

Kleinrock

QUEUEING SYSTEMS

Volume I: Theory



Wiley-
Interscience

QUEUEING SYSTEMS

VOLUME J: THEORY

Leonard Kleinrock

*Professor
Computer Science Department
School of Engineering and Applied Science
University of California, Los Angeles*

A Wiley-Interscience Publication

John Wiley & Sons

New York • Chichester • Brisbane • Toronto

"Ah, 'All things come to those who wait.'
They come, but *often* come too late."

From Lady Mary M. Currie: *Tout Vient à Qui Sait Attendre* (1890)

27947

Copyright © 1975, by John Wiley & Sons, Inc.

All rights reserved. Published simultaneously in Canada.

Reproduction or translation of any part of this work beyond that permitted by Sections 107 or 108 of the 1976 United States Copyright Act without the permission of the copyright owner is unlawful. Requests for permission or further information should be addressed to the Permissions Department, John Wiley & Sons, Inc.

Library of Congress Cataloging in Publication Data:

Kleinrock, Leonard.

Queueing systems.

"A Wiley-Interscience publication."

CONTENTS: v. 1. Theory.

1. Queueing theory. I. Title.

T57.9.K6 **519.8'2** 74-9846

ISBN 0-471-49110-1

13 1415

Preface

How much time did you waste waiting in line this week? It seems we cannot escape frequent delays, and they are getting progressively worse! In this text we study the phenomena of standing, waiting, and serving, and we call this study *queueing theory*.

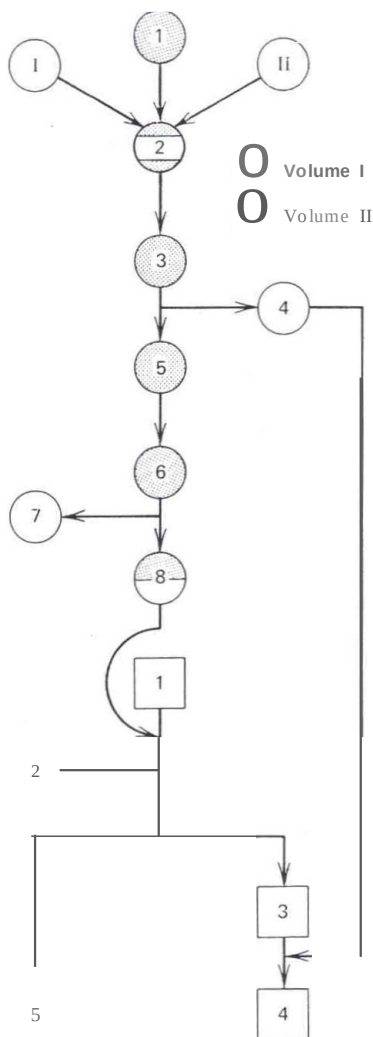
Any system in which arrivals place demands **upon** a finite-capacity resource may be termed a queueing system. In particular, if the arrival times of these demands are unpredictable, or if the size of these demands is unpredictable, then conflicts for the use of the resource will arise and queues of waiting customers will form. The lengths of these queues depend upon two aspects of the flow pattern: first, they depend upon the *average rate* at which demands are placed upon the resource; and second, they depend upon the *statistical fluctuations* of this rate. Certainly, when the average rate exceeds the capacity, then the system breaks down and unbounded queues will begin to form; it is the effect of this average overload which then dominates the growth of queues. However, even if the average rate is less than the system capacity, then here, too, we have the formation of queues due to the statistical fluctuations and spurts of arrivals that may occur; the effect of these variations is greatly magnified when the average load approaches (but does not necessarily exceed) that of the system capacity. The simplicity of these queueing structures is deceptive, and in our studies we will often find ourselves in deep analytic waters. Fortunately, a familiar and fundamental law of science permeates our queueing investigations. This law is the conservation of flow, which states that the rate at which flow increases within a system is equal to the difference between the flow rate into and the flow rate out of that system. This observation permits us to write down the basic system equations for rather complex structures in a relatively easy fashion.

The purpose of this book, then, is to present the theory of queues at the first-year graduate level. It is assumed that the student has been exposed to a first course in probability theory; however, in Appendix II of this text we give a probability theory refresher and state the basic principles that we shall need. It is also helpful (but not necessary) if the student has had some exposure to transforms, although in this case we present a rather complete

transform theory refresher in Appendix I. The student is advised to read both appendices before proceeding with the text itself. Whereas our material is presented in the language of mathematics, we do take great pains to give as informal a presentation as possible in order to strike a balance between the abstractions usually encountered in such a study and the basic need for understanding and applying these tools to practical systems. We feel that a satisfactory middle ground has been established that will neither offend the mathematician nor confound the practitioner. At times we have relaxed the rigor in proofs of uniqueness, existence, and convergence in order not to cloud the main thrust of a presentation. At such times the reader is referred to some of the other books on the subject. We have refrained from using the dull "theorem-proof" approach; rather, we lead the reader through a natural sequence of steps and together we "discover" the result. One finds that previous presentations of this material are usually either too elementary and limited or far too elegant and precise, and almost all of them badly neglect the applications; we feel that the need for a book such as this, which treads the boundary inbetween, is necessary and useful. This book was written over a period of five years while being used as course notes for a one-year (and later a two-quarter) sequence in queueing systems at the University of California, Los Angeles. The material was developed in the Computer Science Department within the School of Engineering and Applied Science and has been tested successfully in the most critical and unforgiving of all environments, namely, that of the graduate student. This text is appropriate not only for computer science departments, but also for departments of engineering, operations research, mathematics, and many others within science, business, management and planning schools.

In order to describe the contents of this text, we must first describe the very convenient shorthand notation that has been developed for the specification of queueing systems. It basically involves the three-part descriptor $A/B/m$ that denotes an m -server queueing system, where A and B describe the interarrival time distribution and the service time distribution, respectively. A and B take on values from the following set of symbols whose interpretation is given in terms of distributions within parentheses: M (exponential); E_r (r -stage Erlangian); H_r (R -stage hyperexponential); D (deterministic); G (general). Occasionally, some other specially defined symbols are used. We sometimes need to specify the system's storage capacity (which we denote by K) or perhaps the size of the customer population (which we denote by M), and in these cases we adopt the five-part descriptor $A/B/m/K/M$; if either of these last two descriptors is absent, then we assume it takes on the value of infinity. Thus, for example, the system $D/M/2/20$ is a two-server system with constant (deterministic) interarrival times, with exponentially distributed service times, and with a system storage capacity of size 20.

