,” *Compu. Commun. Rev.* **5**: 28–42 (April 1975).

Rom, R., and M. Sidi, *Multiple Access Protocols: Performance and Analysis*, Springer-Verlag, 1990.

Saadawi, T. N., and A. Ephremedes, “Analysis, stability and optimization of slotted ALOHA with a finite number of buffered users,” *IEEE Trans. Automatic Control* **AC-26**(3): 680–689 (June 1981).

Tobagi, F. A., “Carrier sense multiple access with message based priority functions”, *IEEE Trans. Commun.* **COM-30**(1): January 1982.

Tsybakov, B. S., “Survey of USSR contributions to random multi-access communications,” *IEEE Trans. Inform. Theory* **IT-31**(2): 143–166 (March 1985).

EXERCISES

5.1 Voice signals are encoded at a rate of 56 kbps.

(a) How many ATM cells are required to encode a 100-ms talk spurt?

(b) If talk spurts are exponentially distributed, what is the probability distribution of the number of cells in a talk spurt?

5.2 Suppose that messages arrive at a constant rate of R per second and that the number of packets in a message follows the binomial distribution.

(a) Find the probability-generating function for the number of packets that arrive in a T-second interval.

(b) What is the average number of arrivals in T seconds?

5.3 Suppose that random-length messages arrive at a constant rate of R messages per second. Suppose also that the number of packets per message is geometrically distributed.

(a) Find the generating function for the number of packets in the buffer.

(b) What is the mean number of packets in the buffer if we have ATDM where a single packet is removed at each time slot?

5.4 Repeat Exercise 5.3 for the case of Poisson arrival of messages.

5.5 Consider asynchronous TDM with a slight twist. Suppose that data traffic shares the line with higher-priority traffic. Such traffic may be, for example, voice. Because of this other traffic, the line is not always available. A crude way of calculating the effect on data traffic is to assume that a packet is transmitted in a slot with probability $p < 1$, with independence from slot to slot. Notice that for $p = 1$ we have the same case as before.

(a) Set up the embedded Markov chain for the number of packets in the buffer.

(b) Calculate the probability-generating function for the number of packets in the buffer.

(c) Find the mean number of packets in the buffer.

5.6 Suppose that we have an N source model for ATDM. Each of the N sources generates packets independently. The probability that a source will generate a packet in T seconds is P.

(a) Find the probability-generating function of the arrivals.

(b) Find the steady-state probability-generating function for the number in the system assuming an infinite buffer.

(c) Write down the load condition for the existence of a steady-state solution for the finite-source model.

5.7 Suppose that we have an ATDM system in which the buffer can hold 5 bytes. Suppose also that there are two sources, each of which may generate a message consisting of one byte independently in a slot with probability $\frac{1}{4}$.

(a) What are the buffer occupancy probabilities?

(b) What is the throughput?

5.8 Consider a synchronous TDM system in which 10 sources share the same line equally. Compare the average number in the buffer when there are one and two dedicated slots per user per frame (i.e., $b = 1,2$). Assume that messages are geometrically distributed with parameter P. P should be varied over a range of value i from $1/(1-P) < 1$ to $1/(1-P) > 2$. (You will need a computer to do this problem.)

5.9 Consider an ATM system in which there are two classes of messages. The first class is control messages that are a constant 512 bits long. The second class consists of data messages, which are uniformly distributed between 1024 and 4096 bits. Assume that there are five control messages for every information message.

(a) What is the probability distribution of the number of bits required to transmit a message?

(b) What is the probability-generating function of the number of bits in a message?

(c) The messages are transmitted over a line that operates at a rate of 148.608 Mbps (OC-3). What is the average message transmission time?

5.10 Consider a STDM system in which 2 slots are dedicated to each user in a frame. Assume 10 users. Each user generates messages at a Poisson rate of 1.5 per frame. Find the average delay in terms of slot-tines.

5.11 Suppose that the terminals in a slotted ALOHA system have a good estimate for the number of backlogged terminals in the system. A terminal decreases the collision probability by choosing the retransmission probability to be inversely proportional to the number of backlogged terminals. Reproduce Figure 5.12 for this case.

5.12 Suppose that n terminals in a slotted ALOHA system have messages. Assume that newly arrived messages are not transmitted until all n messages have been transmitted.

(a) Show that the average number of slots until one of these messages is transmitted is $1/[nP(1-p)^{n-1}]$, where P is the probability of transmitting in a slot.

(b) What is the average number of slots required to transmit all n messages?

6

THE M/G/1 QUEUE: IMBEDDED MARKOV CHAINS

6.1 THE M/G/1 QUEUE[1]

In the previous chapter, we analyzed systems by a Markov chain imbedded at fixed intervals. In this chapter, we shall study a technique whereby a Markov chain is imbedded at randomly spaced points. The terms *imbedded Markov chain* and *semi-Markov* are applied to this technique. We apply the technique to the study of the M/G/1. The same approach is applied to the G/M/1 queue.

The challenge in studying the M/G/1 queue by means of Markov chains is to find a sequence of points in time, which allows the state of the system to be described by a minimal number of parameters. At a random point in time the remaining service time must be considered as well as the number of messages in the system. Since the exponential distribution is memoryless, this complication is not present for the M/M/1 queue, for example. The solution to the problem is to imbed the Markov chain at the points where a message departs the system. At the point of message departure, no new message transmission has begun; consequently, all that is required to predict the evolution of the system is the number of messages in the system. Implicit here is the memorylessness of the Poisson arrival process.

As we shall see, the primary mathematical techniques that we use in the analysis of the imbedded Markov chain are the Laplace transform of the probability density function of continuous random variables and the probability generating function of discrete random variables. As we have seen in Chapter 2, for both of these, numbers

[1] The definitive source is Cohen (1969).

Modeling and Analysis of Telecommunications Networks, by Jeremiah F. Hayes and
Thimma V. J. Ganesh Babu
ISBN 0-471-34845-7 © 2004 John Wiley & Sons, Inc.

on performance, such as moments and probabilities, are obtained by the evaluation of derivatives.

6.1.1 Imbedded Markov Chains[2]

We focus now on the number of messages in the queue at the time instants when a transmission is completed. Notice that these points have the unique property of being independent of message transmission time since no message transmission is in progress. Suppose that the ith departing message leaves behind a non-empty system with N_i messages. At this point, transmission of the $(i+1)$st message begins. While this is in progress, messages continue to arrive. We denote the number of such arrivals by A_{i+1}. The number of messages in the system at the next departure is given by

$$N_{i+1} = N_i - 1 + A_{i+1}; \quad N_i \geq 1 \tag{6.1}$$

If the departure of the ith message leaves the system empty, a slightly different equation describes the process. The $(i+1)$st departing message arrived to an empty queue, consequently it leaves behind only those messages that arrived during its service period. We have

$$N_{i+1} = A_{i+1}; \quad N_i = 0 \tag{6.2}$$

Note that, in both equations, (6.1) and (6.2), the number of new arrivals equals the number of arrivals during a message transmission time. The state dynamics can be summarized in the following equation:

$$N_{i+1} = N_i - U(N_i) + A_{i+1} \tag{6.3}$$

where the unit step

$$U(x) = \begin{cases} 0, & x \leq 0 \\ 1, & x > 0 \end{cases}$$

Of course, this is the same as (5.4) derived in Chapter 5 for TDM. There are several differences:

1. Currently, the imbedded points are randomly spaced, whereas in the previous chapter, there were at the slot boundaries.
2. We are now dealing with variable-length messages, whereas previously we treated fixed-length packets.
3. Finally, A_i is the number of messages that arrive during a message transmission, not the packet arrivals during a slot time.

[2] The analysis of the M/G/1 queue by means of the imbedded Markov chain is due to Kendall (1953).

The dynamics of the number of messages in the system embodied in (6.3) do not depend on the order in which messages are served. Thus, for example, we may have first-come-first-served (FCFS), last-come-first-served (LCFS), or random order of service (RA).

If the average arrival rate of message arrivals during a message transmission is less than one, then the system is stable. It can be shown that the distribution of the number of messages in the system at the departure epochs is equal to the distribution of the number at the arrival epoch for Poisson arrivals. As we have seen, it is also equal to the number seen by a random observer (PASTA—Section 3.3.6).

We proceed to find this steady-state distribution beginning with a calculation of mean values. The steps here are almost the same as that in Section 5.4, so we can be more succinct. Under the assumption that a steady-state solution exists, we can show that

$$P_0 = 1 - \rho \tag{6.4}$$

where ρ is the average number of message arriving during a message transmission time, $\rho = \bar{A}$. As noted above, there are two random quantities involved in A_{i+1}: the duration of the message and the number of arrivals while a message is being transmitted. We compute $\bar{A}$ by first conditioning on the message length and then averaging over the message length:

$$\bar{A} = E[A_{i+1}] = \int_0^\infty E[A_{i+1}/\text{message length} = t]m(t)dt = \int_0^\infty \lambda t m(t)dt = \lambda\bar{M} \tag{6.5}$$

where $m(t)$ is the probability density of message length and $\bar{M}$ denotes the mean time required to transmit a message. In the analysis of the M/M/1 queue (Section 3.4), we had $\bar{M} = 1/\mu$ and $\rho = \lambda/\mu$. Therefore if we define the intensity as $\rho = \lambda\bar{M}$, we are consistent with earlier results.

We can derive the probability-generating function for the number of messages in the system at the departure points in exactly the same way as in Section 5.3.2. The result is written

$$N(z) = \frac{P_0(1-z)A(z)}{A(z) - z} \tag{6.6}$$

In this case, $A(z)$ is the probability-generating function for the number of messages arriving in a message transmission time. In Chapter 3 we found a particular relationship between the probability generating function of the number of arrivals over a random time interval and the Laplace transform for the interval [see (3.38)]. In this case the random interval is the duration of a message whose density function $m(t)$ has the Laplace transform $\mathcal{M}(s)$. From this relationship, we have in the present case

$$A(z) = \int_0^\infty e^{-\lambda t(1-z)} m(t)dt = \mathcal{M}(\lambda(1-z)) \tag{6.7}$$

Differentiating $A(z)$ with respect to z and setting $z = 1$, we find

$$A'(z)|_{z=1} = \bar{A} = \lambda\bar{M} = \rho \tag{6.8}$$

$$A''(z)|_{z=1} = \lambda^2 M''(0) = \lambda^2\bar{M}^2 \tag{6.9}$$

The mean-square value of the number of arrivals is

$$\bar{A}^2 = \lambda^2\bar{M}^2 + \lambda\bar{M}$$

where $\bar{M}^2$ is the mean-square message transmission time. As in the derivation of (5.21), after successive differentiations, we obtain

$$\bar{N} = N'(1) = \frac{(1-\rho)A'(1)}{1-A'(1)} + \frac{A''(1)}{2[1-A'(1)]} \tag{6.10}$$

The next step is to substitute (6.8) and (6.9) into (6.10) to obtain

$$\bar{N} = \rho + \frac{\lambda^2\bar{M}^2}{2[1-\rho]} \tag{6.11}$$

Applying Little's formula, $\bar{D} = \bar{N}/\lambda$ we have the celebrated Pollaczek–Khinchin[3] formula for average delay in the M/G/1 queue.

$$\bar{D} = \bar{M} + \frac{\lambda\bar{M}^2}{2(1-\rho)} \tag{6.12}$$

The form of this equation exhibits the components of total delay. $\bar{M}$ is the average time to transmit a message, and the second term in (6.12) is the time spent in the queue waiting for service. When $\lambda \to 0$, $\bar{D} \to \bar{M}$ and queueing delay is negligible. As λ increases, the second component, which accounts for queueing delay, predominates. As ρ approaches 1, the queueing delay increases without bound. As pointed out above, the dynamics of the number of messages in the system does not depend on the order in which messages are served. Accordingly, the average number in the system and the average delay are the same for FCFS, LCFS, or random order of service.

6.1.2 Distribution of Message Delay: FCFS

As we have seen in Section 3.3.1, the Poisson arrival of messages leads to a simple relation between the transform of the distribution of message delay and the generating function of the number of messages in the system. If service is FCFS, then the number of messages in the system in the steady state will have the same

[3] For further details, see Khinchin (1932) and Pollaczek (1903); see also Section 6.1.3 (below).

distribution as the number of messages that arrived while the departing customer was in the system. As in the calculation of $A(z)$, we have a Poisson arrival during a random interval. For convenience, we repeat equation (3.38):

$$N(z) = \mathcal{D}(\lambda(1-z)) \tag{6.13}$$

where $\mathcal{D}(s)$ is the Laplace transform of the density function of delay. By substituting (6.7) into (6.6) and from (6.13), we have

$$\mathcal{D}(\lambda(1-z)) = \frac{(1-\rho)(1-z)\mathcal{M}(\lambda(1-z))}{\mathcal{M}(\lambda(1-z)) - z} \tag{6.14}$$

Substituting $s = \lambda(1-z)$, we obtain

$$\mathcal{D}(s) = \frac{s(1-\rho)\mathcal{M}(s)}{s - \lambda + \lambda\mathcal{M}(s)} \tag{6.15}$$

As seen in the M/M/1 queue, the total delay is the sum of service time and queueing delay. Furthermore, these random variables are independent of one another. If $\mathcal{Q}(s)$ is the Laplace transform of the density function of queueing delay, then we have

$$\mathcal{D}(s) = \mathcal{M}(s)\mathcal{Q}(s) \tag{6.16}$$

Comparing equations (6.15) and (6.16), we see that

$$\mathcal{Q}(s) = \frac{s(1-\rho)}{s - \lambda + \mathcal{M}(s)} \tag{6.17}$$

We remind the reader that the transform results for delay hold for first-come first-served service. Regardless of the service discipline, the mean delay and the mean number of messages in the system can be related through Little's formula. We have

$$\bar{N} = \lambda\bar{D}$$

The same result holds when we consider only those in the queue

$$\bar{Q} = \lambda\bar{D}_Q$$

It is instructive to check these results with those obtained previously for the M/M/1 queue. For an exponentially distributed service time, we have the transform

$$M(s) = E[e^{-sM}] = \int_0^\infty e^{-st}\left[\mu e^{-\mu t}\right]dt = \frac{\mu}{s+\mu} \tag{6.18}$$

Substituting into (6.7) and successively differentiating yields

$$A(z) = \frac{\mu}{\mu - \lambda(z-1)}$$

$$A'(z)|_{z=1} = \left.\frac{\lambda\mu}{[\mu - \lambda(z-1)]^2}\right|_{z=1} = \rho$$

$$A''(z)|_{z=1} = \left.\frac{2\lambda^2\mu}{[\mu - \lambda(z-1)]^3}\right|_{z=1} = 2\rho^2$$

From Equations (6.15) and (6.18), we have

$$D(s) = \frac{\mu - \lambda}{s + \mu - \lambda} \tag{6.19}$$

Equation (6.19) is the Laplace transform of an exponentially distributed random variable with mean value $1/(\mu - \lambda)$. This checks with results obtained in Chapter 3. [See Equation (3.39).]

Example 6.1. Moments of Delay We consider the example of the Poisson arrival of messages, which have the Erlang 4 distribution $M(s) = (\mu/(s+\mu))^4$. We can find the mean and the mean-square value of message delay for FCFS from (6.15). From the associated *Maple spreadsheet*, we have, respectively, for the mean and mean-square value of delay $2(-2\mu + 3\lambda)/(\mu(-\mu + 4\lambda))$ and $20(2\lambda^2 + \mu^2 - 2\lambda\mu)/(\mu(-\mu + 4\lambda))^2$, where λ is the message arrival rate. Representative plots are shown.

Applications to Data Transmission An important special case is that of deterministic service: the M/D/1 queue. If the service duration is equal to m with probability 1, the Laplace transform for the packet length is then $M(s) = e^{-ms}$. Substituting into (6.7) and (6.6), and then in (6.15) we have

$$N(z) = \frac{(1-\rho)(z-1)e^{\rho(z-1)}}{z - e^{\rho(z-1)}} \tag{6.20}$$

and

$$D(s) = \frac{s(1-\rho)e^{-sm}}{s - \lambda + \lambda e^{-sm}} \tag{6.21}$$

where $\rho = \lambda m$.

The mean-square value of the service duration is simply m^2 and the expressions for the mean number in the system and the mean delay are found by substitution into (6.11) and (6.12):

$$\bar{N} = \rho + \frac{\lambda^2 m^2}{2(1-\rho)} = \rho + \frac{\rho^2}{2(1-\rho)} = \frac{2\rho - \rho^2}{2(1-\rho)} \tag{6.22}$$

$$\bar{D} = m + \frac{\lambda m^2}{2(1-\rho)} = \frac{m(2-\rho)}{2(1-\rho)} \tag{6.23}$$

As mentioned earlier, in computer communications networks an important application of the M/D/1 queue is packet transmission. Quite often it is assumed that fixed-length packets arrive at a node for transmission over a synchronous communications link. If the number of bits in each packet is b bits and if the transmission line rate is R bps, then $m = b/R$. Assuming that the arrival process of packets is Poisson, the average number of packets in the transmit buffer and the average delay are as described in Equations (6.22) and (6.23) under the assumption of no retransmission. It is interesting to compare the results for the M/D/1 queue with those of the M/M/1 queue. From Equation (3.35), the average delay for the M/M/1 queue is given by

$$\bar{D} = \frac{\bar{M}}{1-\rho}$$

where $\bar{M} = 1/\mu$. In Figure 6.1 delays normalized to message length are plotted as a function of load. The calculations are carried out on the associated spreadsheet. As is evident in Figure 6.1, the average delay for the M/M/1 queue is always larger than that of the M/D/1 queue. This difference may be attributed to the variability of the exponential distribution.

The M/G/1 model serves to describe the performance of a number of systems where messages arrive at a Poisson rate to a service facility. The challenge in applying the model lies in finding the probability distribution for the service time. We now give two examples of such an application.

Example 6.2. Frequency-Division Multiplexing A relatively simple application of the M/G/1 queue is frequency-division multiplexing (FDM) and frequency-division multiple access (FDMA),[4] where a portion of the frequency spectrum is allocated to a source. In doing this allocation, it is necessary to allow a guard space for filtering. Depending on the technology deployed, the available bandwidth translates to a particular transmission rate according to the efficiency of the technique in bits per hertz. We denote the resulting transmission rate as R_F bits per second. Messages generated by the sources sharing the medium have a random number of bits, denoted by the random variable B. The mean and the mean-square values of the time required to transmit are $\bar{B}/R_F$ and $\bar{B}^2/R_F^2$, respectively. If messages arrive at a Poisson rate, the average delay of a message in being transmitted over the common medium is found by substitution into (6.12):

$$\bar{D} = \frac{\bar{B}}{R_F} + \frac{\lambda \bar{B}^2/R_F^2}{2(1-\rho)}$$

[4] The distinction between FDM and FDMA lies in the location of the sources. In the former, they are in the same place; in the latter, they are geographically dispersed.

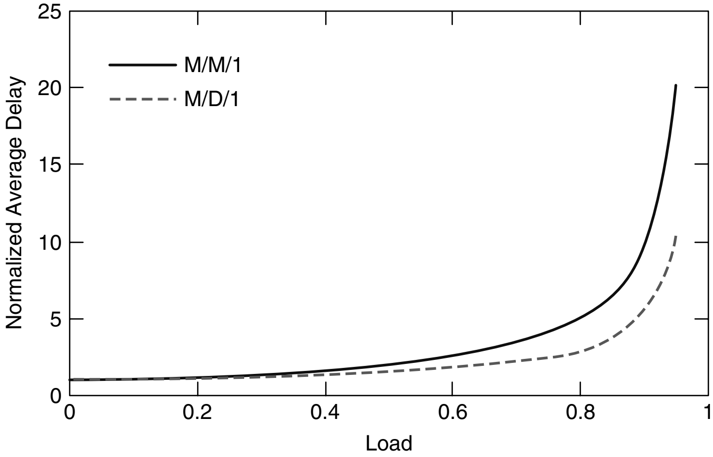


Figure 6.1 M/D/1 and M/M/1 queues.

On the associated Excel spreadsheet, the average delay is found for a message whose length may be described as having a constant component for overhead of 20 bytes and a variable component whose length is a binomial distribution in the number of bytes. The maximum is 200 bytes, and the average is 50 bytes; thus, the binomial parameters are $N = 200$ and $p = 0.25$.

Example 6.3. Stop-and-Wait Protocol The send-and-wait protocol (SAW) is a simple, widely used error control technique. Messages are accompanied by redundant bits, which allow errors in transmission to be detected. If a message is received with no error detected and there is enough storage for the received message in the receiver, an acknowledgment (ACK) is sent to the transmitting node. When the receiver detects an error in a packet, it is dropped. If the transmitting node does not receive an acknowledgment within a specified timeout interval, the message is retransmitted. No messages are transmitted during the timeout interval. Until an acknowledgment is received, the packet is held in a transmit buffer.

We now consider an analysis of the stop-and-wait protocol under the assumption of a random transmission delay in the channel. This random delay may be ascribed to random propagation delay in a physical channel or to the random message length. We model this system as an M/G/1 queue. The difficult part is the derivation of the probability distribution of the time for the completion of message transmission, which begins when a message begins transmission and ends when the receiver has acknowledged the message. It may include a number of retransmissions. This is depicted in Figure 6.2 for the case of three transmissions before an ACK is received. Let $B(t)$ denote the probability distribution of delay until an acknowledgment is received:

$$B(t) \overset{\Delta}{=} P[\text{ACK delay} \leq t]$$

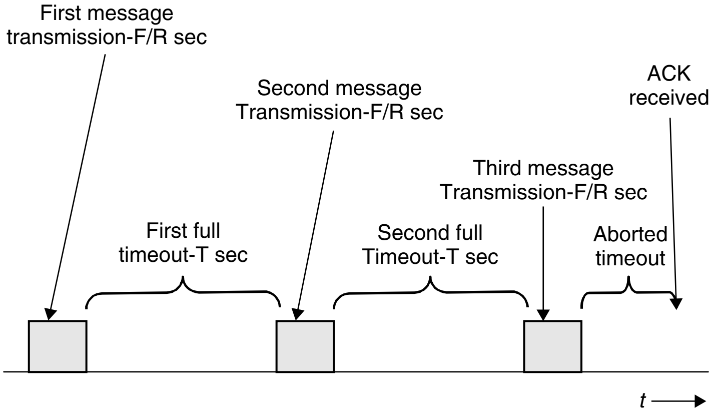


Figure 6.2 Stop-and-wait example with three transmissions.

The density function is denoted $b(t) = dB(t)/dt$. This delay includes delay in the forward path and processing as well as delay of the ACK in the reverse path. Let Q be the probability that the receiver accepts a message. If there is no loss of the ACK in the feedback channel, the probability of receiving, an ACK in the timeout interval, T, is

$$G = QB(T)$$

In writing this equation, we assume that the ACK delay is independent of the event of a loss. The event of the loss of an ACK can be taken into account simply by modifying Q. If the successive transmissions are independent, then the number of transmissions required to get a message through to the receiver is geometrically distributed:

$$P(n \text{ transmissions}) = (1 - G)^{n-1}G; \quad n = 1, 2, \ldots$$

Each transmission, except the last, is made up of the packet transmission time and a timeout interval T. In the last transmission the arrival of an ACK terminates the timeout interval.

We calculate the total time required to transmit a message, with retransmissions and timeouts, by first conditioning on the number of transmissions. Note that successive packet transmissions are of the same-length message; therefore, the time required for N message transmissions has Laplace transform

$$E[e^{-sNP}] = \mathcal{P}(Ns)$$

where P is the time required for a single transmission, and where $\mathcal{P}(s)$ is the Laplace transform of the density of the message transmission time. Similarly, the time required for $N-1$ fixed-length timeouts has Laplace transform $e^{-s(N-1)T}$. By definition, the final timeout interval is ended by an ACK. For this interval, we have

$$P[\text{ACK arrival} \leq t/\text{ACK arrival} \leq T] = \frac{B(t)}{B(T)}, \quad 0 \leq t \leq T$$

The Laplace transform of the probability density of this quantity is

$$\mathcal{R}(s) = \frac{\int_0^T b(t)e^{-st}dt}{B(T)}$$

The message transmission times and the timeout intervals are independent. Thus the Laplace transform of the sum is the product of the individual transforms. Averaging over the number of transmissions we have for the Laplace transform of the density of the total time required to transmit a packet

$$\mathcal{M}(s) = \sum_{n=1}^{\infty} (1-G)^{n-1} G\mathcal{P}(ns)e^{-sT(n-1)}\mathcal{R}(s)$$

This message transmission time may be substituted into Equation (6.15) to find the delay of a packet. We assume the packet to remain in the transmitter until acknowledged.

Let us consider a special case of the preceding. Suppose that the duration of a packet is a constant F bits. If the line transmission rate is R bps, we have

$$\mathcal{P}(s) = e^{-(F/R)s}$$

Suppose further that the ACK delay is a constant equal to the timeout T, we have

$$B(t) = u(t-T)$$

Substituting, we find that

$$\mathcal{M}(s) = \sum_{n=1}^{\infty} (1-G)^{n-1} Ge^{-snF/R}e^{-snT} = \frac{Ge^{-s(F/R+T)}}{1-(1-G)e^{-s(F/R+T)}}$$

The average and the mean-square transmission times are

$$\bar{M} = \frac{F/R+T}{G}$$

$$\bar{M^2} = \frac{(F/R+T)^2(2-G)}{G^2}$$

as may be verified on the associated Maple spreadsheet. These may be substituted into Equation (6.23) to find the average delay. As mentioned above, we assume that a packet remains in the transmitter until acknowledged. If, on the other hand, we assume that a packet is successfully transmitted when the receiver accepts it, we subtract average delay in the feedback channel to find the average delay under this assumption.

These results can be further refined if we assume that the only reason a packet needs to be retransmitted is because of error in the forward channel. Suppose further that errors occur independently from bit to bit. If a packet consists of F bits, then the probability of packets having at least one error is

$$1 - Q = 1 - (1 - B_E)^F \cong FB_E, \quad B_E \ll 1$$

where B_E is the probability of a single bit error. In doing this calculation, it is assumed that there are enough parity bits in the packet so that any likely errors can be detected. We have then that $G = 1 - FB_E$, the probability of message transmission being completed. On the associated Excel spreadsheet, calculations are carried out for transmission over a DS-1 line.

Example 6.4. Head of Line Problem We now consider the *head-of-the-line problem* (Mehmet Ali et al. 1985), a phenomenon that occurs in switches in which messages are queued at input ports (see Fig. 6.3). As messages arrive to input ports, they are switched on a FCFS basis to output ports with only one message from a particular input port switched at a time to a single output port. On its accession to the head of an input port queue, a message indicates to the switch its output port destination. If there are prior outstanding requests for that same port, these requests are served on a FCFS basis. We imagine that the requests for a particular output port are placed in a virtual queue for that output port. Note that a single virtual queue need hold no more than N requests and all virtual queues may hold no more than N requests, one for each of the input queues with messages.

We analyze the performance of the switch by means of the results on the M/G/1 queue that we have just derived. Messages arrive to each of the input ports at a Poisson rate averaging λ messages/s. Message length distributions are assumed to be the same for all input ports. For reasons that will become clear as we proceed

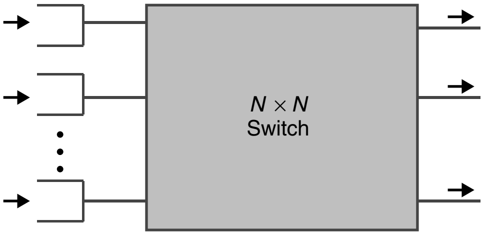


Figure 6.3 $N \times N$ input queued switch.

through the analysis, we assume that all messages have an exponentially distributed length. It is also assumed that the traffic through the switch is symmetric in that a message from any input port is switched to any one of the output ports with equal probability $1/N$.

Without loss of generality, we focus on input port 1. The key to the analysis is the service time of a message, which is the time interval between its requests to the switch to be transferred to an output port until it is actually transmitted through the switch. In this time interval the switch may be dealing with prior requests from other ports. In order to simplify the analysis, we assume that the switch processing time is negligible. The service time of the message is denoted as M and the time required to transmit a message once it has been given access to an output port is denoted as T. The load at an input port is then $\rho = \lambda\bar{M}$. Thus, the probability of a port having at least one message is $1 - P_0 = \sum_{i=1}^{\infty} P_i = \rho$. We emphasize that, from symmetry, this is the same for all queues. The probability that any input port has a request for a particular output port is

$$R = \frac{\rho}{N}$$

In order to carry the analysis forward, we assume that the queues operate independently of one another. The probability of k prior requests in the virtual queue is

$$W_k = \binom{N-1}{k} R^k (1-R)^{N-1-k}; \quad k = 0, 1, \ldots, N-1$$

We further simplify by assuming that messages are exponentially distributed; otherwise, we must distinguish between messages in the process of being transmitted and messages just waiting in the virtual queue. The Laplace transform for an exponentially distributed random variable with mean $\bar{T}$ is $1/(\bar{T}s + 1)$. Conditioned on the existence of k prior messages, the Laplace transform for the service time of a message is $1/(\bar{T}s + 1)^{k+1}$. Notice that we have included the transmission time of the message itself. Averaging over k, we find the Laplace transform for the service time of a message:

$$\mathcal{M}(s) = \sum_{k=0}^{N-1} \binom{N-1}{k} R^k (1-R)^{N-1-k} / (\bar{T}s + 1)^{k+1} = \frac{\left(\frac{R}{(\bar{T}s+1)} + 1 - R\right)^{N-1}}{\bar{T}s + 1} \tag{6.24}$$

It remains to find R in this equation. In order to do this, we find the mean service time of a message by differentiating $\mathcal{M}(s)$ with respect to s and setting $s = 0$. We find

$$\bar{M} = \bar{T}(1 + R(N-1)) = \bar{T}\left(1 + \frac{\rho(N-1)}{N}\right) \tag{6.25}$$

Multiplying both sides of (6.25) by λ and setting $\rho = \bar{M}\lambda$, we solve for ρ:

$$\rho = \frac{\bar{T}\lambda}{1 - \bar{T}\lambda(1 - 1/N)} \cong \frac{\bar{T}\lambda}{1 - \bar{T}\lambda}; \quad N \gg 1 \tag{6.26}$$

As we have seen $R = \rho/N$. The expression in (6.24) can now be substituted into (6.15) to find the Laplace transform for delay.

If we differentiate (6.24) twice and set $s = 0$, the second moment of message transmission time can be found. On the associated Maple spreadsheet, we find

$$\bar{M}^2 = \bar{T}[R^2(N-1)(N-2) + 2 + 4R(N-1)] \tag{6.27}$$

For stability, we must have $\rho < 1$, which from (6.26) implies that $\bar{T}\lambda < \frac{1}{2}$. This illustrates the *head-of-the-line problem*. The switch serves only the message at the front of queue. These messages may be blocked because of requests from other input ports even though there may be messages in the same input port queue, that are destined for free output ports. Suppose that instead, all the arriving messages were routed immediately to a queue at an output port (output port queueing); the constraint on stability would be $\bar{T}\lambda < 1$. Output port queueing requires that the switch operate N times faster, so improved performance comes at a price. By substituting (6.25) and (6.27) into (6.23), we find the average delay. The average delay is shown as a function of $\bar{T}\lambda$ on the associated spreadsheet. As we see, the system saturates at $\bar{T}\lambda = 0.5$.

6.1.3 Residual Life Distribution: Alternate Derivation of the Pollaczek–Khinchin Formula

In this section we shall rederive the *Pollaczek–Khinchin* formula (6.12) using an alternative approach to that above. Our motivation is to introduce the concept of residual life. As a preliminary to this derivation, we derive the residual life of a distribution. In order to illustrate the concept that is involved, we start with a paradox. Suppose that we are told that the interval between buses on a particular route is exponentially distributed with a mean value of 15 min. As a mathematical model, we assume that the buses come and go instantaneously so that one must be waiting for a bus in order to catch it. Since the exponential distribution is memoryless, the waiting time for the next bus is exponentially distributed with mean value 15 min. Now, from the reversibility of the process, we know that the time since the last bus is also exponentially distributed with mean 15 min; consequently, the mean duration of the interbus interval to which we have arrived is 30 min. This seems to conflict with the assumption that the mean inter–arrival time of buses is 15 min. The answer lies in the manner that we have selected a sample from the distribution of inter–arrival times. The arrival process tends to choose the longer samples. In theory, the interval between arrivals can be any length from microseconds to years. It is unlikely that one would choose the former, should it occur. The opposite is true of the latter.

Applying a well-known attribute of point processes, we recognize that the distribution of intervals in which messages arrive and the general distribution of intervals are not the same since, as pointed out above, an arriving message selects longer cycles. The following heuristic derivation illustrates the selection process. Suppose, for the moment, that intervals may assume only L discrete values, $W_1, W_2, \ldots, W_L$ with probabilities $\omega_1, \omega_2, \ldots, \omega_L$, respectively. Now, assume that we have observed K intervals. Let K_j be the number of observed intervals, which are of duration W_j. Then W_jK_j is the duration of all such intervals in the time duration of the process is observed. $\sum_{j=1}^{L} W_jK_j$ is the duration of the K intervals that have been observed. The proportion of all of the intervals of duration W_i is

$$r_i = \frac{W_iK_i}{\sum_{i=1}^{K} W_iK_i} = \frac{W_i(K_i/K)}{\sum_{i=1}^{K} W_i(K_i/K)}$$

This is also the relative frequency of message arrival in an interval W_i. Observing over longer and longer intervals, $K \rightarrow \infty$, we have $K_j/K \rightarrow \omega_j$, the probability of an interval of duration W_i is

$$r_i = \frac{W_i\omega_i}{\sum_{j=1}^{K} W_i\omega_i} = \frac{W_i\omega_i}{\bar{W}} \tag{6.28}$$

where $\bar{W}$ is the mean duration of an interval. We carry this result over to the continuous case in the obvious way. Now, we assume that the interval durations are independent and identically distributed continuous random variables with probability density function $\omega(t)$. The density function of the duration of the interval in which a message arrives is

$$r(x) = \frac{x\omega(x)}{\bar{W}}; \quad x \geq 0 \tag{6.29}$$

From the properties of the Poisson arrival process, we know that an arrival in a given interval is uniformly distributed over that interval. Let the random variable U indicate the time from the arrival to the end of the interval (see Fig. 6.4).

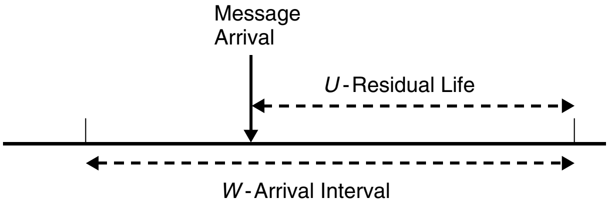


Figure 6.4 Residual life.

Conditioned on the duration of the selected interval, the probability density function of U is

$$u(t/V = x)dt = \begin{cases} \dfrac{dt}{x}; & 0 \leq t \leq x \\ 0; & \text{otherwise} \end{cases}$$

Now averaging over V, we have

$$u(t)dt = \frac{1}{\bar{W}}\int_t^\infty \frac{1}{x} x\omega(x)dx\, dt = \frac{1}{\bar{W}}\int_t^\infty \omega(x)dx\, dt = \frac{1 - W(t)}{\bar{W}}dt; \quad t \geq 0 \qquad (6.30)$$

The function $u(t)$ is a valid probability density since $w(t) \geq 0$ and

$$\int_0^\infty u(t)dt = \int_0^\infty dt \frac{1}{\bar{W}}\int_t^\infty dx\ \omega(x) = \frac{1}{\bar{W}}\int_0^\infty dx\ \omega(x)\int_0^x dt = \frac{\bar{W}}{\bar{W}} = 1$$

The variable U is known as the *residual life* or the *forward recurrence time* of the interval (see Fig. 6.4). We measured from the arrival of the message to the end of the interval, but from the symmetry of the model we could just as well have measured from the beginning to the arrival. We return now to the exponential distribution as an example. If the mean value is $1/\mu$, the probability density function is $\omega(t) = \mu e^{-\mu t}$; $t \geq 0$; consequently, the distribution function is $U(t) = 1 - e^{-\mu t}$; $t \geq 0$. Substituting into (6.30), we find that the density function for the residual life is

$$u(t) = \mu e^{-\mu t}; \quad t \geq 0$$

Thus, the residual life distribution for the exponential case is the same distribution. In view of the memoryless property of the exponential distribution, this is just what we would expect.

The Laplace transform of $u(t)$ is given by

$$U(s) = \frac{1}{\bar{W}}\int_0^\infty dt\, e^{-st}\int_t^\infty dx\, \omega(x) = \frac{1}{\bar{W}}\int_0^\infty dx\, \omega(x)\int_0^t dt\, e^{-st} = \frac{1 - W(s)}{s\bar{W}} \qquad (6.31)$$

The mean value of the residual life is given by

$$\bar{U} = \int_0^\infty t\, u(t)dt = \int_0^\infty dt\, t\, \frac{1}{\bar{W}}\int_t^\infty dx\, \omega(x) = \frac{1}{\bar{W}}\int_0^\infty dx\, \omega(x)\int_0^x t\, dt = \frac{\bar{W}^2}{2\bar{W}} \qquad (6.32)$$

We shall be using this last result immediately.

Consider the arrival of messages at a Poisson rate with average value λ to a server with an arbitrary distribution of service times. Assume also that there is no limitation on the number of messages that can be stored. We first derive an expression for the average time that the message spends in the queue, which we designate as $\bar{D}_Q$. There are two components to this delay: (1) the time until the server is free, $\bar{D}_Q^{(1)}$, and (2) the time required for all the messages in the queue to be served, $\bar{D}_Q^{(2)}$. We deal first with the former component. If the server is occupied, the mean time remaining until it is free is just the mean of residual life of the server distribution, $\bar{M^2}/2\bar{M}$, where and $\bar{M^2}$ is the mean-squared service time. In Chapter 3, we saw that the probability that the server is occupied is $\rho = 1 - P_0$, where $\rho = \lambda\bar{M}$ and $\bar{M}$ is the mean service time. By the PASTA property, this is the probability that an arrival sees an occupied server. We average over the probability that the server is occupied to find the first component of delay:

$$\bar{D}_Q^{(1)} = \rho\frac{\bar{M^2}}{2\bar{M}} = \frac{\lambda\bar{M^2}}{2} \tag{6.33}$$

We invoke Little's formula to find the second component. The average number of message in the queue is given by

$$\bar{P}_Q = \lambda\bar{D}_Q \tag{6.34}$$

since $\bar{D}_Q$ is the average mount of the message spends in the queue. The average time required to transmit these messages is

$$\bar{D}_Q^{(2)} = \bar{M}\lambda\bar{D}_Q = \rho\bar{D}_Q \tag{6.35}$$

Summing the two components, we have from (6.33) and (6.35)

$$\bar{D}_Q = \bar{D}_Q^{(1)} + \bar{D}_Q^{(2)} = \rho\bar{D}_Q + \frac{\lambda\bar{M^2}}{2}$$

Solving for $\bar{D}_Q$, we find

$$\bar{D}_Q = \frac{\lambda\bar{M^2}}{2(1-\rho)} \tag{6.36}$$

Now the delay of a message is composed of the message transmission time as well as the queueing delay. Adding this factor gives the complete message delay and the Pollaczek–Khinchin formula [Eq. (6.12)].

6.1.4 Variation for the Initiator of a Busy Period

A variation of the imbedded Markov chain model is applicable in several communications contexts. Consider the situation where a message arriving at an empty queue receives service different from those arriving at a nonempty queue (Welch 1964). Such a message initiates what is called a "busy period." Again we imbed the Markov chain at successive departure epochs and consider the state defined by the number of messages in the system. When the queue is not empty, the equation for the successive states is that given in (6.1). Recall that in this equation A_{i+1} is the average number of messages that arrive during the transmission time of a message. Now if the transmission time of a message that arrives to an empty queue is different from those that arrive to a nonempty queue, a change is necessary in Equation (6.2). We write $N_{i+1} = \tilde{A}_{i+1}$, indicating the difference in the number of arrivals from the case of a nonempty queue. We can write this in a form analogous to (6.1):

$$\begin{aligned} N_{i+1} &= N_i - U(N_i) + A_{i+1}U(N_i) + \tilde{A}_{i+1}[1 - U(N_i)] \\ &= N_i - U(N_i) + \tilde{A}_{i+1} + (A_{i+1} - \tilde{A}_{i+1})U(N_i) \end{aligned} \tag{6.37}$$

Taking expectations of both sides of (6.37) and assuming equilibrium we find

$$1 - P_0 = \frac{E(\tilde{A})}{1 - E(A) + E(\tilde{A})} \tag{6.38}$$

where P_0 is the probability of the queue being empty. The terms $E(\tilde{A})$ and $E(A)$ are the average number of messages to arrive during the extraordinary and the ordinary message transmission times, respectively. Let $\tilde{M}$ denote the duration of a message that arrives to an empty queue. We have

$$E(\tilde{A}) = \lambda E(\tilde{M})$$

and

$$E(A) = \lambda E(M)$$

Using the same approach that led to (6.15), we can find the probability-generating function for the number of messages in the system. For convenience, we drop dependence on i under the assumption that a steady state has been reached. From (6.37), we have

$$\begin{aligned} P(z) = E(z^N) = E(z^{N-U(N)+AU(N)+\tilde{A}[1-U(N)]}) = P_0E(z^{\tilde{A}}) + \sum_{i=1}^{\infty} P(N = i)z^{i-1+A} \\ = P_0\tilde{A}(z) + z^{-1}\left[\sum_{i=0}^{\infty} P(N = i)z^i - P_0\right]A(z) \end{aligned} \tag{6.39}$$

where $\tilde{A}(z)$ and $A(z)$ are the probability-generating functions of message arrivals during the appropriate message transmission times. Solving Equation (6.39) for $P(z)$, we find for the probability-generating function for the number of messages in the system

$$P(z) = \frac{P_0[A(z) - z\tilde{A}(z)]}{A(z) - z} \tag{6.40}$$

Let $\tilde{M}(s)$ and $M(s)$ denote the Laplace transforms of message transmission time densities for the empty and non-empty queues, respectively. It can be shown that the Laplace transform for message delay is given by

$$D(s) = \frac{P_0[\lambda M(s) - (\lambda - s)\tilde{M}(s)]}{\lambda M(s) + s - \lambda} \tag{6.41}$$

From these transforms, moments can be found in the usual fashion. From (6.38) and (6.41), it can be shown that the average message delay is given by

$$\bar{D} = \frac{E(\tilde{M})}{1 - \lambda[E(M) - E(\tilde{M})]} + \frac{\lambda E(M^2)}{2[1 - \lambda E(M)]} + \frac{\lambda[E(\tilde{M}^2) - E(M^2)]}{2\{1 - \lambda[E(M) - E(\tilde{M})]\}} \tag{6.42}$$

Example 6.5. Satellite Channel Access We illustrate the utility of these results by means of a system in which a number of geographically dispersed terminals contend for access to a satellite channel. A message arriving to an empty terminal immediately transmits a connect request. This is repeated every T seconds until access is granted. We assume that the number of contending terminals is such that a request is granted with probability P independently on each try. Once a terminal gains access to the channel, it may transmit as many messages that it has in its buffer plus new arrivals. We may assume messages to be of constant length. When all of these messages have been transmitted, the terminal must relinquish access to the channel.

The first message of a busy period requires $W + M$ seconds for transmission, where W is the time required to access the channels. Since we have independence from trial to trial, we have a geometric distribution:

$$P(W = kT) = (1 - P)^k P; \quad k = 0, 1, \ldots$$

The mean and the mean-square values of W are, respectively, $\bar{W} = [T(1 - P)]/P$ and $\bar{W}^2 = [T^2(1 - P)]/P^2$. Since the message itself is of constant length, $E[M] = M$ and $E[M^2] = M^2$, the mean and the mean-square value of the service time for the initiating message are, respectively

$$E[\tilde{M}] = E[M] + E[W] = M + \frac{T(1 - P)}{P}$$

and

$$E[\tilde{M}^2] = E[(M + W)^2] = E[M^2] + 2E[WM] + E[W^2]$$
$$= M^2 + \frac{2MT(1-P)}{P} + \frac{T^2(1-P)}{P^2}$$

These expressions may be substituted into (6.42) to find the average delay.

Expressions for $P(z)$ and $\mathcal{D}(s)$ can be found easily enough. The Laplace transform for the channel access time is given by

$$W(s) = E[e^{-sW}] = \sum_{k=0}^{\infty} e^{-skT}(1-P)^k P = \frac{P}{1-(1-P)e^{-sT}}$$

Since message lengths are constant, $\mathcal{M}(s) = e^{-sM}$, and we have for the Laplace transform of the initiating message

$$\tilde{M}(s) = \mathcal{M}(s)\mathcal{W}(s) = \frac{e^{-sM}P}{1-(1-P)e^{-sT}}$$

The PGF for the arrivals is given by the familiar expressions

$$A(z) = M(\lambda(1-z))$$
$$\tilde{A}(z) = \tilde{M}(\lambda(1-z))$$

The appropriate substitutions can be made. On the associated Maple spreadsheet, these formulas are evaluated to find the probability of more than two messages in the queue. We plot this quantity as a function of message arrival probability.

6.1.5 Busy Period of the M/G/1 Queue

As in the preceding sections, we consider the situation in which an M/G/1 queue models the arrival of messages at a buffer for multiplexing on a transmission line. The content of the buffer in bits as a function of time is shown by example in Figure 6.5. Messages arrive at times marked “a” and depart at times marked “d.” At these arrival epochs the content of the buffer increases instantaneously by a random amount, which is the length of the message. In the intervals between message arrivals, the buffer content in bits declines at a steady rate according to the rate of the transmission line. At the final departure of the busy period, the buffer is empty.

A derivation of the mean duration of a busy period can be carried out by segmenting messages into what are called *generations*. The first generation is simply the message that initiates the busy period. The second generation is all the messages that arrive while the first generation is being transmitted. In general, the *i*th

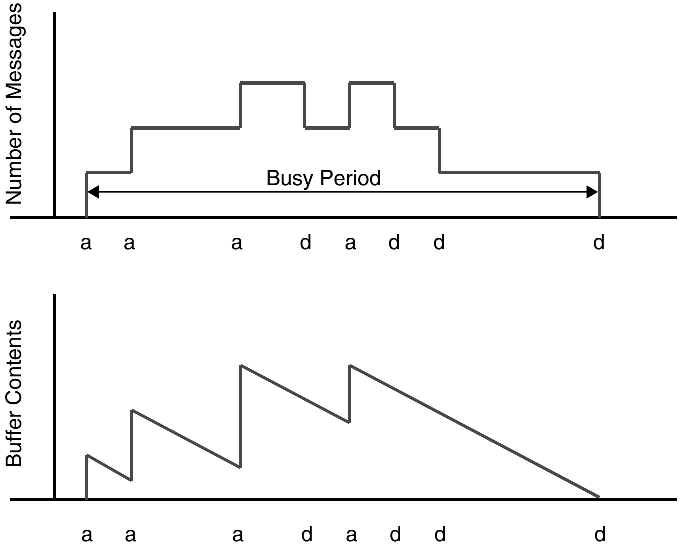


Figure 6.5 Buffer contents as a function of time.

generation is all the messages that arrive while the $(i - 1)$st generation messages are being transmitted. Let $\bar{N}_i$ denote the average number of messages in the ith generation. Recall that in the previous chapters we defined a quantity ρ to be the average number of message arrivals during the transmission time of a message. If the $(i - 1)$st generation had N_{i-1}, messages the average number of arrivals in that generation is $\bar{N}_{i-1}\rho$ we have $\bar{N}_1 = 1, \bar{N}_2 = \rho, \ldots, \bar{N}_i = \bar{N}_{i-1}\rho$.

The average number of messages in a busy period is found by summing over all generations

$$\bar{N}_{BP} = \sum_{i=1}^{\infty} \bar{N}_i = \sum_{i=1}^{\infty} \rho^{i-1} = \frac{1}{1-\rho} \tag{6.43}$$

Since message lengths are independent of the number of messages in the queue, the average duration of a busy period is

$$\overline{BP} = \frac{\bar{M}}{1-\rho} \tag{6.44}$$

where $\bar{M}$ is the average duration of a message.

An implicit equation for the Laplace transform of the density function of the busy period can be derived. The key to this derivation is the memoryless property of the Poisson arrival process. We begin by focusing on the number of messages that arrive during the time that the initial message is being transmitted. Two observations allow us to proceed: (1) different numbers of message arrivals constitute disjoint events; and (2) each message generates a busy period, which has the same probability density function. We begin by using the law of total probability to decompose the event $\{t < BP \leq t + dt\}$ into a set of disjoint events according to the number of arrivals during the first generation. The probability density function of the busy period may then be written

$$\begin{aligned} b(t)dt &= P(t < BP \leq t + dt) \\ &= \sum_{n=0}^{\infty} P(t < BP \leq t + dt, n \text{ arrivals in first generation}) \end{aligned}$$

where BP denotes the duration of the busy period. Define M to be the time required to transmit the initial message of the busy period. For each possible value of the message transmission time, we have a disjoint event. These events are summed through integration over the message length distribution with variable of integration u:

$$b(t)dt = \sum_{n=0}^{\infty} \int_0^t P(u < M \leq u + du, n \text{ arrivals in } M, t < BP \leq t + dt) \tag{6.45}$$

We now rewrite the integrand in Equation (6.45) by conditioning on M and n:

$$\begin{aligned} &P(u < M \leq u + du, n \text{ arrivals in } M, t < BP \leq t + dt) \\ &\quad = P(u < M \leq u + du) \times P(n \text{ arrivals in } M/u < M \leq u + du) \\ &\qquad \times P(t < BP \leq t + dt/n \text{ arrivals in } M, u < M \leq u + du) \\ &\quad = [m(u)du]\left[\frac{e^{-\lambda u}(\lambda u)^n}{n!}\right] \\ &\qquad \times P(t < BP \leq t + dt/n \text{ arrivals in } M, u < M \leq u + du) \end{aligned} \tag{6.46}$$

In order to find an expression for the rightmost term here, we bring into play observation 2 above by assuming that the busy period initiated by each of the n messages has probability density $b(t)$. This allows us to write

$$P(t < BP \leq t + dt) = b^{(n)}(t - u)dt \tag{6.47}$$

where the superscript (n) denotes n-fold convolution. Restating (6.47), we have established that each arrival during the first generation generates its own busy period. These busy periods have had the same distribution as each other and as the overall busy period. The sum of the newly generated busy periods and the busy period and the initial message transmission time must equal the overall busy period.

From Equations (6.45)–(6.47), we have

$$b(t) = \sum_{n=0}^{\infty} \int_0^t m(u) \frac{e^{-\lambda u}(\lambda u)^n}{n!} b^{(n)}(t-u)du \tag{6.48}$$

where $m(u)$ is the probability density of the time required to transmit a message. We take the transform of both sides of (6.48), to find

$$\begin{aligned}
\mathcal{B}(s) &= \int_0^{\infty} dt\, e^{-st} \sum_{n=0}^{\infty} \int_0^t m(u) \frac{e^{-\lambda u}(\lambda u)^n}{n!} b^{(n)}(t-u)du \\
&= \int_0^{\infty} du \sum_{n=0}^{\infty} \frac{e^{-\lambda u}(\lambda u)^n}{n!} m(u) \int_u^{\infty} dt\, e^{-st} b^{(n)}(t-u) \\
&= \int_0^{\infty} du\, m(u) e^{-su} \sum_{n=0}^{\infty} \frac{e^{-\lambda u}(\lambda u)^n}{n!} \mathcal{B}^n(s)
\end{aligned}$$

A final step based on the definition of the exponential gives

$$\mathcal{B}(s) = \int_0^{\infty} du\, m(u) e^{-u(\lambda + s - \lambda \mathcal{B}(s))} = \mathcal{M}(\lambda + s - \lambda \mathcal{B}(s)) \tag{6.49}$$

where $\mathcal{B}(s)$ is the Laplace transform for the busy period and $\mathcal{M}(s)$ is the Laplace transform for the message transmission time.

In (6.49) we have an implicit expression for the busy period of an M/G/1 queue. The equation is self-consistent inasmuch as

$$\mathcal{B}(0) = \mathcal{M}(\lambda - \lambda \mathcal{B}(0)) = \mathcal{M}(0) = 1$$

The mean duration of a busy period can be found by differentiation:

$$\mathcal{B}'(0) = \mathcal{M}'(0)(1 - \lambda \mathcal{B}'(0))$$

Therefore $\overline{BP} = \bar{M}/(1-\rho)$, which checks (6.44). Note that the mean value of the busy period depends on the service time only through the mean value of the service time.

Higher moments of the busy period can be found by repeated differentiation. For example, for the second moment we have

$$\mathcal{B}''(0) = \mathcal{M}''(0)(1 - \lambda \mathcal{B}'(0))^2 - \lambda \mathcal{M}'(0) \mathcal{B}''(0)$$

Rearranging terms, we find that

$$\overline{BP^2} = B''(0) = \frac{(1+\lambda\overline{BP})^2\bar{M}^2}{1-\rho} = \frac{\bar{M}^2}{(1-\rho)^3} \qquad (6.50)$$

Because of the implicit form of (6.49), finding explicit expressions for the density function of the busy period is difficult.[5]

Example 6.6 On the associated Maple spreadsheet, we derive expressions for the mean and the mean-square values of the duration of a busy period. An upper bound, $\overline{BP} + 3 \times \sqrt{\overline{BP^2} + (\overline{BP})^2}$, is also plotted. On the associated Excel spreadsheet, we compute the same 3σ upper bound on the duration of the busy period.