This is Volume I (Theory) of a two-volume series, the second of which is devoted to computer applications of this theory. The text of Volume I (which consists of four parts) begins in Chapter I with an introduction to queueing systems, how they fit into the general scheme of systems of flow, and a discussion of how one specifies and evaluates the performance of a queueing system. Assuming a knowledge of (or after reviewing) the material in Appendices I and II, the reader may then proceed to Chapter 2, where he is warned to take care! Section 2.1 is essential and simple. However, Sections 2.2, 2.3, and 2.4 are a bit "heavy" for a first reading in queueing systems, and it would be quite reasonable if the reader were to skip these sections at this point, proceeding directly to Section 2.5, in which the fundamental birth-death process is introduced and where we first encounter the use of a-transforms and Laplace transforms. Once these *preliminaries* in Part I are established one may proceed with the *elementary queueing theory* presented in Part II. We begin in Chapter 3 with the general equilibrium solution to birth-death processes and devote most of the chapter to providing simple yet important examples. Chapter 4 generalizes this treatment, and it is here where we discuss the method of stages and provide an introduction to networks of Markovian queues. Whereas Part II is devoted to algebraic and transform oriented calculations, Part III returns us once again to probabilistic (as well as transform) arguments. This discussion of *intermediate queueing theory* begins with the important M/G/I queue (Chapter 5) and then proceeds to the dual G/M/I queue and its natural generalization to the system G/M/m (Chapter 6). The material on collective marks in Chapter 7 develops the probabilistic interpretation of transforms. Finally, the *advanced material* in Part IV leads us to the queue G/G/I in Chapter 8; this difficult system (whose mean wait cannot even be expressed simply in terms of the system parameters) is studied through the use of the spectral solution to Lindley's integral equation. An approximation to the precedence structure among chapters in these two volumes is given below. In this diagram we have represented chapters in Volume I as numbers enclosed in circles and have used small squares for Volume II. The shading for the Volume I nodes indicates an appropriate amount of material for a relatively leisurely first course in queueing systems that can easily be accomplished in one semester or can be comfortably handled in a one-quarter course. The shading of Chapter 2 is meant to indicate that Sections 2.2-2.4 may be omitted on a first reading, and the same applies to Sections 8.3 and 8.4. A more rapid one-semester pace and a highly accelerated one-quarter pace would include all of Volume I in a single course. We close Volume I with a summary of important equations, developed throughout the book, which are grouped together according to the class of queueing system involved; this list of results then serves as a "handbook" for later use by the reader in concisely summarizing the principal results of this text. The results



are keyed to the page where they appear in order to simplify the task of locating the explanatory material associated with each result.

Each chapter contains its own list of references keyed alphabetically to the author and year; for example, [KLEI 74] would reference this book. All equations of importance have been marked with the symbol - , and it is these which are included in the summary of important equations. Each chapter includes a set of exercises which, in some cases, extend the material in that chapter; the reader is urged to work them out.

the face of the real world's complicated models, the mathematicians proceeded to advance the field of queueing theory rapidly and elegantly. The frontiers of this research proceeded into the far reaches of deep and complex mathematics. It was soon found that the really interesting models did not yield to solution and the field quieted down considerably. It was mainly with the advent of digital computers that once again the tools of queueing theory were brought to bear on a class of practical problems, but this time with great success. The fact is that at present, one of the few tools we have for analyzing the performance of computer systems is that of queueing theory, and this explains its popularity among engineers and scientists today. A wealth of new problems are being formulated in terms of this theory and new tools and methods are being developed to meet the challenge of these problems. Moreover, the application of digital computers in solving the equations of queueing theory has spawned new interest in the field. It is hoped that this two-volume series will provide the reader with an appreciation for and competence in the methods of analysis and application as we now see them.

I take great pleasure in closing this Preface by acknowledging those individuals and institutions that made it possible for me to bring this book into being. First, I would like to thank all those who participated in creating the stimulating environment of the Computer Science Department at UCLA, which encouraged and fostered my effort in this direction. Acknowledgment is due the Advanced Research Projects Agency of the Department of Defense, which enabled me to participate in some of the most exciting and advanced computer systems and networks ever developed. Furthermore, the John Simon Guggenheim Foundation provided me with a Fellowship for the academic year 1971-1972, during which time I was able to further pursue my investigations. Hundreds of students who have passed through my queueing-systems courses have in major and minor ways contributed to the creation of this book, and I am happy to acknowledge the special help offered by Arne Nilsson, Johnny Wong, Simon Lam, Fouad Tobagi, Farouk Kamoun, Robert Rice, and Thomas Sikes. My academic and professional colleagues have all been very supportive of this endeavour. To the typists I owe all. By far the largest portion of this book was typed by Charlotte LaRoche, and I will be forever in her debt. To Diana Skocypec and Cynthia Ellman I give my deepest thanks for carrying out the enormous task of proofreading and correction-making in a rapid, enthusiastic, and supportive fashion. Others who contributed in major ways are Barbara Warren, Jean Dubinsky, Jean D'Fucci, and Gloria Roy. I owe a great debt of thanks to my family (and especially to my wife, Stella) who have stood by me and supported me well beyond the call of duty or marriage contract. Lastly, I would certainly be remiss in omitting an acknowledgement to my ever-faithful dictating machine, which was constantly talking back to me.

LEONARD KLEINROCK

March, 1974

Contents

VOLUME I

PART I: PRELIMINARIES

Chapter 1 Queueing Systems	3
1.1. Systems of Flow .	3
1.2. The Specification and Measure of Queueing Systems	8
Chapter 2 Some Important Random Processes	10
2.1. Notation and Structure for Basic Queueing Systems	10
2.2. Definition and Classification of Stochastic Processes	19
2.3. Discrete-Time Markov Chains	26
2.4. Continuous-Time Markov Chains .	44
2.5. Birth-Death Processes.	53

PART II: ELEMENTARY QUEUEING THEORY

Chapter 3 Birth-Death Queueing Systems in Equilibrium	89
3.1. General Equilibrium Solution	90
3.2. $M/M/I$: The Classical Queueing System .	94
3.3. Discouraged Arrivals	99
3.4. $M/M/\infty$: Responsive Servers (Infinite Number of Servers)	101
3.5. $M/M/m$: The m-Server Case.	102
3.6. $M/M/I/K$: Finite Storage	103
3.7. $M/M/m/m$: m-Server Loss Systems .	105
3.8. $M/M/I/IM$: Finite Customer Population-Single Server	106
3.9. $M/M/roIM$: Finite Customer Population- "Infinite" Number of Servers	107
3.10. $M/M/m/K/M$: Finite Population, m-Server Case, Finite Storage	108

Chapter 4	Markovian Queues in Equilibrium	I 15
4.1.	The Equilibrium Equations .	115
4.2.	The Method of Stages- Erlangian Distribution E.	119
4.3.	The Queue $M/Er/1$	126
4.4.	The Queue $ErlM/I$	130
4.5.	Bulk Arrival Systems	134
4.6.	Bulk Service Systems	137
4.7.	Series-Parallel Stages: Generalizations	139
4.8.	Networks of Markovian Queues	147

PART III: INTERMEDIATE QUEUEING THEORY

Chapter 5	The Queue $M/G/I$	167
5.1.	The $M/G/I$ System	168
5.2.	The Paradox of Residual Life: A Bit of Renewal Theory .	169
5.3.	The Imbedded Markov Chain	174
5.4.	The Transition Probabilities .	177
5.5.	The Mean Queue Length.	180
5.6.	Distribution of Number in System .	191
5.7.	Distribution of Waiting Time	196
5.8.	The Busy Period and Its Duration .	206
5.9.	The Number Served in a Busy Period	216
5.10.	From Busy Periods to Waiting Times	219
5.11.	Combinatorial Methods	223
5.12.	The Takács Integrodifferential Equation .	226
Chapter 6	The Queue $G/M/m$	241
6.1.	Transition Probabilities for the Imbedded Markov Chain ($G/M/m$)	241
6.2.	Conditional Distribution of Queue Size .	246
6.3.	Conditional Distribution of Waiting Time	250
6.4.	The Queue $G/M/I$	251
6.5.	The Queue $G/M/m$	253
6.6.	The Queue $G/M/2$	256
Chapter 7	The Method of Collective Marks	261
7.1.	The Marking of Customers	261
7.2.	The Catastrophe Process	267

PART IV: ADVANCED MATERIAL

Chapter 8 The Queue $G/G/1$	275
8.1. Lindley's Integral Equation	275
8.2. Spectral Solution to Lindley's Integral Equation	283
8.3. Kingman's Algebra for Queues	299
8.4. The Idle Time and Duality	304
Epilogue	319
 Appendix I: Transform Theory Refresher: z-Transform and Laplace Transform	
1.1. Why Transforms?	321
1.2. The z-Transform	327
1.3. The Laplace Transform	338
1.4. Use of Transforms in the Solution of Difference and Differential Equations	355
 Appendix II: Probability Theory Refresher	
II.1. Rules of the Game	363
11.2. Random Variables	368
11.3. Expectation	377
11.4. Transforms, Generating Functions, and Characteristic Functions	381
11.5. Inequalities and Limit Theorems	388
11.6. Stochastic Processes	393
 <i>Glossary of Notation</i>	396
 <i>Summary of Important Results</i>	400
 <i>Index</i>	411

VOLUME 1/

Chapter 1 A Queueing Theory Primer

1. Notation
2. General Results
3. Markov, Birth-Death, and Poisson Processes
4. The $M/M/1$ Queue
5. The MIM/m Queueing System
6. Markovian Queueing Networks
7. The $M/G/1$ Queue
8. The $GIM/1$ Queue
9. The GIM/m Queue
10. The $G/G/1$ Queue

Chapter 2 Bounds, Inequalities and Approximations

1. The Heavy-Traffic Approximation
2. An Upper Bound for the Average Wait
3. Lower Bounds for the Average Wait
4. Bounds on the Tail of the Waiting Time Distribution
5. Some Remarks for GIG/m
6. A Discrete Approximation
7. The Fluid Approximation for Queues
8. Diffusion Processes
9. Diffusion Approximation for $MIG/1$
10. The Rush-Hour Approximation

Chapter 3 Priority Queueing

1. The Model
2. An Approach for Calculating Average Waiting Times
3. The Delay Cycle, Generalized Busy Periods, and Waiting Time Distributions
4. Conservation Laws
5. The Last-Come- First-Serve Queueing Discipline

6. Head-of-the-Line Priorities
7. Time-Dependent Priorities
8. Optimal Bribing for Queue Position
9. Service-Time-Dependent Disciplines

Chapter 4 Computer Time-Sharing **and** Multiaccess Systems

1. Definitions and Models
2. Distribution of Attained Service
3. The Batch Processing Algorithm
4. The Round-Robin Scheduling Algorithm
5. The Last-Come-First-Serve Scheduling Algorithm
6. The FB Scheduling Algorithm
7. The Multilevel Processor Sharing Scheduling Algorithm
8. Selfish Scheduling Algorithms
9. A Conservation Law for Time-Shared Systems
10. Tight Bounds on the Mean Response Time
11. Finite Population Models
12. Multiple-Resource Models
13. Models for Multiprogramming
14. Remote Terminal Access to Computers

Chapter 5 Computer-Communication Networks

- I. Resource Sharing
2. Some Contrasts and Trade-Offs
3. Network Structures and Packet Switching
4. The ARPANET-An Operational Description of an Existing Network
5. Definitions, the Model, and the Problem Statements
6. Delay Analysis
7. The Capacity Assignment Problem
8. The Traffic Flow Assignment Problem
9. The Capacity and Flow Assignment Problem
10. Some Topological Considerations-Applications to the ARPANET
- II. Satellite Packet Switching
12. Ground Radio Packet Switching

Chapter 6 Computer-Communication Networks
Measurement, Flow Control and ARPANET Traps

1. Simulation and Routing
2. Early ARPANET Measurements
3. Flow Control
4. Lockups, Degradations and Traps
5. Network Throughput
6. One Week of ARPANET Data
7. Line Overhead in the ARPANET
8. Recent Changes to the Flow Control Procedure
9. The Challenge of the Future

Glossary

Summary of Results

Index

QUEUEING SYSTEMS

VOLUME I: *THEORY*

PART I

PRELIMINARIES

It is difficult to see the forest for the trees (especially if one is in a mob rather than in a well-ordered queue). Likewise, it is often difficult to see the impact of a collection of mathematical results as you try to master them; it is only after one gains the understanding and appreciation for their application to real-world problems that one can say with confidence that he understands the use of a set of tools.