6.2 THE G/M/1 QUEUE

The imbedded Markov chain approach can be applied to the obverse of the M/G/1 queue, the G/M/1 queue. As the notation indicates, there is a single exponential server and the intervals between arrivals are a sequence of independent random variables having a general distribution (Nelson 1995). Further, there is an infinite buffer. The salient difference in the analysis is that the imbedded points are the arrival times of messages rather than departure epochs. The equation describing the number of messages in the system at the imbedded points is similar to the corresponding equation for the M/G/1 queue given by (6.3). In the present case we deal with departures between successive arrivals, rather that arrivals between successive departures; accordingly, the state at successive arrival points is given by

$$N_{i+1} = N_i + 1 - D_{i+1} \qquad (6.51)$$

where N_i is the number of messages in the system found by the ith arrival and D_{i+1} is the number of departures from the system between the ith and $(i+1)$st arrivals. Clearly, $D_{i+1} \leq N_i + 1$. Note that, if there are no departures, there is simply an increase of one message.

Assuming equilibrium has been reached, the steady-state probabilities of the number of messages encountered on arrival obey the equation

$$Q_i = \sum_{j=i-1}^{\infty} Q_j p_{ji}; \quad i = 0, 1, \ldots \qquad (6.52)$$

where Q_i is the probability that i messages are in the queue when there is an arrival, and p_{ji}; $i, j = 0, 1, \ldots$ indicates the transition probability to j messages given

[5] In Example 6.8 below, we calculate the mean-square busy period. The method can be extended to find higher-order moments easily enough. The reader is referred to Takacs (1967).

that there are i messages in the system at successive arrival times. The lower limit in (6.52) is due to the fact that there can be an increase of at most one at the next arrival.

We now proceed to derive the transition probabilities for the imbedded chain, p_{ji}. The service times are exponentially distributed with mean $1/\mu$; therefore, the number of departures up to the contents of the system can be described as a Poisson random variable. Conditioned on the inter–arrival time interval, $A_{i+1} = t$, we may write

$$P(N_{i+1} = k + 1 - j/N_i = k, \quad A_{i+1} = t) = \frac{e^{-\mu t}(\mu t)^j}{j!}; \quad k \geq 1,\ j = 0, 1, \ldots$$

Averaging over the inter–arrival time, we find

$$\begin{aligned} p_{k,k+1-j} &= P(N_{i+1} = k + 1 - j/N_i = k) \\ &= \int_0^\infty \frac{e^{-\mu t}(\mu t)^j}{j!} a(t)dt; \quad k \geq 1,\ j = 0, 1, \ldots \end{aligned} \tag{6.53}$$

where $a(t)$ indicates the probability density for the interval between arrivals. This considers arrivals to a nonempty system. (As we shall see momentarily, we do not need an expression for arrivals to an empty system.) Substituting (6.53) into (6.52), we find

$$Q_k = \sum_{m=k-1}^{\infty} Q_m \int_0^\infty \frac{e^{-\mu t}(\mu t)^{m+1-k}}{(m + 1 - k)!} a(t)dt \tag{6.54}$$

A relation for Q_0 is supplied by the normalizing condition

$$\sum_{k=0}^{\infty} Q_i = 1 \tag{6.55}$$

Now, suppose that we *conjecture* that the solution to (6.54) and (6.55) is given by

$$Q_k = (1 - \sigma)\sigma^k; \quad k = 0, 1, \ldots \tag{6.56}$$

where σ is a constant to be determined. Substituting (6.56) into (6.54) and changing the order of integration and summation, we find

$$\sigma^k = \int_0^\infty \sum_{m=k-1}^{\infty} \sigma^m \frac{e^{\mu t}(\mu t)^{m+1-k}}{(m + 1 - k)!} a(t)dt; \quad k = 1, 2, \ldots \tag{6.57}$$

The summation yields

$$\sigma^k = \int_0^\infty \sigma^{k-1} e^{-\mu t(1-\sigma)} a(t) dt \tag{6.58}$$

Finally, we cancel like factors and define $\mathcal{A}(s)$, the Laplace transform of $a(t)$. We then have the following implicit equation for σ

$$\sigma = \int_0^\infty \exp(-\mu t(1-\sigma)) a(t) dt = \mathcal{A}(\mu(1-\sigma)) \tag{6.59}$$

In general, it is necessary to use numerical techniques to solve (6.59).

The form of (6.56) is that of a geometric distribution; accordingly, the mean number of messages encountered by an arriving message is given by

$$\bar{N}_A = \frac{\sigma}{1-\sigma}$$

The average time to service a message is $1/\mu$. The resulting delay is the time required to transmit the encountered messages as well as the arriving message. The average delay is then

$$\bar{D} = \frac{1}{\mu(1-\sigma)} \tag{6.60}$$

Example 6.7. Constant Service Time An interesting example is the D/M/1 queue, constant arrival rate and exponentially distributed service. With a constant inter–arrival time a, the constant σ is the solution of

$$\sigma = \exp(-(1-\sigma)a\mu) = \exp\left(\frac{-(1-\sigma)}{\rho}\right) \tag{6.61}$$

where $\rho = 1/a\mu$ is the average number of departures during an inter–arrival time interval. This, of course, is the system load. Solving (6.61) is a straightforward exercise. In general, it can be shown that when $\rho < 1$, there is a solution inside the unit circle $|\sigma| < 1$. We do the calculation on the associated Excel spreadsheet. For a particular value of ρ, we carry out the iteration $\sigma_{i+1} = \exp(-(1-\sigma_i)/\rho)$. The results are shown on Figure 6.6, where $\bar{D}$, normalized to the message length, is shown as a function of ρ. Also shown is the same curve for the M/D/1 queue. As we see, the delay for the M/D/1 queue is the larger of the two, indicating that the variation in the arrival process has greater effect than variation in the service time.

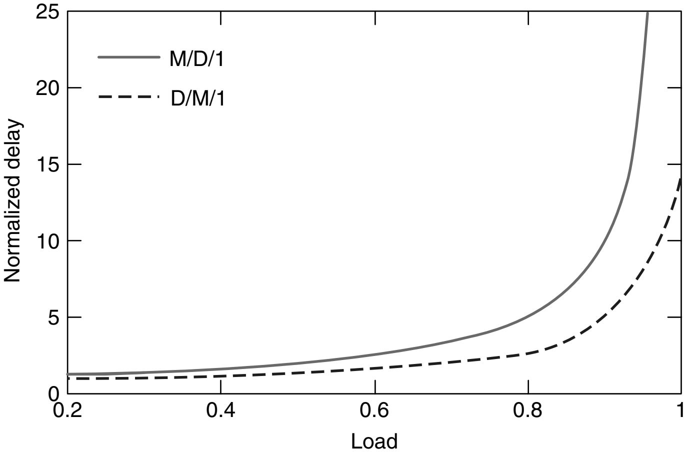


Figure 6.6 Comparison of M/D/1 and D/M/1 queues.

6.3 PRIORITY QUEUES

In this chapter, we study an extension of the material in Section 6.1, which is effected by allocating priorities among different classes of users. The basic assumptions are the same: Poisson arrival and general distributions of service times by a single server.

In considering priorities among queues, three types of priority disciplines are more frequently encountered: *preemptive resume*, *preemptive nonresume*, *and nonpreemptive.*[6] Two classes of M/G/1 queues may illustrate all three of these disciplines. Messages arrive for priority class 1 at a rate of λ_1 messages per second and for priority class 2, at a rate of λ_2 messages per second. These messages are transmitted over a transmission line, which we speak of as the *server*. Each class is stored in a separate buffer, and the server shuttles between the queues for each priority class according to the service discipline. The presence of the server at a buffer means that messages are being transmitted at a constant rate. For the *preemptive resume* discipline, the server switches from the class 2 buffer immediately on arrival of a class 1 message. The server remains at the class 1 buffer until it is empty, whereupon it returns to the class 2 buffer and continues where it left off. Note that there is no loss of work in that no part of the message is transmitted more than once.

In the *preemptive nonresume* discipline there is a loss of work. As in the previous case the server interrupts work on a class 2 message immediately on the arrival of a class 1 message. The server remains with the class 1 messages for the duration of a busy period. However, on return to buffer 2 the buffer begins all over again at the

[6] For a complete treatment of priority queues, see Jaiswal (1968). See also Cohen (1969) and Cox and Smith (1961).

beginning of the message that was left behind in the class 2 buffer. In the case of exponential distributions, the delay and the number in the queue are the same for both the resume and the nonresume disciplines. This is not true for general message length distributions since they have memory.

The third discipline, which may be appropriate in a communications context, is *nonpreemptive*. In this case the server occupied with a class 2 message switches to a newly arrived class 1 message only after the entire message has been transmitted. This is a work-preserving discipline since on return to the class 2 buffer an entirely new message is transmitted. In the nonpreemptive discipline there is no possibility of just part of a message being transmitted.

6.3.1 Preemptive Resume Discipline

General Distribution of Server Absence In some ways the preemptive resume discipline is easiest to analyze. For a class 1 message, the presence of a lower priority class is unseen since the server is always available to it. The equations governing delay and buffer occupancy for class 1 messages are the same as for an M/G/1 queue with a message arrival rate of λ_1 messages per second. Of course, the situation is different for class 2, which experiences service interruptions when a class 1 message enters the system. The time interval during which the server is available to class 2 is exponentially distributed with mean $1/\lambda_1$, since it terminates when there is an arrival for class 1. The duration of this interruption is the busy period of an M/G/1 queue with message arrival rate λ_1, per second; however, in this chapter, we shall consider a general distribution for server absence. This makes the results more widely applicable.[7]

Analysis of the queueing of class 2 messages is carried out using the M/G/1 model with the variation for the initiator of a busy period studied in Section 6.1.4 above. When a class 2 message arrives to an empty buffer, the server may or may not be available depending on the presence of class 1 messages in the system; consequently, the service time is the sum of the message transmission time and the time the message waits for the return of the server. In contrast, for a message that has waited in the queue, the server is immediately available when the message reaches the head of the line.

The probability-generating function for the number of messages in the system is given by (6.40). The important quantities in this equation are the message arrival processes during the two types of service intervals, encountered by initiating and queued messages, respectively. We begin by deriving the probability distribution for each type of service intervals. Let us begin with the service time of the message, which has queued for service. The key consideration here is the service interruptions. The situation is depicted in Figure 6.7, where arrivals and departures of class 2 messages are denoted "a" and "d," respectively. When a class 1 message arrives to the system, transmission ceases for the class 2 messages. When all the

[7] The analysis of an M/G/1 queue with an intermittently available server is from Avi-Itzhak and Naor (1963).

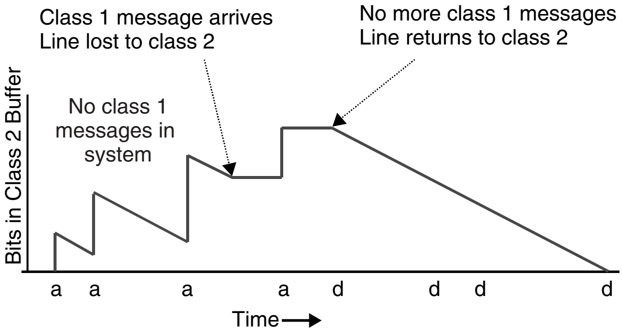


Figure 6.7 Transmission of class 2 messages.

class 1 messages have left the system, transmission resumes for class 2 messages. Note that departures of class 2 messages occur only when the line is available.

In Figure 6.8 an anatomy of the transmission of a class 2 message is shown. The line is available to class 2 messages for an interval in which there are no class 1 messages in the system. An important observation is that the sum of all of these intervals must add up to the total message length M_2. As indicated, transmission of class 2 messages is interrupted when class 1 messages arrive. The durations of the interruptions are independent identically distributed random variables, which we denote as $F_1, F_2, \ldots, F_N$. We may write the total transmission time as

$$T = M_2 + \sum_{i=1}^{N} F_i \tag{6.62}$$

where N is the number of service interruptions.

The key to the calculation of the probability distribution of T is the probability distribution of N. Assume for the moment that the duration of the class 2 message is M_2 seconds. The number of service interruptions is Poisson with average $\lambda_1 M_2$.

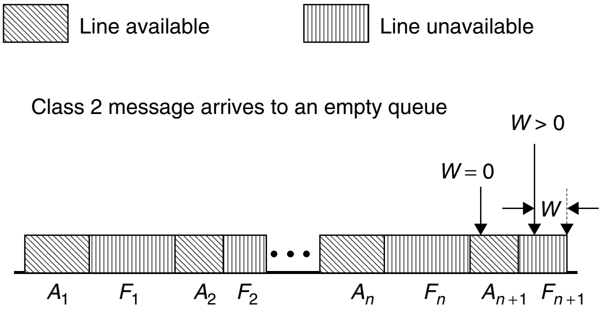

Line available
Line unavailable
Class 2 message arrives to an empty queue
W > 0
W = 0
W
A1 F1 A2 F2 An Fn An+1 Fn+1

Figure 6.8 Line availability—class 2.

This can be seen by observing that the crosshatched intervals (line available) in Figure 6.8 are each exponentially distributed with mean $1/\lambda_1$ since they are ended by an arrival of a class 1 message. Furthermore, the sum of all the intervals is M_2 seconds. Recall that for Poisson arrivals, the inter–arrival distribution time is exponentially distributed. Taking expectations in (6.62), we have

$$\bar{T} = \bar{M}_2 + \bar{N}\bar{F} = \bar{M}_2 + \lambda_1 \bar{M}_2 \bar{F} \tag{6.63}$$

where $\bar{M}_2$ is the mean message length for priority class 2 and $\bar{F}$ is the mean duration of the service interruption.

These considerations lead to an expression for the density function of the message transmission time. We begin by finding the probability density of message transmission conditioned on the assumptions that the class 2 message is u seconds long and that there are $N = n$ interruptions. Suppose that the total transmission time is t seconds long. The sum of the n service interruptions must be $t - u$ seconds in duration. But, the durations of service interruptions are independent identically distributed random variables. The probability density function of the sum is the n-fold convolution of the density function for F. We have

$$P(t < T \leq t + dt/M_2 = u, N = n \text{ interruptions}) = f^{(n)}(t-u)dt$$

where $f(t)$ is the probability density function for F. Averaging over the number of interruptions, we obtain

$$P(t < T \leq t + dt/M_2 = u) = \sum_{n=0}^{\infty} \frac{(\lambda t)^n e^{-\lambda t}}{n!} f^{(n)}(t-u)dt$$

Finally, averaging over the message length M_2, we have

$$t(t)dt = P(t < T \leq t + dt) = \int_0^{\infty} m_2(u) \sum_{n=0}^{\infty} \frac{(\lambda t)^n e^{-\lambda t}}{n!} f^{(n)}(t-u)du\,dt \tag{6.64}$$

where $m_2(t)$ is the probability density function of the class 2 message. A more compact expression is obtained if we calculate the Laplace transform of $t(t)$, denoted $\mathcal{T}(s)$

$$\mathcal{T}(s) = \int_0^{\infty} e^{-st} t(t)dt = \int_0^{\infty} e^{-st} \left[\int_0^{\infty} m_2(u) \sum_{n=0}^{\infty} \frac{(\lambda t)^n e^{-\lambda t}}{n!} f^{(n)}(t-u)du \right] dt \tag{6.65}$$

We change the order of integration and let $v = t - u$. Recall that n-fold convolution in time means a n-fold product of transforms

$$T(s) = \int_0^{\infty} du\, m_2(u) e^{-\lambda_1 u} \sum_{n=0}^{\infty} \frac{(\lambda_1 u)^n}{n!} e^{-su} \mathcal{F}^n(s) = \int_0^{\infty} du\, m_2(u) e^{-\lambda_1 u} e^{-su} e^{\lambda_1 u \mathcal{F}(s)}$$

Finally we obtain

$$\mathcal{T}(s) = \mathcal{M}_2(\lambda_1 + s - \lambda_1 \mathcal{F}(s)) \tag{6.66}$$

where $\mathcal{M}_2(s)$ is the Laplace transform of $m_2(t)$, the density transform of class 2 messages, and $\mathcal{F}(s)$ is the Laplace transform for the transmission interruptions, F.

Number of Messages in System As noted earlier, the difference between queued messages and messages that are first in a busy period is that the latter may find the server occupied when they arrive. We now quantify the amount of time such a message must wait until the server is available. The situation is illustrated in Figure 6.8. We consider the arrival of a class 2 message to a system where there are no other class 2 messages. The interval between the class 2 buffer emptying and the arrival of a class 2 message is exponentially distributed with mean $1/\lambda_2$ since message arrival is Poisson. The line becomes alternately available and unavailable according to the arrivals and departures of class 1 messages. We denote the duration of available and occupied periods by A_i and F_i, respectively, $i = 1, 2, \ldots$. Since an available period is terminated by the arrival of a class 1 message, it is exponentially distributed with mean $1/\lambda_1$. Recall that the distribution of the interval F_i where the server is unavailable has probability density function $f(t)$.

Let W denote the time between the arrival of a class 2 message and the availability of the server. The random variable W is equal to zero if message arrival is during one of the intervals, $A_1, A_2, \ldots$. Also, let Γ denote the time of arrival of a class 2 message. Summing over disjoint events, we find that

$$P(W = 0) = \sum_{n=0}^{\infty} P\left(\sum_{i=0}^{n} (A_i + F_i) < \Gamma \le \sum_{i=0}^{n} (A_i + F_i) + A_{i+1}\right) \tag{6.67}$$

where, for consistency, we have $F_0 = A_0 = 0$. The memoryless property of the exponential random variable Γ allows us to write for each term in the sum in (6.67)

$$P\left(\sum_{i=0}^{n} (A_i + F_i) < \Gamma \le \sum_{i=0}^{n} (A_i + F_i) + A_{i+1}\right)$$
$$= \prod_{i=0}^{n} P(A_i < \Gamma) P(F_i < \Gamma) P(\Gamma \le A_{n+1}) \tag{6.68}$$

We now deal with each of these terms individually. Since Γ and the periods of line availability are exponentially distributed, we can write

$$P(A_i < \Gamma) = \int_0^{\infty} dt\, \lambda_1 e^{-\lambda_1 t} \int_t^{\infty} d\tau\, \lambda_2 e^{-\lambda_2 \tau} = \frac{\lambda_1}{\lambda_1 + \lambda_2}$$

$$P(A_{i+1} \geq \Gamma) = \int_0^\infty dt\, \lambda_2 e^{-\lambda_2 t} \int_t^\infty d\tau\, \lambda_1 e^{-\lambda_1 \tau} = \frac{\lambda_2}{\lambda_1 + \lambda_2}$$

$$P(F_i < \Gamma) = \int_0^\infty dt f(t) \int_t^\infty d\tau\, \lambda_2 e^{-\lambda_2 \tau} = \mathcal{F}(\lambda_2)$$

where $\mathcal{F}(\lambda_2)$ is the Laplace transform of $f(t)$ evaluated at λ_2. Substituting into (6.68), we find

$$P(W=0) = \sum_{n=0}^{\infty} \left(\frac{\lambda_1}{\lambda_1 + \lambda_2} \right)^n \mathcal{F}^n(\lambda_2) \left(\frac{\lambda_2}{\lambda_1 + \lambda_2} \right) = \frac{\lambda_2}{\lambda_1 + \lambda_2 - \lambda_1 \mathcal{F}(\lambda_2)} \tag{6.69}$$

The case for $W > 0$ is only slightly more complicated. Message arrival is during one of the periods of server unavailability $F_1, F_2, \ldots$. Again we sum over disjoint events. For $t > 0$

$$w(t)dt = P(t < W \leq t + dt) = \sum_{n=0}^{\infty} P\left(t < \sum_{i=0}^{n+1} (A_i + F_i) - \Gamma \leq t + dt \right) \tag{6.70}$$

From the memoryless property of Γ, we have

$$\begin{aligned} &P\left(t < \sum_{i=0}^{n+1} (A_i + F_i) - \Gamma \leq t + dt \right) \\ &\quad = \left\{ \prod_{i=0}^{n} P(A_i < \Gamma) P(F_i < \Gamma) \right\} P(A_{n+1} < \Gamma) P(t + \Gamma < F_{n+1} \leq t + \Gamma + dt) \\ &\quad = \left(\frac{\lambda_1}{\lambda_1 + \lambda_2} \right)^{n+1} \mathcal{F}^n(\lambda_2) \int_0^\infty dr \lambda_2 e^{-\lambda_2 r} f(r+t) dt \end{aligned} \tag{6.71}$$

The reader will note that the integral here is just a convolution. After substituting (6.71) into (6.70), we have

$$w(t)dt = \frac{\lambda_1}{\lambda_1 + \lambda_2 - \lambda_1 \mathcal{F}(\lambda_2)} \int_t^\infty d\tau\, \lambda_2 e^{-\lambda_2(\tau - t)} f(\tau) dt; \quad t > 0 \tag{6.72}$$

This gives the value of $w(t)$ for $t > 0$. The density function at $t = 0$ can be represented by a Dirac delta function with weight equal to $P(W = 0)$ given by (6.67). It is a straightforward exercise to find the Laplace transform of $w(t)$. From (6.69) and (6.72), we have

$$\mathcal{W}(s) = \frac{\lambda_2}{\lambda_1 + \lambda_2 - \lambda_1 \mathcal{F}(\lambda_2)} + \frac{\lambda_2 \lambda_2 [\mathcal{F}(s) - \mathcal{F}(\lambda_2)]}{[\lambda_1 + \lambda_2 - \lambda_1 \mathcal{F}(\lambda_2)](\lambda_2 - s)} \tag{6.73}$$

By differentiating $\mathcal{W}(s)$ and setting $s = 0$, we find

$$\bar{W} = -\left.\frac{d\mathcal{W}(s)}{ds}\right|_{s=0} = \frac{1+\lambda_1\bar{F}}{\lambda_1+\lambda_2-\lambda_1\mathcal{F}(\lambda_2)} - \frac{1}{\lambda_2} \tag{6.74}$$

The time spent waiting for the server to become available is independent of the time required to actually transmit the message. The Laplace transform of the time spent in the system by a class 2 message arriving at an empty buffer is given by the product

$$\tilde{\mathcal{T}}(s) = \mathcal{W}(s)\mathcal{T}(s) \tag{6.75}$$

where $\mathcal{T}(s)$ and $\mathcal{W}(s)$ are as given by (6.66) and (6.73), respectively. During the time intervals T and W, class 2 messages arrive at a Poisson rate with mean λ_2 messages per second. We use the same notation as in Section 6.1.4 inasmuch as $A(z)$ and $\tilde{A}(z)$ denote, respectively, the number of messages arriving during an ordinary service time and during a service time that begins a busy period. The probability-generating functions $A(z)$ and $\tilde{A}(z)$ may be written directly

$$A(z) = \mathcal{T}(\lambda_2(1-z)) \tag{6.76}$$

$$\tilde{A}(z) = \tilde{\mathcal{T}}(\lambda_2(1-z)) = \mathcal{W}(\lambda_2(1-z))\mathcal{T}(\lambda_2(1-z)) \tag{6.77}$$

These can be substituted into (6.40) to give the PGF of the number of messages in the system, $P(z)$:

$$P(z) = \frac{P_0\mathcal{T}(\lambda_2(1-z))[1-z\mathcal{W}(\lambda_2(1-z))]}{\mathcal{T}(\lambda_2(1-z))-z} \tag{6.78}$$

Message Delay Since under the preemptive resume discipline the class 1 message is unaffected by the class 2 traffic, its delay is simply that of an M/G/1. If class 2 messages are served in order of arrival, then the Laplace transform of the probability density of delay is given by

$$\mathcal{D}(s) = P\left(1-\frac{s}{\lambda_2}\right) \tag{6.79}$$

From (6.78) and (6.79) we have for the Laplace transform of the probability density of message delay

$$\mathcal{D}(s) = \frac{1-\bar{A}}{1-\bar{A}+\bar{\tilde{A}}}\frac{(\lambda_2-s)\mathcal{W}(s)\mathcal{T}(s)-\lambda_2\mathcal{T}(s)}{\lambda_2-s-\lambda_2\mathcal{T}(s)} \tag{6.80}$$

where $\bar{A}$ and $\bar{\tilde{A}}$ are the average number of message arrivals during ordinary service T and first-in-line service $T + W$, respectively. From (6.63), we have

$$\bar{A} = \lambda_2\bar{M}_2 + \lambda_2\lambda_1\bar{M}_2\bar{F} = \rho_2(1+\lambda_1\bar{F}) \tag{6.81}$$

Since $\bar{\bar{A}} = \lambda_2(\bar{T} + \bar{W})$, we have $1 - \bar{A} + \bar{\bar{A}} = 1 + \lambda_2\bar{W}$. From (6.74), it can be shown that

$$1 - \bar{A} + \bar{\bar{A}} = \frac{\lambda_2 + \lambda_1\lambda_2\bar{F}}{\lambda_1 + \lambda_2 - \lambda_1 \mathcal{F}(\lambda_2)} \tag{6.82}$$

Substituting (6.76), (6.77), and (6.82) into (6.80), we find after some manipulation

$$\mathcal{D}(s) = (1 - P - \rho_2)\frac{\mathcal{T}(s)(s + \lambda_1(1 - \mathcal{F}(s)))}{s - \lambda_2(1 - \mathcal{T}(s))} \tag{6.83}$$

where $P = \lambda_1\bar{F}/(1 + \lambda_1\bar{F})$. Note that in the derivation of (6.83) we have eliminated the troublesome term $\mathcal{T}(\lambda_2)$. From (6.83), the moments of delay can be calculated by the usual process of differentiation and setting $s = 0$. For the average delay, we have

$$\bar{D} = \left.\frac{d\mathcal{D}(s)}{ds}\right|_{s=0} = \frac{\bar{M}_2}{1 - P} + \frac{\lambda_2\overline{M_2^2}}{2(1 - P - \rho_2)(1 - P)} + \frac{\lambda_1(1 - P)^2\overline{F^2}}{2(1 - P - \rho_2)} \tag{6.84}$$

These results apply directly to priority queues where the transmission of a class 2 message is interrupted by the arrival of a class 1 message. (See Fig. 6.8 and the related discussion.) The duration of the interruption is the busy period of an M/G/1 queue. Substituting for $\bar{F}$ the mean duration of the busy period given by (6.44), we find that $P = \lambda_1\bar{M}_1 = \rho_1$. $\overline{F^2}$ is the means square value of a busy period given by (6.50). Substituting these expressions we find for the average delay

$$\bar{D} = \frac{\bar{M}_2}{1 - \rho_1} + \frac{\lambda_2\overline{M_2^2}}{2(1 - \rho_1 - \rho_2)(1 - \rho_1)} + \frac{\lambda_1\overline{M_1^2}}{2(1 - \rho_1 - \rho_2)(1 - \rho_1)} \tag{6.85}$$

Each individual term in (6.85) can be interpreted in terms of the mechanisms of the priority queueing process. The first term is simply the time required to transmit a class 2 message. The factor $1 - \rho_1$ in the denominator is due to interruptions by the class 1 messages. The second term may be interpreted as the queueing time of a class 2 message. If λ_2 is equal to zero, then this term is zero. Also, if λ_1 is equal to zero, then this term is equal to the queueing time of an M/G/1 queue with the appropriate load factors. Finally, the last term accounts for the fact that the server is not always available when a class 2 message arrives. If the duration of the interruption of service decreases, so does the magnitude of this term.

Example 6.8 On the associated Maple spreadsheet, we have calculated expressions for the mean-square delay for an arbitrary message distribution and an arbitrary duration of service interruption for the lower-priority message. We assume that the message length distribution is the same for both priority classes. As part of the calculation, we find the expressions for the mean and the mean square busy period. [Compare with (6.44) and (6.50).]

6.3.2 L-Priority Classes

The analysis of the preemptive resume priority with two priority classes easily generalizes to any number of classes in which the message transmission times and the message arrival rates need not be the same for all classes. Consider the jth class of a total of L classes. Because of the preemptive discipline the classes with lower priority, $j+1, j+2, \ldots, L$ have no effect on the jth class. With the arrival of a message from class $1, 2. \ldots, j-1$ the server immediately leaves off serving a class j message. The aggregate arrival rate of the higher-class messages is Poisson at average rate $\lambda_h = \sum_{i=1}^{j-1} \lambda_i$, where λ_i is the arrival rate of class i messages. The distribution of message lengths can be calculated as a weighted sum. Let $M_i(s)$; $i = 1, 2, \ldots, j-1$ be the Laplace transform of the density of the message lengths in class i. The Laplace transform of the density of the messages, which interrupt service to a class i message, is given by

$$M_h(s) = \frac{\sum_{i=1}^{j-1} \lambda_i M_i(s)}{\sum_{i=1}^{j-1} \lambda_i} \tag{6.86}$$

As far as class j is concerned, all the messages in the higher classes act in the same fashion; consequently, average delay can be found from (6.83)–(6.85) with the obvious substitutions.

When the preemptive nonresume discipline is used, the class with the highest priority is unaffected. However, under the same loading condition the delay for lower priority classes is much higher than for the preemptive resume discipline. The crux of the difference lies in the time required to transmit a message. In the nonresume case the beginning of the message is retransmitted each time a higher-priority class message enters the system. Aside from this difference, the analysis of message delay proceeds in very much the same fashion as in the previous case. Note that for exponentially distributed message lengths, the resume and the nonresume disciplines have the same statistical behavior.

Application to Local-Area Networks with the Ring Topology The ring topology is frequently used in local-area networks. In this section we shall use our results on priority queues to analyze the performance of a particular approach to multiplexing traffic onto the ring, called *demand multiplexing* (Hayes and Sherman 1971). The basic structure is depicted on Figure 6.9. Terminals are connected to a transmission line, which is formed into a ring. In early implementations, the terminals were associated by a T1 line. Flow on the line was organized into fixed-length frame or slots. As indicated, these slots circulate in a clockwise fashion. There is circuitry at each terminal, which allows the terminal time to read bits on the line and to insert data in the form of a fixed-length packet. At the beginning of each slot is a single marker bit (M), which indicates whether the slot is occupied. The terminal may insert its data into an empty slot. Along with the data, source and destination addresses (ADD) are sent. A destination terminal monitors the line and picks off data slots addressed to it, thereby emptying a slot. The same terminal may or may not

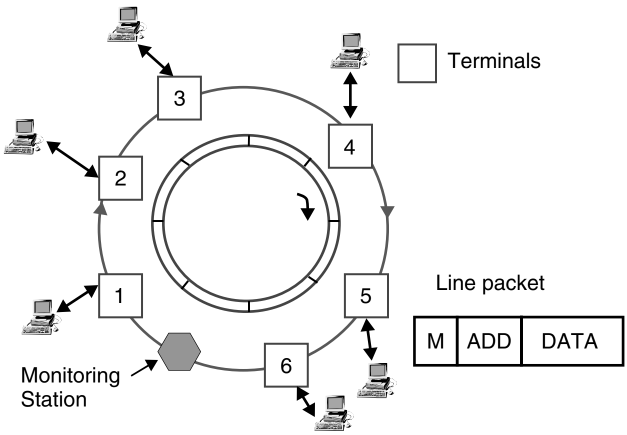


Figure 6.9 Demand multiplexing ring.

insert data into the empty slot, depending on implementation. It may happen that the destination terminal for a packet fails to remove it from the line. In order to prevent this, each slot is tagged with a "history" bit. This bit is initially set to "0." It is set to "1" when it passes a monitoring terminal on the line. If the monitoring terminal sees a slot with a "1," it empties it.

Let us consider the flow in a data collection system. For example, in Figure 6.9 terminals 1–5 transmit all their data to terminal 6, which has no outbound traffic. Priority is given to upstream traffic; thus, for example, traffic from terminals 1 and 2 has priority over traffic from terminals 3, 4, and 5. We may apply the preceding analysis over preemptive priority queueing with a smoothing approximation. Because of the slotted nature of message flow on the line, periods of server availability and unavailability are discrete rather than continuous random variables. However, we assume that the lengths of messages and queueing times are long compared to the slot times, so that this effect can be neglected.[8] Under this assumption, the period of unavailability of the line is the busy period of an M/G/1 queue, which is fed by the total traffic from the upstream traffic. The idle period of the line terminates with the arrival of an upstream message. Since these arrivals are a Poisson processes, the duration of idle period is exponentially distributed with mean value $1/\sum_{i=1}^{j} \lambda_i$, where λ_i; $i = 1, 2, \ldots, j$ is the flow into the upstream terminals.

In the case of general distributions of traffic flow where every terminal transmits to every other terminal, the probability distributions of busy and idle periods are difficult to determine and approximations are necessary. The validity of these approximations has been certified by means of simulation (Anderson et al. 1972). If an idle period is terminated by the arrival of a message, then the duration of the idle period is taken to be exponentially distributed. We also assume that all idle periods are due to message arrival and are therefore exponentially distributed. Finally, we

[8] The discrete case is treated for a data collection system by Spragins (1977).

assume that the periods where the line is not available take the form of a busy period. Let λ_{ij}; $i, j = 1, 2, \ldots, N$ denote the average arrival rate of messages at terminal i that are destined for terminal j. We assume that $\lambda_{ii} = 0$, that is, no flow from a terminal to itself. Notation is simplified, with no loss of generality, if we focus on the performance at terminal N. The average rate of message flow from terminal i through terminal N is $\sum_{j=1}^{i-1} \lambda_{ij}$. Summing over all i gives the total flow through terminal N:

$$\Lambda_N = \sum_{i=2}^{N-1} \sum_{j=1}^{i-1} \lambda_{ij} \tag{6.87}$$

If we assume that slots carrying packets addressed to terminal N are available for use by terminal N, the durations of line idle periods as seen by terminal N have mean $1/\Lambda_N$. If traffic is symmetric, that is

$$\lambda_{ij} = \frac{\lambda}{N-1}; \quad i, j = 1, 2, \ldots, N, \; i \neq j, \; \lambda_{ii} = 0$$

then the line flow seen by each of terminals is the same and is equal to

$$\Lambda_1 + \Lambda_2 + \cdots + \Lambda_N = \frac{N\lambda}{2} - \lambda$$

These are the flows of the higher priority messages.

We begin the calculation of message transmission by calculating the duration of a slot on the line. The accessing technique requires that each packet be accompanied by address bits and bits indicating occupancy and history. If the number of information bits in a packet is I, the minimum packet size is then $I + [\log_2 N]^+ + 2$, where $[x]^+$ is the smallest integer greater than x. Of course, one would not set a slot exactly to this length so that the number of terminals on the ring can be increased. If there are a maximum of 1000 terminals, then an addressing overhead of 20 bits would be enough. If R is the bit rate, the time in seconds required to transmit a single packet in slot duration is

$$T = \frac{I + O}{R} \tag{6.88}$$

where O is the total overhead.

Now, let us assume that the probability distribution of the number of bits in message is given by $B(k))$; $k = 1, 2, \ldots$. Since each line slot, which has duration T, holds I information bits, the probability distribution of the number of slot times required to transmit a message is

$$M_l = P(M = (l+1)T) = \sum_{k=lI+1}^{(l+1)I} B(k); \quad l = 0, 1, 2, \ldots \tag{6.89}$$

The mean and the mean-square message transmission times are, respectively, given by

$$\bar{M} = T\sum_{l=0}^{\infty} lM_l \tag{6.90}$$

and

$$\bar{M^2} = T^2\sum_{l=0}^{\infty} l^2M_l \tag{6.91}$$

where T and M_l are given by (6.88) and (6.89), respectively.

The calculation of average delay is a straightforward application of the results that we have derived. In (6.85), we make the following substitutions: $\lambda_1 = \lambda N/2 - \lambda$, $\lambda_2 = \lambda$, $\bar{M}_1 = \bar{M}_2 = \bar{M}$, and $\bar{M_1^2} = \bar{M_2^2} = \bar{M^2}$.

We should also account for processing at each station along the route around the loop. In each station the packet header must be examined. We allow $(2 + [\log_2 N]^+)/R$ seconds for this. In the symmetric case a packet travels halfway around the loop on the average, and this factor would add an average delay of $(N/2)(2 + [\log_2 N]^+)/R$ seconds to the delay in reaching a destination terminal. In applying the results of the preemptive priority analysis, we recognize a particular point that may lead to serious inaccuracy: The transmission of a final packet of a message cannot be interrupted by the arrival of a higher-class message. We take this into account by replacing $\bar{M}/(1-\rho_1)$ by $(\bar{M} - T)/(1-\rho_1)$. Summing the various components, we find the average delay of a message in the symmetric case to be

$$\bar{D} = \frac{\bar{M} - T}{1-\rho_1} + \frac{\bar{M^2}[N\lambda/2]}{2(1-\rho_1-\rho_2)(1-\rho_1)} + \frac{N}{2}\frac{2+[\log_2 N]^+}{R} \tag{6.92}$$

where $\rho_1 = (\lambda N/2 - \lambda)\bar{M}$ and $\rho_2 = \lambda\bar{M}$.

Example 6.9 On the accompanying Excel spreadsheet, we consider an example where a number of terminals share a transmission line. We assume that each slot on the line can carry 48 octets of information, which is $I = 384$ bits. This is accompanied by five octets of overhead so that the total length of a slot is 424 bits (an ATM cell!). Suppose that the lengths of messages in bits are geometrically distributed, $B(k) = (1-\beta)\beta^{k-1}$; $k = 1, 2, \ldots$, with a mean of $1/(1-\beta)$. We can now find the mean and mean-square times required to transmit a message. From (6.89), we have

$$M_l = \sum_{k=lI+1}^{(l+1)I} (1-\beta)\beta^{k-1} = (1-\beta)\beta^{lI}\sum_{k=0}^{I-1}\beta^k = (1-\beta^I)\beta^{lI}; \quad l = 0, 1, 2, \ldots$$

Thus, the number of slots required to transmit a message is geometrically distributed with mean $1/(1-\beta^I)$. From Table 2.1, the mean-square number of slots required to transmit a message is $2/((1-\beta^l)^2)$. On the spreadsheet, the numbers are worked out for a particular example.

6.3.3 Nonpreemptive Priorities

The nonpreemptive priority discipline is quite useful in modeling communications systems. One can conceive of situations where one type of traffic would have priority over another but without the urgency that would necessitate the interruption of a message transmission. In most practical cases of interest such an interruption would require additional protocol bits to indicate where one message left off and the other began. Therefore, from the point of view of minimizing overhead, it would be advantageous to complete transmission of a message once begun.

A complicating factor in the analysis of nonpreemptive priority queues is that there is interaction between all priority levels. For example, suppose that a message from the highest-priority class arrives in the system to find a lower-priority class message being transmitted. Even if no messages of the same class are in the system, there is a delay until the lower-class message has completed transmission. This delay is affected by the length of the lower-class message. The probability of a lower-class message being in service is also affected by the relative arrival rates of all classes. This stands in contrast to the preemptive priority discipline where higher-class messages are oblivious to lower-class traffic. Because of this interaction, it is necessary to consider no less than three priority classes. The middle class is affected by both lower and higher-priority classes.

As in most of our work on M/G/1 queues, we shall rely on the imbedded Markov chain to carry out the analysis.[9] We assume that messages from all three classes arrive independently at a Poisson rate with average λ_k; $k = 1, 2, 3$, respectively. Again the Markov chain is imbedded at service completion times. However, since there are three distinct classes of messages, we find it necessary to distinguish among the imbedded points as to which class completed service. This is indicated by the term *k class epoch*, where $k = 1$, 2, or 3. Let N_{ik} be the number of class k messages in the system at the ith departure epoch. As in the previous cases, we can express N_{ik} in terms of the number in the system at the previous epoch and the number of new arrivals. Suppose that the $(i+1)$st departure epoch is class 1. The impact of this is that a class 1 message has departed and that new messages of all three types have arrived while this message was being transmitted. We define $A_{jk}, j, k = 1, 2, 3$ as the number of class j messages to arrive during the transmission of a class k message. For simplicity of notation, we have dispensed with any reference to the departure time in A_{jk}.

In writing equations for buffer occupancy, we recognize that there are four disjoint events that cover all the possibilities, of interest. In the first of these the server is occupied with a class 1 message $\{N_{i1} > 0\}$. In the second the server is occupied with a class 2 message, $\{N_{i1} = 0, N_{i2} > 0\}$. The server is occupied with a

[9] The analysis of the nonpreemptive discipline presented here is from Cox and Smith (1961).

class 3 message in the third event. $\{N_{i1} = N_{i2} = 0, N_{i3} > 0\}$. The final event is an empty system: $\{N_{i1} = N_{i2} = N_{i3} = 0\}$.

The priority system has the same probability of being empty as a queue with the same loading but without priorities. This follows from the fact that in either case the server is never idle. From a simple argument using Little's formula, we have for the probability of the system being empty

$$\Pi_0 \stackrel{\Delta}{=} P[N_{i1} = N_{i2} = N_{i3} = 0] = 1 - \lambda_1 \bar{M}_1 - \lambda_2 \bar{M}_2 - \lambda_3 \bar{M}_3 \qquad (6.93)$$

where $\bar{M}_k$, $k = 1, 2, 3$ is the average length of a class k message. As in earlier work we define the traffic intensity as

$$\rho \stackrel{\Delta}{=} \lambda_1 \bar{M}_1 + \lambda_2 \bar{M}_2 + \lambda_3 \bar{M}_3 \qquad (6.94)$$

This quality can also be viewed as the portion of the time that there is at least one message among the three queues. Now considering the dynamics of the process over an interval, which is long enough for equilibrium to prevail. The probability of the departure of a class k message is equal to the probability of a class k message arriving to a nonempty buffer. We find

$$\begin{aligned}
\Pi_1 &\stackrel{\Delta}{=} P[N_{i1} > 0] = \frac{\lambda_1 \rho}{\lambda} \\
\Pi_2 &\stackrel{\Delta}{=} P[N_{i1} = 0, N_{i2} > 0] = \frac{\lambda_2 \rho}{\lambda} \\
\Pi_3 &\stackrel{\Delta}{=} P[N_{i1} = N_{i2} = 0, N_{i3} > 0] = \frac{\lambda_3 \rho}{\lambda}
\end{aligned} \qquad (6.95)$$

Now we consider the number of messages of each class that are in the system at the imbedded points. We may write for $N_{i1} > 0$

$$\begin{aligned}
N_{i+11} &= N_{i1} - 1 + A_{1,1} \\
N_{i+12} &= N_{i2} + A_{2,1} \\
N_{i+13} &= N_{i3} + A_{3,1}
\end{aligned} \qquad (6.96)$$

If the $(i+1)$st departure is class 2, that is, $\{N_{i1} = 0, N_{i2} > 0\}$, we have

$$\begin{aligned}
N_{i+11} &= A_{1,2} \\
N_{i+12} &= N_{i2} - 1 + A_{2,2} \\
N_{i+13} &= N_{i3} + A_{3,2}
\end{aligned} \qquad (6.97)$$

In considering a class 3 epoch, we recognize that the ith departure must have left the system devoid of class 1 and 2 messages: $\{N_{i1} = N_{i2} = 0, N_{i3} > 0\}$. We have

$$\begin{aligned} N_{i+11} &= A_{3,1} \\ N_{i+12} &= A_{2,3} \\ N_{i+13} &= N_{i3} - 1 + A_{3,3} \end{aligned} \tag{6.98}$$

The final equation is obtained by considering the situation when the ith departure leaves the system completely empty $N_{i1} = N_{i2} = N_{i3} = 0$. The total message arrival rate is λ. The probability that the next message to arrive is of class 1, 2, or 3 is λ_1/λ_2, λ_2/λ or λ_3/λ, respectively. Of course, these are the probabilities of the message that departs next. At the $(i+1)$st departure all messages in the system arrive during the transmission time of this message. Thus with probability λ_k/λ we have

$$N_{i+1,l} = A_{lk}, \quad k, l = 1, 2, 3 \tag{6.99}$$

We recognize that (6.96)–(6.99) are generalizations of the imbedded Markov chain equation for a single class of customers given in (6.3).

We calculate the two-dimensional probability-generating functions of $N_{i+1,1}$ and $N_{i+1,2}$ by conditioning on these four events. From (6.96)–(6.99) we have

$$\begin{aligned} E[z_1^{N_{i+1,1}} z_2^{N_{i+1,2}}] &= \Pi_1 E[z_1^{N_{i1}-1+A_{11}} z_2^{N_{i2}+A_{21}} / N_{i1} > 0] \\ &\quad + \Pi_2 E[z_1^{A_{12}} z_2^{N_{i_2}-1+A_{22}} / N_{i1} > 0, N_{i2} > 0] \\ &\quad + \Pi_3 E[z_1^{A_{13}} z_2^{A_{23}} / N_{i1} = N_{i2} = 0, N_{i3} > 0] + \Pi_0 \sum_{k=1}^{3} \left(\frac{\lambda_k}{\lambda}\right) \\ &\quad \times E[z_1^{A_{1k}} z_2^{A_{2k}} / N_{i1} = N_{i2} = N_{i3} = 0] \end{aligned} \tag{6.100}$$

We follow the same steps as in the one-dimensional case to solve for the probability-generating function. The number of message arrivals is independent of the number of messages in the system. The fact that message arrivals are Poisson over a random interval leads to a familiar relationship. We have

$$\begin{aligned} E[z_1^{A_{11}} z_2^{A_{21}}] &= \int_0^\infty \sum_{m=0}^{\infty} z_1^m \frac{(\lambda_1 t)^m}{m!} e^{-\lambda_1 t} \sum_{n=0}^{\infty} z_2^n \frac{(\lambda_2 t)^n}{n!} e^{-\lambda_2 t} m_1(t) dt \\ &= M_1(\lambda_1(1 - z_1) + \lambda_2(1 - z_2)) \end{aligned}$$

where $m_1(t)$ and $M_1(s)$ are, respectively, the probability density function and the Laplace transform of the density function for a class 1 message. Similar relationships can be found for the other message classes:

$$E[z_1^{A_{12}} z_2^{A_{22}}] = M_2(\lambda_1(1 - z_1) + \lambda_2(1 - z_2))$$

$$E[z_1^{A_{13}} z_2^{A_{23}}] = M_3(\lambda_1(1 - z_1) + \lambda_2(1 - z_2))$$

Again by conditioning on events, we may write

$$\begin{aligned} E[z_1^{N_{i+1,1}} z_2^{N_{i+1,2}}] &= \Pi_1 E[z_1^{N_{i+1,1}} z_2^{N_{i+1,2}} / N_{i+1,1} > 0] \\ &\quad + \Pi_2 E[z_2^{N_{i+1,2}} / N_{i+1,1} = 0, N_{i+1,2} > 0] + 1 - \Pi_1 - \Pi_2 \end{aligned}$$

Assuming equilibrium has been attained, we define the probability-generating functions

$$G_1[z_1, z_2] \overset{\Delta}{=} E[z_1^{N_1} z_2^{N_2} / N_1 > 0] = \sum_{i=1}^{\infty} \sum_{j=0}^{\infty} z_1^i z_2^j P(N_1 = i, N_2 = j / N_1 > 0)$$

and

$$G_2[z_2] \overset{\Delta}{=} E[z_2^{N_2} / N_1 = 0, N_2 > 0] = \sum_{j=1}^{\infty} z_2^j P(N_2 = j / N_1 = 0, N_2 > 0)$$

Substituting into (6.100), we obtain

$$\begin{aligned} &\Pi_1 G_1(z_1, z_2) + \Pi_2 G_2(z_2) + 1 - \Pi_1 - \Pi_2 \\ &\quad = \Pi_1 z_1^{-1} G_1(z_1, z_2) M_1(\lambda_1(1 - z_1) + \lambda_2(1 - z_2)) \\ &\qquad + \Pi_2 z_2^{-1} G_2(z_2) M_2(\lambda_1(1 - z_1) + \lambda_2(1 - z_2)) \\ &\qquad + \Pi_3 M_3(\lambda_1(1 - z_1) + \lambda_2(1 - z_2)) \\ &\qquad + \Pi_0 \sum_{k=1}^{3} \frac{\lambda_k}{\lambda} M_k(\lambda_1(1 - z_1) + \lambda_2(1 - z_2)) \end{aligned} \tag{6.101}$$

By setting $z_1 = z_2 = 1$, in (6.101), we see that there is self-consistency inasmuch as we can derive $1 = 1$ if $G_1(1, 1) = G_2(1) = 1$. (See the associated Maple spreadsheet.)

In order to find the moments we derive two equations from $\partial G_1(z_1, z_2)/\partial z_1|_{z_1=1}$ and $\partial G_1(z_1, z_2)/\partial z_2|_{z_2=1}$. These equations can be solved to corroborate the relation $\Pi_j = \lambda_j \rho/\lambda$; $j = 1, 2, 3$. (This corresponds to demonstrating that $P_0 = 1 - \rho$ in the one-dimensional case.)

By differentiating successively with respect to z_1 and with respect to z_2, we find $\partial^2 G_1(z_1, z_2)/\partial^2 z_1\big|_{\substack{z_1=1,\\ z_2=1}}$, $\partial^2 G_1(z_1, z_2)/\partial z_1 \partial z_2\big|_{\substack{z_1=1,\\ z_2=1}}$ and $\partial^2 G_1(z_1, z_2)/\partial^2 z_2\big|_{\substack{z_1=1,\\ z_2=1}}$ which yield the following three equations, respectively

$$\begin{aligned}
\bar{N}_{11} &= 1 + \frac{\lambda_1 \sum_{k=1}^{3} \lambda_k \bar{M}_k^2}{2\rho(1-\lambda_1\bar{M}_1)} \\
\bar{N}_{12} &= \frac{\rho\lambda_2\bar{M}_1(\bar{N}_{11}-1) + \rho\lambda_2\bar{M}_2(\bar{N}_{22}-1) + \lambda_2\sum_{k=1}^{3}\lambda_k\bar{M}_k^2}{\rho(1-\lambda_1\bar{M}_1)} \\
\bar{N}_{22} &= 1 + \frac{2\rho\lambda_1\bar{M}_2\bar{N}_{21} + \lambda_2\sum_{k=1}^{3}\lambda_k\bar{M}_k^2}{2\rho(1-\lambda_2\bar{M}_2)}
\end{aligned} \tag{6.102}$$

where

$$\begin{aligned}
\bar{N}_{11} &= \left.\frac{\partial G_1(z_1, z_2)}{\partial z_1}\right|_{z_1=z_2=1} = E[N_1/N_1 > 0] \\
\bar{N}_{12} &= \left.\frac{\partial G_1(z_1, z_2)}{\partial z_2}\right|_{z_1=z_2=1} = E[N_2/N_1 > 0] \\
\bar{N}_{22} &= \left.\frac{\partial G_2(z_2)}{\partial z_2}\right|_{z_2=1} = E[N_2/N_1 = 0, N_2 > 0]
\end{aligned}$$

Solving for $\bar{N}_{22}$, we find that

$$\bar{N}_{22} = 1 + \frac{\lambda_2\sum_{k=1}^{3}\lambda_k\bar{M}_k^2}{2\rho(1-\lambda_1\bar{M}_1-\lambda_2\bar{M}_2)(1-\lambda_1\bar{M}_1)}$$

Both $\bar{N}_{11}$ and $\bar{N}_{22}$ given in (6.102) represent similar quantities, the expected number of messages of class 1 and class 2, respectively, given that one is beginning transmission. The average numbers of messages that have arrived during the queueing time of the message to be transmitted are $\bar{N}_{11} - 1$ and $\bar{N}_{22} - 1$ for class 1 and class 2, respectively. All of these late-arriving messages arrive to a nonempty system with probability ρ. Since messages that arrive to an empty system suffer no queueing delay, we have for the average queueing delay for class 1 and class 2 messages, respectively:

$$\begin{aligned}
\bar{Q}_1 &= \frac{\rho}{\lambda_1}(\bar{N}_{11}-1) = \frac{\sum_{k=1}^{3}\lambda_k\bar{M}_k^2}{2(1-\lambda_1\bar{M}_1)} \\
\bar{Q}_2 &= \frac{\rho}{\lambda_2}(\bar{N}_{22}-1) = \frac{\sum_{k=1}^{3}\lambda_k\bar{M}_k^2}{2(1-\lambda_1\bar{M}_1)(1-\lambda_1\bar{M}_1-\lambda_2\bar{M}_2)}
\end{aligned}$$

Class 3 is found by lumping classes 1 and 2 into a single higher class:

$$\bar{Q}_3 = \frac{\sum_{k=1}^{3} \lambda_k \bar{M}_k^2}{2(1 - \lambda_1 \bar{M}_1 - \lambda_2 \bar{M}_2)(1 - \lambda_1 \bar{M}_1 - \lambda_2 \bar{M}_2 - \lambda_3 \bar{M}_3)} \tag{6.103}$$

These equations admit easy generalizations to a larger set of classes. Consider class j of L classes where $j > 2$ and $L > 3$. All the classes $1, 2, \ldots, j - 1$ can be lumped together into the higher class, and we have

$$\bar{Q}_j = \frac{\sum_{k=1}^{L} \lambda_k \bar{M}_k^2}{2\left(1 - \sum_{k=1}^{j-1} \lambda_k \bar{M}_k)(1 - \sum_{k=1}^{j} \lambda_k \bar{M}_k\right)} \quad j = 1, 2, \ldots, L \tag{6.104}$$

The average delay of a class j message including transmission time is

$$\bar{D}_j = \bar{M}_j + \bar{Q}_j \quad j = 1, 2, \ldots, L \tag{6.105}$$

Example 6.10 On the associated Maple spreadsheet the equations for average delay are worked out for the case of constant-length messages.

Example 6.11. Buffer Insertion Rings There is an obvious difficulty with the multiplexing technique examined in Example 6.9. Since messages consisting of several packets can be interrupted in the middle of a transmission, it is necessary to transmit addressing information with each packet. Also, there is a random reassembly time of a message at the destination. Adequate storage must be provided at destinations only so that incomplete messages can be buffered during assembly. An alternative technique, which avoids these problems, is a form of nonpreemptive priority in which messages cannot be interrupted once transmission has begun. The technique is called *buffer insertion* [Hafner et al. (1974)]. The flow into a terminal consists of two streams: locally generated traffic and through-line traffic from other terminals. Depending on implementation, one or the other of these is given *nonpreemptive priority*. While a message from one of these streams is being transmitted, messages from the other are buffered. After the transmission of a message, any messages of higher priority stored in the terminal are transmitted. When all higher-priority messages are transmitted, then any lower-priority messages present are transmitted. This alternation is indicated by the switch shown in Figure 6.10. In the case of symmetric traffic, it is reasonable to assume that the traffic already on the line has priority. However, for the data collection ring more equitable service may be obtained if locally generated traffic has priority.

The average delay of a message in a buffer insertion system can be found by the application of the results of the nonpreemptive queue analysis and Little's formula (Bux et al. 1983). We assume that the message arrival rate at each terminal follows

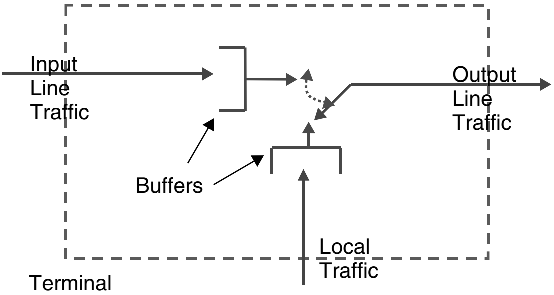


Figure 6.10 Buffer insertion.

an independent Poisson distribution. The average arrival rate for terminal i is λ_i messages per second. A final assumption is that the message lengths for each terminal are independent identically distributed (iid) random variables with mean and mean-square values $\bar{M}$ and $\bar{M}^2$, respectively. The basic principle of the analysis can be explained by considering terminals 1 and 2 of a data collection ring in which terminals transmit all their messages to the nth terminal. We are interested in the average delay of a message that starts at terminal 1, where it suffers the average delay of an M/G/1 queue with message arrival rate λ_1. We indicate this delay by $D(\lambda_1)$. The difficulty is in calculating the delay in passing through terminal 2. If terminal 2 is empty, it goes on without delay to the rest of the system. However, even if line traffic has priority over locally generated traffic, the presence of messages at terminal 2 will cause delay to terminal 1 messages. The trick is to find the average of this second component of delay. Consider the aggregate system formed by terminals 1 and 2, where, for the moment, we assume no delay between terminals 1 and 2. We may analyze this system by imbedding a Markov chain at the points when messages depart over the output line between terminals 2 and 3. Once transmission of a message from terminal 1 or 2 has begun, it continues unhindered. Messages arrive at the system at a Poisson rate with average $\lambda_1 + \lambda_2$ messages per second. When a transmission is completed, a new message is chosen for transmission according to established priorities. When a message arrives at an empty system, it is immediately transmitted over the output line. Considering all of these elements, it is clear that the total number of messages at terminals 1 and 2 is governed by the equation same as that governing the M/G/1 queue [see (6.3)]. We conclude that the average number of messages at terminals 1 and 2 is given by (6.11) with $\lambda = \lambda_1 + \lambda_2$. From Little's formula, the average delay of a message entering the aggregate system is given by (6.12) with $\lambda_1 + \lambda_2$. We indicate this overall average delay by $\bar{D}(\lambda_1 + \lambda_2)$. The average delays in terminals 1 and 2 are denoted $\bar{D}(\lambda_1)$ and $\bar{D}_2$, respectively. From these quantities the average number of messages in the system may be calculated two ways. From the application of Little's formula, we have

$$(\lambda_1 + \lambda_2)\bar{D}(\lambda_1 + \lambda_2) = \lambda_1\bar{D}(\lambda_1) + (\lambda_1 + \lambda_2)\bar{D}_2$$

We emphasize that $\bar{D}_2$ is the average delay for all messages passing through terminal 2. This delay may be decomposed into two components according to the origin of messages. The average delay of messages generated at terminal 2 can be calculated from known results since message generation is a Poisson process. If line traffic has priority, the average delay of such messages is, from (6.104) and (6.105)

$$\bar{D}_{22} = \bar{M} + \frac{(\lambda_1 + \lambda_2)\bar{M}^2}{2(1 - \rho_1 - \rho_2)(1 - \rho_1)} \tag{6.106}$$

where $\rho_1 = \lambda_1 \bar{M}$ and $\rho_2 = \lambda_2 \bar{M}$. If locally generated traffic has priority, the average delay is

$$\bar{D}_{22} = \bar{M} + \frac{(\lambda_1 + \lambda_2)\bar{M}^2}{2(1 - \rho_2)} \tag{6.107}$$

Once again, we proceed by using Little's formula. We calculate the average number of messages in terminal 2 from two approaches. The result is

$$(\lambda_1 + \lambda_2)\bar{D}_2 = \lambda_1 \bar{D}_{21} + \lambda_2 \bar{D}_{22}$$

The average delay of a message in transit through terminal 2 is given by solving for $\bar{D}_{21}$

$$\begin{aligned} \bar{D}_{21} &= \frac{(\lambda_1 + \lambda_2)}{\lambda_1}\bar{D}_2 - \frac{\lambda_2}{\lambda_1}\bar{D}_{22} \\ &= \frac{\lambda_1 + \lambda_2}{\lambda_1}\left[\bar{D}(\lambda_1 + \lambda_2) - \frac{\lambda_1}{\lambda_1 + \lambda_2}\bar{D}(\lambda_1)\right] - \frac{\lambda_2}{\lambda_1}\bar{D}_{22} \end{aligned} \tag{6.108}$$

where $\bar{D}(\lambda)$ is the average delay of a message in an M/G/1 queue with message arrival rate λ, and $\bar{D}_{22}$, the delay at terminal 2 of a message generated at terminal 2, is given by (6.106) or (6.107), depending on priority. Finally, we have the following equation for the total delay of a message from terminal 1:

$$\bar{D}_1 = \bar{D}(\lambda_1) + \bar{D}_{21} = \frac{\lambda_1 + \lambda_2}{\lambda_1}\bar{D}(\lambda_1 + \lambda_2) - \frac{\lambda_2}{\lambda_1}\bar{D}_{22}$$

The result generalizes easily for the case of a data collection ring of N terminals where all messages have the same destination, terminal N. Consider the transient delay for a message passing through the ith terminal. For all intents and purposes, the first $i - 1$ terminals act like a single M/G/1 queue. The transit delay through terminal i is given by (6.106) or (6.107), depending on priority, and (6.108) with λ_1 replaced by $\sum_{j=1}^{i-1} \lambda_j$ and λ_2 by λ_i, where λ_i is the arrival rate of messages to the ith terminal. A message starting at terminal i suffers transit delay at terminals $i+1$, $i+2, \ldots, N-1$. In addition to the transit delay there is a delay at the terminal where the message is generated. Depending on the priority, this initial delay is given by (6.106) or (6.107). For the first terminal in the data collection ring the initial delay formula degenerates to that of the M/G/1 queue.

In the case of more symmetric traffic the foregoing analysis cannot be used to yield exact results; approximations must be employed. We use the equations that we have derived with λ_1 replaced by the number of messages per second passing through the terminal under consideration and λ_2 by the generation rate of local messages. The average rate of messages passing through the Nth terminal is given by (6.87). For the case of symmetric traffic, the rate of messages flowing through a terminal is given by $N\lambda/2 - \lambda$. The rate of locally generated traffic is λ. These quantities are used in (6.106)–(6.108) to find queueing and transit delay.

The calculation of message transmission time and processing delay proceeds in the same manner as in the previous case [see (6.88)–(6.91)]. In the case of buffer insertion, it is not necessary to transmit an address with each packet—one address per message will work. This increases the message transmission time by $[\log_2(N)]^+/R$. In our calculation this is small enough to be ignored.

We point out that a message suffers a transit delay for each terminal it passes through. In the symmetric case, for example, messages pass through $N/2$ terminals on the average. In each of the terminals on the route of a packet, it is necessary to examine the header. This factor is similar to that encountered in the analysis of the previous section, i.e., $(N/2)([\log_2(N)]^+ + 2)/R(N/2)$ seconds on the average, where R is the bit rate on the line.

In order to implement the buffer insertion technique, a certain amount of overhead is necessary. A practical way to implement the buffer insertion technique involves segmenting the flow on the line into slots. User messages are segmented into fixed-length packets that fit into these slots. As in the case of the other multiplexing techniques considered in this chapter, we mark the beginning of the slot to indicate occupancy. Also, there is a bit to indicate whether a slot contains the last packet of a message. As in the previously considered technique, the time required to transmit the data in a message is given by (6.88). In this case it is not necessary to transmit addresses along with each packet, and the duration of a slot is given by

$$T = \frac{P+2}{R} \text{ seconds}$$

where P is the number of information bits in a slot. We may assume that the address information that must accompany each message may be packaged in a special packet consisting of $2[\log_2 N]^+$ bits. The mean and mean-square values of the time required to transmit a message can be calculated from these considerations in a straightforward fashion.

6.4 POLLING

In this section we derive performance results for the polling technique that is used to grant access to a number of geographically dispersed terminals.[10] The particular analysis relies heavily on the results that we have just obtained for the M/G/1 queue.

6.4.1 Basic Model: Applications

In the model of polling systems, a number of independent sources share a common facility, usually a transmission line; however, unlike priority queues, the sharing is equal. In order to explain polling, we use the commutator analogy where a server cycles among source buffers (see Fig. 6.11). Messages arrive at a set of queues. In most cases of interest we take the arrival process to be Poissonian. The server goes from queue to queue in some prescribed order, pausing to remove messages from each of the queues. A salient feature of the model is that the amount of time spent by the server at a queue depends on the number of messages in the queue when the server arrives. This leads to stochastic dependencies among the queues. A second important factor is walktime[11] or overhead. After a server leaves a queue and before it begins work on the next queue, there is a period during which there is a walktime and the server remains idle. In most cases of interest this walktime between queues is a constant. However, analyses can be carried out under more general assumptions. In carrying out the analyses of polling systems, there are two basic quantities of interest: the time required by the server to complete a cycle through all queues, and the delay of a message in obtaining and completing service.

A common realization of the polling model is *rollcall polling*, which is illustrated for the tree topology in Figure 6.12 . Each terminal connected to the tree is assigned a unique address. These addresses are broadcast in sequence over the common line by the central processor. After each broadcast the processor pauses and waits for a response from the terminal address. If the terminal has a message, the polling cycle is interrupted while the message is transmitted. When the terminal has been served, the polling cycle resumes where it left off. The time required to serve a particular terminal depends on the number of messages in a terminal buffer and their duration. The overhead or walktime consists of the amount of time required to broadcast an address and to wait for a reply. Since the central processor is receiving from a number of different terminals over different lines with different transfer characteristics, part of this listening time may be used in adjusting equalizers and acquiring phase and timing where necessary.

The same polling model applies to the *token-passing* [12] technique, which is part of standards for local-area networks (LANs.) In the IEEE 802.5 standard, terminals are connected to a transmission line that assumes the ring topology (see Fig. 6.12).

[10] The definitive reference on the performance of polling systems is Takagi (1986).

[11] The term walktime has its origin in machine repair problems, which are analyzed by polling models (Mack et al. 1957, Kaye 1972). An exposition of this work is contained in Hayes (1984).

[12] The token-passing technique in a ring system seems to be the earliest accessing technique in a local-area network environment (Farmer and Newhall 1969). See also Bux et al. (1981).

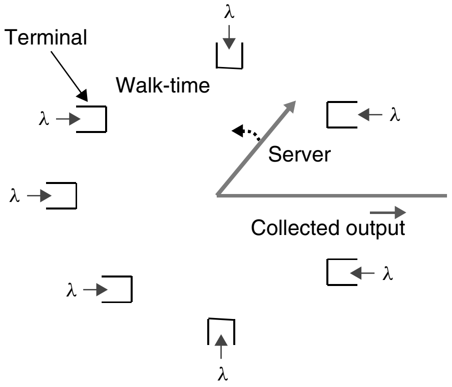


Figure 6.11 Generic polling model.

Transmission around the ring is in one direction. Randomly arriving messages are multiplexed on the line when a terminal is granted access to the line. The discipline used to share access is that only one terminal at a time is allowed to put messages on the line. These messages are presumably addressed to one or more terminals on the line. The messages are removed from the line by the addressed terminal. When a terminal has sent all that it is going to send, an “end of message” character is appended to the transmission. This character indicates that another terminal having a message can seize control of the line. First priority is granted to the terminal immediately downstream. If this terminal has no message, it appends a “no message” character to the transmission from the upstream terminal. In terms of

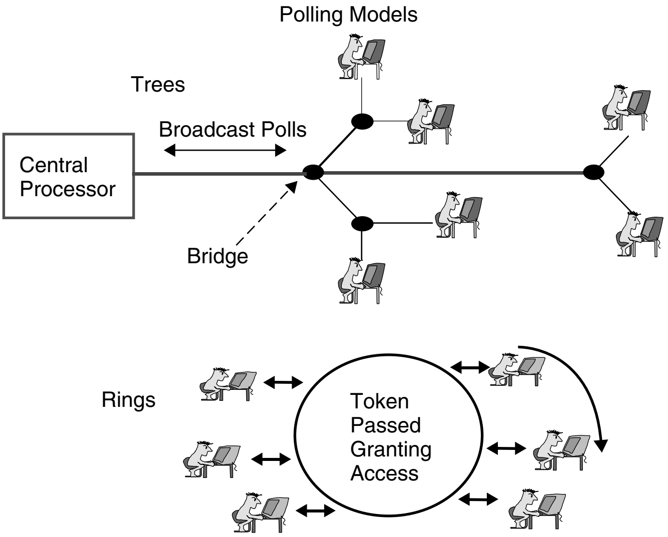


Figure 6.12 Polling applications.

mathematical modeling, this system has the same characteristics as the previous polling model. The time that a terminal has access to the line is random, depending on the message arrival process. The time required to transmit an "end of message" and "no message" characters is overhead.

The token-passing technique is also applicable to the bus topology, which is the subject of the IEEE 802.4. In this case the token is passed from terminal to terminal in some preassigned order without intervention of a central processor. Again the same model applies.

6.4.2 Average Cycle Time

An expression for the average cycle time, which is the time required for the server to visit each of the queues once. We assume that the system is symmetric so that the arrival rate to each of the N terminals is λ messages per second. In the analysis that follows, we will be concerned with the overall message arrival rate to the system, which we denote as $\Lambda = N\lambda$. The average amount of work that enters the system during an average cycle is

$$\bar{W}_E = \Lambda\bar{M}\bar{T}_C = \rho\bar{T}_C$$

where $\bar{T}_C$ is the average cycle time and $\bar{M}$ the average transmission time of a message. All the work must be done while the server is at a queue; accordingly, for a symmetric system, $1/N$ of the work must be done at each queue. $\bar{S}$, the average amount of time that the server spends at a queue, is then given by

$$\bar{S} = \frac{\rho\bar{T}_C}{N} \tag{6.109}$$

Now the average cycle time is the sum of the walk time and the server visit time

$$\bar{T}_C = N\bar{W} + N\bar{S} \tag{6.110}$$

where $\bar{W}$ is the average time required to "walk" between queues. Substituting (6.109) into (6.110) and solving for $\bar{T}_C$, we have

$$\bar{T}_C = \frac{N\bar{W}}{1-\rho} \tag{6.111}$$

6.4.3 Average Delay: Exhaustive, Gated, and Limited Service

In this section we will derive expressions for the average delay for three polling disciplines, *exhaustive* service, *gated* service, and *limited service.*[13] Messages with an arbitrary distribution of lengths arrive to each of the N terminal queues at a

[13] The analysis of average delay presented here has been distilled from Boxma and Groenendijk (1987), Fuhrmann (1984, 1985), Fuhrmann and Cooper (1985).

Poisson rate. As in the previous subsection, we assume that the system is symmetric so that the arrival rate to each terminal is λ messages per second and the total arrival rate is $\Lambda = N\lambda$. For *exhaustive service* the server transmits all the messages that it finds at a queue plus any that arrive while it is at the queue; when the server departs, the queue is empty. For *gated service*, the server transmits only the messages that it finds on arrival to the queue. Any message that arrives while the server is at the queue is held over until the next visit of the server. In the *limited service* discipline, the server removes up to a fixed maximum number from the queue in any visit. In our analysis, we will consider the maximum to be one so that questions relating to gating do not arise. In all three cases the server that arrives to an empty queue moves on immediately. The walktime between terminal pairs is assumed to be a random variable, which is independent from pair to pair. In most cases of interest the walktime is a constant. Our basic approach in calculating average delay is the calculation of the average number of messages in all N terminals simultaneously and the application of Little's formula.

We may view the polling system as a particular sort of M/G/1 queue: one in which the server serves messages in a peculiar order, serving a set of the messages a terminal and moving on to serve a set of messages at another terminal. Another peculiarity is that, after serving a terminal's group of messages, the server is idle, on vacation, as it were, while walking from one terminal to another. In order to calculate the average number of messages in the system, we study a different model, which we call the *mimic* system. The mimic system consists of a *single* M/G/1 queue in which messages are served according to the last-come first served (LCFS) discipline. We assume that the arrivals to the mimic system are *exactly* the same as those to the polling system. Finally, it is assumed that the server in the mimic system goes on vacation at *exactly* the same time as the server in the polling system walks from one queue to another. Work in terms of message transmission times arrives to each system at exactly the same rate. Since the servers work at the same rate in both systems, the amount of work remaining in each system at any point in time is the same for both systems. The difference between the polling systems and the mimic resides in the order in which messages are served. However, as we have seen, order of service has no effect when calculating average delay. The virtue of the mimic system is that the LCFS discipline makes it relatively easy to count the average number of messages. Once we know the average number of messages in the system, we use Little's formula to find average delay.

Server Vacations We begin by deriving the average number of messages left behind by a message departing from the mimic system. For the purposes of analysis, we place message arrivals into two classes (see Fig. 6.13). *Initiating* messages are those that arrive while the server is on vacation. *Following* messages arrive while the server is available. Let us begin the explanation with the server on vacation. Keep in mind that the discipline is LCFS. After the vacation is over, the server begins with the last arriving message. Messages that arrive during this service are served immediately. We conceive of each initiating message as being the first generation of a virtual busy period. The following messages form successive generations of

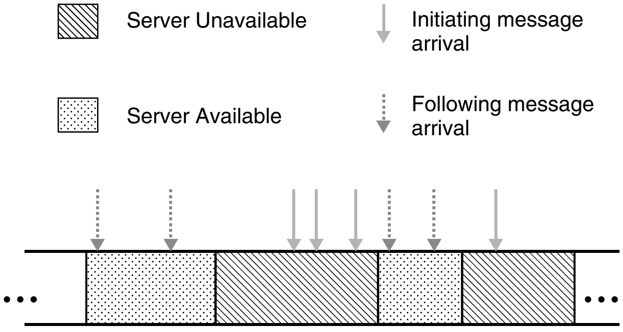


Figure 6.13 Message arrival.

particular busy periods. Thus each arrival belongs to one and only one busy period. After the busy period of an initiating message is finished, the server goes on to the next-to-last (penultimate) arriving initiating message. If the server goes on vacation before the end of a particular busy period, work on the busy period is suspended until all busy periods engendered by the arrivals during the vacation have been transmitted. Note that, if the system is empty, the server will not be available since it is cycling from terminal to terminal.

A tagged departing message leaves behind messages, which may be placed in three categories:

- Messages in its own busy period
- Messages that arrived during the same vacation as the initiator of its busy period
- Messages that were in the system at the beginning as the tagged vacation

Consider the *first* of these, the messages in the tagged message's busy period. In fact, if we consider *only* the initiating message and the resulting following messages that are part of the same *virtual busy period*, the number of messages left behind by a departing message is governed by (6.3) with the resulting PGF given by $N(z)$ in (6.6). The arrival rate in these equations is the total arrival rate to the system, Λ messages per second. The average number of such messages is given by (6.10). Note that this accounting holds even though service may have been interrupted by a server vacation. Arrivals during this interruption are simply not part of the tagged busy period since they do not arrive while members of the busy period are being served.

The *second* category is initiating messages, which arrive in the same server vacation before the one that initiated the tagged busy period discussed above. (We call these arrivals during the *tagged vacation*.) These are the messages that arrive during the residual life of the vacation period. Recall that, in Section 6.1.3, we derived the probability distribution for the residual life. Now, define the discrete random variable L as the number of arrivals during the residual life of a vacation time. From the relationship that we have established between the Laplace transform for an interval and the probability generation function of the number of Poisson arrivals over the interval [see (6.13) and (6.31)], we have the following equation for the probability-generating function for the number of initiating messages remaining after the departure of a message in a virtual busy period:

$$L(z) = U(\Lambda(1-z)) = \frac{1 - W(\Lambda(1-z))}{\Lambda(1-z)\bar{W}} \tag{6.112}$$

Since the L and the number of messages left behind in a virtual busy period are independent random variables, the number of messages initiating and following in the first two categories left behind by a departing message has the PGF

$$T(z) = L(z)P(z) \tag{6.113}$$

where $P(z)$ is as given by (6.6). Differentiating with respect to z and setting $z = 1$, we have for the average in the first two categories left behind by the departing tagged message

$$\bar{T} = T'(0) = \frac{\bar{N} + \Lambda\bar{V}^2}{2\bar{V}} \tag{6.114}$$

where $\bar{V}^2$ is the mean-square vacation time and $\bar{N}$ is as given by (6.10).

We are now in a position to find the average number of messages in the *third* category, the messages in the system when the tagged vacation began. We assume that the number of messages that arrive during a vacation is independent of the number of messages in the system when the vacation began. This is certainly the case if the vacation were the walktime between queues in a polling system. Let J be the random variable indicating the number of messages in the system at the beginning of the tagged vacation. Because of LCFS, all arrivals after the tagged message are gone from the system. Further, all messages present when the vacation began are still in the system when the tagged message departs. Because of the assumption that the number of arrivals during a vacation is independent of the number present at the beginning of the vacations, the PGF of the total number of remaining messages is given by

$$Q(z) = J(z)T(z) \tag{6.115}$$

where $T(z)$ is given by (6.113) and $J(z)$ is the PGF of J. By differentiating and setting $z = 1$, we have the average number of messages in the system after the departure of the tagged message. Substituting from (6.114), we have

$$\bar{Q} = \bar{N} + \frac{\Lambda\bar{W}^2}{2\bar{W}} + \bar{J} \tag{6.116}$$

The quantity $\bar{J}$ remains to be determined. As we shall see, its value depends on the type of service that the system uses.

The preceding results, culminating in (6.116), apply to a single M/G/1 queue with the LCFS discipline and a server that takes vacations. We are now in a position to apply these results specifically to polling systems. The server going on vacation corresponds to the walktime of the server. As stated above, the mimic system and the polling system have the same amount of work remaining at any point in time even though the two systems serve different messages at any point in time. Because of this equality in work the mimic system and the polling system have the same average number of messages at any point in time as given by (6.116). Since the arrival rates are the same, by Little's formula, the average delay is the same. As mentioned above, the utility of the mimic system is that it is relatively easy to calculate the average number of messages that it contains and it has the same average number of messages as the polling system.

We now define the following random variables: (1) R_S is the number of messages that remain in a queue after the *server* departs from that particular queue and (2) R_M is the number of messages that remain in any one of the queues after an *arbitrary message* departs from that queue. Recall that $\bar{Q}$, given by (6.116), is the average number of messages in the entire system when a message departs the system. From symmetry arguments, we have that

$$\bar{Q} = N\overline{R_M} \tag{6.117}$$

We now derive the relationship between $\bar{R}_S$ and $\bar{J}$, the average number of messages in the entire system when the server leaves any one of the queues. If the server is departing from queue i, the number of messages remaining in queue $(i - j)\mathrm{mod}\, N$ is the sum of those at server departure plus new arrivals. We have

$$\overline{R_S} + j\frac{\Lambda}{N}\frac{\overline{T_C}}{N} = \bar{R}_S + \frac{j\Lambda\overline{T_C}}{N^2}; \quad j = 0, 1, \ldots, N-1$$

We have used the symmetry of the arrival process here. Summing over $j = 0, 1, \ldots, N-1$ and substituting (6.111), we have

$$\bar{J} = N\overline{R_S} + \frac{(N-1)\Lambda\overline{T_C}}{2N} = N\overline{R_S} + \frac{(N-1)\Lambda\bar{W}}{2(1-\rho)} \tag{6.118}$$

The differences among the three different polling disciplines rests on the term $\bar{R}_S$, the average number of message left behind at a queue when the server departs. We begin with the exhaustive service case, which is the simplest since $\bar{R}_S = 0$. From (6.116) and (6.118), the average number of messages remaining after a departure of a tagged message is

$$\overline{Q_E} = \rho + \frac{\Lambda^2\bar{M}^2}{2(1-\rho)} + \frac{\Lambda\bar{W}^2}{2\bar{W}} + \frac{(N-1)\Lambda\bar{W}}{2(1-\rho)} \tag{6.119}$$

From Little's formula, the average delay in the case of exhaustive service is

$$\overline{D_E} = \bar{M} + \frac{\Lambda \bar{M^2}}{2(1-\rho)} + \frac{\bar{W^2}}{2\bar{W}} + \frac{(N-1)\bar{W}}{2(1-\rho)} \tag{6.120}$$

For the case of gated service the number of messages that remain at a queue when the server departs are only those that arrive during the server visit, S. From (6.109) we then obtain

$$\overline{R_S} = \bar{S}\left(\frac{\Lambda}{N}\right) = \frac{\Lambda\rho\overline{T_C}}{N^2} = \frac{\Lambda\rho\bar{W}}{N(1-\rho)} \tag{6.121}$$

and we have for the average delay in a gated system

$$\overline{D_G} = \bar{M} + \frac{\Lambda \bar{M^2}}{2(1-\rho)} + \frac{\bar{W^2}}{2\bar{W}} + \frac{(N-1+2\rho)\bar{W}}{2(1-\rho)} \tag{6.122}$$

[see (6.116) and (6.118)].

Finally, we consider limited service where only one message is removed from a queue during the visit of the server. If the queue is not empty, the server remains just long enough to transmit a single message; accordingly, the average duration of a visit of the server to a queue is

$$\bar{S} = (1-q_0)\bar{M} \tag{6.123}$$

where q_0 is the probability that a server finds a queue empty. Substituting (6.109) and (6.111) into (6.123) and solving for q_0, we have

$$q_0 = 1 - \frac{\Lambda\bar{W}}{1-\rho} \tag{6.124}$$

In this case, the departure of a message coincides with the departure of the server; but if a queue is empty, the departure of a server does not coincide with the departure of a message. We then have

$$\overline{R_S} = (1-q_0)\overline{R_M} \tag{6.125}$$

From (6.117), (6.118), and (6.125), we have

$$\bar{J} = (1-q_0)\bar{Q} + \frac{(N-1)\Lambda\bar{W}}{2(1-\rho)} \tag{6.126}$$

Substituting (6.126) into (6.116) and solving for $\bar{Q}$ yields

$$\bar{Q} = \frac{\left\{\rho + \frac{\Lambda^2\bar{M^2}}{2(1-\rho)} + \frac{\Lambda\bar{W^2}}{2\bar{W}} + \frac{(N-1)\Lambda\bar{W}}{2(1-\rho)}\right\}}{q_0} \tag{6.127}$$

Applying Little's formula and (6.124), we have for the average delay in the limited service case

$$\begin{aligned}\overline{D_L} &= \left[\frac{(1-\rho)}{1-\rho-\Lambda\bar{W}}\right]\left[\bar{M}+\frac{\Lambda\bar{M}^2}{2(1-\rho)}+\frac{\bar{W}^2}{2\bar{W}}+\frac{(N-1)\bar{W}}{2(1-\rho)}\right]\\ &= \left[\frac{(1-\rho)}{1-\rho-\Lambda\bar{W}}\right]\overline{D_E} \qquad (6.128)\end{aligned}$$

In summary, we have derived the following equations for the average delay in the cases of exhaustive, gated and limited service, respectively:

$$\begin{aligned}\overline{D_E} &= \bar{M}+\frac{\Lambda\bar{M}^2}{2(1-\rho)}+\frac{\bar{W}^2}{2\bar{W}}+\frac{(N-1)\bar{W}}{2(1-\rho)}\\ \overline{D_G} &= \overline{D_E}+\frac{\rho\bar{W}}{(1-\rho)} \qquad (6.129)\\ \overline{D_L} &= \left[\frac{(1-\rho)}{1-\rho-\Lambda\bar{W}}\right]\overline{D_E}\end{aligned}$$

Recall that the message arrival rate Λ in these equations is the arrival rate to all N stations in the system. We could just as well have written the equations in terms of the arrival rate to each station, $\lambda = \Lambda/N$.

These equations are plotted in Figure 6.14. We have delay as a function of the message arrival rate to a single station for each of the three cases. The particular results are for a ratio of walktime to message length of 0.1. As indicated, for

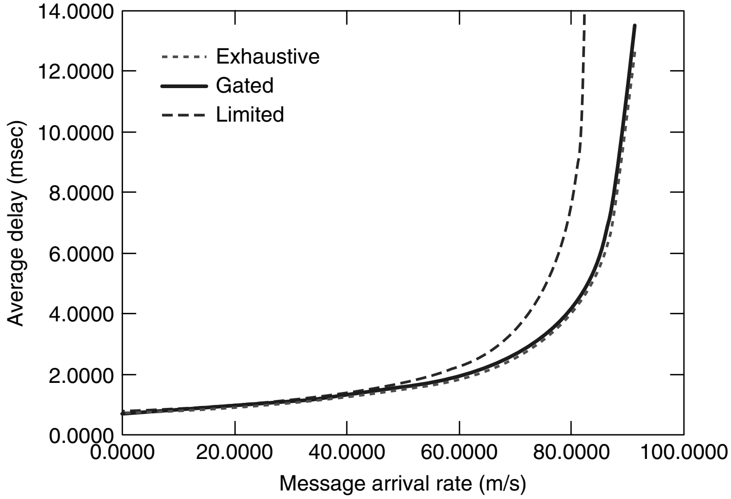


Figure 6.14 Plot showing equations for calculating message delay.

relatively small walktime, there is little difference between gated and exhaustive service. In the recent work by T. Katayama et al. 2003, time limited service polling system has been analyzed for evaluating system delay variance and an optimal assignment of the total time of limited service periods, by using M/G/1 queue model with time-limited service and two types of vacations. The analysis also included the case of customer having its service interrupted due to timer expiration, could be attended by the nonpreemptive discipline.

REFERENCES

Anderson, R. R., J. F. Hayes, and D. N. Sherman, "Simulated performance of a ring switched data network," *IEEE Trans. Commun.* **COM 20**(3): 516–591 (June 1972).

Avi-ltzhak, B., and P. Naor, "Some queueing problems with the service station subject to breakdown," *Oper. Research* **11**(3): 303–320 (1963).

Boxma, O. J., and W. P. Groenendijk, "Pseudo-conservation laws in cyclic-service systems," *J. Appl. Prob.* **24**: 949–964 (1987).

Bux, W., et al., "A reliable token system for local-area communication," *Proc. National Telecommunication Conf.*, New Orleans, Dec. 1983, pp. A2.2.1–A2.2.6.

Cohen, J. W., *The Single-Server Queue*, North-Holland, Amsterdam, 1969.

Cox, D. R., and W. L. Smith, *Queues*, Methuen, New York, 1961.

Farmer, W. D., and E. E. Newhall, "An experimental distributed switching system to handle bursty computer traffic," *Proc. ACM Symp. Problems in Optimization Data Communication Systems*, DP. 1-34, Pine Mountain, GA, 1969.

Fuhrmann, S. W., "A note on the M/G/1 queue with server vacations," *Oper. Research*, **32**(6): 1368–1373 (Nov.–Dec. 1984).

Fuhrmann, S. W., "Symmetric queues served in cyclic order," *Oper. Research Lett.* **4**(3): 139–144 (Oct. 1985).

Fuhrmann, S. W., and R. B. Cooper, "Stochastic decompositions in the M/G/1 queue with generalized vacations," *Oper. Research*, **33**(5): 1117–1129 (Oct. 1985).

Hafner, E. R., Z. Nenadal, and M. Tschranz, "A digital loop communication system" *IEEE Trans. in Commun*, **COM22**(6): 877–881 (June 1974).

Hayes, J. F., *Modeling and Analysis of Computer Communications Networks*, Plenum, 1984.

Hayes, J. F., and D. N. Sherman, "Traffic analysis of a ring-switched data transmission system," *Bell Syst. Tech. J.* **50**(9): 2947–2978 (Nov. 1971).

Hayes, J. F., and D. N. Sherman, "A study of data multiplexing techniques and delay performance," *Bell Syst. Tech. J.* **51**: 1985–2011 (Nov. 1972).

Jaiswal, N. K. *Priority Queues*, Academic Press, New York, 1968.

Katayama, T., et al., "System delay analysis for a non-preemptive time-limited service queueing system with two types of vacations and its application," *18th International Teletraffic Congress*, Vol. 5b, Berlin, Germany, 31 August–5 September 2003, pp. 591–600.

Kaye, A. R., "Analysis of distributed control loop for data transmission," *Proc. Symp. Computer Communications Network Teletraffic*, Polytechnic Institute of Brooklyn, New York, 1972.

Kendall, D. G., “Stochastic processes occurring in the theory of queues and their analysis by the method of the imbedded Markov chain,” *Ann. Math. Stat.*, **24**: 338–354 (1953).

Khinchin, A. Y., “Mathematical theory of stationary queues,” *Mat. Sbornik*, **39**: 73–84 (1932).

Mack, C., T. Murphy, and N. L. Webb, “The efficiency of N machines undirectionally patrolled by one operative when walking and repair times are constant,” *J. Roy. Stat. Soc., Series B* **19**: 166–172 (1957).

Mehmet Ali, M., J. F. Hayes, and A. Elhakeem, “Traffic analysis of a local area network with a star topology,” *IEEE Trans. Commun.* **35**(6): 703–712 (June 1985).

Nelson, R., *Probability, Stochastic Processes and Queueing Theory*, Springer-Verlag, 1995.

Pollaczek, F., “Uber eine Aufgab der Wahrscheinlichkeitstheorie,” *I-II Math. Zeitschrift.,* **32**: 64–100, 729–750 (1903).

Spragins, J. D., “Simple derivation of queueing formulas for loop systems,” *IEEE Trans. Commun.*, **COM-23**: 446–448 (April 1977).

Takacs, L., *Combinatorial Methods in the Theory of Stochastic Processes*, John Wiley, New York, 1967.

Takagi, H., *Analysis of Polling Systems*, MIT Press, 1986.

Welch, P. D., “On a generalized M/G/1 queueing process in which the first customer in a busy period receives exceptional service,” *Oper. Research* **12**: 736–752 (1964).

EXERCISES

6.1 Try to analyze the M/G/1 queue with a finite waiting room using the imbedded Markov chain approach. Why doesn’t it work? Repeat for the M/G/s, with an infinite waiting room $s > 1$ queue.

6.2 For the case of an M/G/1 queue with a FCFS discipline, show that the second moment of delay is given by

$$\bar{D}^2 = \bar{M}^2 + \frac{\lambda \bar{D}\bar{M}^2}{1-\rho} + \frac{\lambda \bar{M}^2}{3(1-\rho)}$$

6.3 Two Poisson streams of message traffic converge on a single queue. The average rates for the streams are given by λ_i, $i = 1, 2$. Assume that each stream has a different distribution of message lengths. Find the transform of the delay distribution for a type 1 message. (*Hint*: Queueing delay is the same for both types of messages.)

6.4 For an M/G/1 queue find the probability-generating function for the number of customers in the queue excluding the customer being served.

6.5 Consider a data communications system using the stop-and-wait protocol. The number of bits in the messages to be transmitted by the system are uniformly distributed between 500 and 1500 bits. Acknowledgment messages are 20 bits in duration. The transmission rates in both the feedforward and the feedback channels are 9600 bps. Assume that the bit error rates in the feedforward and the feedback channels are 10^{-4}, independent from bit to bit.

(a) What is the probability that both the message and its ACK are received correctly?

(b) The timeout interval after transmission is $T = 20$ ms. Assume that the round-trip delay has the following probability distribution function:

$$B(t) = \begin{cases} 0, & t \leq 10\,\text{ms} \\ 1 - e^{-200(t-0.01)}, & t > 10\,\text{ms} \end{cases}$$

What is the total time required to transmit a message on the average?

(c) Plot the average message delay as a function of message arrival rate.

6.6 Consider a packet-switching system in which messages are broken up into fixed-length packets of 1000 bits. The number of packets per message is geometrically distributed with mean value 5. The channel bit rate is 50 kbps. The probability of a packet being in error is 0.01. Each packet is transmitted according to the stop-and-wait protocol. Assume that the timeout interval is 10 ms and that the round-trip delay is a constant 10 ms. Also assume that the probability of an ACK in error is negligible.

(a) What is the average time required to transmit a packet?

(b) What is the average time required to transmit a message?

(c) At what message arrival rate will the system saturate?

(d) What is the average delay of a message at half the arrival rate found in part (c)?

6.7 Repeat Exercise 6.6 under the assumption that the packets contain 5000 bits. Assume that the underlying message lengths have the same geometric distribution in both problems. Also assume that the bit error rate is the same in both problems.

6.8 Data transmission takes place over a satellite channel. The channel is in geostationary orbit and the round-trip delay is approximately 0.25 s. Assume that the messages are a constant 1000 bits long and the channel is 50 kbps in both directions. Suppose that the probability of message error is 0.05. As in previous exercises, the stop-and-wait protocol is used and the timeout interval is equal to the round-trip delay.

(a) Show average message delay as a function of message arrival rate. Suppose that more redundancy is added to messages so that errors may be corrected as well as detected.

(b) Let us assume, for the sake of illustration, that 200 more bits in a message reduce the probability of message error to 0.01. Repeat part (a).

6.9 What is the residual life distribution for a constant interval?

6.10 Assume that the interval between buses at rush hour is exponentially distributed, with an average value of 3 min. Show that the interval in which a randomly arriving customer arrives has a mean value of 6 min and has a two-stage Erlangian distribution.

6.11 Derive an expression for the duration of a busy period that begins with j messages rather than one message as in the foregoing. This is called the j busy period.

6.12 Assume the Poisson arrival of messages at a buffer. If a message arrives to an empty buffer, a training sequence must be transmitted before the message is transmitted. This is necessary for synchronism at the receiver. If a message immediately succeeds another message, we may assume that the receiver is synchronized. Let the training sequence be T seconds in duration.

- **(a)** Find the generating function of the number of messages in the system. You should consider the training sequence as augmenting the duration of messages that are the first in a long string.
- **(b)** Plot average message delay as a function of load for the case of $T = 50$ ms and 1000 bit packets transmitted over a 4800-bps channel.

6.13 Suppose that we have a server that, on completing service to the last message in a busy period, becomes unavailable for a random period of time. In the queueing theory literature, this is known as a "server with vacations." Suppose that the vacation period has probability density $\nu(t)$.

- **(a)** Find the generating function for the number of messages awaiting the server when it returns.
- **(b)** Find the Laplace transform of the probability density of the resulting busy period, assuming that the server finds no messages and thus goes on another vacation. (*Hint*: There is a concentration at zero.)
- **(c)** Calculate the mean of this busy period.

6.14 Constant-length messages arrive at a Poisson rate to a service facility with infinite storage. The server goes on vacation when there are no messages in the system. After a message arrival to an empty system, the server waits an exponentially distributed interval before returning. It serves all messages that it finds on return plus those that arrive while it is serving. Again the server departs when the system is empty.

- **(a)** Write down an expression for the probability-generating function of the number of messages in the system when the server returns.
- **(b)** What is the average?
- **(c)** Write down an expression for the Laplace transform of the time that the server remains in the system.
- **(d)** What is the average?

6.15 Suppose that the inter–arrival time for a G/M/1 queue has an Erlang 2 distribution. Find an expression for the average delay of an arriving message.

6.16 Messages having a constant length arrive at a Poisson rate to a multiplexer with an infinite buffer. Suppose that transmission is interrupted by higher-priority messages arriving at a Poisson rate. These messages also have a constant length but are so long that only one of them can be stored at a time. Write down an expression for the average delay of the lower-priority messages.

6.17 We wish to find the delay of a data message in a system where voice and data share the same line. The arrival of messages and the initiation of calls are independent, but both occur at a Poisson rate. The duration of the voice call is exponentially distributed, and data messages have an arbitrary distribution. Voice calls have preemptive priority over data traffic. As soon as a voice call is initiated, data messages are stored in an infinite buffer. While a call is in progress, new calls are lost.

(a) Write down an expression for the Laplace transform of the probability density of message delay.

(b) What is the average delay of a message?

6.18 Consider the situation where the server breaks down only when it is in the middle of transmitting a message. This means that it is always available to begin transmission.

(a) Assuming a general breakdown distribution, find the Laplace transform of message delay.

(b) Assume the usual Poisson arrival of messages with general distributions and find the average delay of a message.

6.19 Now assume that the server never breaks down once transmission has begun, although it may not be available to a message that arrives to an empty terminal.

(a) Under the usual assumptions on message lengths and arrivals, find the Laplace transform of the delay density.

(b) Calculate the average delay.

6.20 Consider the preemptive nonresume discipline. Again assume Poisson arrivals and general service distributions for messages. What is the probability distribution of the time required to transmit a message?

6.21 In the text, it is stated that a simple argument using Little's formula can be used to prove (6.93). Do it.

6.22 A multipoint private line, such as that illustrated in Figure 6.12, serves 20 terminals. Modems at these terminals transmit at a rate of 2400 bps. For these modems the startup time in which phase and timing are recovered is 20 ms in duration. Access to the line for the terminals is controlled by a central processor by means of rollcall polling. The walktime required to access each terminal is 30 ms. This includes the 20-ms modem startup time.

Assume that the messages to be transmitted are 240 bits in duration and arrive at the terminals at a Poisson rate that is the same for all terminals.

(a) Find the average delay as a function of the message arrival rate.

(b) Assuming that the message arrival rate at each terminal is 0.5 messages per second, find the maximum number of terminals that may share the line under the requirement that the maximum message delay is 1 s.

6.23 An alternative to the 2400-bps modem considered in Exercise 6.22 is a 4800-bps modem. In order to operate at this higher speed, a longer startup time is required because of equalizer training. Assume that the walk time is 50 ms for each terminal. Repeat Exercise 6.22 for this modem, and compare the results. How would the answer to part (b) change if the messages were 480 bits long?

6.24 Suppose that the polling cycle for the system described in Exercise 6.22 is interrupted for one second in order to return messages to the stations. Repeat Exercise 6.22.

6.25 Suppose that 10 independent data sources share a line by means of statistical multiplexing. Assume that the messages are a constant 100 bits in duration and that the linespeed is 9.6 kbps. Assume also that 10 overhead bits are associated with transmission from each of the buffers associated with the data sources. (Find average delay as a function of message arrival rate under the assumption of infinite buffers.)

6.26 Consider a ring system using the token-passing technique to provide access. Suppose that there are 200 terminals sharing a coaxial cable carrying data at a rate of 5 Mbps. Further, suppose that 5-bit intervals are required to transfer the token from one terminal to another. The lengths of the messages to be transmitted are uniformly distributed between 500 and 1500 bits. The traffic is symmetric.

(a) What is the maximum allowable message arrival rate if the average cycle time is to be less than 1 ms?

(b) Calculate the mean delay for the message arrival rate found in part (a).

6.27 Consider the case where 64 stations share the same ring network using demand multiplexing (see Section 6.3.2). The line rate is 1.544 Mbps. Information packets are 512 bits in duration. Assume that the numbers of bits in a message are distributed according to $(1 - \beta)\beta^{k-1}$; $k = 1, 2, \ldots$, where $k = 0.999$. Calculate the average delay for the 1st station, the 32nd station, and the 63rd station in a data collection context.

6.28 What are the conditions for stability for exhaustive, gated, and limited services when there is uniform arrival and unlimited storage at each of N terminals in the system?

7

FLUID FLOW ANALYSIS

7.1 ON–OFF SOURCES

7.1.1 Single Source

The basic characteristic of many classes of traffic that are carried in modern telecommunications networks is burstiness. Essentially, this means the flow out of a traffic source that consists of idle periods interspersed with periods of intense activity. A simple model, that emulates this effect is the elemental ON–OFF source depicted on Figure 7.1. The source alternates between active and idle periods. In the active period, data are emitted at a constant rate R, while in the idle state, no data is generated. The sojourn times in the active and idle states are exponentially distributed with means $1/\alpha$ and $1/\beta$, respectively. The dimension of the flow can be as appropriate to the problem at hand. In keeping with our focus, we shall use cells per second. However it should be emphasized that in this model the flow is treated as a fluid where accumulations in fractions of cells are allowed. In

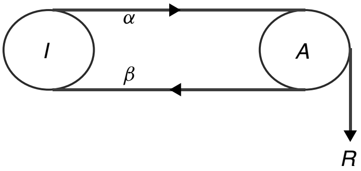


Figure 7.1 ON–OFF process.

Modeling and Analysis of Telecommunications Networks, by Jeremiah F. Hayes and
Thimma V. J. Ganesh Babu
ISBN 0-471-34845-7

applications of the technique, the quantities that are involved are such that this effect is negligible.

We let the realization of the process be denoted as $W(t)$. We can represent the process in terms of the state as

$$W(t) = RS(t) \tag{7.1}$$

where $S(t) = 0$ or 1 as the process is in the idle or active state at time t, respectively. A realization of the source is as shown in Figure 7.2.

The ON–OFF source can be viewed as the simplest realization of a birth and death process—one with a maximum population size of one and birth and death rates given, respectively, by

$$\lambda = \begin{cases} \alpha; S = 0 \\ 0; S \neq 0 \end{cases} \quad \mu = \begin{cases} \beta; S = 1 \\ 0; S \neq 1 \end{cases} \tag{7.2}$$

Substituting into (3.41) and (3.42), we find

$$P(S(t) = 1) = 1 - P(S(t) = 0) = \frac{\alpha}{\alpha + \beta} \tag{7.3}$$

Thus, the mean and the mean-square rate (at which data are produced) are, respectively

$$E(W(t)) = \frac{\alpha R}{\alpha + \beta} \tag{7.4}$$

$$E(W^2(t)) = \frac{\alpha R^2}{\alpha + \beta} \tag{7.5}$$

We are also interested in the autocovariance for the process, which we have defined in Chapter 2 to be $C_W(\tau) = E([W(t) - E(W(t))][W(t + \tau) - E(W(t + \tau))]) =$

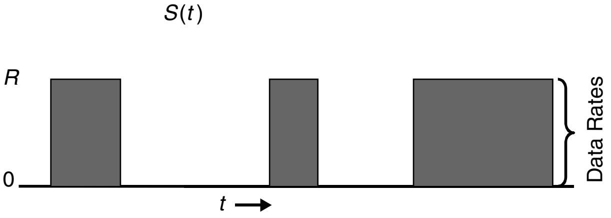


Figure 7.2 Realization of ON–OFF fluid flow.

$E(W(t)W(t+\tau)) - E^2(W(t))$. For our simple ON–OFF process, we have

$$\begin{aligned} E(W(t)W(t+\tau)) &= R^2 P(S(t+\tau)=1, S(t)=1) \\ &= R^2 P(S(t+\tau)=1/S(t)=1)P(S(t)=1) \end{aligned} \tag{7.6}$$

The conditional probability in (7.6) can be found by solving the Kolmogorov differential equation (3.22) for the state probabilities $P_i(t)$; $i = 0, 1$. In vector form, this equation can be written

$$\frac{d}{dt}\mathbf{P}(t) = M\mathbf{P}(t) \tag{7.7}$$

where $\mathbf{P}(t) = \begin{bmatrix} P_0(t) \\ P_1(t) \end{bmatrix}$ and $M = \begin{bmatrix} -\alpha & \beta \\ \alpha & -\beta \end{bmatrix}$. We shall be dealing with this sort of equation in detail later, but for now a simple development will suffice. The eigenvalues of M are $\varepsilon_0 = 0$ and $\varepsilon_1 = -(\alpha+\beta)$. The solution to (7.7) can be shown to be of the form

$$P_i(t) = K_{i0}e^{\varepsilon_0 t} + K_{i1}e^{\varepsilon_1 t} = K_{i0} + K_{i1}e^{-(\alpha+\beta)t}; \quad i = 0, 1 \tag{7.8}$$

where, in general, the constants K_{i0}, K_{i1} are determined from initial conditions. We can take a simpler route to finding these constants by calling into play the properties of autocorrelation functions. From the stationarity of the process we recognize that

$$P(S(t+\tau)=1/S(t)=1) = P_1(\tau) \tag{7.9}$$

and

$$P(S(t)=1) = \alpha/(\alpha+\beta) \tag{7.10}$$

Substituting (7.8) to (7.10) into (7.6), we have

$$E(W(t)W(t+\tau)) = K_1 + K_2 e^{-(\alpha+\beta)|\tau|}$$

K_1 and K_2 are now the constants to be determined. We use the fact that the correlation function is an even function. Note that $\lim_{\tau\to\infty} E(W(t)W(t+\tau)) = E^2(W(t))$. From (7.4) and (7.5), after simple manipulation we have

$$E(W(t)W(t+\tau)) = R^2\left[\left(\frac{\alpha}{\alpha+\beta}\right)^2 + \frac{\alpha\beta}{(\alpha+\beta)^2}e^{-(\alpha+\beta)|\tau|}\right]$$

Thus, we have

$$C_W(\tau) = E(W(t)W(t+\tau)) - E^2(W(t)) = \frac{R^2\alpha\beta}{(\alpha+\beta)^2} e^{-(\alpha+\beta)|\tau|} \tag{7.11}$$

7.1.2 Multiple Sources

More complex bursty sources are composed of N elemental sources operating independently (see Fig. 7.3). Again, the dynamics are that of a birth and death process with a maximum population size of N corresponding to all elemental sources being in the active state. We have studied such a process in Section 3.4.1. When i of the N sources are active, the aggregate birth and death rates are $\lambda_i = (N-i)\alpha$ and $\mu_i = i\beta$, respectively. In the steady state, the probability of i sources being active can be found by substitution into (3.25). Similarly to the derivation of (3.48), we have

$$P_i = \binom{N}{i}\left(\frac{\alpha}{\alpha+\beta}\right)^i \left(\frac{\beta}{\alpha+\beta}\right)^{N-i}; \quad i = 0, 1, \ldots, N \tag{7.12}$$

where P_i is the probability of i of N sources being active in the steady state. Alternatively, this can be seen by noting that we have a binomial distribution with the probability of one of N sources being active $\alpha/(\alpha+\beta)$. The average rate at which data is produced is

$$\overline{W_N} = \frac{NR\alpha}{\alpha+\beta} \tag{7.13}$$

From the properties of the binomial distribution, it follows that the variance of the rate at which data is produced is

$$\sigma^2_{W_N} = \frac{NR^2\alpha\beta}{(\alpha+\beta)^2} \tag{7.14}$$

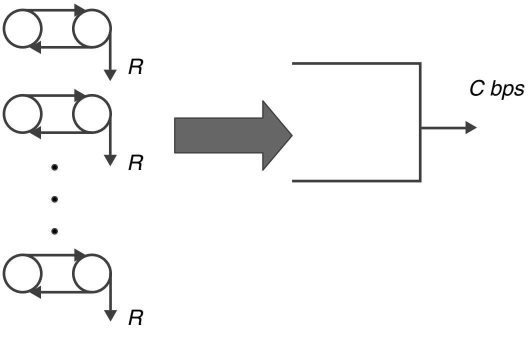


Figure 7.3 Superposition of multiple ON–OFF sources operating independently.

It is not difficult to show that the auto-covariance of N independent sources is the sum of the autocovariances of each of the sources:

$$C_{W_N}(\tau) = \frac{NR^2\alpha\beta}{(\alpha+\beta)^2}e^{-(\alpha+\beta)|\tau|} \tag{7.15}$$

The utility of the autocovariance lies in modeling a physical source of data with the fluid flow model. The mean and the autocorrelation of the physical source are estimated for real data. The parameters α, β and N are chosen so as to match the measured mean and autocorrelation, function as demonstrated in Example 7.1.

Our interest is in provisioning transmission capacity for these kinds of sources. The upper bound on the required transmission rate is NR; thus, all sources active. Now, the probability of this event could be quite small [$i = N$ in (7.12)] and over time considerable capacity would be unused if this much capacity were allocated. On the other hand, the average transmission rate would constitute a lower bound in order for the system to be stable:

$$\frac{RN\alpha}{\alpha+\beta} \leq C \tag{7.16}$$

Provisioning at a rate between the peak and the average would require buffering of cells in order to absorb and smooth temporary flows above the transmission rate. The model is depicted on Figure 7.3 where N sources feed into the buffer which is depleted at a rate C cells per second. The question is how large the buffer should be for given values of C, N, R, $1/\beta$, and $1/\alpha$ in order for the probability of buffer overflow to be acceptably small. The following analysis answers this question.

Example 7.1 The N-source binary ON–OFF model may be used to model video sources. In order to smooth the effect of scene transitions, we use the N binary sources to model M video sources, where $N \gg M \gg 1$. Measurements show that the mean and the autocovariance for the rate generated by the M sources are, respectively (Maglaris 1988)

$$\overline{W_M} = 3.9\,M\,\text{Mbps}$$

and

$$C_M(\tau) = (3.015 \times 10^{12})Me^{-3.9t}$$

The assumptions here are that there are 30 frames per second and 0.25×10^6 pixels per frame. The ratio (N/M) is a parameter of the model. The study shows that the model is sufficiently accurate when $N/M = 20$. The remaining parameters of the fluid flow model can be found by substituting into (7.4), (7.5) and (7.11). We find

$$\begin{aligned}
&\alpha + \beta = 3.9\\
&R\alpha = 3{,}900{,}000 \times M(\alpha + \beta)/N = 15{,}210{,}000 \times M/N\\
&R\beta = 3.0150 \times 10^{12} \times M(\alpha + \beta)^2/N(R\alpha) = 3{,}015{,}000 \times M/N\\
&R = (0.52M(\alpha + \beta)/N + 0.0536M(\alpha + \beta)^2/N(R\alpha))/(\alpha + \beta)\\
&\quad = 4{,}673{,}076.923 \times M/N\\
&\alpha = 15{,}210{,}000/4{,}673{,}076.923 = 3.254814815\\
&\beta = 3.9 - 3.254814815 = 0.645185185
\end{aligned}$$

The autocorrelation is plotted on the associated Excel spreadsheet.