The two chapters contained in this preliminary part are each extreme in opposite directions. The first chapter gives a global picture of where queueing systems arise and why they are important. Entertaining examples are provided as we lure the reader on. In the second chapter, on random processes, we plunge deeply into mathematical definitions and techniques (quickly losing sight of our long-range goals); the reader is urged not to falter under this siege since it is perhaps the worst he will meet in passing through the text. Specifically, Chapter 2 begins with some very useful graphical means for displaying the dynamics of customer behavior in a queueing system. We then introduce stochastic processes through the study of customer arrival, behavior, and backlog in a very general queueing system and carefully lead the reader to one of the most significant results in queueing theory, namely, Little's result, using very simple arguments. Having thus introduced the concept of a stochastic process we then offer a rather compact treatment which compares many well-known (but not well-distinguished) processes and casts them in a common terminology and notation, leading finally to Figure 2.4 in which we see the basic relationships among these processes; the reader is quickly brought to realize the central role played by the Poisson process because of its position as the common intersection of all the stochastic processes considered in this chapter. We then give a treatment of Markov chains in discrete and continuous time; these sections are perhaps the toughest sledding for the novice, and it is perfectly acceptable if he passes over some of this material on a first reading. At the conclusion of Section 2.4 we find ourselves face to face with the important **birth-death** processes and it is here

where things begin to take on a relationship to physical systems once again. In fact, it is not unreasonable for the reader to begin with Section 2.5 of this chapter since the treatment following is (almost) self-contained from there throughout the rest of the **text**. Only occasionally do we find a need for the more detailed material in Sections 2.3 and 2.4. If **the** reader perseveres through Chapter 2 he will have set the stage for the balance of the textbook.

Queueing Systems

One of life's more disagreeable activities, namely, waiting in line, is the delightful subject of this book. One might reasonably ask, "What does it profit a man to study such unpleasant phenomena?" The answer, of course, is that through understanding we gain compassion, and it is exactly this which we need since people will be waiting in longer and longer queues as civilization progresses, and we must find ways to tolerate these unpleasant situations. Think for a moment how much time is spent in one's daily activities waiting in some form of a queue: waiting for breakfast; stopped at a traffic light; slowed down on the highways and freeways; delayed at the entrance to one's parking facility; queued for access to an elevator; standing in line for the morning coffee; holding the telephone as it rings, and so on. The list is endless, and too often also are the queues.

The orderliness of queues varies from place to place around the world. For example, the English are terribly susceptible to formation of orderly queues, whereas some of the Mediterranean peoples consider the idea ludicrous (have you ever tried clearing the embarkation procedure at the Port of Brindisi?). A common slogan in the U.S. Army is, "Hurry up and wait." Such is the nature of the phenomena we wish to study.

1.1. SYSTEMS OF FLOW

Queueing systems represent an example of a much broader class of interesting dynamic systems, which, for convenience, we refer to as "systems of flow." A flow system is one in which some *commodity* flows, moves, or is transferred through one or more finite-capacity *channels* in order to go from one point to another. For example, consider the flow of automobile traffic through a road network, or the transfer of goods in a railway system, or the streaming of water through a dam, or the transmission of telephone or telegraph messages, or the passage of customers through a supermarket checkout counter, or the flow of computer programs through a time-sharing computer system. In these examples the commodities are the automobiles, the goods, the water, the telephone or telegraph messages, the customers, and the programs, respectively; the channel or channels are the road network,

the railway network, the dam, the telephone or telegraph network, the supermarket checkout counter, and the computer processing system, respectively. The “finite capacity” refers to the fact that the channel can satisfy the demands (placed upon it by the commodity) at a finite rate only. It is clear that the analyses of many of these systems require analytic tools drawn from a variety of disciplines and, as we shall see, queueing theory is just one such discipline.

When one analyzes systems of flow, they naturally break into two classes: *steady* and *unsteady* flow. The first class consists of those systems in which the flow proceeds in a predictable fashion. That is, the quantity of flow is exactly known and is constant over the interval of interest; the time when that flow appears at the channel, and how much of a demand that flow places upon the channel is known and constant. These systems are trivial to analyze in the case of a *single channel*. For example, consider a pineapple factory in which empty tin cans are being transported along a conveyor belt to a point at which they must be filled with pineapple slices and must then proceed further down the conveyor belt for additional operations. In this case, assume that the cans arrive at a constant rate of one can per second and that the pineapple-filling operation takes nine-tenths of one second per can. These numbers are constant for all cans and all filling operations. Clearly this system will function in a reliable and smooth fashion as long as the assumptions stated above continue to exist. We may say that the *arrival rate* R is one can per second and the maximum service rate (or *capacity*) C is $1/0.9 = 1.11111 \dots$ filling operations per second. The example above is for the case $R < C$. However, if we have the condition $R > C$, we all know what happens: cans and/or pineapple slices begin to inundate and overflow in the factory! Thus we see that *the mean capacity of the system must exceed the average flow requirements if chaotic congestion is to be avoided*; this is true for all systems of flow. This simple observation tells most of the story. Such systems are of little interest theoretically.

The more interesting case of steady flow is that of a *network* of channels. For stable flow, we obviously require that $R < C$ on each channel in the network. However we now run into some serious combinatorial problems. For example, let us consider a railway network in the fictitious land of Hatafla. See Figure 1.1. The scenario here is that figs grown in the city of Abra must be transported to the destination city of Cadabra, making use of the railway network shown. The numbers on each channel (section of railway) in Figure 1.1 refer to the maximum number of bushels of figs which that channel can handle per day. We are now confronted with the following fig flow problem: How many bushels of figs per day can be sent from Abra to Cadabra and in what fashion shall this flow of figs take place? The answer to such questions of maximal “traffic” flow in a variety of networks is nicely

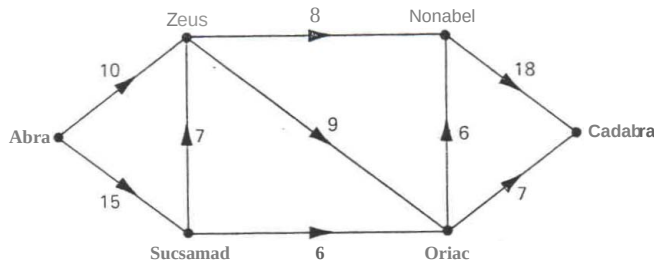


Figure 1.1 Maximal flow problem.

settled by a well-known result in network flow theory referred to as the *max-flow-min-cut* theorem. To state this theorem, we first define a *cut* as a set of channels which, once removed from the network, will separate all possible flow from the origin (Abra) to the destination (Cadabra). We define the *capacity* of such a cut to be the total fig flow that can travel across that cut in the direction from origin to destination. For example, one cut consists of the branches from Abra to Zeus, Sucsamad to Zeus, and Sucsamad to Oriac; the capacity of this cut is clearly 23 bushels of figs per day. The max-flow-min-cut theorem states that the maximum flow that can pass between an origin and a destination is the minimum capacity of all cuts. In our example it can be seen that the maximum flow is therefore 21 bushels of figs per day (work it out). In general, one must consider *all* cuts that separate a given origin and destination. This computation can be enormously time consuming. Fortunately, there exists an extremely powerful method for finding not only what is the maximum flow, but also which flow pattern achieves this maximum flow. This procedure is known as the *labeling algorithm* (due to Ford and Fulkerson [FORD 62]) and is efficient in that the computational requirement grows as a small power of the number of nodes; we present the algorithm in Volume II, Chapter 5.

In addition to maximal flow problems, one can pose numerous other interesting and worthwhile questions regarding flow in such networks. For example, one might inquire into the minimal cost network which will support a given flow if we assign costs to each of the channels. Also, one might ask the same questions in networks when more than one origin and destination exist. Complicating matters further, we might insist that a given network support flow of various kinds. for example, bushels of figs, cartons of cartridges and barrels of oil. This multicommodity flow problem is an extremely difficult one, and its solution typically requires considerable computational effort. These and numerous other significant problems in network flow theory are addressed in the comprehensive text by Frank and Frisch [FRAN 71] and we shall see them again in Volume II, Chapter 5. Network flow theory itself requires methods from graph theory, combinatorial

mathematics, optimization theory, mathematical programming, and heuristic programming.

The *second* class into which systems of flow may be divided is the class of random or stochastic flow problems. By this we mean that the *times* at which demands for service (use of the channel) arrive are uncertain or unpredictable, and also that the *size* of the demands themselves that are placed upon the channel are unpredictable. The randomness, unpredictability, or unsteady nature of this flow lends considerable complexity to the solution and understanding of such problems. Furthermore, it is clear that most real-world systems fall into this category. Again, the simplest case is that of random flow through a *single* channel; whereas in the case of deterministic or steady flow discussed earlier in which the single-channel problems were trivial, we have now a case where these single-channel problems are extremely challenging and, in fact, techniques for solution to the single-channel or single-server problem comprise much of modern queueing theory.

For example, consider the case of a computer center in which computation requests are served making use of a batch service system. In such a system, requests for computation arrive at unpredictable times, and when they do arrive, they may well find the computer busy servicing other demands. If, in fact, the computer is idle, then typically a new demand will begin service and will be **run** until it is completed. On the other hand, if the system is busy, then this job will wait on a queue until it is selected for service from among those that are waiting. Until that job is carried to completion, it is usually the case that neither the computation center nor the individual who has submitted the program knows the extent of the demand in terms of computational effort that this program will place upon the system; in this sense the service requirement is indeed unpredictable.

A variety of natural questions present themselves to which we would like intelligent and complete answers. How long, for example, may a job expect to wait on queue before entering service? How many jobs will be serviced before the one just submitted? For what fraction of the day will the computation center be busy? How long will the intervals of continual busy work extend? Such questions require answers regarding the probability of certain periods and numbers or perhaps merely the average values for these quantities. Additional considerations, such as machine breakdown (a not uncommon condition), complicate the issue further; in this case it is clear that some preemptive event prevents the completion of the job currently in **service**. Other interesting effects can take place where jobs are not serviced according to their order of arrival. Time-shared computer systems, for example, employ rather complex scheduling and servicing algorithms, which, in fact, we explore in Volume II, Chapter 4.

The tools necessary for solving single-channel random-flow problems are

contained and described within queueing theory, to which much of this text devotes itself. This requires a background in probability theory as well as an understanding of complex variables and some of the usual transform-calculus methods; this material is reviewed in Appendices I and II.

As in the case of deterministic flow, we may enlarge our scope of problems to that of *networks* of channels in which random flow is encountered. An example of such a system would be that of a computer network. Such a system consists of computers connected together by a set of communication lines where the capacity of these lines for carrying information is finite. Let us return to the fictitious land of Hatafla and assume that the railway network considered earlier is now in fact a computer network. Assume that users located at Abra require computational effort on the facility at Cadabra. The particular times at which these requests are made are themselves unpredictable, and the commands or instructions that describe these requests are also of unpredictable length. It is these commands which must be transmitted to Cadabra over our communication net as messages. When a message is inserted into the network at Abra, and after an appropriate decision rule (referred to as a routing procedure) is accessed, then the message proceeds through the network along some path. If a portion of this path is busy, and it may well be, then the message must queue up in front of the busy channel and wait for it to become free. Constant decisions must be made regarding the flow of messages and routing procedures. Hopefully, the message will eventually emerge at Cadabra, the computation will be performed, and the results will then be inserted into the network for delivery back at Abra.

It is clear that the problems exemplified by our computer network involve a variety of extremely complex queueing problems, as well as network flow and decision problems. In an earlier work [KLEI 64] the author addressed himself to certain aspects of these questions. We develop the analysis of these systems later in Volume II, Chapter 5.

Having thus classified *systems of flow, we hope that the reader understands where in the general scheme of things the field of queueing theory may be placed. The methods from this theory are central to analyzing most stochastic flow problems, and it is clear from an examination of the current literature that the field and in particular its applications are growing in a viable and purposeful fashion.

- The classification described above places queueing systems within the class of systems of flow. This approach identifies and emphasizes the fields of application for queueing theory. An alternative approach would have been to place queueing theory as belonging to the field of applied stochastic processes; this classification would have emphasized the mathematical structure of queueing theory rather than its applications. The point of view taken in this two-volume book is the former one, namely, with application of the theory as its major goal rather than extension of the mathematical formalism and results.