7.2 INFINITE BUFFERS

7.2.1 The Differential Equation for Buffer Occupancy

In this section we present an analysis of buffer occupancy that was derived by Anik, Mitra, and Sondhi (Anik et al. 1982). As in the case of the BCMP, the work is so well known that it can simply be designated by the authors' initials, AMS. At time t, the state of the system composed of the buffer and the aggregate source is $(A(t), X(t))$, where $A(t)$ indicates the number of active sources and $X(t)$ is the buffer contents at time t. If there are i sources active, the buffer changes at a rate of $Ri - C$ cells per second.[1] We define the probability

$$P_i(t, x) = P[X(t) \le x, A(t) = i]; \quad i = 0, 1, \ldots, N$$

As in the case of birth and death processes, we look at the change in the state in an incremental interval δ. The changes in this interval are a birth or a death plus the increment of the buffer. We write equations analogous to (3.21):

$$\begin{aligned}
P_i(t + \delta, x) = {} & (N - i + 1)\alpha\delta P_{i-1}(t, x - \Delta_{i-1}) + (i + 1)\beta\delta P_{i+1}(t, x - \Delta_{i+1})\\
& + \{1 - [(N - i)\alpha + i\beta]\}\delta P_i(t,x - (iR - C)\delta) \qquad (7.17)
\end{aligned}$$

The salient difference is that we must account for the change in buffer level, as well as the number of active sources. For example, if $i - 1$ sources were active at time t, the buffer must increase by Δ_{i-1} in the interval δ in order for the buffer to have content x at time $t + \delta$. A similar change, Δ_{i+1}, applies if i sources were active. Since these terms vanish in the limiting operation that we perform, we do not dwell on them too much. In order to simplify equations, we set a probability to zero if the number of active sources is outside the range $0 \le i \le N$. Note that the incremental

[1] For a reason that will be apparent shortly, we assume that C is not a multiple of R so that $Ri - C$ cannot be zero.

change in the buffer when i sources are active is $(iR - C)\delta$. After rearranging terms and dividing by δ, we find

$$\begin{aligned}&\frac{P_i(t+\delta, x) - P_i(t, x-(iR-C)\delta)}{\delta}\\ &\qquad = (N-i+1)\alpha P_{i-1}(t, x-\Delta_{i-1}) + (i+1)\beta\, P_{i+1}(t, x-\Delta_{i+1})\\ &\qquad\quad - [(N-i)\alpha + i\beta]P_i(t, x-(iR-C)\delta)\end{aligned} \tag{7.18}$$

The difference from pervious analyses of birth and death processes is that we must now account for incremental changes in the buffer contents. To this end, we write the Taylor series expansion

$$P_i(t, x-(iR-C)\delta) = P_i(t, x) - (iR-C)\delta \frac{\partial P_i(t, x)}{\partial x} + o(\delta) \tag{7.19}$$

This step illustrates the fluid or continuous nature of the buffer contents. If we substitute this expansion (7.19) into (7.18) and let $\delta \to 0$, we derive the partial differential equation

$$\begin{aligned}&\frac{\partial P_i(t, x)}{\partial t} + (iR-C)\frac{\partial P_i(t, x)}{\partial x}\\ &\qquad = (N-i+1)\alpha P_{i-1}(t, x) + (i+1)\beta P_{i+1}(t, x)\\ &\qquad\quad - [(N-i)\alpha + i\beta]P_i(t, x); \quad i = 0, 1, \ldots, N\end{aligned} \tag{7.20}$$

Clearly, the new element is the dynamics of the buffer content as indicated by the differential. Since we are interested in the steady-state solution, $\partial P_i(t, x)/\partial t = 0$, and the partial differential equation of (7.20) becomes the set of ordinary differential equations

$$\begin{aligned}(iR-C)\frac{dF_i(x)}{dx} &= (N-i+1)\alpha F_{i-1}(x) + (i+1)\beta F_{i+1}(x)\\ &\quad - [(N-i)\alpha + i\beta]F_i(x); \quad i = 0, 1, \ldots, N\end{aligned} \tag{7.21}$$

where $F_i(x) = P(X \le x, A(t) = i)$; $i = 0, 1, \ldots, N$ denotes the steady-state probability distribution.

The probability is zero outside the range[2] $0 \le i \le N$. Clearly, $F_i(\infty) = P_i$; $i = 0, 1, \ldots, N$ given by (7.12). A fine point that must be mentioned in connection with (7.17) is that the equation does not hold right on the boundaries $x = 0$ and $x = B$, in the case of a finite buffer of size B.

[2] By stipulating that C cannot be a multiple of R, we ensure that (7.21) is a set of $N + 1$ differential equations; otherwise we have one algebraic equation and N differential equations that can still be solved with more needless trouble.

The matrix version of (7.21) is

$$\frac{d}{dx}\mathbf{F}(x) = D^{-1}M\mathbf{F}(x) \tag{7.22}$$

where $\mathbf{F}(x)^T = [F_0(x), F_1(x), \ldots, F_N(x)]$ (T indicating transpose) and, D^{-1} is the inverse of the diagonal *drift* matrix

$$D = \begin{bmatrix} -C & 0 & \cdots & 0 \\ 0 & R-C & \cdots & 0 \\ & & \ddots & \\ 0 & 0 & & 0 \\ 0 & 0 & \cdots & NR-C \end{bmatrix}$$

The $(N+1) \times (N+1)$ infinitesimal generator matrix M is given by

$$M = \begin{bmatrix} -N\alpha & \beta & 0 & \cdots & 0 & 0 \\ N\alpha & -(N-1)\alpha-\beta & 2\beta & \cdots & 0 & 0 \\ 0 & (N-1)\alpha & -(N-2)\alpha-2\beta & \cdots & 0 & 0 \\ 0 & 0 & (N-2)\alpha & \cdots & 0 & 0 \\ & & & & \vdots & \\ \vdots & \vdots & \vdots & \ddots & (N-1)\beta & \vdots \\ 0 & 0 & 0 & \cdots & -[\alpha+(N-1)\beta] & N\beta \\ 0 & 0 & 0 & \cdots & \alpha & -N\beta \end{bmatrix}$$

Note that M and consequently $D^{-1}M$ have tridiagonal forms and also that the sums of columns are zero. In the sequel, we shall see that this form plays a key role in obtaining a solution.

Let the right eigenvalues and eigenvectors of $D^{-1}M$ be denoted as ε_i; $i = 0, 1, \ldots, N$, and Φ_i; $i = 0, 1, \ldots, N$, respectively:

$$\varepsilon_i \Phi_i = D^{-1}M\Phi_i; \quad i = 0, 1, \ldots, N \tag{7.23}$$

The eigenvalues are column vectors having the components $\Phi_i^T = (\phi_{i0}, \phi_{i1}, \ldots, \phi_{iN})$; $i = 0, 1, \ldots, N$. In the sequel we normalize by letting $\phi_{iN} = 1$; $i = 0, 1, \ldots, N$.

From the theory of linear vector spaces we know that the solution to the differential equations (7.21)–(7.22) is given by

$$\mathbf{F}(x) = \sum_{i=0}^{N} a_i e^{\varepsilon_i x} \Phi_i \tag{7.24}$$

where the constants a_i; $i = 0, 1, \ldots, N$ are determined from the boundary conditions.[3] It may be easier to see what is going on if we rewrite (7.24) as

$$\begin{bmatrix} F_0(x) \\ F_1(x) \\ \vdots \\ F_N(x) \end{bmatrix} = \begin{bmatrix} \sum_{i=0}^{N} a_i e^{\varepsilon_i x} \phi_{i0} \\ \sum_{i=0}^{N} a_i e^{\varepsilon_i x} \phi_{i1} \\ \vdots \\ \sum_{i=0}^{N} a_i e^{\varepsilon_i x} \phi_{iN} \end{bmatrix}$$

$$= \begin{bmatrix} e^{\varepsilon_0 x}\phi_{00} & e^{\varepsilon_1 x}\phi_{10} & \cdots & e^{\varepsilon_N x}\phi_{N0} \\ e^{\varepsilon_0 x}\phi_{01} & e^{\varepsilon_1 x}\phi_{11} & \cdots & e^{\varepsilon_N x}\phi_{N1} \\ \vdots & \vdots & \ddots & \vdots \\ e^{\varepsilon_0 x}\phi_{0N} & e^{\varepsilon_1 x}\phi_{1N} & \cdots & e^{\varepsilon_N x}\phi_{NN} \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ a_N \end{bmatrix} \tag{7.25}$$

Our task now is to find the various quantities in (7.24)–(7.25). We deal with each in sequence: the eigenvalues, ε_i; $i = 0, 1, \ldots, N$, the eigenvectors Φ_i; $i = 0, 1, \ldots, N$ and the coefficients, a_i; $i = 0, 1, \ldots, N$. The first two are determined by the Matrices M and D, while the third is a function of the boundary conditions.

7.2.2 Derivation of Eigenvalues

The properties of the problem are such that an ingenious non-numerical technique for finding the eigenvalues in (7.23) can be used. The first step is writing (7.23) as

$$\begin{aligned} \varepsilon_i(jR - C)\phi_{ij} &= (N - j + 1)\alpha\phi_{ij-1} + (j + 1)\beta\phi_{ij+1} \\ &\quad + [(N - j)\alpha + j\beta]\phi_{ij}; \quad j = 0, 1, \ldots, N \end{aligned} \tag{7.26}$$

We define the generating functions $\Phi_i(z) = \sum_{j=0}^{N} \phi_{ij} z^j$; $i = 0, 1, 2, \ldots, N$. Multiplying both sides of (7.26) by z^j and summing over j, we find after some algebra that

$$\frac{\Phi_i'(z)}{\Phi_i(z)} = \frac{C\varepsilon_i + N\alpha z - N\alpha}{\alpha z^2 + (\varepsilon_i R + \beta - \alpha)z - \beta}; \quad i = 0, 1, \ldots, N \tag{7.27}$$

where the derivative obeys $z\Phi_i'(z) = \sum_{j=1}^{N} j z^j \phi_{ij}$. Since α and β are positive and real, the quadratic equation in z in the denominator of the RHS of (7.27) has the

[3] This relation can be verified by simple substitution. For background on linear vector spaces, see Belman (1962).

distinct real roots

$$r_{i1} = \frac{-(\varepsilon_i R + \beta - \alpha) + \sqrt{(\varepsilon_i R + \beta - \alpha)^2 + 4\alpha\beta}}{2\alpha}; \quad i = 0, 1, \ldots, N$$
$$r_{i2} = \frac{-(\varepsilon_i R + \beta - \alpha) - \sqrt{(\varepsilon_i R + \beta - \alpha)^2 + 4\alpha\beta}}{2\alpha}; \quad i = 0, 1, \ldots, N \tag{7.28}$$

A partial fraction expansion of (7.27) gives

$$\frac{\Phi_i'(z)}{\Phi_i(z)} = \frac{\gamma_{i1}}{z - r_{i1}} + \frac{\gamma_{i2}}{z - r_{i2}}; \quad i = 0, 1, \ldots, N \tag{7.29}$$

with the residues

$$\gamma_{i1} = \frac{\varepsilon_i C + N\alpha r_{i1} - N\alpha}{\alpha(r_{i1} - r_{i2})}; \quad i = 0, 1, \ldots, N$$
$$\gamma_{i2} = N - \gamma_{i1}; \quad i = 0, 1, \ldots, N \tag{7.30}$$

Since r_{i1} and r_{i2} are distinct, γ_{i1} and γ_{i2} are bounded. By differentiation and simple substitution, it can be verified that the solution to (7.29) is

$$\Phi_i(z) = (z - r_{i1})^{\gamma_{i1}}(z - r_{i2})^{N-\gamma_{i1}}; \quad i = 0, 1, \ldots, N \tag{7.31}$$

By the definition of the generating function it is a polynomial in z of degree N; accordingly, γ_{i1} must be an integer in the range $0 \leq i \leq N$.

Up to this point the ordering of the eigenvalues was arbitrary, so we have chosen one that simplifies the notation. The final step in the derivation of the eigenvalues requires substitution into (7.31) i.e., $\gamma_{i1} = k$ and r_{i1}, r_{i2} from (7.28). We gather all terms involving the square root to one side of the equation to obtain $(i - N/2)\sqrt{(\varepsilon_i C + \beta - \alpha)^2 + 4\alpha\beta} = \varepsilon_i C - N\alpha - N(\varepsilon_i R + \beta - \alpha)/2$. We then square both sides in order to write quadratic equations in the eigenvalues

$$A(k)\varepsilon_k^2 + B(k)\varepsilon_k + C(k) = 0; \quad k = 0, 1, 2, \ldots, N$$

where

$$A(k) = R^2\left(\frac{k-N}{2}\right)^2 - \left(\frac{C-RN}{2}\right)^2$$
$$B(k) = 2R(\beta - \alpha)\left(\frac{k-N}{2}\right)^2 + N(\alpha + \beta)\left(\frac{C-RN}{2}\right) \tag{7.32}$$
$$C(k) = (\alpha + \beta)^2\left[\left(\frac{k-N}{2}\right)^2 - \left(\frac{N}{2}\right)^2\right]$$

At this point the eigenvalues can be found simply by substituting into the quadratic equation

$$\varepsilon_k = \frac{-B(k) \pm \sqrt{B(k)^2 - 4A(k)C(k)}}{2A(k)}; \quad k = 0, 1, \ldots, N \tag{7.33}$$

From (7.33), it would seem that there are $2(N+1)$, eigenvalues; however, studying the forms of the coefficients, $A(k)$, $B(k)$ and $C(k)$, shows that values are repeated. Note that the coefficients depend on the index only through the term $(k - N/2)^2$; consequently, there is symmetry about the midpoint. For example, for $N = 9$, $k = 2$ and 7 have the same coefficients for the quadratic equation. Also, we note that $C(k) \leq 0$; accordingly, with this starting point we can study the characteristics of the eigenvalues.

When $k = 0$ or N, $C(k) = 0$ and the quadratic equation has two roots, which we designate as follows:

$$\begin{aligned} \varepsilon_0 &= \frac{N(C(\alpha+\beta) - \alpha RN)}{C(C - RN)} \\ \varepsilon_N &= 0 \end{aligned} \tag{7.34}$$

Note that the numerator and the denominator of ε_0 relate directly to the flow into the buffer. Indeed, flow considerations mandate that this eigenvalue be negative. From (7.16), we see that the numerator must be positive if the system is stable. The denominator must be negative; otherwise, $C > NR$ and the buffer will *never* overflow. This represents a system that is not of interest since it is overprovisioned. Finally, it can be shown that ε_1 is the largest of the negative eigenvalues. This will be demonstrated in an example.

A detailed study of the roots of the quadratic equation shows that there are $N - [C/R]^-$ negative roots and $[C/R]^-$ positive roots, where $[x]$ indicates the largest integer less than x. Also, ε_0 is the largest negative eigenvalue. With the 0 root, we then have exactly $N + 1$ roots corresponding to the $N + 1$ eigenvalues. We demonstrate these results by example as well.

We now apply what we have learned about the eigenvalues to (7.24)–(7.25). As we have seen, one of the eigenvalues is zero, giving a constant term in x in (7.24). Since this is the only term that does not vanish as $x \to 0$, we designate it as $\mathbf{F}(\infty)$. As we have noted in connection with (7.21) that the components of the vector $\mathbf{F}(\infty)$ are P_i; $i = 0, 1, \ldots, N$, that is, the probability that, in the steady state, i sources are active. These values are given by (7.12).

The first case that we consider is that of an infinite buffer. In order for the solution to remain bounded as the buffer size grows, as befits a probability, the coefficients a_i must be zero for the $[C/R]^-$ positive eigenvalues. Since there are $N - [C/R]^- - 1$

negative roots, we have

$$\mathbf{F}(x) = \mathbf{F}(\infty) + \sum_{i=1}^{N-[C/R]^{-}-1} a_i e^{\varepsilon_i x} \varPhi_i \qquad (7.35)$$

Example 7.2 We find the eigenvalues for the following set of parameters: $\alpha = 2$, $\beta = 3$, $N = 5$, $R = 4$, and $C = 10$. The average and the peak flows are 8 and 20, respectively. From (7.12), the steady-state probabilities are found on the associated Excel spreadsheet to be

$$P_0 = 0.0778,\ P_1 = 0.2592,\ P_2 = 0.3456,\ P_3 = 0.2304,$$
$$P_4 = 0.0768,\ P_5 = 0.01024$$

or

$$\mathbf{F}(\infty) = \begin{bmatrix} 0.0778 \\ 0.2592 \\ 0.3456 \\ 0.2304 \\ 0.0768 \\ 0.0102 \end{bmatrix}$$

Substituting into (7.32) and (7.33), we find that the six eigenvalues are $\varepsilon_0 = -0.5$, $\varepsilon_1 = -1.93531$, $\varepsilon_2 = -6.3788$, $\varepsilon_3 = 5.8788$, $\varepsilon_4 = 1.4533$, $\varepsilon_5 = 0$. Note that, as predicted, $N - [C/R]^{-} = 5 - 2 = 3$ of these are negative. Substituting into (7.24), we have

$$\mathbf{F}(x) = (0.7776, 0.2592, 0.3456, 0.2304, 0.0768, 0.01024) + a_0 e^{-0.5x} \varPhi_0$$
$$+ a_1 e^{-1.93531x} \varPhi_1 + a_2 e^{-6.3788x} \varPhi_2 + a_3 e^{1.4533x} \varPhi_3 + a_4 e^{5.8788x} \varPhi_4 + a_5 \varPhi_5$$

In Example 7.3, we will find $\varPhi_0, \ldots, \varPhi_5$ and in Example (7.4), and in $a_0, \ldots, a_5$. We have also worked out the eigenvalues on the associated Matlab program. This will be needed later.

7.2.3 Derivation of the Eigenvectors

In this subsection, we find the remaining quantities in (7.24). By the definition of the generating function the polynomial coefficients give the coefficients. From (7.31)

$$\Phi_i(z) = \sum_{j=0}^{N} \phi_{ij} z^j = (z - r_{i1})^{\gamma_{i1}} (z - r_{i2})^{N-\gamma_{i1}}$$

$$= \sum_{m=0}^{\gamma_{i1}} \binom{\gamma_{i1}}{m} z^m (-r_{i1})^{\gamma_{i1}-m} \sum_{n=0}^{N-\gamma_{i1}} \binom{N-\gamma_{i1}}{n} z^n (-r_{i2})^{N-\gamma_{i1}-n};$$

$$i = 0, 1, \ldots, N$$

After a change of variables in the summation over n, $j = m + n$, we have

$$\Phi_i(z) = \sum_{m=0}^{\gamma_{i1}} \binom{\gamma_{i1}}{m} (-r_{i1})^{\gamma_{i1}-m}$$
$$\times \sum_{j=m}^{N-\gamma_{i1}+m} \binom{N-\gamma_{i1}}{j-m} (-r_{i2})^{N-\gamma_{i1}-j+m} z^j; \quad i = 0, 1, \ldots, N$$

The next step is to change the order of summation. The sketch in Figure 7.4 may help to visualize this operation. (Keep in mind that γ_{i1} is an integer.)

The terms in the summation lie within the parallelogram (0, 0), (0, $N - \gamma_{i1}$), (N, γ_{i1}), (γ_{i1}, γ_{i1}). After this step, the components of the eigenvalues are displayed as coefficients of powers of z:

$$\Phi_i(z) = \sum_{j=0}^{N} \phi_{ij} z^j = \sum_{j=0}^{N} \left[(-1)^{N-j} \sum_{m=0}^{\gamma_{i1}} \binom{\gamma_{i1}}{m} \binom{N-\gamma_{i1}}{j-m} \right.$$
$$\left. \times (r_{i1})^{\gamma_{i1}-m} (r_{i2})^{N-j-\gamma_{i1}+m} \right] z^j; \quad i = 0, 1, \ldots, N \tag{7.36}$$

In writing (7.36), we rely on the fact that the terms $\binom{N-\gamma_{i1}}{j-m} = 0$ within the triangle (0, $N - \gamma_{i1}$), (0, N), (γ_{i1}, N) and $\binom{\gamma_{i1}}{m} = 0$ within the triangle (γ_{i1}, γ_{i1}), (γ_{i1}, N), (N, N).

We have in hand all the quantities necessary for the evaluation of the coefficients ϕ_{ij} by straightforward substitution in (7.36); however, there are further simplifications. Recall that $k = 0$ in (7.32)–(7.33) yielded two eigenvalues that we called ε_0 and ε_1, respectively. The eigenvectors corresponding to each of these

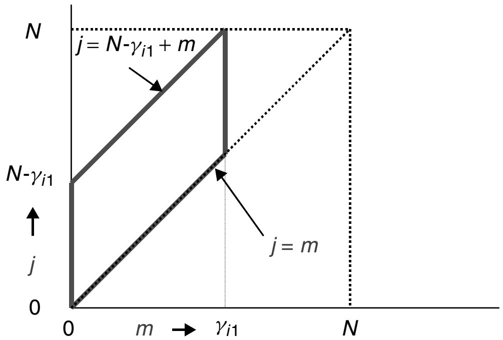


Figure 7.4 Changing the order of summation.

eigenvalues have particularly simple forms. We start with $\varPhi_N$. From (7.28), we have $r_{N1} = 1$ and $r_{N2} = -\beta/\alpha$ for $\varepsilon_N = 0$. Further, note that $\binom{i}{m} = 0$ for $m > i$. Substituting into (7.36), we find

$$\varPhi_N = \left[\left(\frac{\beta}{\alpha}\right)^N, N\left(\frac{\beta}{\alpha}\right)^{N-1}, \ldots, \binom{N}{j}\left(\frac{\beta}{\alpha}\right)^{N-j}, \ldots, 1 \right] \tag{7.37}$$

A second simple result is found by substitution of the eigenvalue ε_0 given by (7.34) into (7.28). We find

$$\begin{aligned} r_{01} &= 1 - \frac{NR}{C} \\ r_{02} &= \frac{\beta}{\alpha(NR/C - 1)} \end{aligned} \tag{7.38}$$

Further, it can be found by straight substitution into (7.30) that $\gamma_{01} = 0$. Thus, from (7.31) and (7.38), we find

$$\varPhi_0(z) = \left(z - 1 + \frac{NR}{C}\right)^N \tag{7.39}$$

From a straightforward application of the binomial theorem, we find that

$$\varPhi_0 = \left[\left(\frac{NR}{C} - 1\right)^N, N\left(\frac{NR}{C} - 1\right)^{N-1}, \ldots, \binom{N}{j}\left(\frac{NR}{C} - 1\right)^{N-j}, \ldots, 1 \right]^T \tag{7.40}$$

Finally, a quantity that we shall use later is the sum of the coefficients of the eigenvector. From (7.39) and the definition of the generating function, we have

$$\sum_{n=0}^{N} \phi_{1n} = \Phi_1(z)|_{z=1} = \left(\frac{NR}{C}\right)^N \tag{7.41}$$

We will return to these quantities presently.

Example 7.3 We continue Example 7.2. Substituting into (7.36), (7.37), and (7.40), we obtain the following results for the eigenvectors in the associated

Matlab program:

$$\Phi_0 = \begin{bmatrix} 1.0 \\ 5.0 \\ 10.0 \\ 10.0 \\ 5.0 \\ 1.0 \end{bmatrix}, \quad \Phi_1 = \begin{bmatrix} -0.095 \\ -0.925 \\ -3.375 \\ -0.505 \\ -2.176 \\ 1.0 \end{bmatrix}, \quad \Phi_2 = \begin{bmatrix} 0.273 \\ 6.706 \\ 54.62 \\ 144.28 \\ -24.39 \\ 1.0 \end{bmatrix},$$

$$\Phi_3 = \begin{bmatrix} 27.85 \\ -452.96 \\ 1786.0 \\ 450.72 \\ 36.89 \\ 1.0 \end{bmatrix}, \quad \Phi_4 = \begin{bmatrix} -80.29 \\ 116.50 \\ 180.18 \\ 79.22 \\ 14.68 \\ 1.0 \end{bmatrix}, \quad \Phi_5 = \begin{bmatrix} 7.59 \\ 25.31 \\ 33.75 \\ 22.5 \\ 7.5 \\ 1.0 \end{bmatrix}$$

7.2.4 Derivation of Coefficients

At this point, we have found the eigenvalues and the eigenvectors, which are necessary for the solution in (7.24). It remains to find the coefficients $\{a_i\}$. This will be done by matching the functions $F_i(x)$ and their derivatives at the boundary $x = 0$. Recall the definition $F_i(x) = P(X \leq x, i$ sources active); accordingly, if the aggregate source is generating data at a rate greater than the output line capacity, the buffer cannot be empty. We have

$$F_i(0) = 0; \quad \left[\frac{C}{R}\right]^- + 1 \leq i \leq N \tag{7.42}$$

Thus, the column vector $\mathbf{F}(0)$ is 0 in places $[C/R]^- + 1 \leq i \leq N$.

Now consider (7.22), which shows the multiplication of $\mathbf{F}(0)$ by the matrix $D^{-1}M$ is equivalent to differentiation. Suppose that successively higher derivatives are taken by successive multiplications of $\mathbf{F}(0)$. On each iteration the particular tridiagonal form of $D^{-1}M$ causes the number of zero places to decrease by one. Therefore, the column vector $(D^{-1}M)^j\mathbf{F}(0)$ has zeros only in places $[C/R]^- + 1 + j \leq i \leq N$. Now, for the last place we may write

$$(D^{-1}M)^j F_N(x)\Big|_{x=0} = 0; \quad j = 0, 1, \ldots, N - \frac{C}{R} - 1 \tag{7.43}$$

With one more iteration, all the places would be nonzero.

Next, considering this Nth component from another point of view from which it is clear that the coefficient of z^N is always 1; consequently, $\phi_{iN} = 1$; $i = 0, 1, \ldots, N$;

accordingly, from (7.24), we have

$$F_N(x) = \left(\frac{\alpha}{\alpha+\beta}\right)^N + \sum_{i=0}^{N-[C/R]^- -1} a_i e^{\varepsilon_i x} \tag{7.44}$$

Recall that only terms with negative eigenvalues are considered in the summation.

Now, having seen the effect of repeated multiplications by $D^{-1}M$, we look at repeated differentiation in (7.22):

$$(D^{-1}M)^j F_N(x) = \sum_{i=0}^{N-[C/R]^- -1} a_i \varepsilon_i^j e^{\varepsilon_i x}; \quad j = 1, 2, \ldots, N - \frac{C}{R} - 1 \tag{7.45}$$

From (7.43)–(7.45), we have the set of simultaneous equations

$$\sum_{i=0}^{N-[C/R]^- -1} a_i \varepsilon_i^j = -\left(\frac{\alpha}{\alpha+\beta}\right)^N \delta_{0j}; \quad j = 0, 1, \ldots, N - \left[\frac{C}{R}\right]^- -1$$

where $\delta_{0j} = 1$ for $j = 0$ and is zero otherwise. In matrix form, this can be written

$$V\mathbf{a} = -\left[\frac{\alpha}{\alpha+\beta}\right]^N \mathbf{e} \tag{7.46}$$

where

$$V = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ \varepsilon_0 & \varepsilon_1 & \cdots & \varepsilon_{N-[C/R]^- -1} \\ {\varepsilon_0}^2 & {\varepsilon_1}^2 & & \varepsilon^2_{N-[C/R]^- -1} \\ \vdots & & & \\ \varepsilon_0^{N-[C/R]^- -1} & \varepsilon_1^{N-[C/R]^- -1} & \cdots & \varepsilon_{N-[C/R]^- -1}^{N-[C/R]^- -1} \end{bmatrix}$$

$$\mathbf{a}^T = (a_0, a_1, a_2, \ldots, a_{N-[C/R]^- -1})$$

and

$$\mathbf{e}^T = (1, 0, 0, \ldots, 0)$$

V is a Vandermonde matrix which is easily shown to have the solution

$$a_j = -\left(\frac{\alpha}{\alpha+\beta}\right)^N \prod_{\substack{i=0 \\ i \neq j}}^{N-[C/R]-1} \frac{\varepsilon_i}{\varepsilon_i - \varepsilon_j}; \quad j = 0, 1, \ldots, N - \frac{C}{R} - 1 \tag{7.47}$$

A good way to use these results is the *survivor function*, which is the probability of the buffer contents being greater than a certain level. This is often used as an approximation for the probability of overflow for a finite buffer:

$$G(x) = P(\text{buffer contents} \geq x) = 1 - \mathbf{U}^T \mathbf{F}(x) \tag{7.48}$$

where $\mathbf{U}^T = (1, 1, \ldots, 1)$. Since $\mathbf{U}^T \mathbf{F}(\infty) = \sum_{i=0}^{N} P_i = 1$, we have from (7.35)

$$G(x) = - \sum_{i=1}^{N-[C/R]^{-}-1} a_i e^{\varepsilon_i x} (\mathbf{U}^T \Phi_i) \tag{7.49}$$

For sufficiently large values of x, the only term in the survivor function that is significant is that corresponding to ε_0, the largest of the negative eigenvalues. In this case, the survivor function of (7.49) can be approximated:

$$G(x) \cong -a_0 e^{\varepsilon_0 x} (\mathbf{U}^T \Phi_0)$$

The term in parentheses is just the sum of the components of Φ_0, which is found from (7.41). Finally, we have

$$G(x) \cong a_0 e^{\varepsilon_0 x} \left(\frac{NR}{C} \right)^N \tag{7.50}$$

The coefficient a_0 is given by (7.47). Substituting into (7.50), we have

$$G(x) = e^{\varepsilon_0 x} \left(\frac{\alpha NR}{(\alpha + \beta) C} \right)^N \prod_{i=1}^{N-[C/R]^{-}-1} \frac{\varepsilon_i}{\varepsilon_i + \varepsilon_0} \tag{7.51}$$

Note that the term in brackets in (7.51) is the load in the system: $\rho = [\alpha NR/(\alpha + \beta) C]$.

Example 7.4 We continue with the same set of parameters as in the previous examples: $\alpha = 2$, $\beta = 3$, $N = 5$, $R = 4$, and $C = 10$. The constant values in (7.24) and (7.25) as found from (7.47) on the associated Matlab program are given by $a_0 = -0.0150$, $a_1 = 0.0051$, $a_2 = -0.000379$. The final solution is then

$$\mathbf{F}(x) = \mathbf{F}(\infty) - 0.00150 e^{-0.5x} \Phi_0 + 0.0051 e^{-1.9353x} \Phi_1 - 0.000379 e^{-6.3788x} \Phi_2$$

We continue with Examples 7.2 and 7.3. Conditioning on two active sources, we have the following equation for the transient solution.

$$\mathbf{F}(x) = \begin{bmatrix} 0.0778 \\ 0.2592 \\ 0.3456 \\ 0.2304 \\ 0.0768 \\ 0.0102 \end{bmatrix} - 0.00150e^{-0.5x} \begin{bmatrix} 1.0 \\ 5.0 \\ 10.0 \\ 10.0 \\ 5.0 \\ 1.0 \end{bmatrix} + 0.0051e^{-1.9353x} \begin{bmatrix} -0.0946 \\ -0.925 \\ -3.375 \\ -5.049 \\ -2.176 \\ 1.0 \end{bmatrix}$$

$$- 0.000379e^{-6.3788x} \begin{bmatrix} 0.273 \\ 6.706 \\ 54.62 \\ 144.28 \\ -24.39 \\ 1.0 \end{bmatrix}$$

Note that the first term here is the steady-state solution. The survivor function is given by

$$G(x) = 0.4794e^{-0.5x} + 0.0542e^{-1.9353x} + 0.0692e^{-6.3788x}$$

It is interesting to compare with the asymptotic term $G(x) \cong 0.4794e^{-0.5x}$, which is obtained by considering only the largest eigenvalue. Both are plotted on the linked Excel spreadsheet. As we see, the asymptotic expression is valid over much of the range of levels. It is certainly a good approximation for buffer overflow.

7.3 FINITE BUFFERS

The salient approximation in the derivation of the previous section is that of an infinite buffer. This approximation is quite useful when the probability of buffer overflow is small. However, there are many systems in which buffer overflow is a significant factor. (One such system is the "leaky bucket," which we study in Section 7.5.) In this case, we must solve all the eigenvalues, positive as well as negative, and eigenvectors of (7.25). As a consequence, it is necessary to find $N + 1$ coefficients, $a_0, a_1, \ldots, a_N$. Note that since the exponential terms do not vanish, we cannot substitute for $\mathbf{F}(\infty)$.

The properties of the infinite buffer case that allowed the simple solution of the previous section do not hold, and we must use numerical techniques to solve equations derived from boundary conditions. From the appropriate boundary conditions, we can find specific values for the elements on the LHS of (7.25). The matrix can then be inverted to find the $\{a_i\}$.

The first set of boundary conditions is the same as that considered in the previous section [see (7.42)]. When the number of active sources is such that $([C/R]^- + 1) \leq i \leq N$, the buffer cannot be empty; consequently, $F_i(0) = 0$. Similarly, when $0 \leq i \leq [C/R]^-$, the buffer cannot be full; that is, there is no probability mass on the boundary, and we can determine the cumulative distribution

$F_i(B) = P_i$ as given in (7.12). Substituting into (7.25), we find

$$\begin{bmatrix} P_0 \\ P_1 \\ \vdots \\ P_{[C/R]^-} \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} = \begin{bmatrix} e^{\varepsilon_0 B}\phi_{00} & e^{\varepsilon_1 B}\phi_{10} & \cdots & e^{\varepsilon_N B}\phi_{N0} \\ \vdots & \vdots & \cdots & \vdots \\ e^{\varepsilon_0 B}\phi_{0[C/R]^-} & e^{\varepsilon_1 B}\phi_{1[C/R]^-} & \ddots & e^{\varepsilon_N B}\phi_{N[C/R]^-} \\ \phi_{0[C/R]^-+1} & \phi_{1[C/R]^-+1} & \cdots & \phi_{N[C/R]^-+1} \\ \phi_{0[C/R]^-+2} & \phi_{1[C/R]^-+2} & \cdots & \phi_{N[C/R]^-+2} \\ \vdots & \vdots & \ddots & \vdots \\ \phi_{0N} & \phi_{1N} & \cdots & \phi_{NN} \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ a_N \end{bmatrix} \tag{7.52}$$

There is a potential numerical problem in solving (7.52) when the buffer B is relatively large. Terms of the form $e^{\varepsilon_i B}$ can cause overflow when the eigenvalue ε_i is positive. One remedy is to solve for the terms $e^{\varepsilon_i B}a_i$ in this case with a suitable rearrangement of terms in (7.52) (Tucker 1988). This term is all that is needed, anyway. It is easiest to illustrate this by the example below.

The survivor function is easily found once these steps have been carried out. From the definition of (7.48), we have

$$G(x) = 1 - \mathbf{U}^T\mathbf{F}(x) = 1 - \sum_{j=0}^{N} F_j(x) = 1 - \sum_{j=0}^{N}\sum_{i=0}^{N} a_i e^{\varepsilon_i x}\phi_{ji} \tag{7.53}$$

Example 7.5 We continue the numerical example considered in the previous example: $\alpha = 2$, $\beta = 3$, $N = 5$, $R = 4$, and $C = 10$. As we have seen in Example 7.2, the eigenvalues are $\varepsilon_0 = -0.5$, $\varepsilon_1 = -1.93531$, $\varepsilon_2 = -6.3788$, $\varepsilon_3 = 5.8788$, $\varepsilon_4 = 1.4533$, $\varepsilon_5 = 0$. Since ε_3 and ε_4 are relatively large, we carry out the rearrangement alluded to above. The desired constant values, $a_1, a_2, \ldots, a_5$, are found by solving the following equation. (See the associated Matlab program.):

$$\begin{bmatrix} 0.0778 \\ 0.2592 \\ 0.3456 \\ 0 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} e^{\varepsilon_0 B} & -0.095e^{\varepsilon_1 B} & 0.273e^{\varepsilon_2 B} & 27.9 & -80.29 & 7.59 \\ 5.0e^{\varepsilon_0 B} & -0.925e^{\varepsilon_1 B} & 6.706e^{\varepsilon_2 B} & -453.0 & 116.50 & 25.31 \\ 10.0e^{\varepsilon_0 B} & -3.375e^{\varepsilon_1 B} & 54.616e^{\varepsilon_2 B} & 1786.0 & 180.18 & 33.75 \\ 10.0 & -5.049 & 114.279 & 450.7e^{-\varepsilon_3 B} & 79.22e^{-\varepsilon_4 B} & 22.5 \\ 5.0 & -2.176 & -24.394 & 36.9e^{-\varepsilon_3 B} & 14.68e^{-\varepsilon_4 B} & 7.5 \\ 1.0 & 1.0 & 1.0 & e^{-\varepsilon_3 B} & e^{-\varepsilon_4 B} & 1.0 \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ a_2 \\ a_3 e^{\varepsilon_3 B} \\ a_4 e^{\varepsilon_4 B} \\ a_5 \end{bmatrix} \tag{7.54}$$

Results of the calculation are shown in Table 7.1 for the buffer sizes of $B = 20$ and $B = 50$.

We can compute the survivor function using Equation (7.54) since we have all the parameters we need.

Table 7.1 Calculations for Finite Buffers

Coefficients	For $B = 20$	For $B = 50$
a_0	$-1.498159088558 \times 10^{-2}$	$-1.498148124350033 \times 10^{-2}$
a_1	$5.120837219566 \times 10^{-3}$	$5.120799742955597 \times 10^{-3}$
a_2	$-3.793212755234 \times 10^{-4}$	$-3.793184994781926 \times 10^{-4}$
a_3	$-4.9787183984357 \times 10^{-62}$	$-3.877193425961592 \times 10^{-145}$
a_4	$8.2854057248979 \times 10^{-21}$	$5.051722668275094 \times 10^{-46}$
a_5	$1.0240074941538 \times 10^{-2}$	$1.024000000002292 \times 10^{-2}$

7.4 MORE GENERAL SOURCES

The analyses of the previous sections are predicated on a birth–death model of the data source. In this section, we introduce a more general model where the underlying Markov chain has a general form in which transition may be from any state to any other state. (Recall that for the birth and death process, transitions are to adjacent states only.) In the more general formulation, the rate generated by a source is governed by one of $N + 1$ underlying state S_i; $i = 0, 1, \ldots, N$. We write $R_i = R(S_i)$; $i = 0, 1, \ldots, N$. The drift matrix D then becomes

$$D = \text{diag}\{R_0, R_1, \ldots, R_N\} \tag{7.55}$$

The transition matrix M in (7.22) has the elements $\{p_{ij}\}$, where

$$p_{ij} = P(\text{transition from state } i \text{ to } j \text{ in } (t, t + \delta)) \tag{7.56}$$

The solution is still given by (7.24); however, it is no longer possible to use the analysis of the preceding sections to find the eigenvalues and eigenvectors in closed form; we must now use numerical techniques to find a solution to the general case. Of course, this approach brings its own problem for large systems. There are models where the task can be simplified. Suppose that the sources can be described as the sum of independent subsources, each with possibly different matrices D and M. A technique based on separability (Stern and Elwaid 1991) can be used to reduce the computational complexity from the order of $\left(\prod N^{(k)}\right)^3$ operations to $\sum (N^{(k)})^3$—a considerable saving in large systems. The analysis can also be extended to systems, which handle real-time as well as jitter-tolerant traffic.

7.5 ANALYSIS: LEAKY BUCKET

The "leaky bucket" (LB) is a means for controlling and smoothing the flow of data into an ATM network. The common form for this system is shown in Figure 7.5. The key component of the system is the token generator, which supplies tokens to a

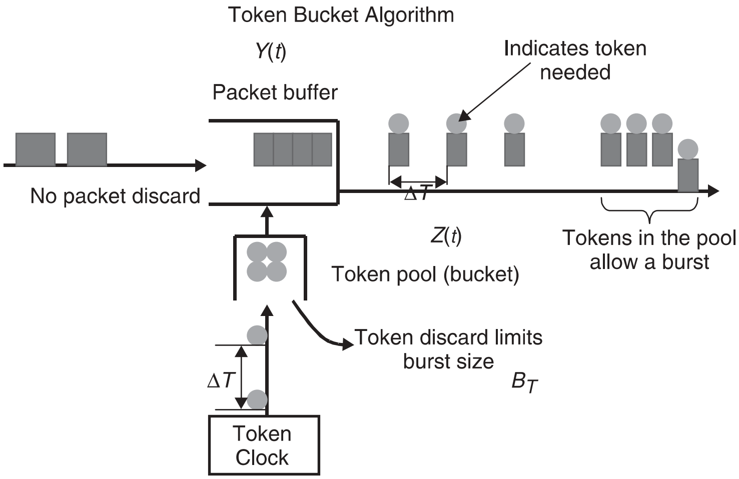


Figure 7.5 Representation of the leaky bucket.

pool at a rate of R_T per second. Data cells arrive at a rate of λ cells per second. In order for a cell to be transmitted to the network, it must be matched with a token obtained from the pool. If no token is available, cells are stored in the cell buffer. The size of the data buffer and the token pool, designated as B_D and B_T, respectively, are key parameters, along with the token generation rate. In the simplest implementation, if a cell or a token arrives to a full storage facility, either buffer or pool, it is lost. An alternative is to mark cells that overflow the cell buffer and to send them, paired with a token. If further congestion is encountered in the ATM network, these marked cells are dropped. The rate at which cells that have been paired with tokens are transmitted to the ATM network is greater than either the peak cell arrival rate or the token generation rate. Were it otherwise, there would be no need for a leaky bucket.

Cells arriving to a nonempty pool are transmitted at their arrival rate until the pool is depleted, at which time they are transmitted at a generation rate of R_T tokens per second. Tokens arriving in excess of this rate are stored and fed out of the buffer at the token generation rate. The maximum burst size is limited to the size of the buffer pool. Longer input bursts are smoothed to the token generation rate. To prevent the buffer from overflowing continually, we must have R_T greater than the average cell arrival rate. Presumably, the token generation rate is less than the peak cell arrival rate.

Let $Y(t)$ and $Z(t)$, respectively, represent the buffer and the pool contents at time t. An interesting feature of the LB operation is that both the data buffer and the token cannot be nonempty at the same time:

$$Y(t)\bullet Z(t) \equiv 0 \tag{7.57}$$

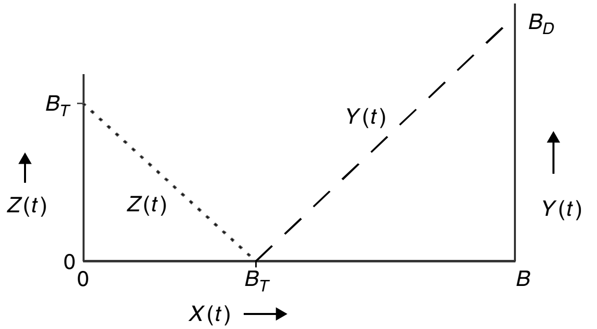


Figure 7.6 Relation among data buffer, $Y(t)$, token pool, $Z(t)$, and virtual buffer.

This is the result of the fact that the line rate is greater than either of the input rates, cell or token. If there were a malfunction resulting in $Y(T) > 0$ and $Z(t) > 0$ occurring simultaneously, both would deplete until one or both were empty.

The leaky bucket can be analyzed using the fluid flow model with a finite buffer which was treated in the immediately preceding section. The relation (7.57) is key to this application. Defining the contents of a *virtual* buffer as $X(t) = Y(t) - Z(t) + B_T$, note that $0 \leq X(t) \leq B_T + B_D = B$ and that both $Y(t)$ and $Z(t)$ can be determined from the value of $X(t)$. A diagram showing the relation among $X(t)$, $Y(t)$, and $Z(t)$ is given in Figure 7.6. The drift of $X(t)$ is given by $d = dX(t)/dt = dY(t)/dt - dZ(t)/dt$. If the data

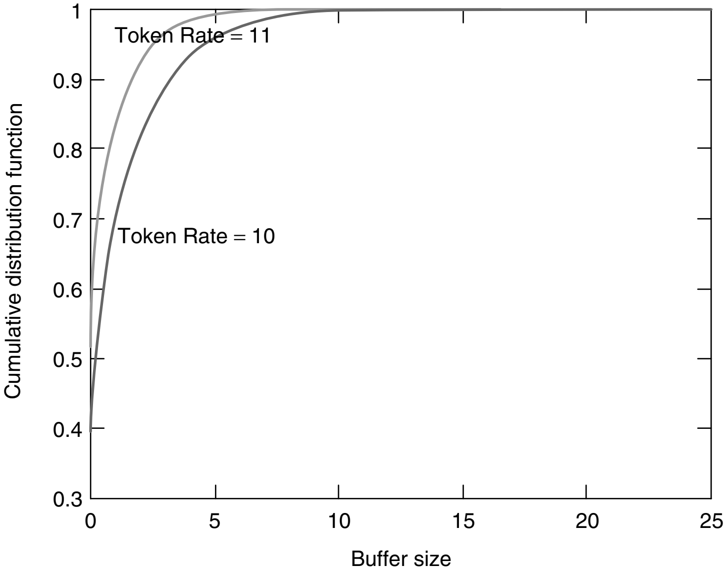


Figure 7.7 Cumulative distribution function versus buffer size.

source consists of N independent ON-OFF sources as in Figure 7.3, then $dY(t)/dt = iR$; $i = 0, 1, \ldots, N$. Since the token generation rate is a constant, R_T, we have

$$d_{ii} = iR - R_T; \quad i = 0, 1, \ldots, N \tag{7.58}$$

Thus, we have same model that was treated in the previous section with R_T replacing C and a maximum buffer size of $B_T + B_D = B$.

The fluid flow model can also be extended to simultaneously handle marked and regular cells (Elwaid and Mitra 1991).

Example 7.6 In this example, we consider same parameters as in the previous example except that now we consider the $R_T = 10$, $B_T = 10$, and $B_D = 15$. We plot the cumulative distribution function against B. We change the token rate to $R_T = 11$ and then plot the cumulative distribution function against B. Clearly, as we can see from Figure 7.7, there is a reduction in buffer occupancy at higher R_T and hence the probability of buffer overflow; in other words, the loss will decrease.

7.6 EQUIVALENT BANDWIDTH

The concept of the *bandwidth* of a signal arose in connection with the transmission of analog signals over radio, telephone, and telegraph systems. These signals can be viewed in either the time or the frequency domain with the Fourier transform acting as the translator. In a strict mathematical sense, real signals cannot be limited in both the time and frequency domains (Papoulis 1962). Thus, since all practical signals have a finite time duration, they must also have energy at all frequencies. However, as a practical matter, limiting signals to a finite portion of the frequency band, provided it is large enough, results in acceptable distortion. The key point for our discussion is that the definition of the bandwidth of a signal implicitly suggests acceptable distortion or error in reproduction.

The same concept applies to the definition of *equivalent bandwidth* for the digital signals that we have been considering. We have developed expressions for the probability of the buffer contents achieving a particular level. This is easily translated into the probability of a buffer level exceeding a particular value. The probability of the contents of an infinite buffer exceeding a particular level forms an upper bound for the probability of overflow of a finite buffer whose size is at that level.

Now, we consider a *quality of service* (QoS) criterion in terms of probability of buffer overflow. The *equivalent bandwidth* is the value of C, the capacity of the output line, which will achieve this QoS. The survivor function, given by (7.49), (7.51), or (7.53), as appropriate, can be set equal to the desired probability of buffer overflow and the required value of C calculated, holding the other parameters fixed. Of course, the expressions on the RHS of (7.49), (7.51), or (7.53) is a very complex function of C; consequently, trial and error is required.

We can use by an approximation based on the asymptotic expression of (7.51) to find an explicit expression for the equivalent bandwidth. An approximation that seems to be valid for a number of cases of interest (Guerin 1991) is to set the coefficient of the exponent equal to one. We then have

$$P_L \cong e^{\varepsilon_0 B} \tag{7.59}$$

where P_L is the probability of loss and B is the buffer size. By substituting the previously derived expression for ε_0 (7.34) and solving for C, we find

$$C = \left(\frac{BN(\alpha+\beta)}{2\ln P_L} + \frac{RN}{2}\right) + \sqrt{\left(\frac{BN(\alpha+\beta)}{2\ln P_L} + \frac{RN}{2}\right)^2 - \frac{BRN^2\alpha}{2\ln P_L}} \tag{7.60}$$

Since $P_L < 1$, the expression under the radical is positive and the whole expression is real and positive. At this point in the text, it should be a simple exercise to plot (7.60) on an Excel spreadsheet.

7.7 LONG-RANGE-DEPENDENT TRAFFIC

7.7.1 Definitions

The detailed empirical study performed on Ethernet LAN traffic (Leland 1994) at the Bellcore Morristown Research and Engineering Center has engendered a great deal of interest in long-range-dependent (LRD) traffic. The study revealed the self-similar nature of this traffic, that is traffic characteristics are the same on all timescales. Later studies of variable-bit-rate (VBR) video traffic (Beran et al. 1995) and Internet packet traffic (Li and Mills 1998, Paxson and Floyd 1995) established the LRD nature of these traffic types. Here and in Chapter 8, we present techniques for modeling this traffic.

We begin our study with a description of the phenomenon. Bear in mind that the traffic is digital, so it is natural to segment the line flow into slots. We represent a random process by means of the random variable, X_t, $t = 0, 1, 2, \ldots$, which is defined as the number of packet/message arrivals or the amount of bits generated in slot t. For the purposes of our explanation, we assume that the slot duration is so small that we can take time to be continuous. We also assume that the process is wide-sense stationary with mean $E(X_t)$ and autocovariance function

$$V(\tau) = \frac{E((X_t - \mu)(X_{t+\tau} - \mu))}{E((X_t - \mu)^2)} \tag{7.61}$$

An LRD process is characterized by an autocovariance function, which is of the form

$$V(\tau) = \psi_{\text{cov}} \tau^{-\beta^l} \tag{7.62}$$

with $0 < \beta^l < 1$.[4]

[4] We are in a dilemma with respect to notation. It has been the practice to use the letter β to describe both self-similar traffic and fluid flow. This is why we use β^l here.

We now present a more precise definition. Let $X^{(m)} = (X_k^{(m)}, k = 0,1,2,\ldots)$ be a new process corresponding to $m = 1, 2, \ldots$, where $X_k^{(m)} = (1/m)\sum_{i=0}^{m-1} X_{km+i}$. It can be shown that each process $X^{(m)}$ is also a wide-sense stationary process. Let $V^{(m)}(j)$ denote its corresponding autocorrelation function. The process X is called *exactly* (*second-order*) *self-similar* with self-similarity parameter $H = 1 - \beta^l/2$, if the processes $X^{(m)}$; $m > 1$ all have the same correlation structure as the original process, $X^{(1)}$, that is, $V^{(m)}(j) = V(j)$ for all $m = 1, 2, \ldots$ and for all j. A process X, asymptotically (second-order) self-similar, if $V^{(m)}(j)$; $m > 1$ agrees asymptotically with $V^{(1)}(j)$. The important consequences of these processes, due to their self-similar (second-order) nature (for both exactly and asymptotically second-order self-similar traffic) can be summarized as follows:

- $\mathrm{Var}(X_k^{(m)}) \sim a_2\tau^{-\beta^l}$ as $\tau \to \infty$ with $0 < \beta^l < 1$.
- Autocovariance functions decay hyperbolically, $V(\tau) \sim \tau^{-\beta^l}$ as $j \to \infty$, which implies a nonsummable autocorrelation function, $\sum_{j=-\infty}^{\infty} r(j) = \infty$.
- The power spectral density function $S(f)$ behaves like that of $1/f$ noise, around the origin: $S(f) \sim f^{-(1-\beta^l)}$ as $f \to 0$.

A process for which $\sum_{j=-\infty}^{\infty} r(j) < \infty$ is called a *short-range-dependent* (SRD) process. The ways in which these processes differ from LRD processes are as follows:

- The variances $\mathrm{Var}(X_k^{(m)})$ decay as m^{-1}
- $r(j)$ decays exponentially: $r(j) \sim a_3 e^{-j}$.
- The power spectral density is bounded at the origin.

LRD traffic can be modeled very closely in terms of first- and second-order statistics, by means of such processes as fractional Gaussian noise (FGN), arithmetic moving averages (ARIMA), or "chaotic maps." These processes can play a role in simulation studies; however, they do not readily lend themselves to the analysis of performance. On the other hand, second-order measures such as the index of dispersion for counts (IDC)

$$\mathrm{IDC}(t) = \frac{\mathrm{Var}(N(t))}{E(N(t))}$$

where $N(t)$ represents the number of arrivals by time t and the index of dispersion for intervals (IDI)

$$\mathrm{IDI}(n) = \frac{\mathrm{Var}\left(\sum_{i=1}^{n} Y_i\right)}{nE^2(Y)}$$

where Y_i represents the interarrival times and Y represents the steady-state interarrival time, random variables showing that standard Markov models of LRD traffic are not accurate over a wide timespan (Leland et al. 1994).

A good description of characterization of arrival processes in terms of these indexes is given in Gusella (1991). The LRD traffic can be modeled using Markov processes over a timescale. [Robert and Le Boudec (1996)] used a Markov modulated chain to model self-similar traffic over finite timescales. The number of arrivals (X_t) given over a given time slot was assumed to be either 0 or 1, and the probability of arrival or no arrival was dependent on the modulator's state ($Y_t = i$, where $i \in 1, 2, \ldots, n$). With a particular structure of transition probability matrix of the n-state discrete-time Markov modulating chain, the model was constructed with two parameters separate from the number of states of the modulating Markov chain that exhibited self-similarity over a finite timescale. In order to evaluate the domain of validity of self-similarity, a quantitative method based on Courtois' (1977) decomposability theory was given. Gallardo et al. (2000) used *α-stable* stochastic processes to model the self-similar nature of traffic on the basis of *α-stable* distributions. Stable distributions are very well suited for approximating the distribution of normalized sums of relatively large numbers of iid random variables. The Gaussian distribution belongs to this family. There are two ways of defining a stable process that is self-similar with stationary increments: linear fractional stable motion (LFSM) and log-fractional stable motion (log-FSM).

In the remainder of this section we present a technique that approximates LRD traffic with a fluid flow model over a predefined timespan. The approach is similar to one in Feldmann and Whitt (1998); however, we match the covariance function, whereas Feldman matched the tail of a probability distribution. In the next chapter, we use an approach of Anderson and Nelson (1998) to match the covariance function. Performance can then be evaluated using the techniques that we have developed in this chapter. Faraj (2000) showed that the technique found good approximations to the performance of LRD traffic.

7.7.2 A Matching Technique for LRD Traffic Using the Fluid Flow Model

We start considering superposition of number of different types of ON–OFF sources. Let us consider d different types of ON–OFF fluid sources, each with parameters of α_i, β_i; $i = 1, 2, \ldots, d$, and let the number of ith traffic types be denoted as N_i. The mean, variance, and covariance functions of superposition of N_i sources are given by (7.13), (7.14), and (7.15), respectively, with appropriate subscripts. Therefore, the total mean arrival rate, variance, and covariance functions of the superposition of N_i sources of d different types of ON–OFF fluid sources, are given by (letting $N_T = \sum_{i=1}^{d} N_i$)

$$\bar{W}_{N_T} = \sum_{i=1}^{d} \frac{N_i R_i \alpha_i}{(\alpha_i + \beta_i)} \tag{7.63}$$

$$\sigma^2_{W_T} = \sum_{i=1}^{d} \frac{N_i R_i^2 \alpha_i \beta_i}{(\alpha_i + \beta_i)^2} \tag{7.64}$$

$$\mathrm{Cov}_{W_T}(\tau) = \sum_{i=1}^{d} \frac{N_i R_i^2 \alpha_i \beta_i}{(\alpha_i + \beta_i)^2} e^{-(\alpha_i+\beta_i)|\tau|} = \sum_{i=1}^{d} C_i(\tau) \tag{7.65}$$

since the covariance function of superposition of independent streams is the sum of covariance functions of individual streams. Let

$$\lambda_i = \alpha_i + \beta_i \tag{7.66}$$

and

$$K_i = \frac{N_i R_i^2 \alpha_i \beta_i}{\lambda_i^2} \tag{7.67}$$

Our objective is to choose the relevant parameters so that the covariance function given in (7.65) approximates the covariance function of long-range-dependent traffic given by (7.62). With a sum of exponentials, we can expect to match over only a certain number of timescales of τ, since the sum of exponentials will have its covariance function decaying exponentially beyond the timescales over which the respective covariance functions are matched. We choose the timescales to be logarithmically related; thus, we match the covariance functions at d points, τ_i; $i = 1, 2, \ldots, d$ such that $\tau_i = 10^{i-1}\tau_1$ for $2 \leq i \leq d$. We also define a scaling factor b such that $1 < b < \tau_{i+1}/\tau_i$ for $1 \leq i \leq (d-1)$. We found satisfactory results when we set $b = 3$.

In order to carry out the fitting algorithm, we assume order-of-magnitude differences among the λ_i values such that $\lambda_1 \gg \lambda_2 \gg \lambda_3 \cdots \lambda_{d-1} \gg \lambda_d$, so that, at τ_i, the terms $\lambda_j \tau_i$ for $i \leq j \leq d$ are small enough such that the sum of exponentials will have contribution only corresponding to the terms of $\lambda_j \tau_i$ for $i \leq j \leq d$, in (7.65). The contributions due to the terms $\lambda_j \tau_i$ for $1 \leq j \leq (i-1)$ are negligible since the exponential contribution due to $\lambda_j \tau_i$ in these cases is of many magnitudes lower than the contribution due to the term $\lambda_i \tau_i$. We use this fact in solving for λ_i and K_i values. Note that in order to ensure this property, we would have to choose τ_i values that are very well separated over a wide range of timescales. We can test this characteristics after we obtain the parameters. The idea is illustrated in Figure 7.7 for $d = 3$. [For purposes of explanation, we define $C_i(\tau) = K_i e^{-\lambda_i \tau}$.] At times τ_3 and $b\tau_3$, only, $C_3(\tau)$ has a significant value. For the other terms, $C_1(\tau)$ and $C_2(\tau)$, the exponents are so large that they can be neglected. At times τ_2 and $b\tau_2$, two terms, $C_2(\tau)$ and $C_3(\tau)$, have significant values, but the values of $C_3(\tau)$ are known from the previous step, so the terms for $C_2(\tau)$ can be found. The process continues in this fashion for all d points.

We write the following relationships as our first step

$$\mathrm{Cov}(\tau_d) = \psi_{\mathrm{cov}} \tau_d^{-\beta^l} = \sum_{i=1}^{d} K_i e^{-\lambda_i \tau_d} \approx K_d e^{-\lambda_d \tau_d} \tag{7.68}$$

where we assume that ψ_{cov} and β^l (which gives the Hurst[5] parameter) are given, so that covariance function values can be calculated. Also, we can write

$$\text{Cov}(b\tau_d) = \psi_{\text{cov}}(b\tau_d)^{-\beta^l} = \sum_{i=1}^{d} K_i e^{-\lambda_i b\tau_d} \approx K_d e^{-\lambda_d b\tau_d} \tag{7.69}$$

Dividing (7.69) by (7.68), taking logarithms, and then simplifying, we get $\lambda_d = [\beta^l \ln(b)]/[(b-1)\tau_d]$. Substituting for λ_d in (7.68) gives K_d. In the next step, we can solve for λ_{d-1} and K_{d-1}, as shown below. Similar to (7.68) and (7.69), we can write

$$\text{Cov}(\tau_{d-1}) = \psi_{\text{cov}}\tau_{d-1}^{-\beta^l} = \sum_{i=1}^{d} K_i e^{-\lambda_i \tau_{d-1}} \approx K_d e^{-\lambda_d \tau_{d-1}} + K_{d-1} e^{-\lambda_{d-1}\tau_{d-1}} \tag{7.70}$$

and

$$\begin{aligned}\text{Cov}(b\tau_{d-1}) = \psi_{\text{cov}}(b\tau_{d-1})^{-\beta^l} &= \sum_{i=1}^{d} K_i e^{-\lambda_i b\tau_{d-1}} \\ &\approx K_d e^{-\lambda_d b\tau_{d-1}} + K_{d-1} e^{-\lambda_{d-1} b\tau_{d-1}}\end{aligned} \tag{7.71}$$

Since we already have λ_d and K_d, we can solve for λ_{d-1} and K_{d-1} using (7.70) and (7.71). Similarly, we can solve for λ_{d-2}, and K_{d-2}, by using the following equations:

$$\text{Cov}(\tau_{d-2}) = \psi_{\text{cov}}\tau_{d-2}^{-\beta^l} = \sum_{i=1}^{d} K_i e^{-\lambda_i \tau_{d-2}} \approx \sum_{i=d-2}^{d} K_i e^{-\lambda_i \tau_{d-2}} \tag{7.72}$$

$$\text{Cov}(b\tau_{d-2}) = \psi_{\text{cov}}(b\tau_{d-2})^{-\beta^l} = \sum_{i=1}^{d} K_i e^{-\lambda_i b\tau_{d-2}} \approx \sum_{i=d-2}^{d} K_i e^{-\lambda_i b\tau_{d-2}} \tag{7.73}$$

Thus, we can solve recursively for λ_j and K_j. The general equations are

$$\text{Cov}(\tau_j) = \psi_{\text{cov}}\tau_j^{-\beta^l} = \sum_{i=1}^{d} K_i e^{-\lambda_i \tau_j} \approx \sum_{i=j}^{d} K_i e^{-\lambda_i \tau_j} \tag{7.74}$$

$$\text{Cov}(b\tau_j) = \psi_{\text{cov}}(b\tau_j)^{-\beta^l} = \sum_{i=1}^{d} K_i e^{-\lambda_i b\tau_j} \approx \sum_{i=j}^{d} K_i e^{-\lambda_i b\tau_j} \tag{7.75}$$

At the last step, we get λ_1 and K_1.

Until this point, we have only two parameters corresponding to all d different types of ON–OFF sources. However, the parameters we require are N_i, α_i, β_i, and R_i.

[5] The Hurst parameter, $H = 2 - 2\beta$, commemorates the classic work on self-similar traffic (Hurst (1956)).

To simplify the requirements, we assume that N_i values are given beforehand. Thus, we need to solve for only $3d$ variables, but from the method described above, we could solve for only $2d$ variables. Since we already made the following assumption $\lambda_1 \gg \lambda_2 \gg \lambda_3 \cdots \lambda_{d-1} \gg \lambda_d$, in order to satisfy this relation, we also assume the following: $\alpha_1 \gg \alpha_2 \gg \alpha_3 \cdots \alpha_{d-1} \gg \alpha_d$ and $\beta_1 \gg \beta_2 \gg \beta_3 \cdots \beta_{d-1} \gg \beta_d$. Satisfying the relation $\alpha_1 \gg \alpha_2 \gg \alpha_3 \cdots \alpha_{d-1} \gg \alpha_d$, we choose $\alpha_i = 10^{-(i-1)}\alpha_1$ for $2 \le i \le d$. From (7.67), we can write $R_i = \sqrt{K_i\lambda_i^2/N_i\alpha_i\beta_i}$. Substituting for R_i, in $N_iR_i\alpha_i/(\alpha_i + \beta_i)$ and using $\beta_i = \lambda_i - \alpha_i$, we can write (7.63), using $\alpha_i = 10^{-(i-1)}\alpha_1$, as

$$\bar{W}_{N_T} = \sum_{i=1}^{d} \sqrt{\frac{K_iN_i\alpha_1}{(10^{i-1}\lambda_i - \alpha_1)}} \tag{7.76}$$

Since N_i, λ_i, and K_i values are known for $1 \le i \le d$, we can solve for α_1 using the nonlinear equation (7.76) and hence α_i for $2 \le i \le d$, and β_i, and R_i for $1 \le i \le d$. We use Newton's method to find the root of (7.76), and we start α_1 with a fraction of λ_1 that we obtained from the previous steps. We illustrate this method using the following example.

Example 7.7 Consider the following parameters. The parameters corresponding to ON–OFF sources are $\bar{W}_{N_T} = 5.5$, $d = 4$, $N_1 = 3$, $N_2 = 4$, $N_3 = 2$, $N_4 = 3$; and the parameters corresponding to the long-range-dependent correlation function are $\psi_{\text{cov}} = 2$, $H = 0.75$, $b = 3$, and $\tau_1 = 2$. The associated Matlab program uses the preceding method to solve for the parameters. The results obtained using this program are as shown.

Source Type	Number of ON–OFF Sources (Given)	α_i	β_i	R_i
1	3	0.13363107	0.57261274	2.5531718495
2	4	0.01336310	0.02990933	0.8384286644
3	2	0.00133631	0.00184958	0.5819627876
4	3	0.00013363	0.00000369	0.8655721989

When we compute the total average bit rate using these parameters, we get 5.499999999999974. When we plot the covariance function value up to the timescales we considered, we find that both the curves are almost overlapping.

REFERENCES

Anderson. A. T., and B. F. Nelson, "A Markovian approach for modeling packet traffic with long-range dependence," *IEEE J. Select. Areas Commun.*, **16**(5): 719–732 (June 1998).

Anik, D., D. Mitra, and M. M. Sondhi, "Stochastic theory of a data handling system with multiple sources," *Bell Syst. Tech. J.*, **61**: 1871–1894 (Oct. 1982).

Belman, R., *Introduction to Matrix Analysis*, McGraw-Hill, New York, 1962.

Beran, J. et al., "Long-range dependence in variable-bit-rate video traffic," *IEEE Trans. Commun.*, 1566–1579 (Feb.–March–April 1995).

Berger, A. W., "Performance analysis of a rate control throttle where tokens and jobs queue," *IEEE J. Select. Areas Commun.*, **SAC-9**: 165–170 (1991).

Courtois, P. J., *Decomposability*, ACM Monograph Series, 1977.

Elwaid, A. I., and D. Mitra, "Analysis and design of rate-based congestion control of high-speed networks, I: Stochastic fluid models, access regulation," *Queueing Syst.*, **9**: 29–64 (1991).

Faraj, R., *Modeling and Analysis of Self-Similar Traffic in ATM Networks*, Ph.D. thesis, Concordia Univ., 2000.

Feldmann, A., and W. Whitt, "Fitting mixture of exponentials to long-tail distributions to analyze network performance models," *Perform. Eval.*, **31**: 245–279 (1998).

Gallardo, J. R. et al., "Use of *α-stable* self-similar stochastic processes for modeling traffic in broadband networks," *Perform. Eval.*, **40**: 71–98 (2000).

Guerin, R. et al., "Equivalent capacity and its application to bandwidth allocation in high speed networks," *IEEE J. Select. Areas Commun.*, **9**(7): 968–981 (Sept. 1991).

Gusella, R., "Characterizing the variability of arrival processes with indexes of dispersion," *IEEE J. Select. Areas Commun.*, **9**(2) (Feb. 1991).

Hurst, H. E., "Methods of using long-term storage capacity of reservoirs," *Trans. Am. Soc. Civil Eng.*, **116**(Part 1): 770–799 (1956).

Leland, W. E. et al., "On the self-similar nature of Ethernet traffic (extended version)," *IEEE/ACM Trans. Network*, **2**, 1–15 (Feb. 1994).

Li, Q., and D. L. Mills, "On the long-range dependence of packet round-trip delays in Internet," *International Conf. Communications*, Montreal, 1998, pp. 1185–1191.

Maglaris, B. et al., "Performance models of statistical multiplexing in packet video communications," *IEEE Trans. Commun.*, **36**(7): 834–844 (July 1988).

Papoulis, A., *The Fourier Integral and its Applications*, McGraw-Hill, New York, 1962.

Paxson, V., and S. Floyd, "Wide area traffic: The failure of Poisson modeling," *IEEE/ACM Trans. Network.*, **3**(3): 226–244 (June 1995).

Robert, S., and J. Y. Le Boudec, "On a Markov modulated chain exhibiting self-similarities over a finite time-scale," *Perform. Eval.*, **27–28**: 159–173 (1996).

Sidi, M., W. Z. Liu, I. Cidon, and I. Gopal, "Congestion control through input rate regulation," *Proc. GLOBECOM'89*, Dallas, Nov. 1989, pp. 1764–1768.

Tucker, R. B. F., "Accurate method for analysis of a packet-speech multiplexer with limited delay," *IEEE Trans. Commun.*, **36**(4): 479–483 (April 1988).

Stern, T. E., and A. I. Elwaid, "Analysis of separable Markov-modulated rate models for information-handling systems," *Adv. Appl. Prob.*, **23**: 105–139 (1991).

Turner, J. S., "New directions in communications (or which way to the information age?)" *IEEE Commun. Mag.*, **23**(10): 8–15 (Oct. 1986).

EXERCISES

7.1 Carry out the same steps as in example 7.2 for the following set of parameters $N = 16$, $\beta = 0.6$, $\alpha = 0.4$, $R = 10$ and $C = 100$.

7.2 Repeat example 7.2 for the set of parameters in example 7.1 with $N/M = 20$ and the capacity of the output line $C = 0.5$ Mbps.

7.3 Comment on the survivor function as an approximation to buffer overflow probability. Upper or lower bound? How tight?

7.4 Repeat example 7.3 for the set of parameters in exercise 7.1.

7.5 Repeat example 7.3 for the set of parameters in example 7.1 with $N/M = 20$ and the capacity of the output line $C = 0.5$ Mbps.

7.6 Repeat example 7.4 for the set of parameters in exercise 7.1.

7.7 Repeat example 7.4 for the set of parameters in example 7.1 with $N/M = 20$ and the capacity of the output line $C = 0.5$ Mbps.

7.8 Repeat example 7.5 for the set of parameters in exercise 7.1.

7.9 Repeat example 7.5 for the set of parameters in example 7.1 with $N/M = 20$ and the capacity of the output line $C = 0.5$ Mbps.

7.10 For the following set of parameters $\alpha = 4$, $\beta = 8$, $N = 6$, $B_T = 15$, $B_D = 25$ and $R_T = 4$ plot the cumulative distribution function as a function of the token rate in a leaky bucket.

7.11 For the set of parameters derived in example 7.7, repeat exercise 7.5.

7.12 Repeat example 7.6 with $N = 8$, $R = 5$, $\alpha = 4$, $\beta = 8$, $R_T = 18$, $B_T = 20$ and $B_D = 30$. Also compare the distribution function with $R_T = 28$.

7.13 Starting from (7.74) and (7.75) prove that, the quantities λ_j and K_j are given y,

$$\lambda_j = \frac{\ln\left(b^{-\beta^l} - (c'/(g' - e'))\right)^{-1}}{\tau_j(b-1)}$$

$$K_j = \frac{c'}{b^{-\beta^l} e^{-\lambda_j \tau_j} - e^{-\lambda_j b \tau_j}}$$

where $c' = \sum_{i=j+1}^{d} K_i(e^{-\lambda_i b \tau_j} - b^{-\beta^l} e^{-\lambda_i \tau_j})$, $e' = \sum_{i=j+1}^{d} K_i e^{-\lambda_i b \tau_j}$, and $g' = \psi_{\mathrm{cov}} \tau_j^{-\beta}$.

8

THE MATRIX GEOMETRIC TECHNIQUES

8.1 INTRODUCTION

In the previous chapter, we presented a queuing analysis for a non-Poisson arrival model. In the present chapter, we consider an alternative model for non-Poisson traffic, which is far more flexible in that it allows a much wider range of models. We also present an algorithmic technique for finding the properties of the queues engendered by the arrival process. The technique, which is called *matrix analytic* or *matrix geometric analysis*, was pioneered by Marcel Neuts (Neuts 1981, 1989) and has undergone continual development by Neuts and his students. As we shall see, the algorithm has several steps leading to the final result. A virtue of the technique is that each of the intermediate results has probabilistic significance, which gives insight into the model. The technique has two basic forms, M/G/1 and G/M/1. The nomenclature reflects the fact that the structure of the respective models is the multidimensional analog to the single-dimensional models treated in Chapter 6. Since it is more useful in a telecommunications context, we shall study the M/G/1 paradigm.

8.2 ARRIVAL PROCESSES

In this section we consider the arrival processes that can be handled by the matrix analytic technique. In this and in succeeding sections of the chapter, we shall focus on the *Markov modulated Poisson process* (MMPP), which is both widely used and a straightforward extension of the Poisson arrival process treated in Chapter 3. This process is presented in detail in the next subsection. In order to illustrate the power

Modeling and Analysis of Telecommunications Networks, by Jeremiah F. Hayes and
Thimma V. J. Ganesh Babu
ISBN 0-471-34845-7 © 2004 John Wiley & Sons, Inc.

of the matrix analytic technique, we show in the subsequent subsections that the MMPP is a special case of the *batch Markov arrival process* (BMAP).

8.2.1 The Markov Modulated Poisson Process (MMPP)

The basic characteristic of the matrix analytic technique is that an underlying L-state Markov chain drives the arrival process. Conforming to the literature, we call this state of the process its *phase*. The state of arrival process at time t is given by $(N(t), \Phi(t))$, where $N(t)$ is the population or the number of arrivals up to time t and $\Phi(t)$ is the phase of the underlying Markov chain at time t. For the MMPP, and for each of the L phases, the message generation rate is different, λ_i; $i = 1, 2, \ldots, L$, respectively. Note that when $L = 1$, we have the ordinary Poisson arrival process. In an incremental interval the underlying Markov chain goes from phase n to phase k with probability $r_{nk}\delta$; $n, k = 1, 2, \ldots, L$. If the underlying Markov chain is in phase j, there is an arrival in the incremental interval, with probability $\lambda_j\delta$. We define the state transition probability as

$$Q_{ij}(n, t) = P(N(t) = n, \Phi(t) = j/N(0) = 0, \Phi(0) = i) \tag{8.1}$$

We now derive the *Kolmogorov forward equation* for the MMPP. The approach should be familiar by now. (See Sections 3.3.2 and 4.4.2, for example.) Essentially, we look at the transitions in the system state from time t to time $t + \delta$. The process can arrive at state $(N(t) = n, \Phi(t) = j)$ in one of two ways: by a message arrival or a phase change: The probabilities are as shown in (8.1).

$$\begin{aligned} Q_{ij}(n, t + \delta) = {} & \underbrace{\left[1 - \left(\sum_{\substack{k=1 \\ k \neq j}}^{L} r_{jk} + \lambda_j\right)\delta\right] Q_{ij}(n, t)}_{\text{neither message arrival nor phase transition}} + \underbrace{\lambda_j \delta Q_{ij}(n-1, t)}_{\text{message arrival}} \\ & + \underbrace{\left(\sum_{\substack{k=1 \\ k \neq j}}^{L} r_{kj}\delta\right) Q_{ik}(n, t)}_{\text{phase transition}}; \quad n = 0, 1, 2, \ldots; \; j = 1, 2, \ldots, L \end{aligned} \tag{8.2}$$

We subtract $Q_{ij}(n, t)$ from both sides, divide by δ, and let $\delta \to 0$ to get the *Kolmogorov forward equation*

$$\begin{aligned} \frac{dQ_{ij}(n,t)}{dt} = {} & -Q_{ij}(n, t)\left[\sum_{\substack{k=1 \\ k \neq j}}^{L} r_{jk} + \lambda_j\right] + \lambda_j Q_{ij}(n-1, t) \\ & + \sum_{\substack{k=1 \\ k \neq j}}^{L} r_{kj} Q_{ik}(n, t); \quad n = 0, 1, 2, \ldots; \; j = 1, 2, \ldots, L \end{aligned} \tag{8.3}$$

Equation (8.3) can be written in matrix form. We begin by defining the matrix $\mathbf{Q}(n, t)$ as the $(L \times L)$ matrix whose (i, j) elements are $\{Q_{ij}(n, t)\}$. This matrix describes all the phase transitions when the population size is n. Further, we define the matrix with L rows

$$\mathsf{Q}(t) = [\mathbf{Q}(0, t), \mathbf{Q}(1, t), \ldots, \mathbf{Q}(n, t), \ldots]$$

or

$$\mathsf{Q}(t) = \begin{bmatrix} Q_{11}(0, t) & Q_{12}(0, t) & \cdots & Q_{1L}(0, t) & Q_{11}(1, t) & Q_{12}(1, t) & \cdots & Q_{1L}(1, t) & \cdots \\ Q_{21}(0, t) & Q_{22}(0, t) & \cdots & Q_{22}(0, t) & Q_{21}(1, t) & Q_{22}(1, t) & \cdots & Q_{22}(1, t) & \cdots \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \ddots \\ Q_{L1}(0, t) & Q_{L2}(0, t) & \cdots & Q_{LL}(0, t) & Q_{L1}(1, t) & Q_{L2}(1, t) & \cdots & Q_{LL}(1, t) & \cdots \end{bmatrix}$$

with this definition we have

$$\frac{d\mathsf{Q}(t)}{dt} = \mathsf{Q}(t)\Gamma \tag{8.4}$$

where the infinitesimal generator matrix[1] is given by

$$\Gamma = \begin{bmatrix} D_0 & D_1 & 0 & \cdots & 0 & \cdots \\ 0 & D_0 & D_1 & \cdots & 0 & \cdots \\ 0 & 0 & D_0 & \cdots & 0 & \cdots \\ \vdots & \vdots & \vdots & \ddots & \vdots & \cdots \\ 0 & 0 & 0 & \cdots & D_0 & \cdots \\ \vdots & \vdots & \vdots & \cdots & \vdots & \ddots \end{bmatrix} \tag{8.5}$$

with D_0 defined as

$$D_0 = \begin{bmatrix} -\sum_{j=2}^{L} r_{1j} - \lambda_1 & r_{12} & \cdots & r_{1L} \\ r_{21} & -\sum_{\substack{j=1 \\ j \neq 2}}^{L} r_{2j} - \lambda_2 & \cdots & r_{2L} \\ \vdots & \vdots & \ddots & \vdots \\ r_{L1} & r_{L2} & \cdots & -\sum_{j=1}^{L-1} r_{Lj} - \lambda_L \end{bmatrix}$$

[1] In order to conform with usage in the literature, in this chapter the infinitesimal generator matrix is written in such a way that the rows sum to zero. In Chapters 3 and 4, the columns of the infinitesimal generator matrix sum to zero.

and D_1 as

$$D_1 = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_L \end{bmatrix}.$$

Example 8.1 Consider the two-phase MMPP process shown in Figure 8.1. We have the parameters $r_{12} = 1$, $r_{21} = 2$, $\lambda_1 = 10$, $\lambda_2 = 5$. By simple substitution, we have

$$D_0 = \begin{bmatrix} -11 & 1 \\ 2 & -7 \end{bmatrix} \quad \text{and} \quad D_1 = \begin{bmatrix} 10 & 0 \\ 0 & 5 \end{bmatrix}.$$

The infinitesimal generator matrix is

$$\Gamma = \begin{bmatrix} -11 & 1 & 10 & 0 & 0 & 0 & \cdots \\ 2 & -7 & 0 & 5 & 0 & 0 & \cdots \\ 0 & 0 & -11 & 1 & 10 & 0 & \cdots \\ 0 & 0 & 2 & -7 & 0 & 5 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}$$

8.2.2 The Batch Markov Arrival Process

We now go to the next level of generality in the arrival process. For the MMPP the arrival is a single message. Now, for the *batch Markov arrival process* (BMAP) the arrival process is compound in the sense of the discussion in Section 5.2.2 in that each arrival contains a random number of messages with probabilities $a_1^j, a_2^j, \ldots, a_M^j$; $j = 1, 2, \ldots, L$, where M is the maximum batch size. Note that we allow the batch distribution to vary with the underlying phase. As in the MMPP, the arrival

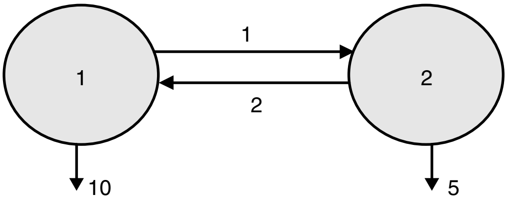


Figure 8.1 Two-phase MMPP.

processes is driven by an underlying L-state Markov chain. It can be shown that the form of the infinitesimal generator matrix for the BMAP process is written as

$$\Gamma = \begin{bmatrix} D_0 & D_1 & D_2 & \cdots & D_M & 0 & \cdots \\ 0 & D_0 & D_1 & \cdots & D_{M-1} & D_M & \cdots \\ 0 & 0 & D_0 & \ddots & \vdots & \vdots & \ddots \\ 0 & 0 & 0 & \ddots & D_1 & D_2 & \cdots \\ 0 & 0 & 0 & \cdots & D_0 & D_1 & \ddots \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots \end{bmatrix} \tag{8.6}$$

where

$$D_0 = \begin{bmatrix} -\sum_{j=2}^{L} r_{1j} - \lambda_1 & r_{12} & \cdots & r_{1L} \\ r_{21} & -\sum_{\substack{j=1 \\ j\neq 2}}^{L} r_{2j} - \lambda_2 & \cdots & r_{2L} \\ \vdots & \vdots & \ddots & \vdots \\ r_{L1} & r_{L2} & \cdots & -\sum_{j=1}^{L-1} r_{Lj} - \lambda_L \end{bmatrix}$$

and

$$D_i = \begin{bmatrix} a_i^1\lambda_1 & 0 & \cdots & 0 \\ 0 & a_i^2\lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & a_i^L\lambda_L \end{bmatrix}; \quad i = 1, 2, \ldots, M$$

The elements of the matrices D_i; $i = 0, 1, 2, \ldots, M$ are $(L \times L)$ matrices indicating the rate of phase transitions for the arrival of a batch of i messages. The subscripts indicate a batch arrival of size i. The offset of D_i from the diagonal ensures that there is a proper jump in the population size. A salient feature of the BMAP is that the arrival process is independent of the state of the system. This property ensures *spatial homogeneity*, a concept that we will encounter later. This stands in contrast to the pure birth process, for example, where the arrival rate depends on the number of messages in the system (see Section 3.2.5).

For both the MMPP and the BMAP, we define the generating function

$$D(z) = \sum_{i=0}^{M} z^i D_i \tag{8.7}$$

The $(L \times L)$ matrix

$$D = D(1) = \sum_{i=0}^{M} D_i = \begin{bmatrix} -\sum_{j=2}^{L} r_{1j} & r_{12} & \cdots & r_{1L} \\ r_{21} & -\sum_{\substack{j=1 \\ j \neq 2}}^{L} r_{2j} & \cdots & r_{2L} \\ \vdots & \vdots & \ddots & \vdots \\ r_{L1} & r_{L2} & \cdots & -\sum_{j=1}^{L-1} r_{Lj} \end{bmatrix} \tag{8.8}$$

distills the phase information. The row vector $\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots, \pi_L)$, which is the solution to

$$\boldsymbol{\pi} D = 0 \tag{8.9}$$

with the normalizing condition $\sum_{i=1}^{L} \pi_i = 1$, gives the steady-state probabilities of the phases of the arrival process. In the literature the column vector of all ones (1s), denoted **e**, is used to express the normalizing condition as well as other relations; thus, $\boldsymbol{\pi}\mathbf{e} = 1$.

Having focused on the underlying phase, we turn to the arrival process. Averaging over the batch arrival, excluding the phase transitions, the fundamental arrival rate for the process is given by

$$\bar{\lambda} = \boldsymbol{\pi} \sum_{i=1}^{M} i D_i \mathbf{e} \tag{8.10}$$

Note that $\bar{\lambda}$ is a scalar quantity.[2] If a message has a constant transmission time of m seconds, the load is given by

$$\rho = \bar{\lambda} m \tag{8.11}$$

[2] In going though the text, it may be instructive for the reader to work out the dimensions of the equations that will be presented. For example, we have

$$\underbrace{\bar{\lambda}}_{(1\times 1)} = \underbrace{\boldsymbol{\pi}}_{(1\times L)} \underbrace{\sum_{i=1}^{N} i D_i}_{(L\times L)} \underbrace{\mathbf{e}}_{(L\times 1)}$$

Example 8.2 Consider Example 8.1. The solution to (8.9) with the normalizing condition is $\boldsymbol{\pi} = \left(\frac{2}{3} \quad \frac{1}{3}\right)$. The average arrival rate is given by

$$\bar{\lambda} = \left(\frac{2}{3} \quad \frac{1}{3}\right)\begin{bmatrix} 10 & 0 \\ 0 & 5 \end{bmatrix}\begin{bmatrix} 1 \\ 1 \end{bmatrix} = 8\tfrac{1}{3}$$

Example 8.3 Now consider an example where batches are of size 1 and 2. Assume the following set of parameters: $r_{12} = 1$, $r_{21} = 2$, $\lambda_1 = 9$, $\lambda_2 = 3$, $a_1^1 = \frac{2}{3}$, $a_2^1 = \frac{1}{3}$, $a_1^2 = \frac{5}{6}$, $a_2^2 = \frac{1}{6}$. The infinitesimal generator matrix is given by

$$\Gamma = \begin{bmatrix} -10 & 1 & 6 & 0 & 3 & 0 & \cdots \\ 2 & -5 & 0 & 2.5 & 0 & 0.5 & \cdots \\ 0 & 0 & -10 & 1 & 6 & 0 & \cdots \\ 0 & 0 & 2 & -5 & 0 & 2.5 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}$$

Again, the steady-state phase probability is $\boldsymbol{\pi} = \left(\frac{2}{3} \quad \frac{1}{3}\right)$. From (8.10), we have for the average message arrival rate

$$\bar{\lambda} = \boldsymbol{\pi}\left\{\begin{bmatrix} 6 & 0 \\ 0 & 3 \end{bmatrix} + 2\begin{bmatrix} 2.5 & 0 \\ 0 & 0.5 \end{bmatrix}\right\}\begin{bmatrix} 1 \\ 1 \end{bmatrix} = 8\tfrac{1}{3} \text{ messages per second}$$

See the associated Matlab program for details.

8.2.3 Further Extensions

In the models that we have seen to this point, the arrival process and the phase transition process are independent. The theory proceeds in the same fashion if this restriction is relaxed. We can have a realization of a BMAP in which the process stays in a phase for an exponentially distributed amount of time whereupon there is simultaneously a batch arrival of messages and a phase transition. In this case, there is a term in (8.2) indicating simultaneous message arrival and phase transition, in addition to the previously discussed arrival types.

Another useful realization is one that can have arbitrary intervals between arrivals. We use the method of stages, which was presented in Section 3.4 to implement the inter–arrival time (see Fig. 3.22). There are *K* phases to the process, corresponding to *K* stages in Figure 3.22.

8.2.4 Solutions of Forward Equation for the Arrival Process

We now return to the general form of BMAP. From Equation (8.4), this can be reformulated as

$$\frac{d\mathbf{Q}(k,t)}{dt} = \sum_{i=0}^{k} \mathbf{Q}(i, t)D_{k-i}; \quad k = 0, 1, 2, \ldots \tag{8.12}$$

with the initial condition of an empty system

$$\mathbf{Q}(0, 0) = I$$

Define the generating function

$$\mathsf{Q}(z, t) = \sum_{i=0}^{\infty} \mathbf{Q}(i, t) z^i \tag{8.13}$$

We take z transforms of both sides of (8.12) to obtain

$$\frac{d\mathsf{Q}(z,t)}{dt} = \mathsf{Q}(z, t) D(z) \tag{8.14}$$

The solution to this equation is simply given by

$$\mathsf{Q}(z, t) = e^{D(z)t} = e^{\sum_{i=0}^{M} z^i D_i t} = \prod_{i=0}^{M} e^{z^i D_i t} \tag{8.15}$$

The evolution of the empty state has particular interest later in the text. For $z = 0$ in (8.15), we have

$$\mathsf{Q}(0, t) = e^{D_0 t} \tag{8.16}$$

We emphasize that this is the probability of an empty system at time t. The probabilities for higher occupancy levels can be found from the following relation, which follows easily from (8.13):

$$\mathbf{Q}(n, t) = \frac{1}{n!} \frac{d^n \mathsf{Q}(z, t)}{dz^n} \bigg|_{z=0} \tag{8.17}$$

In the case of the MMPP $\mathbf{Q}(n, t)$ has a particularly simple form since $D_n = 0$; $n \geq 2$. From (8.15)–(8.17), we have

$$\mathbf{Q}(n, t) = \frac{1}{n!} \frac{d^n e^{(D_0 + D_1 z)t}}{dz^n} \bigg|_{z=0} = \frac{1}{n!} e^{D_0 t} (D_1 t)^n \tag{8.18}$$

Note the similarity of form with the Poisson process. In Section 3.6.2, the one-dimensional version is presented.

Example 8.4 Consider the submatrices of Example 8.1 for the MMPP. The solution given by (8.16) is

$$\mathbf{Q}(0, t) = e^{D_0 t} = \begin{bmatrix} 0.00014 & 0.00029 \\ 0.00058 & 0.00130 \end{bmatrix}^t$$

Applying (8.18) shows that

$$\mathbf{Q}(n, t) = \frac{1}{n!}\begin{bmatrix} 0.00014 & 0.00029 \\ 0.00058 & 0.00130 \end{bmatrix}^t \left(\begin{bmatrix} 10 & 0 \\ 0 & 5 \end{bmatrix} t \right)^n$$

Thus, for $n = 2$ and $t = 0.25$, we have

$$\mathbf{Q}(2, 0.25) = \begin{bmatrix} 0.2179 & 0.0219 \\ 0.1752 & 0.1421 \end{bmatrix}$$

This means that the probability that there are two messages in the system and that the system is in phase 2 having started in phase 1 is 0.0219.

8.3 IMBEDDED MARKOV CHAIN ANALYSIS

8.3.1 Revisiting the M/G/1 Queue

In Chapter 6, the M/G/1 queue was analyzed by means of a Markov chain imbedded at departure epochs. We now present an analysis that is different from the one in Chapter 6 with the objective of introducing an approach that can be used for the arrival processes considered in this chapter. We focus on the general case where the first message of a busy period has a transmission time different from those that have waited in the queue for service (see Section 6.1.4). Equation (6.37) gives the dynamics of the process. The number of messages in the system after the $(i + 1)$st departure is written in terms of the number in the system at the ith departure and the number of arrivals between departures

$$\begin{aligned} N_{i+1} &= N_i - U(N_i) + A_{i+1}U(N_i) + B_{i+1}[1 - U(N_i)] \\ &= N_i - U(N_i) + B_{i+1} + (A_{i+1} - B_{i+1})U(N_i) \end{aligned} \tag{8.19}$$

where B_{i+1} and A_{i+1} are, respectively, the number of Poisson arrivals during an initiating message transmission time and a queued message transmission time. Alternatively, we can display system dynamics in the form of the state transition matrix discussed Chapter 2. As in Section 6.1, we focus on the imbedded points. The state transition matrix is given by

$$P(N_{i+1} = k/N_i = j) = \begin{cases} P(B_{i+1} = k); & j = 0 \\ P(A_{i+1} = k - j + 1); & j > 0 \\ 0; & j > k + 1 \end{cases} \tag{8.20}$$

A key observation in connection with (8.20) is that the number of messages in the system at successive imbedded points can increase by any number but may decrease by only one at most. This is indicated by the third relation on the RHS of (8.20). The state transition matrix can be written in the canonical form

$$\mathsf{M} = \begin{bmatrix} b_0 & b_1 & b_2 & b_3 & \cdots \\ a_0 & a_1 & a_2 & a_3 & \cdots \\ 0 & a_0 & a_1 & a_2 & \cdots \\ 0 & 0 & a_0 & a_1 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix} \tag{8.21}$$

where $a_k = P(A_i = k)$ and $b_k = P(B_i = k)$. Since arrivals are Poisson we have

$$a_k = \int_0^\infty \frac{(\lambda t)^k}{k!} e^{-\lambda t} m(t) dt \tag{8.22}$$

where $m(t)$ is the probability distribution of the message transmission time. For initiating message transmission, b_k is calculated by replacing $m(t)$ in (8.22) by $\tilde{m}(t)$, the probability density for the initiating message. The form of the matrix M is characteristic of the M/G/1 paradigm. The lower left triangle is all 0 because the number of messages can decrease by at most one from one imbedded point to another. This property of the process plays an important role in the development of the next section. For constant-length messages, we have simply $a_k = (\lambda m)^k e^{-\lambda m}/k!$.

We can use the preceding to derive an alternative, recursive approach to the solution of the M/G/1 queue. This approach will serve as a model for the development later. As in Chapter 4, we define the steady-state probabilities $\mathbf{P} = (P_0, P_1, \ldots, P_n, \ldots)$ for the number of messages left behind by a departing message. These can be found as the solution to

$$\mathbf{P} = \mathbf{P}\mathsf{M} \tag{8.23}$$

Writing out the components, we have

$$\begin{aligned} P_0 &= P_0 b_0 + P_1 a_0 \\ P_1 &= P_0 b_1 + P_1 a_1 + P_2 a_0 \\ &\vdots \\ P_k &= P_0 b_k + \sum_{j=1}^{k+1} P_j a_{k+1-j} \\ &\vdots \end{aligned} \tag{8.24}$$

From (8.24), we have the following recursive equation:

$$P_k = \frac{P_{k-1} - P_0 b_{k-1} - \sum_{j=1}^{k-1} P_j a_{k-j}}{a_0}; \quad k = 1, 2, \ldots \tag{8.25}$$

Thus, the probabilities $P_1, \ldots, P_n, \ldots$ can be expressed in terms of the probability of the system being empty, P_0. In the case of the simple M/G/1 system, we have $P_0 = 1 - \rho$.

8.3.2 The Multidimensional Case

In this subsection we derive expressions related to the number of arrivals during a service time. These expressions are the multidimensional analogs of the terms a_k and b_k in the one-dimensional case [see (8.22)]. The expressions play a similar, but more complex, role in the multidimensional case.

We begin by defining the $(L \times L)$ matrices of *probability mass functions*,[3] $\mathbf{A}_0(t), \mathbf{A}_1(t), \ldots, \mathbf{A}_n(t), \ldots$ and $\mathbf{B}_0(t), \mathbf{B}_1(t), \ldots, \mathbf{B}_n(t), \ldots$ with elements

$$\begin{aligned}\{A_n(t)\}_{ij} = P(&\text{departure time} \le t, n \text{ arrivals, arrival phase} \\ &= j/\text{previous departure at } t = 0, \text{ system not empty, arrival phase} = i)\end{aligned} \tag{8.26}$$

and

$$\begin{aligned}\{B_n(t)\}_{ij} = P(&\text{departure time} \le t, n \text{ arrivals, arrival phase} \\ &= j/\text{previous departure at } t = 0, \text{ system empty, arrival phase} = i)\end{aligned} \tag{8.27}$$

We also define the corresponding density functions $\mathbf{a}_i(t) = d\mathbf{A}_i(t)/dt$ and $\mathbf{b}_i(t) = d\mathbf{B}_i(t)/dt$. The first of these quantities can be written in terms of the arrival process $\mathbf{Q}(n, t)$ as

$$\mathbf{A}_n(t) = \int_0^t \mathbf{Q}(n, \tau) m(\tau) d\tau \tag{8.28}$$

as the probability of n messages remaining at time t, together with phase transitions. We define

$$\mathbf{A}_n = \mathbf{A}_n(\infty) \tag{8.29}$$

[3] *Probability mass functions* are defined as nondecreasing functions taking on values between 0 and 1 but not necessarily assuming 0 at $-\infty$ or 1 at $+\infty$. The emphasis here and later should serve to remind the reader that we are dealing with $(L \times L)$ matrices embodying phase transitions.

The density function is[4]

$$\mathbf{a}_n(t) = \frac{d\mathbf{A}_n(t)}{dt} = \mathbf{Q}(n, t)m(t) \tag{8.30}$$

We take the Laplace transform for $\mathbf{a}_n(t)$:

$$\mathsf{A}_n(s) = \int_0^\infty e^{-st}\mathbf{a}_n(t)dt = \int_0^\infty e^{-st}\mathbf{Q}(n, t)m(t)dt \tag{8.31}$$

Taking the z-transform, we have

$$\mathsf{A}(z, s) = \sum_{n=0}^{\infty} z^n \mathsf{A}_n(s) \tag{8.32}$$

From (8.15) and (8.31), it follows easily that

$$\mathsf{A}(z, s) = \int_0^\infty e^{-st}\mathsf{Q}(z, t)m(t)dt = \int_0^\infty e^{-st}e^{D(z)t}m(t)dt \tag{8.33}$$

The $(L \times L)$ matrix

$$\mathbf{A} = \mathsf{A}(1, 0) = \sum_{n=0}^{\infty} \mathsf{A}_n(0) = \int_0^\infty e^{Dt}m(t)dt \tag{8.34}$$

embodies the phase changes during a message transmission. For $\boldsymbol{\pi}$ given by (8.9), we have, by repeated application of (8.9)

$$\begin{aligned}\boldsymbol{\pi}\mathbf{A} &= \boldsymbol{\pi}\int_0^\infty \sum_{j=0}^{\infty} \frac{(\mathbf{D}(t))^j}{j!} m(t)dt \\ &= \boldsymbol{\pi}(\mathbf{D})^0 \int_0^\infty m(t)dt + \int_0^\infty \mathbf{D}t \sum_{i=1}^{\infty} \frac{(\mathbf{D}(t))^{j-1}}{j!} m(t)dt = \boldsymbol{\pi}\end{aligned} \tag{8.35}$$

We get particularly simple expressions when the arrival process is an MMPP and the service time is constant $m(t) = \delta(t - m)$. From (8.18), (8.28), and (8.29), we have

$$\mathbf{A}_n = \mathbf{Q}(n, m) = \frac{e^{D_0 m}(D_1 m)^n}{n!} \tag{8.36}$$

[4] The reader is asked to bear with a notational problem. Previously, the a and b symbols indicated scalar quantities, as in (8.25), whereas in (8.30), we have vectors, as indicated by the boldface type. The problem is short-lived since the vector quantities do not appear later.

This is the matrix of phase transitions during a message transmission and n arrivals. Further, from (8.33)

$$\mathsf{A}(z, 0) = \mathsf{A}(z) = e^{(D_0+D_1 z)m} \tag{8.37}$$

$$\mathbf{A} = \mathsf{A}(1, 0) = e^{Dm} = e^{(D_0+D_1)m} \tag{8.38}$$

We now derive a similar expression for the case where the previously departing message leaves an empty system. After the departure of the last message of the previous busy period, the system remains empty until time $\tau \leq t$ with probability $e^{D_0\tau}$ along with phase transitions. The idle period ends with an arrival of a batch of $i \leq n + 1$ messages. The service time of the first message in the subsequent busy period ends at time t. The density function of this latter interval is $\tilde{\mathbf{a}}_n(t)$. During this service time, there are further arrivals. Putting these components together, we find for $\mathbf{b}_n(t)$

$$\mathbf{b}_n(t) = \sum_{i=1}^{n+1} \int_0^t d\tau\, e^{D_0\tau} D_i \tilde{\mathbf{a}}_{n-i+1}(t - \tau) \tag{8.39}$$

where, similar to (8.30), the probability density function $\tilde{\mathbf{a}}_n(t)$ can be written

$$\tilde{\mathbf{a}}_n(t) = \mathbf{Q}(n, t)\tilde{m}(t) \tag{8.40}$$

where $\sim$ indicates that the first message of the busy period may be different from other messages. We carry this notation through all of the development in this chapter. We take the Laplace transform of $\mathbf{b}_n(t)$.

$$\begin{aligned}
\mathsf{B}_n(s) &= \int_0^\infty dt\, e^{-st}\mathbf{b}_n(t) \\
&= \int_0^\infty dt\, e^{-st} \sum_{i=1}^{n+1} \int_0^t d\tau\, e^{D_0\tau} D_i \tilde{\mathbf{a}}_{n-i+1}(t - \tau) \\
&= \sum_{i=1}^{n+1} \int_0^\infty d\tau\, e^{D_0\tau} D_i e^{-s\tau} \int_0^\infty du\, e^{-su}\tilde{\mathbf{a}}_{n-i+1}(u) \\
&= [sI - D_0]^{-1} \sum_{i=1}^{n+1} D_i \tilde{\mathbf{A}}_{n-i+1}(s)
\end{aligned} \tag{8.41}$$

We take z transforms of both sides to get

$$\mathsf{B}(z, s) = \sum_{n=0}^{\infty} z^n \mathsf{B}_n(s) = [sI - D_0]^{-1} \sum_{n=0}^{\infty} z^n \sum_{i=1}^{n+1} D_i \tilde{\mathbf{A}}_{n-i+1}(s)$$
$$= z^{-1}[sI - D_0]^{-1}[D(z) - D_0]\tilde{\mathsf{A}}(z, s) \tag{8.42}$$

By setting $s = 0$ in $\mathsf{B}(z, s)$, we average over the departure time:

$$\mathsf{B}(z, 0) = \mathsf{B}(z) = -z^{-1}D_0^{-1}[D(z) - D_0]\tilde{\mathsf{A}}(z, 0) = -z^{-1}D_0^{-1}[D(z) - D_0]\tilde{\mathsf{A}}(z) \tag{8.43}$$

The coefficients of $\mathsf{B}(z)$ give the probabilities of the number of messages remaining after a departure

$$\mathbf{B}_k = -D_0^{-1} \sum_{l=0}^{k} D_{l+1} \tilde{\mathsf{A}}_{k-l}(0) \tag{8.44}$$

[see (8.32)]. The negative sign on the RHS of (8.44) should pose no difficulty since the diagonal elements of D_0 are themselves negative. When the arrival process is a MMPP and there is a constant message transmission time, we have

$$\mathsf{B}(z) = -z^{-1}D_0^{-1}D_1 z e^{(D_0 + D_1 z)m} \tag{8.45}$$

and from (8.44)

$$\mathbf{B}_k = -D_0^{-1}D_1 \tilde{\mathsf{A}}_k(0) \tag{8.46}$$

Example 8.5 We return to the MMPP of Example 8.1. We assume constant message lengths, $m(t) = \delta(t - 0.03)$ and $\tilde{m}(t) = \delta(t - 0.05)$, respectively.

From (8.18) and (8.36) we have

$$\mathbf{A}_n = \mathbf{Q}(n, 0.03) = \frac{\begin{bmatrix} 0.04213 & 0.02203 \\ 0.04405 & 0.13023 \end{bmatrix} \left(\begin{bmatrix} 10.0 & 0 \\ 0 & 5.0 \end{bmatrix} 0.03 \right)^n}{n!}$$

Recognizing the different service time for the initiator of a busy period, we have

$$\tilde{\mathbf{A}}_n = \mathbf{Q}(n, 0.05) = \frac{\begin{bmatrix} 0.06643 & 0.00705 \\ 0.01410 & 0.03464 \end{bmatrix} \left(\begin{bmatrix} 10.0 & 0 \\ 0 & 5.0 \end{bmatrix} 0.05 \right)^n}{n!}$$

From (8.45)

$$\mathbf{B}_n = -\frac{\begin{bmatrix} -11 & 1 \\ 2 & -7 \end{bmatrix}\begin{bmatrix} 0.06643 & 0.00705 \\ 0.01410 & 0.03464 \end{bmatrix}\left(\begin{bmatrix} 10.0 & 0 \\ 0 & 5.0 \end{bmatrix} 0.05\right)^n}{n!}$$

We now turn to the solution to the M/G/1 queue with BMAP input. Define

$$\mathsf{M} = \begin{bmatrix} \mathbf{B}_0 & \mathbf{B}_1 & \mathbf{B}_2 & \mathbf{B}_3 & \cdots \\ \mathbf{A}_0 & \mathbf{A}_1 & \mathbf{A}_2 & \mathbf{A}_3 & \cdots \\ 0 & \mathbf{A}_0 & \mathbf{A}_1 & \mathbf{A}_2 & \cdots \\ 0 & 0 & \mathbf{A}_0 & \mathbf{A}_1 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}$$

[see (8.21)]. We also define

$$\mathsf{P} = [\mathbf{P}_0, \mathbf{P}_1, \ldots, \mathbf{P}_n, \ldots]$$

as the solution to the equation

$$\mathsf{P} = \mathsf{PM} \tag{8.47}$$

A straightforward calculation gives the vector version of (8.24):

$$\begin{aligned} \mathbf{P}_0 &= \mathbf{P}_0\mathbf{B}_0 + \mathbf{P}_1\mathbf{A}_0 \\ \mathbf{P}_k &= \mathbf{P}_0\mathbf{B}_k + \sum_{j=1}^{k+1} \mathbf{P}_j\mathbf{A}_{k-j+1}; \quad k = 1, 2, \ldots \end{aligned} \tag{8.48}$$

A simple rearrangement of terms gives a recursive equation for the level probabilities in terms of the zero vector, $\mathbf{P}_0$:

$$\begin{aligned} \mathbf{P}_1 &= \mathbf{P}_0(I - \mathbf{B}_0)\mathbf{A}_0^{-1} \\ \mathbf{P}_k &= \left(\mathbf{P}_{k-1} - \mathbf{P}_0\mathbf{B}_{k-1} - \sum_{j=1}^{k-1} \mathbf{P}_j\mathbf{A}_{k-j}\right)\mathbf{A}_0^{-1}; \quad k = 2, 3, \ldots \end{aligned} \tag{8.49}$$

In Section 8.3.3 the basic properties of Markov chains are used to find $\mathbf{P}_0$ and the remaining probabilities.

Later, we will find the moment-generating function for the levels to be useful in calculating moments. Before proceeding, we pause for definitions and relations. Define $\mathsf{P}(z) = \sum_{k=0}^{\infty} z^k\mathbf{P}_k$. Multiplying both sides of (8.49) by powers of z and summing, we find that

$$\mathsf{P}(z) = \mathbf{P}_0\mathsf{B}(z) + z^{-1}[\mathsf{P}(z) - \mathbf{P}_0]\mathsf{A}(z) \tag{8.50}$$

After rearranging terms and substituting (8.43), we have

$$\mathsf{P}(z)[zI - \mathsf{A}(z)] = -\mathbf{P}_0 D_0^{-1} D(z)\mathsf{A}(z) \qquad (8.51)$$

These terms can be substituted directly into (8.49) to find the probability of the various level probability vectors in terms of the empty probability vector.

8.3.3 Application of Renewal Theory

In this subsection we complete the final step in the solution to the $M/G/1$ queue with the BMAP input: the derivation of an expression for $\mathbf{P}_0$. The derivation applies a basic result in *renewal theory.* The busy period of the $M/G/1$, which we studied in Section 6.1.5, exemplifies the *first passage time*, which is a basic concept in renewal theory. The duration of a busy period is the interval between the system's departure of the system from the empty state and its first return to the empty state after serving messages.

The Fundamental Period The state of the $M/G/1$ queue is the pair $(i, j;\ i = 0, 1, 2, \ldots, j = 1, 2, \ldots, L)$, where i indicates the number of messages in the system, and j the phase of the arrival process. We define the L-dimensional level vector, as a vector of all states at level i and any of the L phases. An important property of the processes following the $M/G/1$ paradigm is what is called *left skip-free for levels.* Since there is a single server, messages depart singly, levels move downward one step at a time.[5] A number of messages may arrive between imbedded points; accordingly, the level may jump several steps. Left skip-free for levels implies that, in going from level $i + r$ to i, the process must visit each of the levels $i + r - 1, i + r - 2, \ldots, i + 1$ at least once. Figure 8.2 illustrates the concept of the first passage for $r = 3$.