1.2. THE SPECIFICATION AND MEASURE OF QUEUEING SYSTEMS

In order to completely specify a queueing system, one must identify the stochastic processes that describe the arriving stream as well as the structure and discipline of the service facility. Generally, the arrival process is described in terms of the probability distribution of the *interarrival times* of customers and is denoted $A(t)$, where*

$$A(t) = P[\text{time between arrivals} \leq t] \quad (1.1)$$

The assumption in most of queueing theory is that these interarrival times are independent, identically distributed random variables (and, therefore, the stream of arrivals forms a stationary renewal process; see Chapter 2). Thus, only the distribution $A(t)$, which describes the time between arrivals, is usually of significance. The second statistical quantity that must be described is the amount of demand these arrivals place upon the channel; this is usually referred to as the *service time* whose probability distribution is denoted by $B(x)$, that is,

$$B(x) = P[\text{service time} \leq x] \quad (1.2)$$

Here service time refers to the length of time that a customer spends in the service facility.

Now regarding the structure and discipline of the service facility, one must specify a variety of additional quantities. One of these is the extent of *storage capacity* available to hold waiting customers and typically this quantity is described in terms of the variable K ; often K is taken to be infinite. An additional specification involves the *number of service stations* available, and if more than one is available, then perhaps the distribution of service time will differ for each, in which case the distribution $B(x)$ will include a subscript to indicate that fact. On the other hand, it is sometimes the case that the arriving stream consists of more than one identifiable *class* of customers; in such a case the interarrival distribution $A(t)$ as well as the service distribution $B(x)$ may each be characteristic of each class and will be identified again by use of a subscript on these distributions. Another important structural description of a queueing system is that of the queueing *discipline*; this describes the order in which customers are taken from the queue and allowed into service. For example, some standard queueing disciplines are first-come-first-serve (FCFS), last-come-first-serve (LCFS), and random order of service. When the arriving customers are distinguishable according to groups, then we encounter the case of *priority* queueing disciplines in which priority

* The notation $P[A]$ denotes, as usual, the "probability of the event A ,"

among groups may be established. A further statement regarding the *availability* of the service facility is also necessary in case the service facility is occasionally required to *pay* attention to other tasks (as, for example, its own breakdown). Beyond this, queueing systems may enjoy customer behavior in the form of *defections* from the queue, *jockeying* among the many queues, *balking* before entering a queue, *bribing* for queue position, *cheating* for queue position, and a variety of other interesting and not-unexpected humanlike characteristics. We will encounter these as we move through the text in an orderly fashion (first-come-first-serve according to page number).

Now that we have indicated how one must specify a queueing system, it is appropriate that we identify the measures of performance and effectiveness that we shall obtain by analysis. Basically, we are interested in the *waiting time* for a customer, the *number of customers* in the system, the *length of a busy period* (the continuous interval during which the server is busy), the *length of an idle period*, and the current *work backlog* expressed in units of time. All these quantities are random variables and thus we seek their complete probabilistic description (i.e., their probability distribution function). Usually, however, to give the distribution function is to give more than one can easily make use of. Consequently, we often settle for the first few moments (mean, variance, etc.).

Happily, we shall begin with simple considerations and develop the tools in a straightforward fashion, paying attention to the essential details of analysis. In the following pages we will encounter a variety of simple queueing problems, simple at least in the sense of description and usually rather sophisticated in terms of solution. However, in order to do this properly, we first devote our efforts in the following chapter to describing some of the important random processes that make up the arrival and service processes in our queueing systems.

REFERENCES

- FORD 62 Ford, L. R. and D. R. Fulkerson, *Flows in Networks*, Princeton University Press (Princeton, N.J.), 1962.
- FRAN 71 Frank, H. and I. T. Frisch, *Communication, Transmission, and Transportation Networks*, Addison-Wesley (Reading, Mass.), 1971.
- KLEI 64 Kleinrock, L., *Communication Nets; Stochastic Message Flow and Delay*. McGraw-Hill (New York), 1964, out of print. Reprinted by Dover (New York), 1972.

Some Important Random Processes*

We assume that the reader is familiar with the basic elementary notions, terminology, and concepts of probability theory. The particular aspects of that theory which we require are presented in summary fashion in Appendix II to serve as a review for those readers desiring a quick refresher and reminder; it is recommended that the material therein be reviewed, especially Section 11.4 on transforms, generating functions, and characteristic functions.

Included in Appendix " are the following important definitions, concepts, and results:

Sample space, events, and probability.

- Conditional probability, statistical independence, the law of total probability, and Bayes' theorem.

A real random variable, its probability distribution function (PDF), its probability density function (pdf), and their simple properties.

- Events related to random variables and their probabilities.
 - Joint distribution functions.
 - Functions of a random variable and their density functions.
 - Expectation.
 - Laplace transforms, generating functions, and characteristic functions and their relationships and properties.^t
 - Inequalities and limit theorems.
- Definition of a stochastic process.

2.1. NOTATION AND STRUCTURE FOR BASIC QUEUEING SYSTEMS

Before we plunge headlong into a step-by-step development of queueing theory from its elementary notions to its intermediate and then finally to some advanced material, it is important first that we understand the basic

* Sections 2.2, 2.3, and 2.4 may be skipped on a first reading.

^t Appendix I is a transform theory refresher. This material is also essential to the proper understanding of this text.

structure of queues. Also, we wish to provide the reader a glimpse as to where we are heading in this journey.

It is our purpose in this section to define some notation, both symbolic and graphic, and then to introduce one of the basic stochastic processes that we find in queueing systems. Further, we will derive a simple but significant result, which relates some first moments of importance in these systems. In so doing, we will be in a position to define the quantities and processes that we will spend many pages studying later in the text.

The system we consider is the very general queueing system $G/G/m$; recall (from the Preface) that this is a system whose interarrival time distribution $A(t)$ is completely arbitrary and whose service time distribution $B(x)$ is also completely arbitrary (all interarrival times and service times are assumed to be independent of each other). The system has m servers and order of service is also quite arbitrary (in particular, it need not be first-come-first-serve). We focus attention on the flow of customers as they arrive, pass through, and eventually leave this system: as such, we choose to number the customers with the subscript n and define C_n as follows:

$$C_n \text{ denotes the } n\text{th customer to enter the system} \quad (2.1)$$

Thus, we may portray our system as in Figure 2.1 in which the box represents the queueing system and the flow of customers both in and out of the system is shown. One can immediately define some random processes of interest. For example, we are interested in $N(t)$ where *

$$N(t) \triangleq \text{number of customers in the system at time } t \quad (2.2)$$

Another stochastic process of interest is the unfinished work $V(t)$ that exists in the system at time t , that is,

$$\begin{aligned} V(t) &\triangleq \text{the unfinished work in the system at time } t \\ &\triangleq \text{the remaining time required to empty the system of all} \\ &\quad \text{customers present at time } t \end{aligned} \quad (2.3)$$

Whenever $V(t) > 0$, then the system is said to be busy, and only when $V(t) = 0$ is the system said to be idle. The duration and location of these busy and idle periods are also quantities of interest.



Figure 2.1 A general queueing system.

- The notation $\triangleq$ is to be read as "equals by definition."

The details of these stochastic processes may be observed first by defining the following variables and then by displaying these variables on an appropriate time diagram to be discussed below. We begin with the definitions. Recalling that the n th customer is denoted by C_n , we define his arrival time to the queueing system as

$$\tau_n \triangleq \text{arrival time for } C_n \quad (2.4)$$

and further define the interarrival time between C_{n-1} and C_n as

$$\begin{aligned} t_n &\triangleq \text{interarrival time between } C_{n-1} \text{ and } C_n \\ &= \tau_n - \tau_{n-1} \end{aligned} \quad (2.5)$$

Since we have assumed that all interarrival times are drawn from the distribution $A(t)$, we have that

$$P[t'' \leq t] = A(t) \quad (2.6)$$

which is independent of n . Similarly, we define the service time for C_n as

$$x_n \triangleq \text{service time for } C_n \quad (2.7)$$

and from our assumptions we have

$$P[X_n \leq x] = B(x) \quad (2.8)$$

The sequences $\{t_n\}$ and $\{x_n\}$ may be thought of as input variables for our queueing system; the way in which the system handles these customers give rise to queues and waiting times that we must now define. Thus, we define the waiting time (time spent in the queue)* as

$$w_n \triangleq \text{waiting time (in queue) for } C_n \quad (2.9)$$

The total time spent in the system by C_n is the sum of his waiting time and service time, which we denote by

$$\begin{aligned} s_n &\triangleq \text{system time (queue plus service) for } C_n \\ &= w_n + x_n \end{aligned} \quad (2.10)$$

Thus we have defined for the n th customer his arrival time, "his" interarrival time, his service time, his waiting time, and his system time. We find

* The terms "waiting time" and "queueing time" have conflicting definitions within the body of queueing-theory literature. The former sometimes refers to the total time spent in the system, and the latter then refers to the total time spent on queue; however, these two definitions are occasionally reversed. We attempt to remove that confusion by defining waiting and queueing time to be the same quantity, namely, the time spent waiting in queue (but not being served); a more appropriate term perhaps would be "wasted time". The total time spent in the system will be referred to as "system time" (occasionally known as "flow time").

expedient at this point to elaborate somewhat further on notation. Let us consider the interarrival time I_n once again. We will have occasion to refer to the limiting random variable i defined by

$$\bar{i} \triangleq \lim_{n \rightarrow \infty} I_n \quad (2.11)$$

which we denote by $I_n \rightarrow i$. (We have already required that the interarrival times I_n have a distribution independent of n , but this will not necessarily be the case with many other random variables of interest.) The typical notation for the probability distribution function (PDF) will be

$$P[t_n \leq t] = A_n(t) \quad (2.12)$$

and for the limiting PDF

$$P[i \leq l] = A(l) \quad (2.13)$$

This we denote by $A_n(l) \rightarrow A(l)$; of course, for the interarrival time we have assumed that $A_n(l) = A(l)$, which gives rise to Eq. (2.6). Similarly, the probability density function (pdf) for t_n and i will be $a_n(l)$ and $a(l)$, respectively, and will be denoted as $a_n(t) \rightarrow a(t)$. Finally, the Laplace transform (see Appendix II) of these pdf's will be denoted by $A_n^*(s)$ and $A^*(s)$, respectively, with the obvious notation $A_n^*(s) \rightarrow A^*(s)$. The use of the letter A (and a) is meant as a cue to remind the reader that they refer to the interarrival time. Of course, the moments of the interarrival time are of interest and they will be denoted as follows*:

$$E[t_n] \triangleq \bar{t}_n \quad (2.14)$$

According to our usual notation, the mean interarrival time for the limiting random variable will be given] by $\bar{i}$ in the sense that $\bar{t}_n \rightarrow \bar{i}$. As it turns out $\bar{i}$, which is the average interarrival time between customers, is used so frequently in our equations that a *special* notation has been adopted as follows:

$$\bar{i} \triangleq \frac{1}{\lambda} \quad (2.15)$$

Thus λ represents the *average arrival rate* of customers to our queueing system. Higher moments of the interarrival time are also of interest and so we define the k th moment by

$$E[\bar{i}^k] \triangleq \bar{i}^k \triangleq a_k \quad k = 0, 1, 2, \dots \quad (2.16)$$

* The notation $E[\quad]$ denotes the expectation of the quantity within square brackets. As shown, we also adopt the overbar notation to denote expectation.

† Actually, we should use the notation $\bar{i}$ with a tilde and a bar, but this is excessive and will be simplified to $\bar{i}$. The same simplification will be applied to many of our other random variables.