We now return to the concept to the *virtual busy period* in the $M/G/1$ paradigm, which we saw in Section 6.5.3 in connection with polling. Suppose that the state of the system is at level i. Suppose also that there is a batch arrival of r messages with the arrival process moving to phase j. Our interest is in the interval between this batch arrival and the return to level i. Note that there may be any number of arrivals in this interval, which the system processes along with the initial set of r messages. We define $T(i + r, j;\ i, k)$ as the time required to pass from state $(i + r, j)$ to state (i, k) for the first time. Note that the beginning and ending phases of the arrival process are stipulated here. In the case of the simple $M/G/1$ queue, $T(i + r, j;\ i, k)$ is just the duration of an r busy period, that is, a busy period that starts with r messages. We define the quantity $V(i + r, j;\ i, k)$ as the number of transitions that take place during this first passage. Of course, the minimum number for this quantity is r, corresponding to a straight downward passage, with no upward movement.

[5] We have used this property in Sections 5.3.1 and 5.5.2.

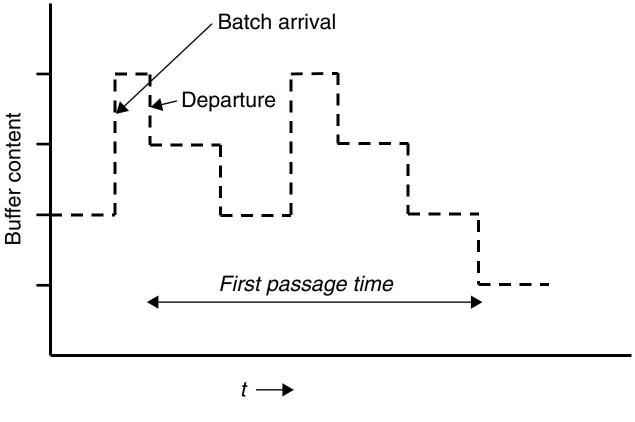


Figure 8.2 Illustration of first passage time.

The processes that we consider are *spatially homogeneous* inasmuch as $T(i+1, j; i, k)$ is the same for all $i > 0$; in other words, the time required to decrease one level is independent of the starting point. This time required to decrease one level is known as the *fundamental period.* Because of spatial homogeneity, the transition time between any pair of levels is simply the sum of fundamental periods. We may write this as

$$
\begin{aligned}
T(i+r, j; i, k) &= \sum_{n=1}^{r} T(i+n,\ \phi_{i+n};\ i+n-1,\ \phi_{i+n-1}) \\
V(i+r, j; i, k) &= \sum_{n=1}^{r} V(i+n,\ \phi_{i+n};\ i+n-1,\ \phi_{i+n-1})
\end{aligned}
\tag{8.52}
$$

where ϕ_{i+r-n}; $n = 0, 1, \ldots, r$ is the sequence of phases encountered in passing through the sequence of levels, with $\phi_i = j$ and $\phi_{i+r} = k$. The quantity that ties the parts of a path together is the sequences of phases since the duration of a fundamental period may depend on the initial and final phases. However, if we condition on the sequence of phases, the Markovian nature of the process ensures that the sequence of transition times and number of transitions are mutually independent. We define the joint probability of the duration of a fundamental period and the number of transitions in that fundamental period as $G_{jk}(m, t) = P(V(i+1, j; i, k) = m,\ T(i+1, j; i, k) \leq t)$. The corresponding density function is $g_{jk}(m, t) = dG_{jk}(m, t)/dt$. We define the $(L \times L)$ matrix $\mathbf{g}(m, t) = \{g_{jk}(m, t)\}$. Note that the row and column of this matrix indicate, respectively, the initial and ending phases of the fundamental period.

By means of the conditioning on phase that we mentioned above, we are able to construct the generalized busy period as the sum of independent random variables, which are the single steps down. As we have seen in Chapter 2, the density function of the sum of two independent random variables is found by a convolution operation.

We define the matrix convolution of two of these matrices as the matrix $g^{(2)}(x, m)$ with elements

$$\begin{aligned}\{g_{jk}^{(2)}(m, t)\} &= \mathbf{g}(m; x)^{*}\mathbf{g}(m; x)\\ &= \left\{\sum_{n=0}^{\infty}\sum_{l=1}^{L}\int_{0}^{\infty} d\tau g_{jl}(m-n, t-\tau)g_{lk}(n, \tau)\right\}\end{aligned} \tag{8.53}$$

where * denotes convolution. Notice the second summation here averages out the arrival phase. By repeated convolution, we find the probability for the r-step transition probability:

$$\mathbf{g}^{(r)}(m; x) = \underbrace{\mathbf{g}(m; x)^{*}\mathbf{g}(m; x)^{*}\cdots{}^{*}\,\mathbf{g}(m; x)}_{r \text{ times}} \tag{8.54}$$

We now write the relations between the arrival matrices and the probability distributions for the generalized busy period. There is only one transition between imbedded points if there are no arrivals during the first service time of the busy period; accordingly

$$\mathbf{G}(1, t) = \mathbf{A}_0(t) \tag{8.55}$$

where $\mathbf{A}_n(t)$ is as defined in (8.28). If there are n arrivals during the first service time of the busy period, each of these arrivals engenders a virtual busy period. Averaging over the arrival distribution, we have

$$\mathbf{g}(m, t) = \sum_{n=1}^{\infty}\mathbf{a}_n(t)^{*}\mathbf{g}^{(n)}(m-1, t) \tag{8.56}$$

where $\mathbf{a}_n(t)$ is as defined in (8.30). The next step in our derivation is to take transforms. We define the simultaneous z transforms in k and the Laplace transform in t as

$$\mathsf{G}(z, s) = \sum_{m=0}^{\infty} z^m \int_{0}^{\infty} e^{-st}\mathbf{g}(m, t)dt \tag{8.57}$$

Since convolutions imply products in the z and s domains, (8.57) is expressed using (8.56) as

$$\mathsf{G}(z, s) = z\sum_{n=0}^{\infty}\mathbf{A}_n(s)\mathsf{G}^n(z, s) \tag{8.58}$$

where the Laplace transform is given by $\mathsf{A}_n(s) = \int_0^\infty e^{-st}\mathbf{a}_n(t)dt$. Now, we let $z = 1$ and $s = 0$ in (8.58) to obtain

$$\mathbf{G} = \mathsf{G}(1, 0) = \sum_{n=0}^{\infty} \mathsf{A}_n(0)\mathsf{G}^n(1, 0) = \sum_{n=0}^{\infty} \mathbf{A}_n\mathbf{G}^n \tag{8.59}$$

For constant message transmission time and MMPP arrival, we obtain

$$\mathbf{G} = \sum_{n=0}^{\infty} \frac{\mathbf{G}^n e^{D_0 m}(D_1 m)^n}{n!} = e^{(D_0+\mathbf{G}D_1)m} \tag{8.60}$$

[see (8.32) and (8.37)]. It is not difficult to show that the solution to (8.59) is a stochastic matrix: $\mathbf{Ge} = \mathbf{e}$. We define the vector $\mathbf{g}$ as

$$\mathbf{gG} = \mathbf{g} \tag{8.61}$$

with the normalization $\mathbf{ge} = 1$.

Example 8.6 We do an example for the results derived in this subsection for the MMPP case introduced in Example 8.1. On the accompanying Matlab program, the iteration in (8.60) is carried out. With a threshold of 10^{-30}, we get the solution

$$\mathbf{G} = \begin{bmatrix} 0.97130 & 0.0287 \\ 0.05740 & 0.9426 \end{bmatrix}$$

in a single iteration. The solution to (8.61), gives a value of $\mathbf{g} = [0.6667 \quad 0.3333]$ for this value of $\mathbf{G}$. If we increase to message length from $m = 0.03$ to 0.1 s, the load goes to 0.833 and the solution is

$$\mathbf{G} = \begin{bmatrix} 0.7351 & 0.2649 \\ 0.2279 & 0.7721 \end{bmatrix}$$

after 21 iterations.

Steady-State Queue Length Distribution at Message Departure The solution of the nonlinear matrix equation of (8.59) is the key to the matrix analytic technique. In general, it is solved numerically using iteration where one starts with an estimate of $\mathbf{G}$ that is substituted into the RHS of (8.59) to obtain a refined estimate of $\mathbf{G}$ and so on.

Assuming, for the moment, that the solution $\mathsf{G}(z, s)$ is available, the next step in the solution is to find the matrix $\mathbf{K}(m, t)$ whose (j, k) elements are the probabilities that the process starting in state $(0, j)$ (empty system) returns to $(0, k)$ in m transitions, in time less than or equal to t. By definition, the system states at an empty level. We derive the state transition matrix for the phases over a busy period. We write this as

$$K_{jk}(m, t) = P(V(0, j;\ 0, k) = m,\ T(0, j;\ 0, k) \leq t) \tag{8.62}$$

Contrast this with the definition of $G_{jk}(m, t)$.

The corresponding density function is $k_{ik}(m, t) = dK_{ik}(m, t)\,dt$. We define the simultaneous z and Laplace transforms:

$$\mathsf{K}(z\ ,s) = \sum_{m=0}^{\infty} z^m \int_0^{\infty} e^{-st}\mathbf{k}(m,\ t)dt \tag{8.63}$$

As in the derivation of (8.58), we have the sum of independent random variables. In the present case of an initial arrival to an empty system, we find

$$\mathsf{K}(z,\ s) = z\sum_{n=0}^{\infty} \mathsf{B}_n(s)\mathsf{G}^n(z,\ s) \tag{8.64}$$

where $\mathsf{B}_n(z,\ s) = \int_0^{\infty} e^{-st}\mathbf{b}_n(t)dt$. From the definition in (8.42), we have

$$\mathsf{K}(z,\ s) = z\sum_{n=0}^{\infty} \mathsf{B}_n(s)\mathsf{G}^n(z,\ s) = z\mathsf{B}(\mathsf{G}(z,\ s),\ s) \tag{8.65}$$

In Appendix 8A, it is shown that

$$\mathsf{K}(z,\ s) = [sI - D_0]^{-1}[D(\mathsf{G}(z,\ s)) - D_0] \tag{8.66}$$

where $D(\mathsf{G}(z,\ s)) = \sum_{n=0}^{\infty} D_n\mathsf{G}^n(z,\ s)$. For constant service time and MMPP arrival, this reduces to

$$\mathbf{K}(z,\ s) = [sI - D_0]^{-1}D_1\mathsf{G}(z,\ s) \tag{8.67}$$

Again setting $z = 1$ and $s = 0$ in (8.66), we find

$$\mathbf{K} = \mathsf{K}(1,\ 0) = -D_0^{-1}[D(\mathbf{G}) - D_0] = I - D_0^{-1}D(\mathbf{G}) \tag{8.68}$$

where $\mathbf{G}$ is the solution to (8.59). The $(L \times L)$ matrix $\mathbf{K}$ gives the phase transition probabilities in going from an empty system to an empty system. For an MMPP system with constant service time, we have the simple result

$$\mathbf{K} = -D_0^{-1}D_1\mathbf{G} \tag{8.69}$$

In this section we will use the results of the previous sections to calculate $\mathbf{P}_0$, the probability vector for an empty system. We use a basic result from renewal theory, which says that *the steady-state probability of any state is the inverse of the average number of transitions before a first return to the state*. The jth component of the column vector

$$\boldsymbol{\nu} = \left.\frac{d\mathsf{K}(z,\ 0)}{dz}\right|_{z=1}\mathbf{e} \tag{8.70}$$

gives the average number of transitions in traveling from phase j at zero level to a zero level. In Appendix 8A, it is shown that

$$\boldsymbol{\nu} = -D_0^{-1}[\mathbf{D} - \mathbf{D}(\mathbf{G}) + \mathbf{D}'(1)\mathbf{eg}][I - \mathbf{A} + (\mathbf{e} - \boldsymbol{\alpha})\mathbf{g}]^{-1}\mathbf{e} \tag{8.71}$$

where

$$\boldsymbol{\alpha} = \mathbf{e}\bar{\lambda}\bar{M} + (\mathbf{e}\boldsymbol{\pi} + D)^{-1}(\mathbf{A} - I)D'(1)\mathbf{e}$$

with $\mathbf{g}$ is as given by (8.61). For the case of MMPP and constant service time, these equations simplify to

$$\boldsymbol{\nu} = -D_0^{-1}D_1[I - \mathbf{G} + \mathbf{eg}][I - \mathbf{A} + (\mathbf{e} - \boldsymbol{\alpha})\mathbf{g}]^{-1}\mathbf{e} \tag{8.72}$$

where

$$\boldsymbol{\alpha} = \mathbf{e}\bar{\lambda}m + (\mathbf{e}\boldsymbol{\pi} + D)^{-1}(\mathbf{A} - I)D_1\mathbf{e}$$

The steady-state distribution of phases on a first return-to-zero level is a L-dimensional row vector, which is the solution of the equation

$$\boldsymbol{\kappa}\mathbf{K} = \boldsymbol{\kappa} \tag{8.73}$$

with the normalization condition $\boldsymbol{\kappa}\mathbf{e} = 1$. The average number of these transitions, averaged over the phases, is then $\boldsymbol{\kappa\nu}$. The probability of the zero level is then $1/\boldsymbol{\kappa\nu}$. The probability vector $\mathbf{P}_0$ can be found by weighing, according to the distribution of the phases. We find

$$\mathbf{P}_0 = \frac{\boldsymbol{\kappa}}{\boldsymbol{\kappa\nu}} \tag{8.74}$$

By substituting into (8.49), the steady-state distribution of the queue lengths can be found.

In evaluating Equation (8.49) numerically, it is frequently found to be unstable, as it involves subtractions. Therefore, an alternative method that is numerically stable suggested by Ramaswami (1988) is commonly used. Using this method, we can calculate $\mathbf{P}_k$ values for $k \geq 1$, using the following recursive relations

$$\mathbf{P}_k = \mathbf{P}_0\overline{B_k} - \sum_{j=1}^{k-1}\mathbf{P}_j\overline{A_{k+1-j}})(I - \overline{A_1})^{-1}; \quad k = 1, 2, 3, \ldots \tag{8.75}$$

where $\overline{B}_k = \sum_{i=k}^{\infty} B_i G^{i-k}$ and $\overline{A}_k = \sum_{i=k}^{\infty} A_i G^{i-k}$ for $k \geq 0$.

Example 8.7 We continue the previous example. Recall that we found

$$\mathbf{G} = \begin{bmatrix} 0.9713 & 0.0287 \\ 0.0574 & 0.9426 \end{bmatrix}$$

and the solution to (8.61), which gives a value of $\mathbf{g} = [0.6667 \quad 0.3333]$. The next step is to calculate K, which, in this case, is given by (8.69). For $m = 0.03$ s

$$\mathbf{K} = \begin{bmatrix} 0.9104 & 0.0896 \\ 0.3011 & 0.6989 \end{bmatrix}$$

(Note that, as it should be, the matrix is stochastic.) The solution to (8.73) is $\boldsymbol{\kappa} = [0.7706 \quad 0.2294]$. The value of $\boldsymbol{\nu}$ as determined in (8.72) is

$$\boldsymbol{\nu} = \begin{bmatrix} 1.3843 \\ 1.2568 \end{bmatrix}$$

Finally, we have from (8.74), $\mathbf{P}_0 = [0.5687 \quad 0.1693]$.

We also perform the same computation with $m = 0.05$ s. We get $\mathbf{P}_0 = [0.3942 \quad 0.1613]$. In the second case, the load is 41.67% compared to the load of 25% for the case of $m = 0.03$ s. Therefore, we see that the probability of the system being empty is lower for higher load. Finally, we substitute into (8.49) to find the steady-state distribution of queue lengths at message departure times. For example, we get $\mathbf{P}_1 = [0.252 \quad 0.0498]$, $\mathbf{P}_2 = [0.0925 \quad 0.0097]$, and so on. We have used the recursion proposed by Ramaswami (1988).

The matrix-geometric technique that we have discussed so far has been applied to a wide range of queueing models, especially in the context of high speed multiplexing and switching networks in a bursty traffic environment. For the case of analysis of low-priority bursty traffic sharing capacity with high-priority bursty traffic as applied to the case of WATM uplink multiplexing system see Ganesh Babu et al. 2001. This analysis has also been applied to the case of output port of the switch after modeling the arrival process to the output port through the switch (see Ganesh Babu et al. 2003).

8.3.4 Moments at Message Departure

Equations (8.50) and (8.51) can be used to find the moments of the number of messages in the queue. We begin by noting the identity $\mathsf{P}(1)\mathbf{e}\boldsymbol{\pi} = \boldsymbol{\pi} = \boldsymbol{\pi}[I - \mathbf{A} + \mathbf{e}\boldsymbol{\pi}]$, where $\boldsymbol{\pi}$ is as given in (8.35). This identity follows from $\mathsf{P}(1)\mathbf{e} = 1$ and $\boldsymbol{\pi}\mathbf{e} = 1$. Setting $z = 1$ in (8.51) and adding the identity to both sides, we find that

$$\mathsf{P}(1)[I - \mathbf{A} + \mathbf{e}\boldsymbol{\pi}] = -\mathbf{P}_0 D_0^{-1} D\mathbf{A} + \boldsymbol{\pi}[I - \mathbf{A} + \mathbf{e}\boldsymbol{\pi}]$$

Consequently

$$\mathsf{P}(1) = \boldsymbol{\pi} - \mathbf{P}_0 D_0^{-1} D\mathbf{A}[I - \mathbf{A} + \mathbf{e}\boldsymbol{\pi}]^{-1} \tag{8.76}$$

We use this vector immediately. The first moment of the number of messages in the system at message departure is given by $\mathsf{P}^{(1)}(1)\mathbf{e}$, where the superscripts indicate orders of differentiation. This quantity can be found by differentiation of (8.51). The result is

$$\begin{aligned}\mathsf{P}^{(1)}(1)\mathbf{e} = \frac{1}{2(1-\rho)}\{&\mathsf{P}(1)\mathsf{A}^{(2)}(1)\mathbf{e} + \mathsf{U}^{(2)}(1)\mathbf{e}\\ &+2[\mathsf{U}^{(1)}(1) - \mathsf{P}(1)[I - \mathsf{A}^{(1)}(1)]][I - \mathbf{A} + \mathbf{e}\boldsymbol{\pi}]^{-1}\boldsymbol{\alpha}\}\end{aligned} \tag{8.77}$$

and

$$\begin{aligned}&\mathsf{P}^{(2)}(1)\mathbf{e}\\ &= \frac{1}{3(1-\rho)}\left\{\begin{matrix}3\mathsf{P}^{(1)}(1)\mathsf{A}^{(2)}(1)\mathbf{e} + \mathsf{P}(1)\mathsf{A}^{(3)}(1)\mathbf{e} + \mathsf{U}^{(3)}(1)\mathbf{e}+\\ 3[\mathsf{U}^{(2)}(1) + \mathsf{P}(1)\mathsf{A}^{(2)}(1) - 2\mathsf{P}^{(1)}(1)[I - \mathsf{A}^{(1)}(1)]][I - \mathbf{A} + \mathbf{e}\boldsymbol{\pi}]^{-1}\boldsymbol{\alpha}\end{matrix}\right\}\end{aligned} \tag{8.78}$$

where we have defined the term $\mathsf{U}(z) = -\mathbf{P}_0\mathbf{D}_0^{-1}\mathbf{D}(z)\mathsf{A}(z)$.

Example 8.8 Once again, we continue the previous example with $m = 0.05$ s. Corresponding to Equation (8.76), we get $\mathsf{P}(1) = [0.7701 \quad 0.2299]$. Corresponding to Equations (8.77) and (8.78), we get $\mathsf{P}^{(1)}(1)\mathsf{e} = 0.6413$ and $\mathsf{P}^{(2)}(1)\mathbf{e} = 0.3564$. Note that in solving for $[\mathsf{A} \quad \mathsf{A}^{(1)}(1) \quad \mathsf{A}^{(2)}(1) \quad \mathsf{A}^{(3)}(1)]$, we use a recursion suggested in Lucantoni (1991).

8.3.5 Steady-State Queue Length at Arbitrary Points in Time

Since the arrival process is not Poisson we cannot invoke the PASTA property to find the probability distribution of messages at a random point in time. This result is gained by an application of renewal theory, which is beyond the scope of our text; accordingly, we simply give the results without proof. We denote $\hat{\mathsf{P}}(z)$ as the probability distribution of the number of messages in the queue at an arbitrary point in time. It can be shown (Neuts 1989) that the probability generating function is given by

$$\hat{\mathsf{P}}(z) = \bar{\lambda}(z-1)\mathsf{P}(z)D^{-1}(z) \tag{8.79}$$

with the normalization

$$\hat{\mathsf{P}}(1) = \boldsymbol{\pi} \tag{8.80}$$

By equating coefficients of like powers of z, we find for the individual probabilities in terms of the probabilities at message departure:

$$\hat{\mathbf{P}}_0 = -\boldsymbol{\lambda}\mathbf{P}_0\mathbf{D}_0^{-1}$$

$$\hat{\mathbf{P}}_i = \left\{\sum_{j=0}^{i-1}\hat{\mathbf{P}}_j\mathbf{D}_{i-j} - \boldsymbol{\lambda}[\mathbf{P}_{i-1} - \mathbf{P}_i]\right\}[-\mathbf{D}_0^{-1}]; \quad i = 1, 2, \ldots \tag{8.81}$$

Example 8.9 Once again, we continue the previous example with $m = 0.05$ s. Corresponding to Equation (8.80), we get $\hat{\mathbf{P}}_0 = [0.3424 \quad 0.2409]$, $\hat{\mathbf{P}}_1 = [0.2164 \quad 0.0703]$, $\hat{\mathbf{P}}_2 = [0.0783 \quad 0.0136]$ and so on.

8.3.6 Moments of the Queue Length at Arbitrary Points in Time

By successive differentiation of (8.79), the first two moments of the queue length can be found as

$$\hat{\mathsf{P}}^{(1)}(1)\mathbf{e} = \mathsf{P}^{(1)}(1)\mathbf{e} - \frac{1}{2\boldsymbol{\lambda}}\boldsymbol{\pi}\mathbf{D}^{(2)}(1)\mathbf{e} + \frac{\boldsymbol{\pi}\mathbf{D}^{(1)}(1)}{\boldsymbol{\lambda} - \mathsf{P}(1)}(\mathbf{e}\boldsymbol{\pi} + \mathbf{D})^{-1}\mathbf{D}^{(1)}(1)\mathbf{e} \tag{8.82}$$

and

$$\hat{\mathsf{P}}^{(2)}(1)\mathbf{e} = \mathsf{P}^{(2)}(1)\mathbf{e} - \hat{\mathsf{P}}^{(1)}(1)\mathbf{D}^{(2)}(1)\frac{\mathbf{e}}{\boldsymbol{\lambda}} - \frac{1}{3\boldsymbol{\lambda}}\boldsymbol{\pi}\mathbf{D}^{(3)}(1)\mathbf{e}$$
$$- 2\left[\frac{\mathsf{P}^{(1)}(1) - \hat{\mathsf{P}}^{(1)}(1)\mathbf{D}^{(1)}(1)}{\boldsymbol{\lambda} - \boldsymbol{\pi}\mathbf{D}^{(2)}(1)/\boldsymbol{\lambda}}\right](\mathbf{e}\boldsymbol{\pi} + \mathbf{D})^{-1}\mathbf{D}^{(1)}(1)\mathbf{e} \tag{8.83}$$

Example 8.10 Once again, we continue the previous example with $m = 0.05$ s. Corresponding to Equations (8.82) and (8.83), we get $\hat{\mathbf{P}}^{(1)}(1)\mathbf{e} = 0.5914$ and $\hat{\mathbf{P}}^{(2)}(1)\mathbf{e} = 0.6998$.

8.3.7 Virtual Waiting Time

The delay suffered by a message that arrives at a random point in time is called the *virtual waiting time*. It can be shown that the Laplace transform of this delay is given by

$$W_V(s) = s\hat{\mathbf{P}}_0[sI + \mathbf{D}(M(s))]^{-1}\mathbf{e} \tag{8.84}$$

When the arrival process is pure Poisson, (8.84) reduces to (6.15) for the FCFS discipline.

The moments of the virtual waiting time can be found by successive differentiation. Define $V(s) = D(M(s))$ and the following moments obtained through differentiation:

$$
\begin{aligned}
&V^{(1)}(1) = -\bar{M}D^{(1)}(1)\\
&V^{(2)}(1) = (\bar{M})^2 D^{(2)}(1) + \bar{M}^{2} D^{(1)}(1)\\
&V^{(3)}(1) = -(\bar{M})^3 D^{(3)}(1) - 3\bar{M}\bar{M}^{2} D^{(2)}(1) - \bar{M}^{3} D^{(1)}(1) \qquad (8.85)
\end{aligned}
$$

Now, differentiation of (8.84) gives

$$
\begin{aligned}
W_V^{(1)}(0)\mathbf{e} = -\frac{1}{2(1-\rho)}[\rho + (\hat{\mathbf{P}}_0 - \boldsymbol{\pi}\mathsf{V}^{(1)}(1))[\mathbf{e}\boldsymbol{\pi} + D]^{-1}\mathsf{V}^{(1)}(1)\mathbf{e}\\
+ \boldsymbol{\pi}\mathsf{V}^{(2)}(1)\mathbf{e}] \qquad (8.86)
\end{aligned}
$$

and

$$
\begin{aligned}
W_V^{(2)}(0)\mathbf{e} = \frac{1}{3(1-\rho)}[(2W_V^{(1)}(0) + 2W_V^{(1)}(0)\mathsf{V}^{(1)}(1)\\
+ \boldsymbol{\pi}\mathsf{V}^{(2)}(1))[\mathbf{e}\boldsymbol{\pi} + D]^{-1}\mathsf{V}^{(1)}(1)\mathbf{e}\\
- 3W_V^{(1)}(0)\mathsf{V}^{(2)}(1)\mathbf{e} - \boldsymbol{\pi}\mathsf{V}^{(3)}(1)\mathbf{e}] \qquad (8.87)
\end{aligned}
$$

When the arrival process is pure Poisson, (8.86) simply reduces to the Pollaczek–Khinchin formula given by (6.12).

Example 8.11 Once again, we continue the previous example with $m = 0.05$ s. Corresponding to Equations (8.85) and (8.86), we get $W_V^{(1)}(0)\mathbf{e} = -0.0185$ and $W_V^{(2)}(0)\mathbf{e} = 0.0011$.

8.4 A MATCHING TECHNIQUE FOR LRD[6] TRAFFIC

8.4.1 *d* MMPPs and Equivalents

In this section we describe a method of fitting second-order statistics of LRD traffic over several timescales using superposition of MAP (Markov arrival process) models. We start with d two-state MMPPs with following parameter matrices:

$$
\begin{aligned}
D_0^i &= \begin{bmatrix} -(r_{12}^i + \lambda_1^i) & r_{12}^i \\ r_{21}^i & -(r_{21}^i + \lambda_2^i) \end{bmatrix}\\
D_1^i &= \begin{bmatrix} \lambda_1^i & 0 \\ 0 & \lambda_2^i \end{bmatrix}; \quad i = 1, 2, \ldots, d \qquad (8.88)
\end{aligned}
$$

[6] See Section 7.7 for an introduction to long range dependent (LRD) traffic.

In order to simplify expressions, we define $q_i = r_{12}^i + r_{21}^i$ and $p_i = (\lambda_1^i - \lambda_2^i)^2 \times r_{12}^i r_{21}^i/(q_i)^3$. Also, let $N_i(t)$ be the number of arrivals by time t from source i of the d sources. In Appendix 8B, it is shown that variance of $N_i(t)$ is given by[7]

$$\operatorname{Var}(N_i(t)) = (\lambda_i^* + 2p_i)t - \frac{2p_i(1 - e^{-q_i t})}{q_i}; \quad i = 1, 2, \ldots, d \tag{8.89}$$

where $\lambda_i^* = (r_{21}^i \lambda_1^i + r_{12}^i \lambda_2^i)/q_i$ is recognized as the average arrival rate of the ith MMPP. The net arrival rate of superposed process is given by $\lambda^* = \sum_{i=1}^d \lambda_i^*$.

We assume that the time axis is segmented into slots of duration Δt. The covariance function of number of arrivals in two slots separated by $(k - 1)$ slots is given by

$$C_i(k) = \frac{(\lambda_1^i - \lambda_2^i)^2 r_{12}^i r_{21}^i e^{-(q_i(k-1)\Delta t)}}{q_i^4}(1 - 2e^{-(q_i \Delta t)} + e^{-(2q_i \Delta t)}) \tag{8.90}$$

By assuming $q_i \Delta t << 1$, we can express, $[1 - 2e^{-(q_i \Delta t)} + e^{-(2q_i \Delta t)}] = (q_i \Delta t)^2 + O((q_i \Delta t)^2)$. Therefore (8.90) can be written as

$$C_i(k) \approx (\Delta t)^2 p_i q_i e^{-(q_i(k-1)\Delta t)} \tag{8.91}$$

Since the covariance of superposition of independent processes is equal to the sum of individual covariances of the processes, the covariance of the aggregate of d processes is just $\sum_{i=1}^d C_i(k)$.

The fitting algorithm is easier to explain if we replace the two-state MMPP of (8.88) by the sum of an *interrupted Poisson process* (IPP) and a simple Poisson process. The IPP alternates between the ON and OFF states with the same probabilities as the two-state MMPP. In the ON state, the Poisson rate is $(\lambda_1^i - \lambda_2^i)$, where we assume without loss of generality $\lambda_1^i > \lambda_2^i$. The simple Poisson process with arrival rate of λ_2^i. Since the sum of Poisson processes is Poisson, the aggregate of the IPP and the simple Poisson is the exactly same as the two-state MMPP and the covariance function given by (8.90) is the same. Clearly, the superposition of d two-state MMPPs can be considered as superposition of d two-state IPPs and d simple Poisson processes. For simplicity, we will describe this sum of processes as the *aggregate process*. Clearly, the total arrival rate from the aggregate simple Poisson process is just $\lambda_P = \sum_{i=1}^d \lambda_2^i$; consequently, the net arrival rate from the d IPPs is $\lambda^* - \sum_{i=1}^d \lambda_2^i$.

[7] Equation (8.89), as well as subsequent equations, were simplified by a Maple program (see Appendix 8B).

8.4.2 A Fitting Algorithm

In Section 7.7, we stated that the covariance function of LRD traffic is given asymptotically as $\psi_{\text{cov}}k^{-\beta}$, where ψ_{cov} is the variance and β is a parameter that lies between 0 and 1. As we have seen in Section 7.7, this parameter is related to the Hurst parameter H by $\beta^l = 2 - 2H$. Our objective is to fit $\sum_{i=1}^{d} C_i(k)$ to $\psi_{\text{cov}}k^{-\beta^l}$, such that

$$C(k) = \sum_{i=1}^{d} C_i(k) \cong \psi_{\text{cov}}k^{-\beta^l}; \quad 1 \leq k \leq 10^n \tag{8.92}$$

where n denotes the number of timescales over which we want the approximation to hold the covariance function values of $C(k)$ to $\psi_{\text{cov}}k^{-\beta^l}$.

The parameters that are in play in the fitting algorithm are λ_1^i, λ_2^i, r_{12}^i, r_{21}^i; $i = 1, 2, \ldots, d$, totaling $4d$. In order to simplify the algorithm, we make certain assumptions. The dynamics of the LRD process and the aggregate process can be matched over a number of timescales by making appropriate choices of phase transition parameters, r_{1i} and r_{2i}. For the fitting algorithm that we will describe, these parameters are chosen as

$$r_{12}^i = r_{21}^i = a^{1-i}r_{12}^1; \quad i = 1, 2, \ldots, d \tag{8.93}$$

where a, the *logarithmic scaling factor*, controls the range of the timescale. Since a is chosen so that $a >> 1$, the magnitude of the parameters increases with increase of the index i. Since $q_i = r_{12}^i + r_{21}^i$, the exponentials in (8.91) fall off more slowly as i is incremented. This observation is used to model over longer and longer timescales. The concept is basically the same as that of the previous chapter (see Fig. 7.8 and the explanation in text). With this assumption, all the r_{12}^i and r_{21}^i are determined by an initial value and a scaling factor. We begin the algorithm with an assumed value of q_1, which can be adjusted in subsequent iterations. Experience seems to suggest that an initial value of $q_i = 0.8$ leads to a good fit. We also choose an initial value for a that can be modified in the course of the fitting algorithm. Thus, we have reduced $4d$ parameters to $2d$, specifically, λ_1^i, λ_2^i; $i = 1, 2, \ldots, d$.

Fitting Algorithm Steps

Step 1 Given n, the number of timescales over which the fitting is carried out, and d, the number of MMPPs used to approximate the autocovariance function of the LRD process, we initialize the logarithmic scaling factor a in (8.93) as $a = 10^{n/(d-1)}$, which from (8.93) gives the ratio $r_{12}^d/r_{12}^1 = 10^n$. We choose n and d such that $a \geq 5$ or $n/(d-1) \geq 0.7$. because, as seen in (8.91), the covariance function value of IPP will remain fairly constant for all lags up to some k such that $q_i * k \cong 1$ and it rapidly decays by orders of magnitude lower at lags k' where $q_i * k' \cong 5$.

Step 2 Let $\lambda_i^{\text{IPP}} = (\lambda_1^i - \lambda_2^i)$. We outline the method used to calculate λ_i^{IPP} up to a normalizing constant; thus, we first find the quantities $\phi_i = \lambda_i^{\text{IPP}}/\alpha$; $i = 1, 2, \ldots, d$. Setting $\Delta t = 1$ in (8.91), we have

$$C_i(k) = \frac{(\lambda_i^{\text{IPP}})^2}{4} e^{-q_i(k-1)} \tag{8.94}$$

The covariance function is fitted at d different points, each defined by $q_i k_i = 1$; $i = 1, 2, \ldots, d$. Our objective is to fit $k^{-\beta} \cong \sum_{i=1}^{d} C_i(k)$ at these points, and we consider $\psi_{\text{cov}} = 1$. The idea is to first get the shape right, then to adjust the level. As in Section 7.7, the fitting algorithm is based on properties of exponentials. From the definition of q_i, we have $q_i/q_j = a^{i-j}$. Substuting into (8.94), we find

$$C_j(k_i) \cong 0; \quad i > j, \;\; a >> 1 \tag{8.95}$$

Phase 1 We start first with the dth IPP. The matching is in the slot k_d, defined as $k_d = 1/q_d$. Therefore, from (8.94), we can express the covariance function value as, $C_i(k) = [(\lambda_i^{\text{IPP}})^2/4]e^{-a^{d-i}}$. Now, as shown in (8.95), the contribution of other terms at k is negligible; accordingly, we have

$$k_d^{-\beta^l} = \sum_{i=1}^{d} \frac{(\lambda_i^{\text{IPP}})^2}{4} e^{-q_i(k-1)} \cong \frac{\alpha^2}{4}\phi_d^2 e^{-1+q_d} \cong \frac{\alpha^2}{4}\phi_d^2 e^{-1}$$

where α is a constant of proportionality that will be determined for a normalization procedure. The last approximation on the right here is due to $q_d = 1/k_d << 1$. Since λ_i^{IPP} values are represented up to a normalization constant, we solve for ϕ_d to obtain

$$\phi_d^2 \cong \frac{4ea^{-(d-1)\beta^l}}{\alpha^2} \tag{8.96}$$

Phase 2 For the next lower k such that $q_{d-1}k = 1$, where $k = a^{d-2}$ (the logarithmic spacing), we first need to express r_{1i} in terms of $r_{1(d-1)}$. By using the same logic as in phase 1, we can express r_{1i} as $r_{1i} = a^{d-i-1} r_{1(d-1)}$ $i = 1, 2, \ldots, d$. Therefore, the covariance function value at this k, can be given as

$$k_{d-1}^{-\beta^l} = \frac{\alpha^2}{4}\sum_{i=1}^{d} \phi_i^2 e^{-q_i(k_{d-1}-1)} \cong \frac{\alpha^2}{4}[\phi_d^2 e^{-a^{-1}+q_d} + \phi_{d-1}^2 e^{-1+q_{d-1}}]$$

$$\cong \frac{\alpha^2}{4}[\phi_d^2 e^{-a^{-1}} + \phi_{d-1}^2 e^{-1}]$$

where once again the contributions of other terms are negligible for the reason stated in step 1. Solving for $(\phi_{d-1})^2$, we find

$$\phi_{d-1}^2 = \frac{4ek_{d-1}^{-\beta^l}}{\alpha^2} - \phi_d^2 e^{1-a^{-1}} \tag{8.97}$$

We continue in this fashion through the remaining phases to the final phase d.

Phase d Finally, at the last step we have $q_1 k_1 = 1$ where $k = 1$, we have $r_{1i} = a^{1-i} r_{11}$ and following a similar argument, we can write

$$k_1^{-\beta^l} = \frac{\alpha^2}{4}\sum_{i=1}^{d} \phi_i^2 e^{-q_i(k_1-1)} = \frac{\alpha^2}{4}\sum_{i=1}^{d} \phi_i^2 e^{-a^{-(i-1)}+q_i} \cong \frac{\alpha^2}{4}\sum_{i=1}^{d} \phi_i^2 e^{-a^{-(i-1)}} \tag{8.98}$$

Finally, we solve for ϕ_1^2:

$$\phi_1^2 = \frac{4e^{-1}k_1^{-\beta^l}}{\alpha^2} - \sum_{i=1}^{d-1} \phi_i^2 e^{-a^{-(i-1)}} \tag{8.99}$$

It may happen that, at one of the phases, $1 \le k \le d$, we find $\phi_k^2 < 0$. In this case, we can simply set $\phi_k = 0$ and proceed with the calculation of other ϕ_i values. If the number of positive ϕ_i terms are less than some desired d^*, we increment d, thereby decreasing a, and repeat the algorithm starting at phase 1.

Step 3 In the previous step we determined λ_i^{IPP} up to a normalizing constant. Now, we find this normalization constant. For this purpose we use the correlation function value at lag 1, namely, γ, and the value λ^*. The value of ρ is expressed as

$$\gamma = \frac{\sum_{i=1}^{d} C_i(1)}{\sum_{i=1}^{d} \text{Var}(N_i(\Delta t))} \tag{8.100}$$

From (8.89) and (8.90) with $\Delta t = 1$, we have

$$\gamma = \frac{\alpha^2 \sum_{i=1}^{d} \phi_i^2 r_{12}^i r_{21}^i (q_i)^{-4}(1 - e^{-q_i})^2}{\lambda^* + \alpha^2 \sum_{i=1}^{d} 2\phi_i^2 r_{12}^i r_{21}^i (q_i)^{-4}(q_i - (1 - e^{-q_i}))}$$

If $2\rho \sum_{i=1}^{d} \phi_i^2 (q_i)^{-2}(q_i - (1 - e^{-q_i})) > \sum_{i=1}^{d} \phi_i^2 (q_i)^{-2}(1 - e^{-q_i})^2$, we cannot model with the given correlation. In that case we have to lower $r_{1i} = r_{2i} = q_i/2$

and/or lower ρ. Since, $r_{1i} = r_{2i} = q_i/2$, we have

$$\alpha = \sqrt{\frac{4\rho\lambda^*}{\sum_{i=1}^{d}(\phi_i)^2(q_i)^{-2}((1-e^{-q_i})^2 - 2\rho(q_i - (1-e^{-q_i})))}} \tag{8.101}$$

If $\lambda^* \geq \mu \sum_{i=1}^{d} \phi_i[r_{2i}/(r_{1i}+r_{2i})] = \mu \sum_{i=1}^{d}(\phi_i/2)$, we have a feasible solution. In that case, $\lambda_i^{\mathrm{IPP}} = \alpha\phi_i$ and $\lambda_P = \lambda^* - (\alpha/2)\sum_{i=1}^{d}\phi_i$, the total rate of the simple Poisson processes.

On the other hand, if $\lambda^* < \alpha \sum_{i=1}^{d} \phi_i/2$, we do not have a feasible solution. In that case, we can use the following method. Let us fix $p_i = (\alpha^2/4q_i)(\phi_i)^2$ and $q_i = r_{12}^i + r_{21}^i$. Let us assume that all IPPs have the same probability of being in the ON state, and we also assume $\lambda_P = 0$. Therefore, $p_{\mathrm{ON}} = r_{21}^i / r_{12}^i + r_{21}^i$ for all i. Thus, we can write

$$\frac{\alpha^2}{4}(\phi_i)^2 = \frac{r_{12}^i r_{21}^i}{(q_i)^2}(\lambda_i^{\mathrm{IPP}})^2 = p_{\mathrm{ON}}(1-p_{\mathrm{ON}})(\lambda_i^{\mathrm{IPP}})^2$$

Thus

$$\lambda_i^{\mathrm{IPP}} = \frac{\alpha}{2}\phi_i \frac{1}{p_{\mathrm{ON}}}\sqrt{\frac{p_{\mathrm{ON}}}{(1-p_{\mathrm{ON}})}}$$

Since $\lambda_P = 0$, we have

$$\lambda^* = \sum_{i=1}^{d} p_{\mathrm{ON}}\lambda_i^{\mathrm{IPP}} = \sum_{i=1}^{d}\frac{\alpha\phi_i}{2}\sqrt{\frac{p_{\mathrm{ON}}}{(1-p_{\mathrm{ON}})}}$$

From this, we can solve for p_{ON} and therefore, we get

$$p_{\mathrm{ON}} = \frac{(2\lambda^*)^2}{(2\lambda^*)^2 + (\alpha\sum_{i=1}^{d}(\phi_i)^2)}$$

Therefore

$$\lambda_i^{\mathrm{IPP}} = \phi_i \frac{(2\lambda^*)^2 + (\alpha\sum_{i=1}^{d}(\phi_i)^2)}{4\lambda^*\sum_{i=1}^{d}\phi_i} \tag{8.102}$$

Thus we can see that the fractional term on the RHS, ϕ_i, is the normalization constant α. Finally, we need the expressions for r^i_{12} and r^i_{21}:

$$r^i_{12} = \frac{\alpha \sum_{i=1}^{d} (\phi_i)^2}{(2\lambda^*)^2 + (\alpha \sum_{i=1}^{d} (\phi_i)^2)} q_i$$

$$r^i_{21} = \frac{(2\lambda^*)^2}{(2\lambda^*)^2 + (\alpha \sum_{i=1}^{d} (\phi_i)^2)} q_i \tag{8.103}$$

Example 8.12 As an example, we consider $\lambda^* = 8, n = 8, d = 6, \rho = 0.5, H = 0.75, q_1 = 0.8$ as input parameters. We can see that n and d has been chosen such that $a = 10^{1.6} = 39.81$. Running these steps as a Matlab program, we get the following results: the value of λ_P equal to 1.034×10^{-1}. We can verify that $\sum_{i=1}^{6} (\lambda_i^{\text{IPP}}/2) + \lambda_P$ (since $r_{1i} = r_{2i}$) is equal to the net arrival rate λ^*:

i	λ_i^{IPP}	$r^i_{12} = r^i_{21}$
1	9.535	4.0×10^{-1}
2	3.801	1.0×10^{-2}
3	1.505	2.524×10^{-4}
4	0.612	6.34×10^{-6}
5	0.224	1.592×10^{-7}
6	0.117	4.0×10^{-9}

APPENDIX 8A: DERIVATION OF SEVERAL BASIC EQUATIONS USED IN TEXT

In this appendix, we derive a number of relations that are useful in doing the calculations in the text. The manipulations are many and tedious; accordingly, in order to clarify the presentation, we mark the main results with the double angular brackets $\langle\langle\ \rangle\rangle$.

To begin, we prove (8.66): $\mathsf{K}(z, s) = [sI - D_0]^{-1}[D(\mathsf{G}(z, s)) - D_0]$. We begin by substituting (8.42) into (8.65). Note that in writing the term $\mathsf{B}(\mathsf{G}(z, s), s)$, we are careful regarding the order the matrices in the product:

$$\begin{aligned}\mathsf{K}(z, s) &= z\mathsf{B}(\mathsf{G}(z, s), s) = z[sI - D_0]^{-1}[D(\mathsf{G}(z,s)) - D_0]\mathsf{G}^{-1}(z, s)\tilde{\mathsf{A}}(\mathsf{G}(z, s), s)\\ &= z[sI - D_0]^{-1}[D(\mathsf{G}(z,s)) - D_0]\mathsf{G}^{-1}(z,s)\sum_{n=0}^{\infty} \tilde{\mathsf{A}}_n \mathsf{G}^n(z,s)\\ &= [sI - D_0]^{-1}[D(\mathsf{G}(z,s)) - D_0]\end{aligned}$$

After applying (8.58), we have

$$\langle\langle \mathsf{K}(z,s) = [sI - D_0]^{-1}[D(\mathsf{G}(z,s)) - D_0]\rangle\rangle$$

We now derive an expression for the vector $\boldsymbol{\alpha}$, whose jth component is the average number of arrivals during a service that starts with the arrival process in phase j. We begin with the definition

$$\left\langle\left\langle \boldsymbol{\alpha} = \sum_{n=1}^{\infty} nA_n\mathbf{e} = \left.\frac{d(\mathsf{A}(z,0))}{dz}\right|_{z=1}\mathbf{e}\right\rangle\right\rangle$$

[see (8.28)–(8.32)]. From (8.33), we have

$$\begin{aligned}\boldsymbol{\alpha} &= \left.\frac{d\left(\int_0^\infty e^{D(z)t}m(t)dt\right)}{dz}\right|_{z=1}\mathbf{e} \\ &= \int_0^\infty \left.\left(\frac{d\left(\sum_{i=0}^\infty D^i(z)t^i/i!\right)}{dz}\right)\right|_{z=1} m(t)dt\,\mathbf{e} \\ &= \int_0^\infty \sum_{i=1}^\infty \frac{t^i}{i!}\left(\sum_{j=1}^{i} D^{j-1}(1)D'(1)D^{i-j}(1)\right)\mathbf{e}\,m(t)dt\end{aligned}$$

Now, since $D(1)\mathbf{e} = 0$, only the term $i - j = 0$ survives in the second equation, and we have

$$\boldsymbol{\alpha} = \int_0^\infty \sum_{i=1}^\infty \frac{t^i}{i!}D^{i-1}(1)D'(1)\mathbf{e}\,m(t)dt = D^{-1}\int_0^\infty (e^{Dt} - I)D'(1)\mathbf{e}\,m(t)dt \qquad (8.104)$$

From (8.34), we obtain

$$\boldsymbol{\alpha} = D^{-1}(\mathbf{A} - I)D'(1)\mathbf{e} \Rightarrow D\boldsymbol{\alpha} = (\mathbf{A} - I)D'(1)\mathbf{e} \qquad (8.105)$$

Now, we go back to (8.104). Multiply both sides by $\mathbf{e}\boldsymbol{\pi}$ and use the identity $\boldsymbol{\pi}D = 0$ to obtain

$$\mathbf{e}\boldsymbol{\pi}\boldsymbol{\alpha} = \mathbf{e}\int_0^\infty t\boldsymbol{\pi}D'(1)\mathbf{e}\,m(t)dt$$

From the definition in (8.7) and from (8.10), we have

$$\mathbf{e}\boldsymbol{\pi}\boldsymbol{\alpha} = \mathbf{e}\,\bar{\lambda}\int_0^\infty tm(t)dt = \mathbf{e}\,\bar{\lambda}\bar{M} \tag{8.106}$$

Combining (8.105) and (8.106) gives

$$\mathbf{e}\boldsymbol{\pi}\boldsymbol{\alpha} + D\boldsymbol{\alpha} = \mathbf{e}\,\bar{\lambda}\bar{M} + (A - I)D'(1)\mathbf{e}$$

After rearranging terms, we have

$$\boldsymbol{\alpha} = (\mathbf{e}\boldsymbol{\pi} + D)^{-1}\mathbf{e}\,\bar{\lambda}\bar{M} + (\mathbf{e}\,\boldsymbol{\pi} + D)^{-1}(A - I)D'(1)\mathbf{e}$$

This expression is simplified by the relation $(\mathbf{e}\boldsymbol{\pi} + D)\mathbf{e} = \mathbf{e}$ to obtain

$$\langle\langle \boldsymbol{\alpha} = \mathbf{e}\bar{\lambda}\bar{M} + (\mathbf{e}\boldsymbol{\pi} + D)^{-1}(A - I)D'(1)\mathbf{e}\rangle\rangle$$

We now show that $\boldsymbol{\nu} = -D_0^{-1}[D - D(\mathbf{G}) + D'(1)\mathbf{eg}][I - A + (\mathbf{e} - \gamma)\mathbf{g}]^{-1}\mathbf{e}$. We begin with (8.58). Define

$$\left\langle\left\langle \boldsymbol{\gamma} = \frac{d\mathsf{G}(z,0)}{dz}\bigg|_{z=1}\mathbf{e}\right\rangle\right\rangle$$

Taking care with the order of matrix multiplication, we find that

$$\begin{aligned}
\boldsymbol{\gamma} &= \sum_{j=0}^{\infty}\mathbf{A}_j\mathbf{G}^j\mathbf{e} + \sum_{j=0}^{\infty}\mathbf{A}_j\sum_{k=0}^{j-1}\mathbf{G}^k\frac{d\mathsf{G}(z,0)}{dz}\bigg|_{z=1}\mathbf{G}^{j-k-1}\mathbf{e} \\
&= G\mathbf{e} + \sum_{j=0}^{\infty}\mathbf{A}_j\sum_{k=0}^{j-1}\mathbf{G}^k\frac{d\mathsf{G}(z,0)}{dz}\bigg|_{z=1}\mathbf{G}^{j-k-1}\mathbf{e} = \mathbf{e} + \sum_{j=0}^{\infty}\mathbf{A}_j\sum_{k=0}^{j-1}\mathbf{G}^k\frac{d\mathsf{G}(z,0)}{dz}\bigg|_{z=1}\mathbf{e} \\
&= \mathbf{e} + \sum_{j=0}^{\infty}\mathbf{A}_j\sum_{k=0}^{j-1}\mathbf{G}^k\boldsymbol{\gamma}
\end{aligned}$$

In the simplification here we use the relation $\mathbf{Ge} = \mathbf{e}$, which follows from the fact that $\mathbf{G}$ is a stochastic matrix. Thus

$$\boldsymbol{\gamma} = \left[I - \sum_{j=0}^{\infty}\mathbf{A}_j\sum_{k=0}^{j-1}\mathbf{G}^k\right]^{-1}\mathbf{e}$$

We can further simplify this expression with the following manipulation, with **g** as defined in (8.61). (We use (8.59) here).

$$\begin{aligned}\sum_{j=0}^{\infty}\mathbf{A}_j\sum_{k=0}^{j-1}\mathbf{G}^k[I-\mathbf{G}+\mathbf{eg}] &= \sum_{j=0}^{\infty}\mathbf{A}_j\left[\sum_{k=0}^{j-1}\mathbf{G}^k-\sum_{k=1}^{j}\mathbf{G}^k+\sum_{k=0}^{j-1}\mathbf{G}^k\mathbf{eg}\right]\\ &= \sum_{j=0}^{\infty}\mathbf{A}_j\left[I-\mathbf{G}^j+\sum_{k=0}^{j-1}\mathbf{G}^k\mathbf{eg}\right]\\ &= \mathbf{A}-\mathbf{G}+\sum_{j=0}^{\infty}j\mathbf{A}_j\mathbf{eg}=\mathbf{A}-\mathbf{G}+\boldsymbol{\alpha}\mathbf{g}\end{aligned}$$

or

$$\sum_{j=0}^{\infty}\mathbf{A}_j\sum_{k=0}^{j-1}\mathbf{G}^k=(\mathbf{A}-\mathbf{G}+\boldsymbol{\alpha}\mathbf{g})[I-\mathbf{G}+\mathbf{eg}]^{-1}$$

Continuing on, we find

$$\begin{aligned}\left[I-\sum_{\mathrm{j}=0}^{\infty}\mathbf{A}_j\sum_{k=0}^{j-1}\mathbf{G}^k\right]^{-1} &= \left[[(I-\mathbf{G}+\mathbf{eg})-\mathbf{A}+\mathbf{G}-\boldsymbol{\alpha}\mathbf{g}][I-\mathbf{G}+\mathbf{eg}]^{-1}\right]^{-1}\\ &= [I-\mathbf{G}+\mathbf{eg}][I+\mathbf{eg}-\mathbf{A}-\boldsymbol{\alpha}\mathbf{g}]^{-1}\end{aligned}$$

Consequently

$$\boldsymbol{\gamma}=(I-\mathbf{G}+\mathbf{eg})\mathbf{e}[I+\mathbf{eg}-\mathbf{A}-\boldsymbol{\alpha}\mathbf{g}]^{-1}$$

We will use this result in the next derivation.

From (8.70), we have the definition

$$\boldsymbol{\nu}=\left.\frac{d\mathsf{K}(z,\,0)}{dz}\right|_{z=1}\mathbf{e}$$

Further, as we have seen earlier in this appendix, we have

$$\mathsf{K}(z,s)=[sI-D_0]^{-1}[D(\mathsf{G}(z,s))-D_0]$$

Substituting, we find

$$\begin{aligned}
\boldsymbol{\nu} &= \left.\frac{d\mathsf{K}(z,0)}{dz}\right|_{z=1} \qquad \mathbf{e} = -D_0^{-1}\left.\frac{dD(\mathsf{G}(z,0)) - D_0}{dz}\right|_{z=1}\mathbf{e} \\
&= -D_0^{-1}\sum_{j=0}^{\infty} D_j \left.\frac{d\mathbf{G}^j(z,0)}{dz}\right|_{z=1}\mathbf{e} \\
&= -D_0^{-1}\sum_{j=0}^{\infty} D_j \sum_{k=0}^{j-1}\mathbf{G}^k(z,0)\left.\frac{d\mathbf{G}(z,0)}{dz}\right|_{z=1}\mathbf{G}^{j-j-1}(z,0)\mathbf{e}
\end{aligned}$$

Since **G** is stochastic, we do the familiar simplification

$$\boldsymbol{\nu} = -D_0^{-1}\sum_{j=0}^{\infty} D_j \sum_{k=0}^{j-1}\mathbf{G}^k \left.\frac{d\mathbf{G}(z,0)}{dz}\right|_{z=1}\mathbf{e}$$

From the expression that we have derived for $\boldsymbol{\gamma}$, we have

$$\begin{aligned}
\boldsymbol{\nu} &= -D_0^{-1}\sum_{j=0}^{\infty} D_j \sum_{k=0}^{j-1}\mathbf{G}^k\boldsymbol{\gamma} \\
&= -D_0^{-1}\sum_{j=0}^{\infty} D_j \sum_{k=0}^{j-1}\mathbf{G}^k[I - \mathbf{G} + \mathbf{eg}][I - \mathbf{A} + (\mathbf{e} - \boldsymbol{\alpha})\mathbf{g}]^{-1}\mathbf{e} \\
&= -D_0^{-1}\sum_{j=0}^{\infty} D_j[I - \mathbf{G}^j + \sum_{k=0}^{j-1}\mathbf{eg}][I - \mathbf{A} + (\mathbf{e} - \boldsymbol{\alpha})\mathbf{g}]^{-1}\mathbf{e}
\end{aligned}$$

which easily reduces to

$$\langle\!\langle \boldsymbol{\nu} = -D_0^{-1}[D - D(\mathbf{G}) + D'(1)\mathbf{eg}][I - \mathbf{A} + (\mathbf{e} - \boldsymbol{\alpha})\mathbf{g}]^{-1}\mathbf{e} \rangle\!\rangle$$

APPENDIX 8B: DERIVATION OF VARIANCE AND COVARIANCE FUNCTIONS OF TWO-STATE MMPP

Our derivations can be applied for any general BMAP with D_0 and D_1. The problem of extending this to the case of finite D_i is left as an exercise for the readers. Note that for convenience, we do not retain the index i, which is used to represent different two-state MMPPs for representing the LRD traffic over several timescales.

The z transform of the number of arrivals by time t is given by the Equation (8.15):

$$\mathsf{Q}(z,t) = e^{D(z)t} = e^{\sum_{i=0}^{M} z^i D_i t} = \prod_{i=0}^{M} e^{z^i D_i\, t}$$

For the two-phase MMPP process we consider here (see Example 8.1), $D(z)$ for the general BMAP case is given by Equation (8.7). In Equation (8.17) we showed how we can get the probability of having n arrivals by time t, by successively differentiating $\mathsf{Q}(z,t)$ and setting $z = 0$ for the two-phase MMPP process. We can get the moments of number of arrivals by time t, by differentiating $\mathsf{Q}(z,t)$ and setting $z = 1$. Let $M(t)$ be given by

$$\begin{aligned} M(t) &= \left.\frac{d\left(e^{D(z)t}\right)}{dz}\right|_{z=1} = \left.\frac{d\left(\sum_{i=0}^{\infty} D^i(z)t^i/i!\right)}{dz}\right|_{z=1} \\ &= \sum_{i=1}^{\infty}\frac{t^i}{i!}\left(\sum_{j=0}^{i-1} D^j(1)D'(1)D^{i-1-j}(1)\right) \end{aligned} \tag{8.107}$$

Thus, Equation (8.10) can be represented in terms of $M(t)$ as

$$\bar{\lambda} = \boldsymbol{\pi}\, M(t)\mathbf{e}$$

Clearly, for the two-phase MMPP process, we have $D(z) = D_0 + zD_1$ and therefore $D'(1) = D_1$ and $D(1) = D_0 + D_1 = D$. With these simplified representations, we can write $M(t)$ as

$$M(t) = \sum_{i=1}^{\infty}\frac{t^i}{i!}\left(\sum_{j=0}^{i-1} D^j D_1 D^{i-1-j}\right)$$

In order to derive the variance and covariance functions, we need to evaluate both $\boldsymbol{\pi}\, M(t)$ and $M(t)\mathbf{e}$. Since $D\mathbf{e} = 0$, we can write

$$M(t)\mathbf{e} = \sum_{i=1}^{\infty}\frac{t^i}{i!}D^{i-1}D_1\mathbf{e}$$

By multiplying after D^{i-1} by $(D + \mathbf{e}\boldsymbol{\pi})(D + \mathbf{e}\boldsymbol{\pi})^{-1}$, we can write

$$M(t)\mathbf{e} = \sum_{i=1}^{\infty}\frac{t^i}{i!}D^{i-1}(D + \mathbf{e}\boldsymbol{\pi})(D + \mathbf{e}\boldsymbol{\pi})^{-1}D_1\mathbf{e}$$

which can be written as

$$M(t)\mathbf{e} = \sum_{i=1}^{\infty} \frac{t^i}{i!} D^i (D + \mathbf{e}\boldsymbol{\pi})^{-1} D_1 \mathbf{e} + \sum_{i=1}^{\infty} \frac{t^i}{i!} D^{i-1} \mathbf{e}\boldsymbol{\pi} (D + \mathbf{e}\boldsymbol{\pi})^{-1} D_1 \mathbf{e}$$

Since $D\mathbf{e} = 0$ and $\boldsymbol{\pi}(D + \mathbf{e}\boldsymbol{\pi})^{-1} = \boldsymbol{\pi}$, we can finally write

$$M(t)\mathbf{e} = (e^{Dt} - I)(D + \mathbf{e}\boldsymbol{\pi})^{-1} D_1 \mathbf{e} + t\mathbf{e}\boldsymbol{\pi} D_1 \mathbf{e} \tag{8.108}$$

clearly, $\boldsymbol{\pi}\, M(t)\mathbf{e}$ gives Equation (8.10), corresponding to the case of two-phase MMPP, since $\boldsymbol{\pi} D = 0$.

Our next step is to evaluate $\boldsymbol{\pi}\, M(t)$:

$$\boldsymbol{\pi}\, M(t) = \boldsymbol{\pi} \sum_{i=1}^{\infty} \frac{t^i}{i!} \left(\sum_{j=0}^{i-1} D^j D_1 D^{i-1-j} \right)$$

Since $\boldsymbol{\pi} D = 0$, we obtain

$$\boldsymbol{\pi}\, M(t) = \sum_{i=1}^{\infty} \frac{t^i}{i!} \boldsymbol{\pi}\, D_1 D^{i-1}$$

Again, multiplying after D^{i-1} by $(D + \mathbf{e}\boldsymbol{\pi})(D + \mathbf{e}\boldsymbol{\pi})^{-1}$, and simplifying, we get

$$\boldsymbol{\pi}\, M(t) = \boldsymbol{\pi} D_1 (e^{Dt} - I)(D + \mathbf{e}\boldsymbol{\pi})^{-1} + t\boldsymbol{\pi} D_1 \mathbf{e}\boldsymbol{\pi}$$

To find $\mathrm{Var}(N(t))$, where $N(t)$ is as defined in Equation (8.1), we need to find the second moment: $\mathrm{Var}(N(t)) = E(N^2(t)) - (E(N(t)))^2$. Let $M_2(t)$ be defined by

$$M_2(t) = \left. \frac{d^2(e^{D(z)t})}{dz^2} \right|_{z=1} = \left. \frac{d^2\left(\sum_{i=0}^{\infty} D^i(z) t^i / i!\right)}{dz^2} \right|_{z=1}$$

Clearly

$$\boldsymbol{\pi}\, M_2(t)\mathbf{e} = E(N^2(t)) - E(N(t))$$

Therefore, $M_2(t)\mathbf{e}$ is given by

$$M_2(t)\mathbf{e} = 2 \sum_{i=2}^{\infty} \frac{t^i}{i!} \left(\sum_{j=0}^{i-2} D^j D_1 D^{i-2-j} \right) D_1 \mathbf{e} \tag{8.109}$$

Note that in simplifying to Equation (8.109) we would have to differentiate $e^{D(z)t}$ twice and use the fact that $D''(z)|_{z=1} = 0$, since $D(z) = D_0 + zD_1$ for two-phase

MMPP and $D\mathbf{e} = 0$. Multiplying after D^{i-2-j} by $(D + \mathbf{e}\boldsymbol{\pi})(D + \mathbf{e}\boldsymbol{\pi})^{-1}$, we can write

$$\begin{aligned} M_2(t)\mathbf{e} &= 2\sum_{i=2}^{\infty} \frac{t^i}{i!} \left(\sum_{j=0}^{i-2} D^j D_1 D^{i-1-j} \right) (D + \mathbf{e}\boldsymbol{\pi})^{-1} D_1 \mathbf{e} \\ &\quad + 2\sum_{i=2}^{\infty} \frac{t^i}{i!} \left(\sum_{j=0}^{i-2} D^j D_1 D^{i-2-j} \right) \mathbf{e}\boldsymbol{\pi} (D + \mathbf{e}\boldsymbol{\pi})^{-1} D_1 \mathbf{e} \end{aligned} \tag{8.110}$$

We simplify the first part of the RHS of the equation and then the second part. Thus

$$\begin{aligned} &2\sum_{i=2}^{\infty} \frac{t^i}{i!} \left(\sum_{j=0}^{i-2} D^j D_1 D^{i-1-j} \right) (D + \mathbf{e}\boldsymbol{\pi})^{-1} D_1 \mathbf{e} \\ &\quad = 2\sum_{i=2}^{\infty} \frac{t^i}{i!} \left(\sum_{j=0}^{i-1} D^j D_1 D^{i-1-j} - D^{i-1} D_1 \right) (D + \mathbf{e}\boldsymbol{\pi})^{-1} D_1 \mathbf{e} \\ &\quad = 2\left(\sum_{i=1}^{\infty} \frac{t^i}{i!} \left(\sum_{j=0}^{i-1} D^j D_1 D^{i-1-j} \right) - tD_1 \right) (D + \mathbf{e}\boldsymbol{\pi})^{-1} D_1 \mathbf{e} \\ &\qquad - 2\sum_{i=2}^{\infty} \frac{t^i}{i!} D^{i-1} D_1 (D + \mathbf{e}\boldsymbol{\pi})^{-1} D_1 \mathbf{e} \end{aligned}$$

By using the definition of $M(t)$, we can continue

$$\begin{aligned} &= 2(M(t) - tD_1)(D + \mathbf{e}\boldsymbol{\pi})^{-1} D_1 \mathbf{e} \\ &\quad - 2\sum_{i=2}^{\infty} \frac{t^i}{i!} D^{i-1} (D + \mathbf{e}\boldsymbol{\pi})(D + \mathbf{e}\boldsymbol{\pi})^{-1} D_1 (D + \mathbf{e}\boldsymbol{\pi})^{-1} D_1 \mathbf{e} \\ &= 2(M(t) - tD_1)(D + \mathbf{e}\boldsymbol{\pi})^{-1} D_1 \mathbf{e} \\ &\quad - 2\left[\sum_{i=2}^{\infty} \frac{t^i}{i!} D^i (D + \mathbf{e}\boldsymbol{\pi})^{-1} D_1 (D + \mathbf{e}\boldsymbol{\pi})^{-1} D_1 \mathbf{e} \right. \\ &\quad \left. + \sum_{i=2}^{\infty} \frac{t^i}{i!} D^{i-1} \mathbf{e}\boldsymbol{\pi} (D + \mathbf{e}\boldsymbol{\pi})^{-1} D_1 (D + \mathbf{e}\boldsymbol{\pi})^{-1} D_1 \mathbf{e} \right] \\ &= 2(M(t) - tD_1)(D + \mathbf{e}\boldsymbol{\pi})^{-1} D_1 \mathbf{e} \\ &\quad - 2\left[(e^{Dt} - Dt - I)(D + \mathbf{e}\boldsymbol{\pi})^{-1} D_1 (D + \mathbf{e}\boldsymbol{\pi})^{-1} D_1 \mathbf{e} \right] \end{aligned}$$

since $D\mathbf{e} = 0$. Until now, we have simplified the first part of Equation (8.110). Now we have to simplify the second part of the RHS of the Equation (8.110):

$$\begin{aligned}
&2\sum_{i=2}^{\infty}\frac{t^i}{i!}\left(\sum_{j=0}^{i-2} D^j D_1 D^{i-2-j}\right)\mathbf{e}\boldsymbol{\pi}(D+\mathbf{e}\boldsymbol{\pi})^{-1}D_1\mathbf{e}\\
&\quad = 2\sum_{i=2}^{\infty}\frac{t^i}{i!}D^{i-2}(D+\mathbf{e}\boldsymbol{\pi})(D+\mathbf{e}\boldsymbol{\pi})^{-1}D_1\mathbf{e}\boldsymbol{\pi}D_1\mathbf{e}\\
&\quad = 2\sum_{i=2}^{\infty}\frac{t^i}{i!}D^{i-1}(D+\mathbf{e}\boldsymbol{\pi})^{-1}D_1\mathbf{e}\boldsymbol{\pi}D_1\mathbf{e}\\
&\qquad + 2\sum_{i=2}^{\infty}\frac{t^i}{i!}D^{i-2}\mathbf{e}\boldsymbol{\pi}(D+\mathbf{e}\boldsymbol{\pi})^{-1}D_1\mathbf{e}\boldsymbol{\pi}D_1\mathbf{e}\\
&\quad = 2\sum_{i=2}^{\infty}\frac{t^i}{i!}D^{i-1}(D+\mathbf{e}\boldsymbol{\pi})(D+\mathbf{e}\boldsymbol{\pi})^{-2}D_1\mathbf{e}\boldsymbol{\pi}D_1\mathbf{e} + t^2\mathbf{e}\boldsymbol{\pi}D_1\mathbf{e}\boldsymbol{\pi}D_1\mathbf{e}\\
&\quad = 2\sum_{i=2}^{\infty}\frac{t^i}{i!}D^{i}(D+\mathbf{e}\boldsymbol{\pi})^{-2}D_1\mathbf{e}\boldsymbol{\pi}D_1\mathbf{e}\\
&\qquad + 2\sum_{i=2}^{\infty}\frac{t^i}{i!}D^{i-1}\mathbf{e}\boldsymbol{\pi}(D+\mathbf{e}\boldsymbol{\pi})^{-2}D_1\mathbf{e}\boldsymbol{\pi}D_1\mathbf{e} + t^2\mathbf{e}\boldsymbol{\pi}D_1\mathbf{e}\boldsymbol{\pi}D_1\mathbf{e}\\
&\quad = 2(e^{Dt} - Dt - I)(D+\mathbf{e}\boldsymbol{\pi})^{-2}D_1\mathbf{e}\boldsymbol{\pi}D_1\mathbf{e} + t^2\mathbf{e}\boldsymbol{\pi}D_1\mathbf{e}\boldsymbol{\pi}D_1\mathbf{e}
\end{aligned}$$

since $D\mathbf{e} = 0$ and $\boldsymbol{\pi}(D+\mathbf{e}\boldsymbol{\pi})^{-1} = \boldsymbol{\pi}$. Therefore, Equation (8.109) can be written as

$$\begin{aligned}
M_2(t)\mathbf{e} &= 2(M(t) - tD_1)(D+\mathbf{e}\boldsymbol{\pi})^{-1}D_1\mathbf{e}\\
&\quad - 2[(e^{Dt} - Dt - I)(D+\mathbf{e}\boldsymbol{\pi})^{-1}D_1(D+\mathbf{e}\boldsymbol{\pi})^{-1}D_1\mathbf{e}]\\
&\quad + 2(e^{Dt} - Dt - I)(D+\mathbf{e}\boldsymbol{\pi})^{-2}D_1\mathbf{e}\boldsymbol{\pi}D_1\mathbf{e} + t^2\mathbf{e}\boldsymbol{\pi}D_1\mathbf{e}\boldsymbol{\pi}D_1\mathbf{e} \qquad (8.111)
\end{aligned}$$

By using $\boldsymbol{\pi}\ M(t)$ and $\boldsymbol{\pi}D = 0$, we can simplify $\boldsymbol{\pi}\ M_2(t)\mathbf{e}$ to

$$\begin{aligned}
\boldsymbol{\pi}\ M_2(t)\mathbf{e} &= 2\boldsymbol{\pi}D_1(e^{Dt} - I)(D+\mathbf{e}\boldsymbol{\pi})^{-2}D_1\mathbf{e}\\
&\quad + 2t\big(\boldsymbol{\pi}D_1\mathbf{e}\boldsymbol{\pi} - \boldsymbol{\pi}D_1(D+\mathbf{e}\boldsymbol{\pi})^{-1}\big)D_1\mathbf{e} + t^2\boldsymbol{\pi}D_1\mathbf{e}\boldsymbol{\pi}D_1\mathbf{e}
\end{aligned}$$

Finally $\mathrm{Var}(N(t))$ is given by $\boldsymbol{\pi}\, M_2(t)\mathbf{e} + E(N(t)) - (E(N(t)))^2$, where $E(N(t)) = \boldsymbol{\pi} M(t)\mathbf{e} = t\boldsymbol{\pi} D_1\mathbf{e}$, from Equation (8.107). Thus

$$\begin{aligned}\mathrm{Var}(N(t)) &= \boldsymbol{\pi} M_2(t)\mathbf{e} + E(N(t)) - (E(N(t)))^2\\ &= 2\boldsymbol{\pi} D_1(e^{Dt} - I)(D + \mathbf{e}\boldsymbol{\pi})^{-2} D_1\mathbf{e}\\ &\quad + 2t\big(\boldsymbol{\pi} D_1\mathbf{e}\boldsymbol{\pi} - \boldsymbol{\pi} D_1(D + \mathbf{e}\boldsymbol{\pi})^{-1}\big)D_1\mathbf{e} + t\boldsymbol{\pi} D_1\mathbf{e} \qquad (8.112)\end{aligned}$$

Using the accompanying Maple code, we can show that Equation (8.112) simplifies to Equation (8.89).

Our final step is to express the covariance function of Equation (8.90). Previously we derived the first and second moments of $N(t)$. But, now, in trying to determine the covariance function of number of arrivals in different time intervals, we need to identify the intervals themselves. Therefore, as explained earlier in this chapter, we consider a time slot Δt. Thus we want to determine the covariance function of the number of arrivals in time slots that are separated by, say, $(j-1)$ time slots for $j > 1$. In other words, we would like to evaluate the following expression [see Neuts (1979)]:

$$\begin{aligned}&\mathrm{Cov}(N(\Delta t), [N(j\Delta t) - N((j-1)\Delta t)])\\ &\quad = E[N(\Delta t) \times (N(j\Delta t) - N((j-1)\Delta t))]\\ &\qquad - [E(N(\Delta t)) \times E(N(j\Delta t) - N((j-1)\Delta t))]\end{aligned}$$

where

$$\begin{aligned}&E[N(\Delta t) \times (N(j\Delta t) - N((j-1)\Delta t))]\\ &\quad = \boldsymbol{\pi}\frac{d^2}{dz_1\, dz_2}\Big(E[z_1^{N(\Delta t)} z_2^{N(j\Delta t)-N((j-1)\Delta t)}]\Big)\Big|_{z_1=1,z_2=1}\mathbf{e}\\ &\quad = \boldsymbol{\pi}\frac{d^2}{dz_1\, dz_2}(e^{D(z_1)\Delta t} e^{D(j-1)\Delta t} e^{D(z_2)\Delta t})\Big|_{z_1=1,z_2=1}\mathbf{e}\\ &\quad = \boldsymbol{\pi} M(\Delta t) e^{D(j-1)\Delta t} M(\Delta t)\mathbf{e}\end{aligned}$$

From Equation (8.107) and the following derivation for $\boldsymbol{\pi} M(t)$, we can write

$$\begin{aligned}&E[N(\Delta t)(N(j\Delta t) - N((j-1)\Delta t))]\\ &\quad = [(\boldsymbol{\pi} D_1(e^{D\Delta t} - I)(D + \mathbf{e}\boldsymbol{\pi})^{-1}) + (\Delta t\boldsymbol{\pi} D_1\mathbf{e}\boldsymbol{\pi})]e^{D(j-1)\Delta t}\\ &\quad \times [((e^{D\Delta t} - I)(D + \mathbf{e}\boldsymbol{\pi})^{-1} D_1\mathbf{e}) + (\Delta t\mathbf{e}\boldsymbol{\pi} D_1\mathbf{e})]\end{aligned}$$

This expression can be represented as

$$\begin{aligned}
&E[N(\Delta t)(N(j\Delta t) - N((j-1)\Delta t))] \\
&\quad = (\boldsymbol{\pi} D_1(e^{D\Delta t} - I)(D + \mathbf{e}\boldsymbol{\pi})^{-1} e^{D(j-1)\Delta t}(e^{D\Delta t} - I)(D + \mathbf{e}\boldsymbol{\pi})^{-1} D_1 \mathbf{e}) \\
&\qquad + (\boldsymbol{\pi} D_1(e^{D\Delta t} - I)(D + \mathbf{e}\boldsymbol{\pi})^{-1} e^{D(j-1)\Delta t} \Delta t \mathbf{e}\boldsymbol{\pi} D_1 \mathbf{e}) \\
&\qquad + (\Delta t \boldsymbol{\pi} D_1 \mathbf{e}\boldsymbol{\pi} e^{D(j-1)\Delta t}(e^{D\Delta t} - I)(D + \mathbf{e}\boldsymbol{\pi})^{-1} D_1 \mathbf{e}) \\
&\qquad + (\Delta t \boldsymbol{\pi} D_1 \mathbf{e}\boldsymbol{\pi} e^{D(j-1)\Delta t} \Delta t \mathbf{e}\boldsymbol{\pi} D_1 \mathbf{e})
\end{aligned}$$

Let us simplify the four parts of this expression. The expressions $e^{D(j-1)\Delta t}$ and $(D + \mathbf{e}\boldsymbol{\pi})^{-1}$ commute with respect to multiplication. We prove this here. We want to prove that

$$(D + \mathbf{e}\boldsymbol{\pi})^{-1} e^{D(j-1)\Delta t} = e^{D(j-1)\Delta t}(D + \mathbf{e}\boldsymbol{\pi})^{-1}$$

$$(D + \mathbf{e}\boldsymbol{\pi})^{-1} e^{D(j-1)\Delta t} = (D + \mathbf{e}\boldsymbol{\pi})^{-1} \sum_{i=0}^{\infty} \frac{(D(j-1)\Delta t)^i}{i!}$$

We can see that $D(D + \mathbf{e}\boldsymbol{\pi})^{-1} = (D + \mathbf{e}\boldsymbol{\pi})^{-1} D$, because $(D + \mathbf{e}\boldsymbol{\pi})D = D^2$, which implies $D = (D + \mathbf{e}\boldsymbol{\pi})^{-1} D^2$, but $D(D + \mathbf{e}\boldsymbol{\pi}) = D^2$ also, therefore replacing D^2 by $D(D + \mathbf{e}\boldsymbol{\pi})$ in $D = (D + \mathbf{e}\boldsymbol{\pi})^{-1} D^2$, we prove the commutativity of $D(D + \mathbf{e}\boldsymbol{\pi})^{-1} = (D + \mathbf{e}\boldsymbol{\pi})^{-1} D$. Therefore it can be seen clearly that

$$\begin{aligned}
(D + \mathbf{e}\boldsymbol{\pi})^{-1} e^{D(j-1)\Delta t} &= (D + \mathbf{e}\boldsymbol{\pi})^{-1} \sum_{i=0}^{\infty} \frac{(D(j-1)\Delta t)^i}{i!} \\
&= \left(\sum_{i=0}^{\infty} \frac{(D(j-1)\Delta t)^i}{i!} \right) (D + \mathbf{e}\boldsymbol{\pi})^{-1} \\
&= e^{D(j-1)\Delta t}(D + \mathbf{e}\boldsymbol{\pi})^{-1}
\end{aligned}$$

Simplifying the first part, we get

$$\begin{aligned}
&\boldsymbol{\pi} D_1(e^{D\Delta t} - I)(D + \mathbf{e}\boldsymbol{\pi})^{-1} e^{D(j-1)\Delta t}(e^{D\Delta t} - I)(D + \mathbf{e}\boldsymbol{\pi})^{-1} D_1 \mathbf{e} \\
&\quad = \boldsymbol{\pi} D_1(e^{D\Delta t} - I) e^{D(j-1)\Delta t}(e^{D\Delta t} - I)(D + \mathbf{e}\boldsymbol{\pi})^{-2} D_1 \mathbf{e}
\end{aligned}$$

by using commutative property, that we established.

Simplifying the second part, we get

$$\begin{aligned}
&\boldsymbol{\pi} D_1(e^{D\Delta t} - I)(D + \mathbf{e}\boldsymbol{\pi})^{-1} e^{D(j-1)\Delta t} \Delta t \mathbf{e}\boldsymbol{\pi} D_1 \mathbf{e}\\
&\quad = \boldsymbol{\pi} D_1(e^{D\Delta t} - I)(D + \mathbf{e}\boldsymbol{\pi})^{-1} e^{D(j-1)\Delta t} \mathbf{e}\boldsymbol{\pi} D_1 \mathbf{e} \Delta t\\
&\quad = \boldsymbol{\pi} D_1(e^{D\Delta t} - I)(D + \mathbf{e}\boldsymbol{\pi})^{-1} \mathbf{e}\boldsymbol{\pi} D_1 \mathbf{e} \Delta t\\
&\quad = \boldsymbol{\pi} D_1(D + \mathbf{e}\boldsymbol{\pi})^{-1} (e^{D\Delta t} - I) \mathbf{e}\boldsymbol{\pi} D_1 \mathbf{e} \Delta t\\
&\quad = 0
\end{aligned}$$

since $D\mathbf{e} = 0$. Simplifying the third part, we get

$$\begin{aligned}
&\Delta t \boldsymbol{\pi} D_1 \mathbf{e}\boldsymbol{\pi} e^{D(j-1)\Delta t} (e^{D\Delta t} - I)(D + \mathbf{e}\boldsymbol{\pi})^{-1} D_1 \mathbf{e}\\
&\quad = \Delta t \boldsymbol{\pi} D_1 \mathbf{e}\boldsymbol{\pi} (e^{D\Delta t} - I)(D + \mathbf{e}\boldsymbol{\pi})^{-1} D_1 \mathbf{e}\\
&\quad = 0
\end{aligned}$$

since $\boldsymbol{\pi} D = 0$. Simplifying the fourth part, we get

$$\begin{aligned}
\Delta t \boldsymbol{\pi} D_1 \mathbf{e}\boldsymbol{\pi} e^{D(j-1)\Delta t} \Delta t \mathbf{e}\boldsymbol{\pi} D_1 \mathbf{e} &= \Delta t \boldsymbol{\pi} D_1 \mathbf{e}\boldsymbol{\pi} \Delta t \mathbf{e}\boldsymbol{\pi} D_1 \mathbf{e}\\
&= (\Delta t)^2 \boldsymbol{\pi} D_1 \mathbf{e}\boldsymbol{\pi} D_1 \mathbf{e}
\end{aligned}$$

Thus we have

$$\begin{aligned}
&E[N(\Delta t) \times (N(j\Delta t) - N((j-1)\Delta t))]\\
&\quad = \left(\boldsymbol{\pi} D_1(e^{D\Delta t} - I) e^{D(j-1)\Delta t} (e^{D\Delta t} - I)(D + \mathbf{e}\boldsymbol{\pi})^{-2} D_1 \mathbf{e}\right)\\
&\qquad + \left((\Delta t)^2 \boldsymbol{\pi} D_1 \mathbf{e}\boldsymbol{\pi} D_1 \mathbf{e}\right) \qquad (8.113)
\end{aligned}$$

Therefore

$$\begin{aligned}
&\mathrm{Cov}(N(\Delta t), [N(j\Delta t) - N((j-1)\Delta t)])\\
&\quad = E[N(\Delta t) \times (N(j\Delta t) - N((j-1)\Delta t))]\\
&\qquad - [E(N(\Delta t)) \times E(N(j\Delta t) - N((j-1)\Delta t))]\\
&\quad = \boldsymbol{\pi} D_1(e^{D\Delta t} - I) e^{D(j-1)\Delta t} (e^{D\Delta t} - I)(D + \mathbf{e}\boldsymbol{\pi})^{-2} D_1 \mathbf{e} \qquad (8.114)
\end{aligned}$$

since $E(N(\Delta t)) \times E(N(j\Delta t) - N((j-1)\Delta t)) = (\Delta t)^2 \boldsymbol{\pi} D_1 \mathbf{e} \boldsymbol{\pi} D_1 \mathbf{e}$. As in the case of Equation (8.111), we have written the Maple code, which shows that Equation (8.113) simplifies to Equation (8.90).

REFERENCES

Andersen, A. T. et al., "A Markovian approach for modeling packet traffic with long-range dependence," *IEEE J. Select. Areas Commun.* **16**(5) (June 1998).

Daigle, J. N., *Queueing Theory for Telecommunications*, Addison-Wesley, 1992.

Ganesh Babu, T. V. J., T. Le-Ngoc, and J. F., Hayes, "Performance of a priority based dynamic capacity allocation scheme for wireless ATM systems," *IEEE Journal on Selected Areas in Communications* **19**(2): February 2001, 355–369.

Ganesh Babu, T. V. J., T. Le-Ngoc, and J. F., Hayes, "Performance evaluation of a switch using priority based dynamic capacity allocation scheme," *IEEE Transactions on Communications*, **51**(8): 1399–1408, August, 2003.

Lucantoni, D. M., "New results on the single server queue with a batch Markovian arrival process," *Commun. Stat. Stochastic Models*, **7**(1): 1–46 (1991).

Nelson, R., *Probability, Stochastic Processes and Queueing Theory*, Springer-Verlag, 1995.

Neuts, M. F., "A versatile Markovian point process," *J. Appl. Probability* **16**: 764–779 (1979).

Neuts, M. F., *Matrix-Geometric Solutions in Stochastic Models: An Algorithmic Approach*, Johns Hopkins Univ. Press, Baltimore, 1981.

Neuts, M. F., *Structured Stochastic Matrices of the M/G/1 Type and Their Applications*, Marcel Decker, New York, 1989.

Ramaswami, V., "A stable recursion for the steady state vector in Markov chains of M/G/1 types," *Commun. Stat. Stochastic Models* **4**(1): 183–188 (1988).

EXERCISES

8.1 Find the infinitesimal generator matrix for three-phase process shown below in Figure 8.3.

8.2 By deriving the *Kolmogorov forward equation* for the BMAP, show that its infinitesimal generator matrix is given by (8.6).

8.3 **(a)** Assume the compound Poisson arrival process $P(\text{one message}) = 0.3$ and $P(\text{two messages}) = 0.7$, where the message transmission time is given by $m(t) = \delta(t - 0.04)$.

(b) Assume the same phase transition and message arrival rates as in Example 8.1. What are the values of the submatrices $D_0, D_1, \ldots, D_k, \ldots$ in the BMAP formulation?

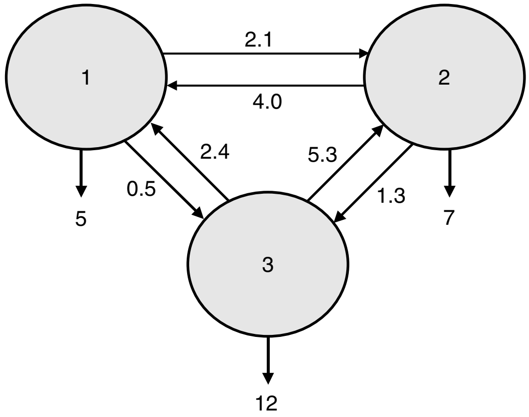


Figure 8.3

8.4 Repeat Exercise 8.3 for the phase transition and message arrival rates of Exercise 8.1.

8.5 **(a)** Find the fundamental arrival rate given by (8.10) for the MMPP process with the phase transition and message arrival rates given in Exercise 8.1.

(b) Find the load for the message length $m(t) = \delta(t - 0.02)$.

8.6 **(a)** Find the fundamental arrival rate given by (8.10) for the process given in Exercise 8.3.

(b) Find the load.

8.7 **(a)** For the process defined in Exercise 8.3, find $\mathbf{Q}(n, t)$ at $n = 0$.

(b) Find $\mathbf{Q}(n, t)$ at $n = 5$ and $t = 0.5$. (See Example 8.4.)

8.8 Repeat Exercise 8.7 for the process defined in Exercise 8.1.

8.9 Find $\mathbf{A}_n$, $\tilde{\mathbf{A}}_n$, and $\mathbf{B}_n$ assuming the same message distribution as in Example 8.3 and the phase transition and message arrival rates of Exercise 8.1.

8.10 Discuss the similarity between (8.58) and (6.49), the expression for the busy period of the M/G/1 queue.

8.11 Show that the solution to (8.59) is a stochastic matrix.

8.12 Show that (8.18) reduces to simple Poisson arrival when $L = 1$.

8.13 Find $\mathbf{G}$ and $\mathbf{g}$ for the process described in Exercise 8.5.

8.14 Repeat Exercise 8.13 for the process described in Exercise 8.3.

8.15 Find the steady-state distribution of queue lengths at message departure for the process described in Exercise 8.5.

8.16 Repeat Exercise 8.15 for the process described in Exercise 8.3.

8.17 Find the moments of queue lengths at message departure for the process described in Exercise 8.5.

8.18 Repeat Exercise 8.17 for the process described in Exercise 8.3.

8.19 Find the steady-state distribution of queue lengths at arbitrary points in time for the process described in Exercise 8.5.

8.20 Repeat Exercise 8.19 for the process described in Exercise 8.3.

8.21 Find the moments of queue lengths at arbitrary points in time for the process described in Exercise 8.5.

8.22 Repeat Exercise 8.21 for the process described in Exercise 8.3.

8.23 Find the virtual waiting time for the process described in Exercise 8.5.

8.24 Repeat Exercise 8.23 for the process described in Exercise 8.3.

8.25 **(a)** Show that the G/M/1 queue has the state transition matrix of the canonical form

$$\boldsymbol{P} = \begin{bmatrix} b_0 & a_0 & 0 & 0 & \cdots \\ b_1 & a_1 & a_0 & 0 & \cdots \\ b_2 & a_2 & a_1 & a_0 & \cdots \\ b_3 & a_3 & a_2 & a_1 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}$$

(b) Find an expression for the elements of the matrix.

8.26 Write a Matlab program to perform matching of LRD traffic to superposition of simple Poisson processes (SPPs) by using the algorithm provided in the text (Section 8.4.2), with the following input parameters: $\lambda^* = 6$, $n = 10$, $d = 6$, $\rho = 0.5$, $H = 0.8$, $q_1 = 0.8$. Verify that the average arrival rate of superposition of the SPPs equals λ^*.

8.27 Repeat exercise 8.26 with the following set of input parameters: $\lambda^* = 34$, $n = 8$, $d = 5$, $\rho = 0.4$, $H = 0.85$, $q_1 = 0.8$.

9

MONTE CARLO SIMULATION

9.1 SIMULATION AND STATISTICS

9.1.1 Introduction

As mentioned in the introductory chapter, simulation plays an important role in project development. In telecommunications systems, Monte Carlo simulation is the most prominent. As the name indicates, repeated probabilistic trials are the basis of this technique. Typically, there are exogenous input random variables that emulate such events as random message arrival and transmission times. Samples of a system's response to these inputs are taken, and estimates of system performance are calculated from these samples. Standard techniques for forming estimates of such quantities as mean response time are used.

9.1.2 Sample Mean and Sample Variance

In this section we review certain basic concepts on sampling and estimation that are relevant to simulation studies. For purposes of explanation, we focus on a particular example that is frequently encountered in telecommunications systems: message delay. Let $D_1, D_2, \ldots, D_n$ indicate a series of measurements of the interval between the arrival of messages to a multiplexer and their departure. We assume *stationarity*; that is, the underlying probability distribution from which these samples are drawn remains the same while the samples are being taken. We assume that the mean of the distribution is μ and its variance is σ^2. For example, in Section 3.4.3, it was shown

Modeling and Analysis of Telecommunications Networks, by Jeremiah F. Hayes and
Thimma V. J. Ganesh Babu
ISBN 0-471-34845-7

that, in an M/M/1 queueing system, the delay has an exponential distribution with mean $1/(\mu - \lambda)$, where $1/\mu$ is the mean message length and λ is the message arrival rate. For the moment, we will also assume that the samples are *independent* so that the value of one sample has no influence on any other sample. As we will see later in this chapter, when we deal with simulation, care must be taken to ensure that samples are independent.

The samples are used to calculate *statistics*, which are defined to be a number that is a function of samples. Statistics are designed to estimate aspects of the underlying probability distribution. The most common statistic is the *sample mean* or *arithmetic average*:

$$\bar{D} = \frac{1}{n}\sum_{i=1}^{n} D_i \tag{9.1}$$

This statistic is designed to estimate the mean value of the underlying distribution. Clearly, $\bar{D}$ is a random variable since it is the sum of random variables; thus, when the samples assume the values $d_1, d_2, \ldots, d_n$, the sample mean has the value $(1/n)\sum_{i=1}^{n} d_i$. As is the case for any random variable, the statistical properties of the sample mean are defined by its underlying probability distribution. The most common attribute of this distribution is the mean value:

$$E(\bar{D}) = E\left(\frac{1}{n}\sum_{i=1}^{n} D_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(D_i) = \frac{1}{n} n\mu = \mu \tag{9.2}$$

Since the mean of the sample mean is equal to the mean of the underlying distribution, the sample mean is called an *unbiased estimate*.

A second important property of the sample mean is its variance. Here we call on the independence assumption, since, as we have seen in Section 2.4.2, the variance of a sum of random variables is the sum of the variances. Therefore, we have

$$\mathrm{Var}(\bar{D}) = \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} D_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} \mathrm{Var}(D_i) = \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n} \tag{9.3}$$

The significance of this result is that the variance decreases as the number of samples increases, which is consistent with common experience.

The glaring flaw in the calculation of the preceding paragraph is that, we seldom know the variance of the underlying distribution. This quantity is estimated by the sample variance, which is defined to be

$$S^2 = \sum_{n=1}^{n} \frac{(D_i - \bar{D})^2}{n - 1} \tag{9.4}$$

The *sample standard deviation* is simply the square root of the sample variance, $S = \sqrt{S^2}$.

The sample variance is an unbiased estimate of the variance of the underlying distribution, since, $E(S^2) = \sigma^2$. This is shown by the following straightforward manipulation

$$
\begin{aligned}
&E\left(\frac{1}{n-1}\sum_{i=1}^{n}(D_i - \bar{D})^2\right) \\
&\quad = \frac{1}{n-1}E\left(\sum_{i=1}^{n}((D_i - \mu) - (\bar{D} - \mu))^2\right) \\
&\quad = \frac{1}{n-1}\left(E\left(\sum_{i=1}^{n}(D_i - \mu)^2\right) - 2E\left((\bar{D} - \mu)\sum_{i=1}^{n}(D_i - \mu)\right) + nE(\bar{D} - \mu)^2\right) \\
&\quad = \frac{1}{n-1}\left(\sum_{i=1}^{n}\underbrace{E[(D_i - \mu)^2]}_{\mathrm{Var}(D_i)} - n\underbrace{E[(\bar{D} - \mu)^2]}_{\mathrm{Var}(\bar{D})}\right) \\
&\quad = \frac{1}{n-1}\sum_{i=1}^{n}(\sigma^2 - n\mathrm{Var}(\bar{D})) = \frac{n\sigma^2 - n(\sigma^2/n)}{n-1} = \sigma^2
\end{aligned}
$$

By some further manipulation, we can find a formula that is a bit easier to calculate than (9.4):

$$
\begin{aligned}
S^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(D_i - \bar{D})^2 = \frac{\sum_{i=1}^{n}(D_i^2 - 2D_i\bar{D} + \bar{D}^2)}{n-1} \\
&= \frac{\sum_{i=1}^{n} D_i^2 - 2\bar{D}\sum_{i=1}^{n} D_i + m(\bar{D})^2}{n-1} = \frac{\sum_{i=1}^{n} D_i^2 - n(\bar{D})^2}{n-1} \qquad (9.5)
\end{aligned}
$$

For the estimates of the mean and the variance to be close to true value, a sufficient number of samples must be taken. This can be seen immediately from (9.3) for the sample mean, where the variance of the estimate of the mean is inversely proportional to the number of samples. Enough samples need to be taken for the sample standard deviation to be sufficiently low for credibility. The Chebyshev inequality, discussed in Section 2.6, would supply probabilistic limits. For many simulation studies, gathering a sufficient number of samples is not difficult.

9.1.3 Confidence Intervals

Given the sample mean and sample variance, we can give a probabilistic bound on the true mean of the distribution. Since the sample mean is the sum of independent random variables, we can evoke the *central-limit theorem* to treat the sample mean as a Gaussian random variable.[1] We also assume that we have enough sample values to take the sample variance to be the true variance. We find an interval such that the sample mean lies in it with probability $1 - \alpha$. From a standard table of normal distributions, we can find $z_{\alpha/2}$ such that

$$P(-z_{\alpha/2} < Z \leq z_{\alpha/2}) = \frac{1}{\sqrt{2\pi}} \int_{-z_{\alpha/2}}^{z_{\alpha/2}} e^{-z^2/2} dz = 1 - \alpha$$

where Z is a normally distributed random variable with mean 0 and variance 1. Now let

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

Substituting, we find

$$P\left(-z_{\alpha/2} < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq z_{\alpha/2}\right) = P\left(\bar{X} - \frac{z_{\alpha/2}\sigma}{\sqrt{n}} < \mu \leq \bar{X} + \frac{z_{\alpha/2}\sigma}{\sqrt{n}}\right) = 1 - \alpha \quad (9.6)$$

Thus, we see that, with probability $1 - \alpha$, the true mean of the distribution lies within the interval $(\bar{X} - z_{\alpha/2}\sigma/\sqrt{n}, \bar{X} + z_{\alpha/2}\sigma/\sqrt{n})$. Standard values for $z_{\alpha/2}$ are shown in Table 9.1.

Example 9.1 On the associated Excel spreadsheet, we simulate the generation of 10,000 samples from the $U(0, 1)$ distribution using the command `rand( )`. These numbers emulate samples drawn from a random variable uniformly distributed in the interval (0, 1). The sample mean and the estimate of the sample mean standard deviation are calculated for batches of 10 samples and 100 samples. The results are used to calculate the 99% confidence intervals, which are plotted in Figure 9.1. Clearly, there is less variation in the statistics when the sample size is increased by a factor of 10. The bounds move closer as the number of samples is increased.

9.1.4 Sample Sizes and Run Times

As (9.3) demonstrates, the variance of the estimate of the mean decreases inversely with the number of samples. Thus, given a confidence interval that we want to attain, we can directly calculate the number of samples that are required. From (9.6), the true mean lies in an interval $(-z_{\alpha/2}\sigma/\sqrt{n}, +z_{\alpha/2}\sigma/\sqrt{n})$ about the sample mean. Given the variance (or its estimate) and the level of confidence that we want, we can reduce this interval to any size just by increasing the number of samples. Of course, this is slow since the interval decreases as the square root of the number of samples.

[1] As a rule of thumb we take a "large number of samples" to be more than 10. For less than 10 samples, the Gaussian assumption is not valid and the *Student t* distribution must be used to find the confidence interval.

Table 9.1 $z_{\alpha/2}$ **Values Corresponding to Given (1 − α)**

$1-\alpha$	$z_{(\alpha/2)}$
0.9	1.645
0.95	1.96
0.99	2.576
0.999	3.291

For example, assume $\sigma = 10$ and that we want to reduce the interval to 0.1. A simple calculation shows that we need approximately 400 samples for 95% confidence. If we reduce the interval to 0.01, we need about 40,000 samples. This consideration motivates the variance reduction techniques that we will consider later.

There is an additional consideration when we consider the samples obtained from simulation. In (9.6) the assumption is that the samples are independent. If there were correlations among the samples, the variance of the sample mean will change. The following simple example illustrates this. We have two samples from a random variable with zero mean. The variance of the sample mean for just two samples is given by

$$E\left(\left(\frac{1}{2}(X_1+X_2)\right)^2\right) = \frac{1}{4}(2\sigma^2 + 2E(X_1X_2)) = \frac{\sigma^2 + E(X_1X_2)}{2}$$

Depending on the sign of $E(X_1X_2)$, the variance of the sample mean is increased or decreased. We will return to this in Section 9.4 when we deal with variance reduction techniques.

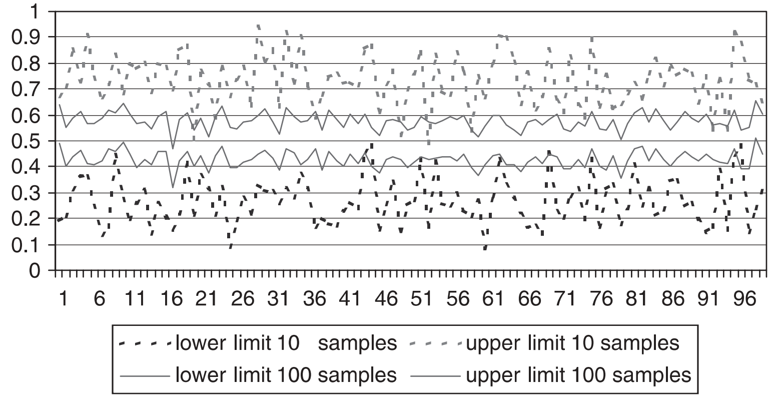


Figure 9.1 99% confidence intervals.

In simulations there is certainly no guarantee that samples are independent. For example, successive measurements of message delay in a queue will be highly correlated. How, then, do we know how long to run a simulation in order to obtain results with a required degree of confidence? For complex systems there are no simple answers, but there are techniques that can help. We look at three of these. One can do a number of independent runs choosing a different seed for each run. The sample mean is calculated by averaging over the sample means obtained for each run. This is called the method of *independent replications*. In another approach, a long sequence of samples is separated into batches whose lengths are such that the correlation between the first sample of a batch and the first sample of the next batch is close to zero. The samples means from the batches are averaged to obtain an overall sample mean. This is the *batch means* method. Finally, one can take advantage of the fact that interval separated by a return of a system to the zero state are independent. This is called the *regenerative* method. For a detailed discussion, the reader is referred to Lavenberg (1983).

9.1.5 Histograms

While the sample mean and the sample variance give useful information about the underlying distribution, in certain situations one wishes direct estimates of the density or the distribution. This is the function of the *histogram*. Consider first a discrete random variable. We assume that n samples are drawn from a distribution. Since the distribution is discrete, each sample must fall in the countable set of values assumed by the random variable. For example, samples from the binomial distribution must fall in the range 0 to N, where N is the number of trials. For the Poisson and the Geometric distribution, any nonnegative integer is a valid sample. Over the range of samples, we count the number of times each possible value appears, $n_0, n_1, \ldots, n_i, \ldots, n_{\max}$ where $n_{\max}$ is the maximum allowed value of the random variable. Notice that $\sum_{i=0}^{\max} n_i = n$, the total number of samples.

Example 9.2 An elementary example is shown on the associated Excel spreadsheet. The command `rand( )` generates a random variable that is uniformly distributed in the interval (0, 1). The operation `5` × `rand( )` + `1` yields a random variable that is uniform in the interval (1, 6).[2] The command `int(5` × `rand( ))` + `1` generates a discrete uniform distribution on the integers 1 to 6. A set of 20 samples is shown in column A of the spreadsheet. The count of each value is shown in columns B–F.

The *relative frequency* of the different sample values is given by

$$f_i = \frac{n_i}{n}; \quad i = 0, 1, 2, \ldots, \max \tag{9.7}$$

These are calculated in cells D1–D6 and plotted in Figure 9.2. This is referred to as the *relative frequency histogram*, also the *probability histogram* since it estimates probability (see below).

[2] Transformations are covered in Section 9.2.2.

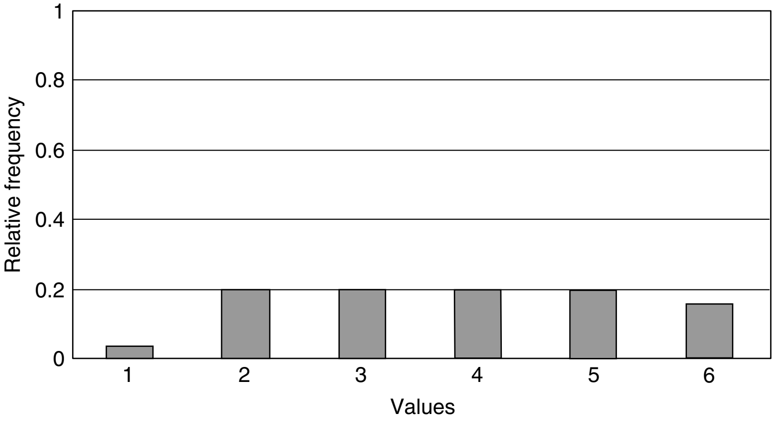


Figure 9.2 Relative frequency histogram.

Histograms are also useful in the estimation of continuous distributions. In this case, the range of possible outcomes is segmented into intervals, and the number of samples falling in each interval is counted. The intervals, called *class intervals* or *cells*, need not be equally spaced. Let the intervals be denoted as (t_{i-1}, t_i); $i = 1, 2, \ldots,$ max, and let the number of sample falling in the ith interval be denoted as n_i; $i = 1, 2, \ldots,$ max. The relative frequency is given by (9.7), where n is once again the total number of samples. For the case of a continuous distribution, we define

$$h_i = \frac{f_i}{t_i - t_{i-1}}; \quad i = 1, 2, \ldots, \text{ max} \tag{9.8}$$

The h_i are plotted against the class intervals to form the relative frequency histogram. The difference in the plots between discrete and continuous random variables is similar to the different ways in which these variables are represented. For a discrete random variable, the definition is a distribution that is a probability, whereas a continuous distribution is defined by a density, that is not a probability. When the class intervals are equally spaced, this distinction becomes less important since the shapes will be unaffected.

Example 9.3 On the associated Matlab program, a histogram for a random variable that is uniformly distributed in the interval $(-3, 8)$, is calculated. The first command in the program creates a vector with 15,000 random numbers uniformly distributed in $(-3, 8)$. The hist command counts the number of samples in each of 100 equal cells and plots them to form the histogram. The results are shown on Figure 9.3. Notice that the bars seem to center on 150, which is the value of the mean number of sample that would fall in a cell for a uniform distribution.

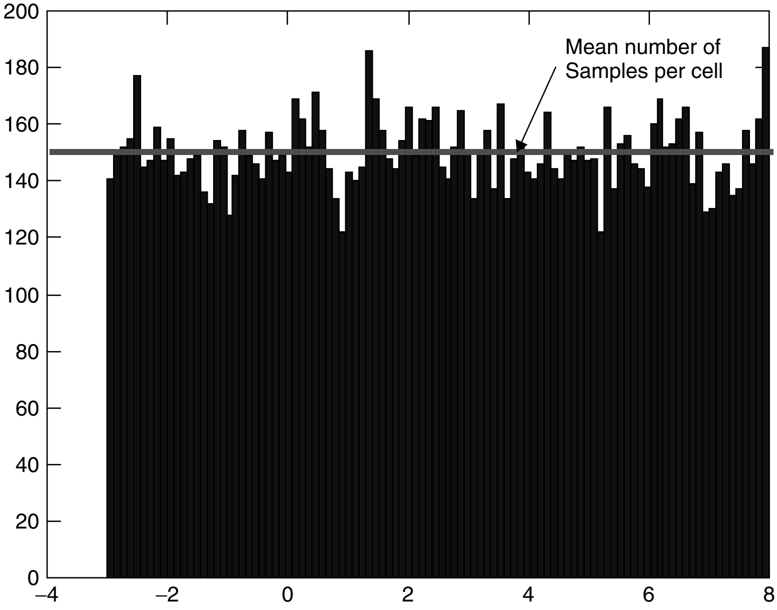


Figure 9.3 Matlab histogram results from the samples of $U(-3{,}8)$ random variable.

There is an aspect of estimation in computing a histogram. To show this in a systematic way, we define *indicator functions* for discrete and continuous random variables for each of the occurrence intervals. Suppose that we have a discrete random variable that can take on the values $1, 2, \ldots, \max$:

$$\text{Discrete case: } I_i(N) = \begin{cases} 1; & \text{for } N = i \\ 0; & \text{otherwise} \end{cases}; \quad i = 1, 2, \ldots, \max$$

It is important to note that $E(I_i(t)) = P(N = i) = P_i$; $i = 1, 2, \ldots, \max$.

Now, suppose that we have a continuous random variable, T. We define indicator functions over a set of intervals $(t_0, t_1), (t_1, t_2), \ldots, (t_{\max-1}, t_{\max})$.

$$\text{Continuous case: } I_i(T) = \begin{cases} 1; & \text{for } t_{i-1} < T \le t_i \\ 0; & \text{otherwise} \end{cases}; \quad i = 1, 2, \ldots, \max$$

In the continuous case, the expected value of the indicator function is

$$E(I_i(t)) = P(t_{i-1} < T \le t_i) = P_i \cong \Delta_i f_T(t_i)$$

where Δ_i is the width of the ith cell i.e., $(t_i - t_{i-1})$ and$f_T(t_i)$ is the value of the density function a point in the ith interval. We gather n samples $X_1, X_2, \ldots, X_n$ and evaluate the indicator function for each sample. Clearly, $n_i = \sum_{j=1}^{n} I_i(X_j)$; $i = 1, 2, \ldots, \max$ and

$$f_i = \frac{1}{n}\sum_{j=1}^{n} I_i(X_j); \quad i = 1, 2, \ldots, \max \tag{9.9}$$

As (9.9) indicates, the frequency of occurrence may be viewed as the sample mean of the indicator function for each interval. The mean of this statistic is

$$E(f_i) = \frac{1}{n}\sum_{j=1}^{n} E(I_i(X_j)) = \frac{1}{n}nP_i = P_i; \quad i = 1, 2, \ldots, \max \tag{9.10}$$

and we have an unbiased estimate of the probability distribution. As we increase the number of samples, the variance of the estimate will decrease [see (9.3)].

Example 9.4 This convergence of the estimate is illustrated by means of the associated Matlab program. The distribution under consideration is $U(0, 1)$. The array `histogr` produced by this program is plotted in Figure 9.4 for four sample sizes 1000, 10,000, 100,000, and 1,000,000. In each of these, the plot is over 100 cells, so the mean for each cell is 0.01. Clearly, as the sample size increases, the accuracy of the estimate of the probability increases.

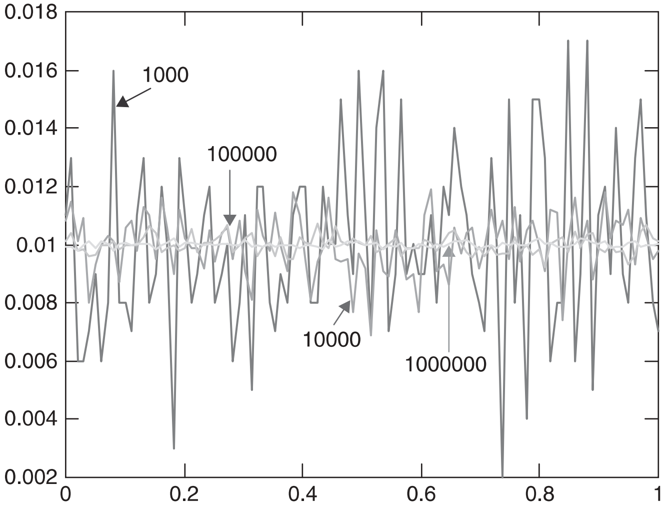


Figure 9.4 Histograms showing the effect of sample size.

9.1.6 Hypothesis Testing and the Chi-Square Test

The accuracy of the estimates can be assessed by a comparison with the underlying distribution. By sampling and computing the indicator for a particular interval, we are conducting a sequence of Bernoulli trials in which a success is counted when the sample falls in the interval in question. We can assess the veracity of an estimate by comparing it to an underlying distribution. The following example serves as an illustration of this point.

Example 9.5 We assume that the underlying distribution is uniform and there are only five cells. The probability of a sample falling in any particular cell is $p = 0.2$. In order to test the plausibility of a sample value, we compute the statistic $P(|S - np| > t)$, where S is the sample value and n is the number of samples. The results, which were worked out on the associated Excel spreadsheet, are shown in Figure 9.5 for $n = 100$. As the sample value varies from the mean, np, the probability decreases. For example, the probability of a sample value being more than 120 or less than 80 is less than 10^{-5}.

The line of reasoning of the preceding paragraph is formalized in *hypothesis testing*. In the case at hand, the null hypothesis is that the underlying distribution is a certain specified distribution. Samples are gathered, and a statistic is calculated as a function of these samples. The probability that this statistic lies outside a certain range is calculated under the assumption that the null hypothesis holds. If this probability exceeds a certain threshold, α, we say that the null hypothesis is rejected at the significance level, α.

Example 9.5 serves as an illustration of the idea of hypothesis testing for a single cell, but the approach would be too complicated for a general distribution. Because it

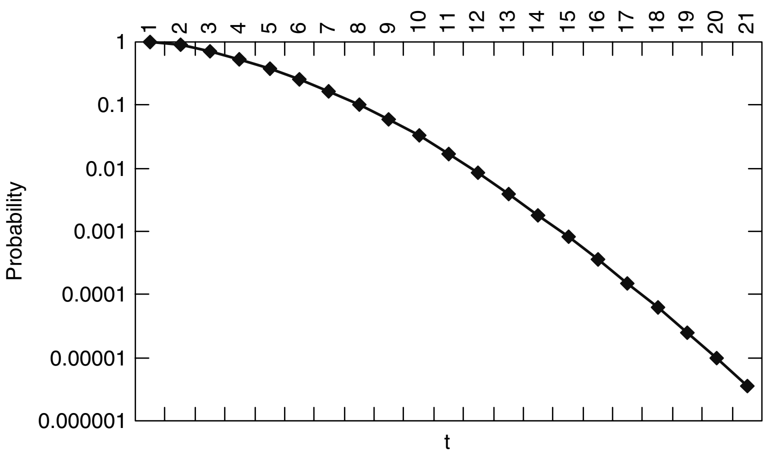


Figure 9.5 Probability of sample values $P\{|S - np| > t\}$.

is more tractable mathematically, the *chi-square test* (χ^2) is used to test the goodness of the fit to a particular distribution over all cells. Suppose that the probability of a sample falling in cell k is P_k; $k = 1, 2, \ldots, M$. If there were n trials, the average number of times that a sample falls in cell k would be nP_k; $k = 1, 2, \ldots, M$. Letting n_i denote the number of times that a sample falls in the ith cell, we compute the statistic

$$\chi^2 = \sum_{i=1}^{M} \frac{(n_i - nP_i)^2}{nP_i} \tag{9.11}$$

Clearly, low values of this statistic indicate a good fit to the hypothesized distribution. In practice, the approximations that we have made here are not very good unless nP_i is greater than 4.[3]

Since n_i is the sum of independent random variables, the central-limit theorem can be invoked whereby n_i is approximated as a Gaussian random variable. Because X^2 is the sum of the squares of M Gaussian random variables, it has the chi-square distribution with $M - 1$ degrees of freedom. As above, given a set of samples $(n_1, n_2, \ldots, n_M)$, we calculate the probability that the statistic is above a threshold.

The chi-square distribution with $(M - 1)$ degrees of freedom has the density function

$$f_X(x) = \frac{x^{(M-3)/2} e^{-x/2}}{2^{(M-1)/2}\Gamma((M-1)/2)}; \quad x \geq 0$$

where $\Gamma(\cdot)$ is the gamma function. The chi-square test is based on the tail of the distribution. We reject the hypothesis if the probability of the measured value is small, that is, if the tail probability

$$\int_{X^2}^{\infty} f_{X^2}(x)dx = \int_{X^2}^{\infty} \frac{x^{(k-2)/2} e^{-x/2}}{2^{k/2}\Gamma(k/2)} dx \tag{9.12}$$

is small, where X^2 is as given in (9.11) and $k = (M - 1)$. The following example quantifies this statement.

Example 9.6 On the associated Excel spreadsheet, examples are worked out. We use the discrete uniform distribution over 1,2,3,4,5. We compute the X^2 statistic for each of four experiments of 100 samples. The probability of exceeding these values is calculated as a chi-square distribution. The results of the first experiment are quite consistent with the hypothesis since with probability 0.99995 the chi-square statistic 0.02 can be exceeded under the hypothesis. As we go on through the other experiments, the outcomes are less and less likely. For the third experiment, less

[3] In addition to the chi-square test, another test for *goodness of fit* is the *Kologorov–Smirnov test*; see Allen (1978), pp. 311–317.

than 5% of experiments would have a statistic of 9.96, and it is reasonable to reject the hypothesis. We speak of this as having a 5% significance level. The fourth experiment has a 1% significance level since it is only with probability 1% that the statistic would exceed 14.4. Prior to the easy availability of computers, the chi-square test was carried out with the assistance of tables.[4] The tables show that the 1% and 5% levels of significance are 9.488 and 13.277, respectively, for 4 degrees of freedom. We will return to the chi-square test after we have seen how to generate random variables.

9.2 RANDOM-NUMBER GENERATION

9.2.1 Pseudorandom Numbers

The first step in a Monte Carlo simulation is the generation of random variables having the appropriate distributions. For example, if the arrival of messages follows the Poisson distribution, then the inter–arrival times of messages are exponentially distributed. As we shall demonstrate, the random variables of interest to us can be generated from random variables uniformly distributed in the interval (0, 1), i.e. $U(0, 1)$.

We first consider how to generate a sequence of independent $U(0, 1)$ random variables. The software must achieve three objectives:

1. The sequence of numbers produced must have the right statistical properties. In this respect, there are two basic criteria:
   a. The sequence of numbers must be uniformly distributed between 0 and 1 with a reasonable degree of approximation. The chi-square test might be used here to evaluate the goodness of fit to a uniform distribution.
   b. The random variables must be independent of one another in the sequence.
2. A second objective of a random-number generator is to simplify debugging by producing a random sequence that can be replicated so that experiments can be replicated.
3. The final objective is simplicity of computation.

In a certain sense, these criteria are self-contradictory. If the numbers in a sequence were truly random, there would be no simple way to reproduce them. We will compute a sequence of *pseudorandom* numbers. These numbers, although not truly random, are *supposed to* have the same statistical properties as a truly random sequence.

A standard approach to the generation of a sequence of uniformly distributed pseudorandom numbers is the *linear congruence generator*. We begin with sequence obeying the iterative relationship

$$X_{i+1} = (aX_i + c) \bmod m; \quad i = 0, 1, 2, \ldots \tag{9.13}$$

[4] For further details, see Walpole and Meyers (1989), p. 702.

where the parameters a, c, and m are determined from number theory in order to give good performance.

Since the calculation of (9.13) is modulo m, the sequence of numbers

$$U_i = \frac{X_i}{m}; \quad i = 1, 2, \ldots$$

lie in the interval (0, 1).

This sequence of numbers is not difficult to calculate. Moreover, once the first number X_0, called the "seed," is given the rest of the sequence is determined. Thus, the sequence is repeatable. It can be shown that the resulting sequence will have the right statistical properties if the parameters a, c, and m in (9.13) are chosen properly. The software that we will be using for our simulation studies will have random number generators, which use algorithms similar to (9.13).

Example 9.7 We illustrate these concepts with a sequence generated by the recursive relationship

$$X_{i+1} = (314{,}159{,}269 \times X_i + 453{,}806{,}245) \bmod 2^{15}; \quad i = 0, 1, 2, \ldots$$

This technique is illustrated on the Excel spreadsheet, where 100 random variables distributed in the interval (0, 1) are shown. One can vary the starting point to see the effect on the sequence of pseudorandom numbers.

9.2.2 Generation of Continuous Random Variables

A key part of simulation is the generation of random variables having a distribution appropriate to the problem at hand. The basic technique involves the transformation of random variables. We continue now the discussion of random variable transformations, which was begun in Section 2.5. If a transformation of a random variable is monotonic, either nondecreasing or nonincreasing, then there is a unique inverse. As we have seen, for the probability distribution and for the probability density, we have, respectively

$$F_Y(y) = P(Y \leq y) = P(g(X) \leq y) = P(X \leq g^{-1}(y)) = F_X(g^{-1}(y))$$

$$f_Y(y) = \frac{dF_X(g^{-1}(y))}{dy} = f_X(g^{-1}(y))\left|\frac{dg^{-1}(y)}{dy}\right| \tag{9.14}$$

where $g^{-1}(y)$ is the inverse of $g(y)$.

Uniform Distribution Perhaps the simplest useful transformation is a linear transformation on a uniformly distributed random variable. We start with a random variable X that is uniformly distributed in $(0, 1)$; thus, $f_X(x) = U(x) - U(x - 1)$, where $U(\cdot)$ denotes the unit step. Now, let $Y = aX + b$. Solving for X, we have $g^{-1}(y) = (y - b)/a$. Substituting into (9.14), we find

$$f_Y(y) = \frac{1}{|a|}\left(U\left(\frac{y-b}{a}\right) - U\left(\frac{y-a-b}{a}\right)\right) \tag{9.15}$$

The distribution function is given by

$$F_Y(y) = \begin{cases} 0; & -\infty < y \leq b \\ y/|a|; & b < y \leq a+b \\ 1; & y > b \end{cases} \tag{9.16}$$

Exponential Distribution A useful example is when U is uniformly distributed between 0 and 1 and the transformation is

$$E = -\ln(1 - U) \tag{9.17}$$

The inverse of the transformation is

$$U = g^{-1}(E) = 1 - e^{-E} \tag{9.18}$$

The distribution of U is given by (9.16) with $a = 0$ and $b = 1$, and is written

$$F_U(t) = \begin{cases} 0; & -\infty < t \leq 0 \\ t; & 0 \leq t \leq 1 \\ 1; & t > 1 \end{cases} \tag{9.19}$$

Substituting into (9.18) in (9.19), we find

$$F_E(y) = 1 - e^{-y}; \quad y \geq 0$$

However, this is an exponentially distributed random variable with mean 1. This can be transformed into an exponentially distributed random variable with mean $1/\mu$ simply by multiplying the random variable by $1/\mu$.

Example 9.8 On the associated Excel spreadsheet we generate 10,000 exponentially distributed random variables with mean 10. As in Example 9.1, the means and the variances are estimated for 10 samples and 100 samples. We have also written a Matlab program, `ex98.m`, which also generates 10,000 exponentially distributed random variables with mean value 10. A histogram of this experiment is shown in Figure 9.6. The histogram is found using the `Hist(y,100)`, which compiles a histogram of the 10,000-element vector into 100 cells. On the same diagram we show the true distribution.

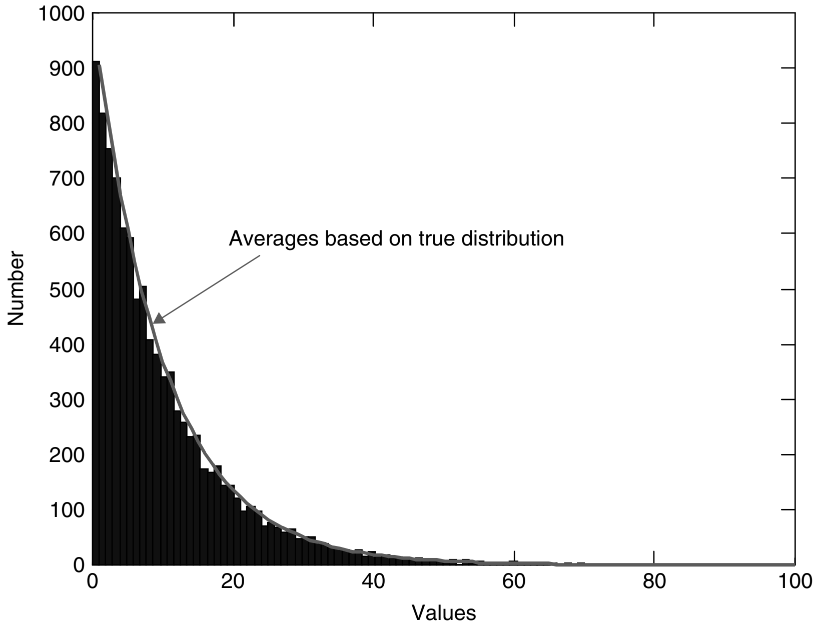


Figure 9.6 Histogram exponential distribution of 10,000 samples: cell size = 1.

The generation of the exponentially distributed random variable here demonstrates the *inversion technique* for generating random variable. Suppose that the desired random variable has the probability distribution function $F_T(t)$. Further, suppose that this is an invertible function with the inverse denoted as $G(x) = F_T^{-1}(x)$. Samples of the desired random variable are found simply by substituting uniformly distributed random variables into $G(x)$. The proof of this is left as an exercise. We demonstrate the technique again by obtaining Gaussian random variables.

Gaussian Distribution Gaussian random variables can be generated by the inversion technique. To show this, we consider an exercise in target shooting for which the error in both the vertical and horizontal directions are independent Gaussian random variables, each with zero mean and variance 1. The joint density is

$$f_{XY}(x, y) = \frac{\exp(-(x^2 + y^2)/2)}{2\pi}; \quad -\infty < x, \ y < \infty \tag{9.20}$$

Now, suppose that we transform the error to polar coordinates.[5] The transformation is

$$\begin{aligned} R &= \sqrt{X^2 + Y^2} \\ \Theta &= \tan^{-1}\frac{Y}{X} \end{aligned} \tag{9.21}$$

[5] The derivation of the Rayleigh distribution may be found in Leon-Garcia (1994), pp. 229–230.

The inverse of the transformation is

$$\begin{aligned} X &= R\cos(\Theta) \\ Y &= R\sin(\Theta) \end{aligned} \tag{9.22}$$

To find the distribution of R and Θ, we compute the Jacobian of the transformation as

$$J(r,\,\theta) = \begin{vmatrix} \dfrac{\partial x}{\partial r} & \dfrac{\partial x}{\partial \theta} \\ \dfrac{\partial y}{\partial r} & \dfrac{\partial y}{\partial \theta} \end{vmatrix} = \begin{vmatrix} \cos(\theta) & -r\sin(\theta) \\ \sin(\theta) & r\cos(\theta) \end{vmatrix} = r \tag{9.23}$$

From (9.20)–(9.23), the joint density of R and Θ is

$$f_{R\Theta}(r,\,\theta) = \frac{r\exp(-\,r^2/2)}{2\pi}; \quad r \geq 0,\ 0 \leq \theta \leq 2\pi \tag{9.24}$$

From (9.22), we see that the marginal density functions are

$$\begin{aligned} f_R(r) &= r\exp\left(\frac{-r^2}{2}\right); \quad r \geq 0 \\ f_\Theta(\theta) &= \frac{1}{2\pi}; \quad 0 \leq \theta \leq 2\pi \end{aligned} \tag{9.25}$$

These are, respectively, the density function for the Rayleigh distribution and the $U(0,\,2\pi)$ distribution. The probability distribution for the Rayleigh distribution is given by

$$F_R(r) = \int_0^r x\exp\left(\frac{-x^2}{2}\right)dx = 1 - \exp\left(\frac{-r^2}{2}\right); \quad r \geq 0 \tag{9.26}$$

Our objective now is to generate a Rayleigh distributed random variable. We begin by inverting the transformation $u = 1 - \exp(-\,r^2/2)$. We find that $r = \sqrt{-2\ln(1-u)}$. Now, suppose that a $U(0,1)$ random variable is transformed according to

$$R = \sqrt{-2\ln(1-U)} \tag{9.27}$$

The probability distribution for R can be verified by straight substitution:

$$\begin{aligned} P(R \leq r) &= P\left(\sqrt{-2\ln(1-U)} \leq r\right) = P\left(U \leq 1 - \exp\frac{-r^2}{2}\right) \\ &= 1 - \exp\left(\frac{-r^2}{2}\right) \end{aligned} \tag{9.28}$$

The last equality of (9.28) follows from the $U(0,1)$ distribution. It is then clear that R is a Rayleigh distributed random variable. From (9.21) and (9.22), the transformations

$$\begin{aligned} X &= \sqrt{-2\ln(1-U_1)}\cos(2\pi U_2) \\ Y &= \sqrt{-2\ln(1-U_1)}\sin(2\pi U_2) \end{aligned} \tag{9.29}$$

yield independent Gaussian random variables, provided that U_1 and U_2 are independent $U(0,1)$ random variables. Since U_1 is uniform $U(0,1)$, the substitution $\ln(U_1)$ would work just as well here.

If indeed U_1 and U_2 are independent and uniform, (9.29) gives true Gaussian random variables even way out on the tails. The only limit is the accuracy of the machine. The same cannot be said for a well-established alternate technique that invokes the central-limit theorem. If we add 12 $U(0,1)$ random variables and subtract the number 6, the result approximates a Gaussian random variable with 0 mean and variance 1.

There is a potential difficulty in practical computation. However, if U_1 and U_2 are pseudorandom numbers generated in sequence from the same seed, correlations can be induced in the Gaussian pair given in (9.29). The remedy is to generate U_1 and U_2 from different seeds.

Example 9.9 On the associated Matlab program, `Gaussian.m`, `file`, we generate 10,000 samples from a Gaussian distribution. Although there are Matlab functions that automatically generate Gaussian random variables, we use (9.29). A histogram of the results is shown in Figure 9.7. Also shown are the average values in cells drawn for the true distribution as calculated on a associated Matlab program `cgauss.m`.

9.2.3 Discrete Random Variables—General Case

There is a generic technique for generating discrete random variables. To illustrate, we begin with Bernoulli random variables. Recall that these assume the values 0 and 1 with probability $1-P$ and P, respectively. As in the continuous case, we use the $U(0,1)$ random variables to generate the Bernoulli random variables. We set a threshold at P. For less than P, the variable is set to 1 and to 0 otherwise.

Example 9.10 An example is shown on the associated Excel spreadsheet. A sequence of $U(0,1)$ random variables is generated and compared to a single threshold. The result is a sequence of Bernoulli random variables.

We can generalize this example to an arbitrary discrete distribution. Let the random variable D have the distribution $P(D=k)=P_k$; $k=1,2,\ldots,n$. The interval $(0,1)$ is divided into n intervals. If the value of a $U(0,1)$ random variable falls in an interval of width P_k, the discrete random variable takes on the value k. The obvious way to implement this is to start with the test $U \leq P_0$, then to $U \leq P_0 + P_1$, and so on. When a test is successful, the sequence is terminated and

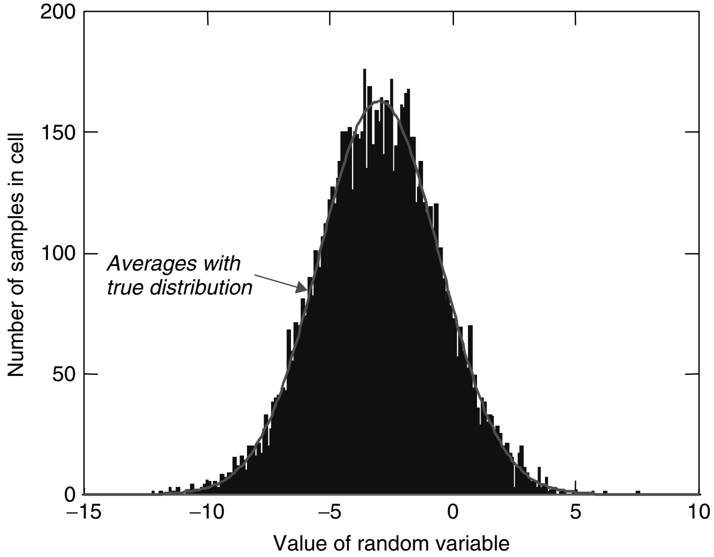


Figure 9.7 Histogram for Gaussian distribution of 10,000 samples: cell size = 0.1

a value of the random variable declared. The associated Matlab program ex910.m examples shows the steps.

Example 9.11 On the attached Excel spreadsheet, we have worked out an example for a discrete random variable, which takes on four values, as shown. As shown in the spreadsheet, we use the device of nested "if" statements. Of course, this is quite awkward for a discrete random variable that takes on a large number of values. On the associated Matlab program, ex911.m., the histogram for an arbitrary discrete distribution, is calculated. The input is the probability vector and the number of trials. Random variables are generated and classified into one of eight bins.

Example 9.12 Binomial and Geometric Distributions In many of the applications the distribution is in one of the standard forms. Prime examples of these are the binomial and the geometric distributions, whose histograms are calculated in the associated Matlab programs, ex912bin.m and ex912geo.m, respectively. The geometric distribution has a significant difference from the binomial distribution inasmuch as it can take on a countably infinite number of values. Certainly, this affects the way samples from the distribution are generated. The results from the Matlab program for the geometric distribution are shown in Figure 9.8. In order to illustrate the effect of sample size, the histograms for 1000 and 10,000 samples are shown. As we see, the latter is close to the true distribution, while there are statistical variations for the 1000-sample case.

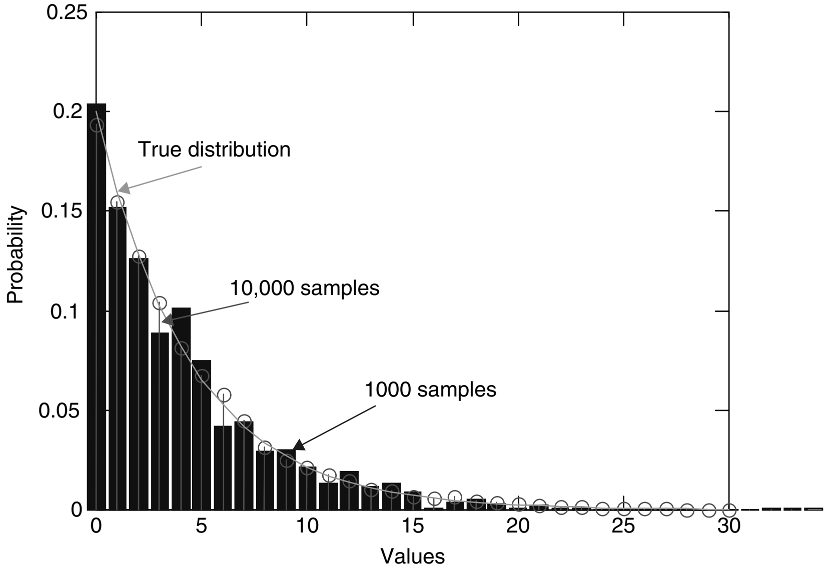


Figure 9.8 Geometric distribution, $P = 0.2$.

9.2.4 Generating Specific Discrete Random Variables

In the previous subsection, a general technique for generating discrete random variables was presented. In this subsection we present techniques that are specific to particular distributions.

Binomial Distribution We consider first the binomial distribution. The generation is based on the fact that a binomially distributed random variable, with parameters n and P, is just the sum of n independent Bernoulli random variables, each with probability P.

Geometric Distribution We generate the geometric distribution by simply conducting sequences of trials. Assume that the probability of success is P and U is a $U(0, 1)$ random variable. We have a success when $U < P$. We tally the number of the trials until the first success. Refer to the associated Matlab program `ex913geo.m`.

Poisson Distribution In Chapter 3, we showed that the intervals between Poisson arrivals are exponentially distributed random variables. We can use this property to generate Poisson random variables. Let e_k; $k = 1, 2, \ldots$ indicate a sequence of iid exponentially distributed random variables, each with mean value $1/\lambda$. The event of k arrivals in the interval $(0, t)$ is given by

$$\sum_{i=1}^{k} e_i \leq t < \sum_{i=1}^{k+1} e_i$$

When the sum of $k+1$ exponential random variables brackets t in this way, we assume that there are k arrivals in t seconds. The number k has a Poisson distribution. From (9.17), we have $e_k = -(1/\lambda)\ln(U_k)$; $k = 1, 2, \ldots,$ where U_k; $k = 1, 2, \ldots$ are a sequence of iid $U(0, 1)$ variables. Substituting and rearranging with exponentiation, we find the equivalent event

$$\prod_{i=1}^{k+1} U_i < \exp(-\lambda t) \le \prod_{i=1}^{k} U_i \tag{9.30}$$

Example 9.13 In this example, we generate binomial, geometric, and Poisson random variables using the properties of each distribution:

- *Binomial*—Matlab program `ex913bin.m`. The inputs are the number of trials, the probability of success, and the sample size.
- *Geometric*—Matlab program `ex913geo.m`. The inputs are the probability of success on a trial and the sample size.
- Poisson Matlab program `ex913pois.m`. The inputs to the program are the average rate λ, the time interval, and the samples size.

Uniform Property of Poisson Arrivals Suppose that a study calls for the simulation of a specified number of Poisson arrivals in a specified interval. We can generate this by means of the uniform property of the Poisson process that was discussed briefly in Section 3.3.1. In many cases, we are interested in the exact times of Poisson arrivals in an interval, as well. In order to generate these arrival times, we use what might be called the *uniform property* of the Poisson arrival. We explain this property by example. Let us consider the event of two successes in five trials. There are $\binom{5}{2} = 10$ ways for this to happen. Specifically, these outcomes are

$$\begin{array}{ccccc}
S & S & F & F & F \\
S & F & S & F & F \\
S & F & F & S & F \\
S & F & F & F & S \\
F & S & S & F & F \\
F & S & F & S & F \\
F & S & F & F & S \\
F & F & S & S & F \\
F & F & S & F & S \\
F & F & F & S & S
\end{array}$$

Each outcome has the same probability, $P^2(1-P)^3$. Note that the successes and failures are evenly spread among the trials. The probability of success on any trial is $2/5$. It is easy to see that the result would be the same for any number of successes in any number of trails; the successes are uniformly spread among the trials. Now, suppose that we go to the limit, $n \to \infty$, $P \to 0$, $nP \to \lambda t$ of trials in an interval $(0, t)$. Clearly the successes are spread uniformly in the interval. It is effectively the same as choosing samples from a $U(0, t)$ distribution.

Example 9.14 In the associated Matlab `ex916.m` program, the inputs are the number of arrivals and the length of the arrival interval. The output is a vector giving the arrival times in the interval.

9.2.5 The Chi-Square Test Revisited

As we have generated random variables in this section, we have plotted their histograms. While the histograms may provide a rough measure of the conformity of the samples to an underlying distribution, we may require something more quantitative. As we have seen, the chi-square test gives such a quantitative measure. The chi-square statistic is given by (9.11). There are M cells, and n_i; $i = 1, 2, \ldots, M$ is the number of times that a sample falls into a particular cell. This is compared to the average number of times that the sample would fall into a cell if the hypothesized underlying distribution were true, nP_i; $i = 1, 2, \ldots, M$. In order for the test to be effective,[6] it is recommended that all of the values of nP_i be approximately equal and that $nP_i \geq 5$. A further but more problematic stipulation is that the number of cells be no more than 30 or 40. In the case of a continuous distribution, the requirement that the nP_i be equal is not too difficult to fulfill, as illustrated in the following example.

The calculations in Example 9.6 rely on the Excel spreadsheet to compute the tail probability (9.12). In the following examples we compute this directly. We make the simplifying assumption that there are an odd number of samples, meaning that the chi-square distribution has an even number of degrees of freedom. This allows a closed-form solution to the integral in (9.12). Of course, this assumption does not affect basic principles.

Example 9.15 We now give an example of a chi-square test that has been set up in such a way that it should fail. We generate a sequence of Rayleigh distributed samples. Then, we test for goodness of fit of these samples to an exponential distribution. The Rayleigh distribution has the probability distribution function

$$F_R(r) = 1 - \exp\left(\frac{-r^2}{2\alpha^2}\right); \quad t \geq 0$$

and the mean value $E(R) = \alpha\sqrt{\pi/2}$. If we set $E(R) = 2$, then $F_R(r) = 1 - \exp\left(-\, r^2/(16/\pi)\right)$. Since the distribution function can be inverted, we can apply the

[6] For more detail on conducting the chi-square test, see Law and Kelton (2000).

inversion technique to generate Rayleigh distributed samples. The function $\sqrt{(16/\pi)\ln(1/(1-U))}$ evaluated for a sequence of independent $U(0, 1)$ variables gives a sequence of Rayleigh random variables.

In order to carry out the chi-square test, we need to segment the exponential distribution into equal probability bins. The probability of falling in the interval (t_1, t_2) is

$$P(t_1 \leq E < t_2) = \int_{t_1}^{t_2} \lambda \exp(-\lambda t) dt = \exp(-\lambda t_1) - \exp(-\lambda t_2)$$

Since, in general, we don't know the mean of the underlying distribution, we estimate it as the sample mean; accordingly, $1/\lambda \cong \sum_{i=1}^{n} S_i/n$, where the S_i; $i = 1, 2, \ldots, n$ are the individual samples. On the associated Matlab program, `ex915.m`, we calculate the chi-square statistic. We use 25 cells and we take 125 samples. The chi-square statistic is 74, which has a probability of 5.25×10^{-7}. Of course, this result reflects the fact that we deliberately chose to mismatch the distributions.

Test for Poisson In Chapter 2, the Poisson distribution was derived as the limit of the binomial distribution. Essentially, this test is based on the uniform property of the Poisson process as discussed in Section 3.3.1. Recall that this means that, given a certain number of arrivals in an interval, each arrival is uniformly distributed in that interval. Thus, if there are n arrivals, $t_1, t_2, \ldots, t_n$, respectively in the interval $(0, T)$, the variables $t_1, t_2, \ldots, t_n$ should be tested for being $U(0, T)$.

Example 9.16 An example is worked out on the associated Matlab program. We use the same set of numbers as in the previous example, 25 cells and 125 samples. Another parameter was the length of the interval, which had no effect. The resulting chi-square statistic was 14.8, which has a probability of 0.9265. Of course, this was a setup, so the result is expected.

9.3 DISCRETE-EVENT SIMULATION

9.3.1 Time-Driven Simulation

There are two different ways to advance time in simulation programs, *time-driven* simulation and *event-driven* simulation. In the first of these, time is segmented into fixed-length slots. As the simulation progresses, there is a clock that advances by the length of a slot. At the end of each slot the system is updated in accordance with events that may have occurred in the slot interval. For example, in a queueing system, new arrivals can be added to buffers and departures removed. If the arrival process is Poisson, the number of new arrivals has a Poisson distribution, whose mean is the average arrival rate times the duration of the slot. After the system is updated for all the events that can occur, the clock advances by one time segment.

Example 9.17 On the associated Matlab program, we simulate an M/M/1 queue. The message transmission time is 2 s. The message arrival rate is varied over the range 0 to 0.475 messages per second to produce loads varying over the range 0 to 0.95. The simulation is over 1000 s for three different runs. The results are as shown in Figure 9.9. Clearly, for the heavy loads, the results are not accurate and more data need to be gathered by longer runs.

9.3.2 Event-Driven Simulation

Although time-driven simulation is often straightforward to program, it may be inefficient. The time slot should be small enough to avoid an excessive number of events that can occur in a slot. The difficulty is that there may be many slots in which nothing happens. In a sense, event-driven simulation skips over these noneventful slots. In event-driven simulation, the clock advances to the time of the next occurrence affecting the system. For example, in an M/M/1, the clock is advanced to the time of the next arrival. If there are messages in the system, this is the

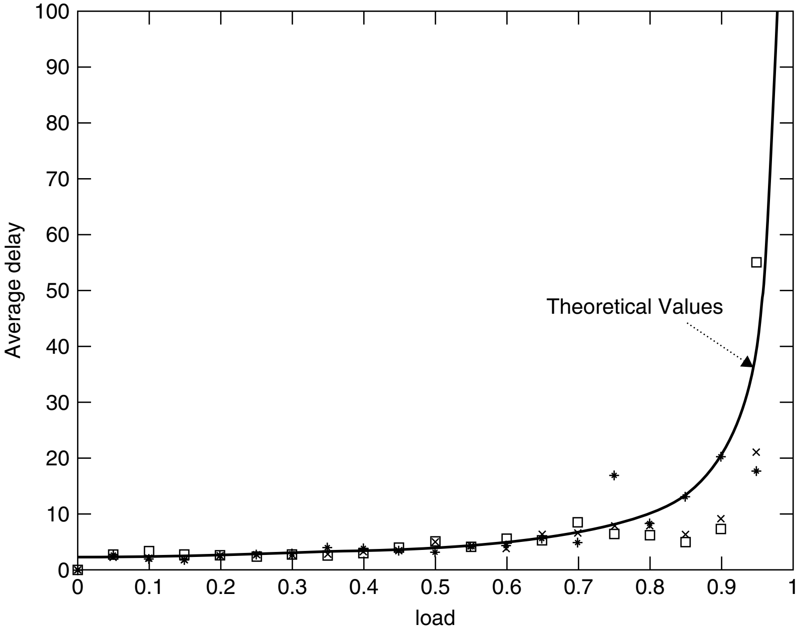


Figure 9.9 Simulation results for M/M/1 queue.

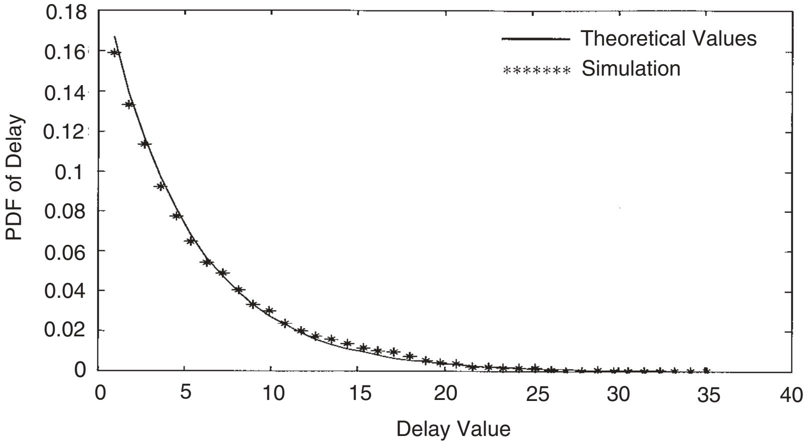


Figure 9.10 Probability density function of M/M/1 delay.

minimum of two exponential random variables, one with mean $1/\lambda$ the mean interarrival time; and the other with mean $1/\mu$, the mean message length. It is not difficult to show that this minimum is exponentially distributed with mean $1/(\lambda + \mu)$.

Example 9.18 On the associated Matlab program, we simulate an M/M/1 queue over an interval of 10^4 s. The message arrival rate is 0.8 messages/s, and the mean message length is 1 s. The average number of arrivals should be 8000; indeed, we measure 7974 message arrivals. A histogram is shown in Figure 9.10. The bin size is 0.9. As we have seen in Section 3.4.3, for the M/M/1 queue, the delay is exponentially distributed with mean $1/(\mu - \lambda)$, where λ is the arrival rate and $1/\mu$ is the mean message length. Also shown is the plot for an exponentially distributed random variable that has mean value 5.

9.4 VARIANCE REDUCTION TECHNIQUES

As we have seen in Section 9.1.4, the variance of an estimate is directly related to simulation runtimes. By reducing the variance of the quantity being estimated, without affecting its mean, we can get more accurate measurements for the same amount of simulation time or achieve a required degree of precision with smaller simulation runs. We will be examining four basic approaches: *common random numbers*, *antithetic variates*, *control variates*, and *importance sampling*. Other techniques that are beyond our scope are *indirect estimation*, *conditioning*, and *stratified sampling* [see Bratley et al. (1987)]

9.4.1 Common Random-Number Technique

The common random-number approach is appropriate when one is comparing two different implementations or approaches. We compare the effect of the service time distribution on message delay in M/G/1 queues. Let D_{1i} and D_{2i} be the ith samples from the two different systems. Define $Z_i = D_{1i} - D_{2i}$. We estimate the mean value of Z in order to assess the effect of the different implementations. The sample mean of the difference is $\bar{Z} = (1/n)\sum_{i=1}^{n} Z_i$. The objective is to reduce the variance of the sample mean. As we have seen, the variance of the sample mean is given by $\sigma_{\bar{Z}}^2 = \sigma_Z^2/n$. Now the variance of Z is given by

$$\begin{aligned}\sigma_Z^2 &= E((Z - E(Z))^2) = E((D_1 - D_2 - E(D_1 - D_2))^2)\\ &= E((D_1 - E(D_1))^2) + E((D_2 - E(D_2))^2) - 2E((D_1 - E(D_1))(D_2 - E(D_2)))\\ &= \sigma_{D_1}^2 + \sigma_{D_2}^2 - 2\text{Cov}(D_1, D_2)\end{aligned}$$

Thus

$$\sigma_{\bar{Z}}^2 = \frac{\sigma_{D_1}^2 + \sigma_{D_2}^2 - 2\text{Cov}(D_1, D_2)}{n} \tag{9.31}$$

When the two random variables, D_1 and D_2, are independent, the covariance is zero and the variance of Z is simply the sum of the variances; however, if a positive correlation can be introduced, the variance of the difference would be reduced.

The positive correlation *may* be introduced if the same random sequence is used to generate samples for both D_1 and D_2. For this to work, the sequence must be synchronized so that the same pseudorandom input affects each system in approximately the same way. Of course, in complex systems, considerable study may be necessary to ensure that the condition holds. We illustrate the approach by simple M/G/1 queues.

Example 9.19 We consider a simple example where the effect of the service time distribution of the first message of a busy period is varied. Two service times, M_1 and M_2, are defined as $P(M_1 = 0.5) = 0.5$, $P(M_1 = 1.0) = 0.5$ and $P(M_2 = 0.5) = 0.944444$, $P(M_2 = 5) = 0.055556$, respectively. The respective mean and mean square values are $\bar{M}_1 = 0.75$, $\bar{M}_1^2 = 0.625$ and $\bar{M}_2 = 0.75$, $\bar{M}_2^2 = 1.625$. Now we consider two systems, In system one, if the queue is empty, the second of these two distributions, M_2, is used for the newly arrived message; otherwise, messages are taken to have the first distribution M_1. In system two, the service distribution is always M_1. We compare the two systems. Note that, we have contrived to keep the mean value the same in both cases so that the load will be the same if the arrival rate is the same. We now run five pairs of simulation runs of 1000 samples each for systems one and two, respectively. We estimate the difference in the mean delay for each pair of runs. If we use the same seed for both system one and two, we obtain successive sample standard deviations of 1.7123, 1.9064, 1.9331, 2.0607 and 1.9152. We repeated the experiments in which the seeds were different for each run in the pair. The sample standard deviations were 2.3512, 2.2705, 2.5682, 2.9589 and

2.7980. In both cases, we could get estimates of the variance of the sample mean by dividing these numbers by $\sqrt{1000}$. Note that, all of the "same seed" results are less than all of those obtained when the seeds were different.

Let us calculate the probability of all of the samples in the first batch being less than all of the sample in the second batch under the assumption that the variance of the underlying distribution is the same for both batches. Under this assumption, the probability of a particular sample from the first batch being less than a particular sample from the second is $1/2$. The probability of a particular one from the first being less than all of the second is 2^{-9}. The probability of all of the first batch being less than all of the second is $2^{-25} \cong 3 \times 10^{-8}$. Certainly, this is a very unlikely event under the assumption; therefore, it is reasonable to conclude that there is a difference between the two underlying variances.

9.4.2 Antithetic Variates

The second approach, antithetic variates, uses a negative correlation to reduce the variance of a set of samples. Consider the antithetic random variables U and $1 - U$, where U is uniformly distributed in (0,1). Clearly, $1 - U$ is uniformly distributed in (0, 1); consequently, it could be used in a simulation to generate random variables. It is also clear that U and $1 - U$ are negatively correlated, that is, when one is large the other is small and vice versa.

Suppose that we do pairs of simulation runs with one run using U to generate random variables and the other using $1 - U$ with both having the same starting seed. We denote the respective sets of samples as $S_1 = S_{11}, S_{12}, \ldots, S_{1n}$ and $S_2 = S_{21}, S_{22}, \ldots, S_{2n}$. In the example below, these are samples of the delay. The sample mean, over all samples is

$$\bar{S} = \frac{1/n \sum_{i=1}^{n} S_{1i} + 1/n \sum_{i=1}^{n} S_{2i}}{2} = \frac{1}{2n} \sum_{i=1}^{n} (S_{1i} + S_{2i}) \tag{9.32}$$

Using the same analysis that led to (9.31), we find that the variance of $S_{1i} + S_{2i}$ is $\sigma_{S_1}^2 + \sigma_{S_2}^2 + 2\mathrm{Cov}(S_1, S_2)$. Assuming independence between pairs of sample, we find that the sample variance is

$$\sigma_{\bar{S}}^2 = \frac{\sigma_{S_1}^2 + \sigma_{S_2}^2 + 2\mathrm{Cov}(S_1, S_2)}{4n} \tag{9.33}$$

Now, if S_1 and S_2 are produced by antithetic variates that are properly synchronized, the covariance, $\mathrm{Cov}(S_1, S_2)$, should be negative and the sample variance will be reduced.

Example 9.20 The technique described here was tested by simulating an M/G/1 queue whose message service time is given by $P(M_1 = 0.5) = 0.5$ and $P(M_1 = 1.0) = 0.5$. Ten runs were carried out, five with the same seed for antithetic variates and five with different seeds. For the same seeds, the sample standard deviations measured were 1.0396, 1.0988, 0.9326, 1.165, and 1.0147. When the seeds were different, the measurements were 1.1961, 0.9547, 1.2604, 1.1041, and 1.1529. If all the numbers were random, the probability of this event would be approximately 3 × 10^{-6}. We can conclude that there is some variance reduction to be gained by this technique.

9.4.3 Control Variates

The third variance reduction technique also relies on correlation to reduce the variance of the estimate. Suppose that we want to estimate the mean of the random variable, X, which has variance σ_X^2. Suppose also that there is another random variable, Y, which is positively correlated with X. In the example with which we have been dealing, X could be the message delay and Y could be the service time of a message. We form the random variable $X_C = X - a(Y - E(Y))$, where a is a positive constant. Clearly, the mean of X_C is the same as that of X, $E(X_C) = E(X) - aE((Y - E(Y))) = E(X)$. The variance is a different matter. A simple calculation shows that

$$\sigma_{X_C}^2 = \sigma_X^2 + a^2\sigma_Y^2 - 2a\mathrm{Cov}(X, Y) \tag{9.34}$$

If $a\sigma_Y^2 < 2\mathrm{Cov}(X, Y)$, there is a reduction in the variance of the estimate. The optimum value of a can be found by differentiating (9.34) with respect to a and setting the result equal to zero. We find

$$a^* = \frac{\mathrm{Cov}(X, Y)}{\sigma_Y^2} \tag{9.35}$$

The larger is the correlation between X and Y, the larger we make a. With this optimum value, the variance is

$$\sigma_{X_C}^2 = \sigma_X^2 - \left[\frac{\mathrm{Cov}(X, Y)}{\sigma_Y}\right]^m \tag{9.36}$$

Clearly, as with all of these variance reduction techniques, application of this technique also requires a detailed knowledge of the simulation model. In an extensive simulation, it may be profitable to run preliminary simulations in order to estimate the correlation between random variables.

Example 9.21 We can illustrate the technique with a simple example based on the M/G/1 queue. The delay of a message consists of two components; the time required to transmit the message and the time that the message waits in the queue. We express this as $D = M + D_Q$. It is not difficult to show that the queueing delay and the message transmission time are independent random variables; consequently, $\sigma_D^2 = \sigma_M^2 + \sigma_{D_Q}^2$. Since we know the variance of the message, we could reduce the variance of the estimate simply by estimating the queueing delay rather than the total delay. Another way to reach this same conclusion is to go through the steps described above. We form $D_C = D - a(M - E(M))$. The correlation between D and M is

$$\begin{aligned}\mathrm{Cov}(D, M) &= E((D - \bar{D})(M - \bar{M})) = E((D_Q - \bar{D}_Q + M - \bar{M})(M - \bar{M})) \\ &= E((M - \bar{M})(M - \bar{M})) = \sigma_M^2\end{aligned}$$

Substituting into (9.36), we find $\sigma_{D_C}^2 = \sigma_D^2 - \sigma_M^2 = \sigma_{D_Q}^2$.

9.4.4 Importance Sampling

It is often the case that the performance measure that is required is the result of a rarely occurring event. The probability of error in a transmission over an optical line is a good example. It may be important to know whether the probability of error is 10^{-9} or 10^{-10}. In either case, the occurrence of an error is a rare event. Another example is the overflow of a buffer. With increasing higher and higher data rates prevalent in the telecommunication network this is an event that could result in the loss of a great deal of user data; consequently, it should be rare. An appropriate question is how rare?

The obvious difficulty with estimating the occurrence of a rare event is gathering enough samples to provide a good estimate. For example, suppose that we want to measure the probability of a random variable, X, falling in the interval (x_1, x_2). In the case of buffer overflow x_1 could be the buffer size and $x_2 = \infty$. The number of samples of the random variable that must be taken until one falls in the interval is governed by the geometric distribution. If p is the probability of the event, the mean number of trials until occurrence is $1/p$; consequently, gathering enough points for an estimate with a small enough confidence may require a prohibitively large number of samples. *Importance sampling*, circumvents the problem by changing the sampled random variable and the sample space to provide the same unbiased estimate with fewer samples.

Let X be a random variable with probability density function $f(x)$. The probability X falling in range (x_{i-1}, x_i) is given by

$$p = \int_{x_1}^{x_2} f(x)dx = \int_{-\infty}^{\infty} I_i(x)f(x)dx = E_f[I_i(X)] \tag{9.37}$$

where $I_i(X)$ is the indicator function given by

$$I_i(X) = \begin{cases} 1; & x_{i-1} \le X \le x_i \\ 0; & \text{otherwise} \end{cases}$$

The subscript f in E_f serves to indicate that the averaging is over the probability density $f(x)$. There are situations where the density function is known, but is so complicated that calculating it by numerical integration is too difficult.[7] An alternative is simulation to estimate p. The problem is that in the interval (x_{i-1}, x_i), the density function $f(x)$ is too small. Now, suppose that we have another density function, $g(x)$, with the property $g(x) > 0; x_1 \le x \le x_2$. We rewrite (9.37) as follows

$$p = \int_{-\infty}^{\infty} I_i(x)\frac{f(x)}{g(x)}g(x)dx = E_g\left[\frac{f(X)}{g(X)}I_i(X)\right] \tag{9.38}$$

As (9.38) indicates the expectation is of the random variable $(f(X)/g(X))I_i(X)$ on the sample space governed by the probability density $g(x)$. Clearly, the expected

[7] See Foschini and Gans, 1998 for an example of this technique.

value of this random variable is p, the parameter we want to estimate. We make the density function $g(x)$ large in the interval so the event $(f(X)/g(X))I_i(X) \in (x_{i-1}, x_i)$ occurs often. Thus, we estimate p by calculating the statistic

$$\hat{p}_N(f') = \frac{\sum_{m=1}^{N} I_m(x_m)f(x_m)/g(x_m)}{N} \tag{9.39}$$

We have given only the barest outline of the importance sampling in here. It can be applied to more general cases where the probability density function is not known, but is indeed the quantity that we want to estimate. Treating this general case is beyond the scope of the text. The reader is referred to P. Bratley et al. 1987, P. Heidelberger, 1995, and P. L'Ecuyer, 2001.

REFERENCES

Allen, A. O., *Probability, Statistics and Queueing Theory*, Academic Press, 1978.

Bratley, P., B. L. Fox, and L. E. Schrage, *A Guide to Simulation*, 2nd ed., Springer-Verlag, 1987.

Foschini, G. J. and M. J. Gans, "On limits of wireless communications in a fading environment when using multiple antennas", *Wireless Communications*, **6**, (1998), pp. 311–335.

Heidelberger, P., "Fast simulation of rare events in queueing and reliability models," *ACM Transactions on Modeling and Computer Simulation*, **5**(1), (Jan. 1995), pp. 43–85.

Lavenberg, A. S., *Computer Performance Modeling Handbook*, Academic Press, 1983.

Law, A. M., and W. D. Kelton, *Simulation Modeling and Analysis*, McGraw-Hill, New York, 2000.

L'Ecuyer, P. and Y. Champoux, "Estimating small cell loss ratios in ATM switches via importance sampling," *ACM Transactions on Modeling and Computer Simulation*, **11**(1), (Jan. 2001), pp. 76–105.

Leon-Garcia, A., *Probability and Random Processes for Electrical Engineering*, Addison-Wesley, 1994.

Pawlikowski, K. F., H.-D. J. Jeong, and J.-S. Ruth Lee, "On credibility of simulation studies of telecommunication networks," *IEEE Commun.*, **40**(1), (Jan. 2002).

Walpole, R. E., and R. H. Meyers, *Probability and Statistics for Engineers and Scientists*, 4th ed., Collier-Macmillan, 1989.

EXERCISES

9.1 Repeat Exercise 9.1 for the case of exponentially distributed random variables with mean 3. Show 95% confidence intervals.

9.2 Repeat Exercise 9.3 for a random variable uniformly distributed in (5, 10). Show the histogram for 50 cells resulting for 1000 and for 10,000 samples.

9.3 Consider a situation where samples are taken in pairs, X_i and Y_i; $i = 1, 2, \ldots, n$. We are told that the means of x_i and y_i are the same. We are also told that pairs are independent of one another, that is, $E((X_i + Y_i)(X_j + Y_j)) = E(X_i + Y_i)\ E(X_j + Y_j)$; $i \neq j$, but that X_i is correlated with Y_i.

- **(a)** Indicate the estimate of the common mean.
- **(b)** Find the confidence interval for the estimate for a given number of samples.
- **(c)** Explain how the number of samples required to achieve a certain confidence interval is a function of the correlation between pairs.

9.4 Generate a binomial distribution for six trials and probability of success 0.3. Perform the chi-square (χ^2) test on the histogram for probability of success 0.8.

9.5 The probability density and the probability distribution of the Pareto distribution are respectively given by

$$f(t) = \begin{cases} \dfrac{\alpha a^\alpha}{t^{\alpha+1}}; & \alpha, a > 0, t \geq a \\ 0; & \text{otherwise} \end{cases}$$

$$F(t) = \begin{cases} 1 - \left(\dfrac{a}{t}\right)^\alpha; & \alpha, a > 0, t \geq a \\ 0; & \text{otherwise} \end{cases}$$

Find the transformation, that generates the Pareto distribution from the $U(0, 1)$ distribution.

9.6 Repeat the previous exercise for the Weibull distribution.

$$f(t) = \begin{cases} \alpha\beta^{-\alpha} t^{\alpha-1} e^{-(t/\beta)^\alpha}; & \alpha, \beta > 0, t \geq 0 \\ 0; & \text{otherwise} \end{cases}$$

$$F(t) = \begin{cases} 1 - e^{-(t/\beta)^\alpha}; & \alpha, \beta > 0, t \geq 0 \\ 0; & \text{otherwise} \end{cases}$$

9.7 Reproduce a spreadsheet similar to the associated sheets 3 and 4 for the case of a Gaussian random variable with mean 4 and variance 3. Show 99.9% confidence intervals.

9.8 Repeat Example 9.9 for the Poisson distribution.

9.9 Using the approach in Example 9.13, generate a traffic stream whose messages arrive at intervals governed by the Pareto distribution.

9.10 Repeat Example 9.15, except test the Rayleigh distributed samples against a Rayleigh distribution.

9.11 In Section 2.7.2, an example of a Markov chain involving a multiplexer was given. Write a Matlab program for simulating this system.

INDEX

Modeling and Analysis of Telecommunications Networks, by Jeremiah F. Hayes and Thimma V. J. Ganesh Babu
ISBN 0-471-34845-7