In this last equation we have introduced the definition a_k to be the k th moment of the interarrival time i ; this is fairly standard notation and we note immediately from the above that

$$\bar{t} = \frac{1}{\lambda} = a_1 \triangleq a \quad (2.17)$$

That is, three *special* notations exist for the mean interarrival time; in particular, the use of the symbol a is very common and various of these forms will be used throughout the text as appropriate. Summarizing the information with regard to the interarrival time we have the following shorthand glossary:

$$\begin{aligned} t_i &= \text{interarrival time between } C_i \text{ and } C_{i-1} \\ t_n &\rightarrow i, \quad An(t) \rightarrow A(t), \quad an(t) \rightarrow a(t), \quad An^*(s) \rightarrow A^*(s) \\ \bar{t}_n &\rightarrow \bar{t} = \frac{1}{\lambda} = a_1 = a, \quad t_n^k \rightarrow t^k = a_k \end{aligned} \quad (2.18)$$

In a similar manner we identify the notation associated with x_n , μ , and s_n as follows:

$$\begin{aligned} x_n &= \text{service time for } C_n \\ x_n &\rightarrow \tilde{x}, \quad B_n(x) \rightarrow B(x), \quad b_n(x) \rightarrow b(x), \quad B_n^*(s) \rightarrow B^*(s) \\ x_n &\rightarrow \bar{x} = \frac{1}{\mu} = b_1 = b, \quad x_n^k \rightarrow x^k = b_k \end{aligned} \quad (2.19)$$

$$\begin{aligned} w_n &= \text{waiting time for } C_n \\ w_n &\rightarrow \tilde{w}, \quad W_n(y) \rightarrow W(y), \quad w_n(y) \rightarrow w(y), \quad W_n^*(s) \rightarrow W^*(s) \\ \bar{w}_n &\rightarrow \bar{w} = W, \quad w_n^k \rightarrow w^k \end{aligned} \quad (2.20)$$

$$\begin{aligned} s_i &= \text{system time for } C_i \\ s_i &\rightarrow \tilde{s}, \quad Sn(Y) \rightarrow S(y), \quad sn(Y) \rightarrow s(y), \quad Sn^*(s) \rightarrow S^*(s) \\ \bar{s}_n &\rightarrow \bar{s} = T, \quad s_n^k \rightarrow s^k \end{aligned} \quad (2.21)$$

All this notation is self-evident except perhaps for the occasional special symbols used for the first moment and occasionally the higher moments of the random variables involved (that is, the use of the symbols λ , a , μ , b , W , and T). The reader is, at this point, directed to the Glossary for a complete set of notation used in this book.

With the above notation we now suggest a *time-diagram notation* for queues, which permits a graphical view of the dynamics of our queueing system and also provides the details of the underlying stochastic processes. This diagram is shown in Figure 2.2. This particular figure is shown for a

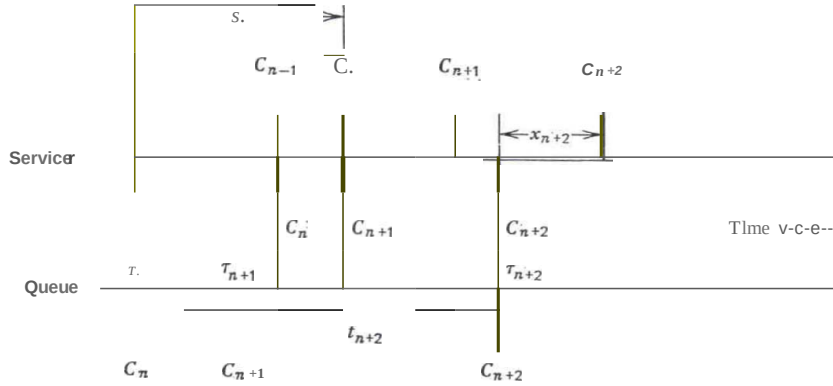


Figure 2.2 Time-diagram notation for queues.

first-come-first-serve order of service, but it is easy to see how the figure may also be made to represent any order of service. In this time diagram the lower horizontal time line represents the queue and the upper horizontal time line represents the service facility; moreover, the diagram shown is for the case of a single server, although this too is easily generalized. An arrow approaching the queue (or service) line from below indicates that an arrival has occurred to the queue (or service facility). Arrows emanating from the line indicate the departure of a customer from the queue (or service facility). In this figure we see that customer C_{n+1} arrives before customer C_n enters service; only when C_n departs from service may C_{n+1} enter service and, of course, these two events occur simultaneously. Notice that when C_{n+2} enters the system he finds it empty and so immediately proceeds through an empty queue directly into the service facility. In this diagram we have also shown the waiting time and the system time for C_n (note that $w_{n+2} = 0$). Thus, as time proceeds we can identify the number of customers in the system $N(t)$, the unfinished work $V(t)$, and also the idle and busy periods. We will find much use for this time-diagram notation in what follows.

In a general queueing system one expects that when the number of customers is large then so is the waiting time. One manifestation of this is a very simple relationship between the mean number in the queueing system, the mean arrival rate of customers to that system, and the mean system time for customers. It is our purpose next to derive that relationship and thereby familiarize ourselves a bit further with the underlying behavior of these systems. Referring back to Figure 2.1, let us position ourselves at the input of the queueing system and count how many customers enter as a function of time. We denote this by $\alpha(t)$ where

$$\alpha(t) \triangleq \text{number of arrivals in } (0, t) \quad (2.22)$$

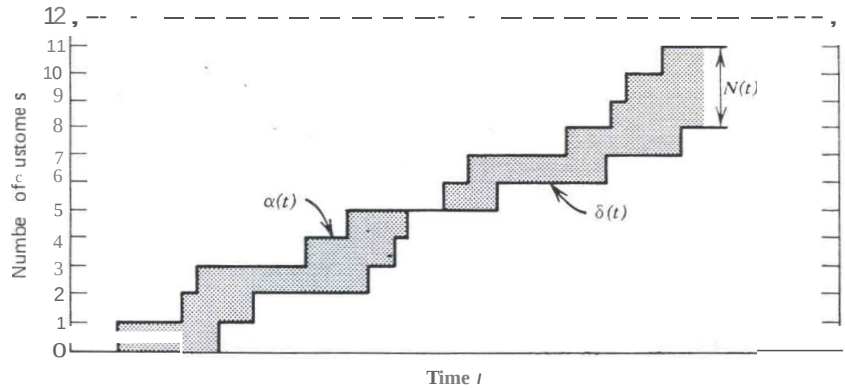


Figure 2.3 Arrivals and departures.

Alternatively, we may position ourselves at the output of the queueing system and count the number of departures that leave; this we denote by

$$\delta(t) \triangleq \text{number of departures in } (0, t) \quad (2.23)$$

Sample functions for these two stochastic processes are shown in Figure 2.3.

Clearly $N(t)$, the number in the system at time t , must be given by

$$N(t) = \alpha(t) - \delta(t)$$

On the other hand, the total area between these two curves up to some point, say t , represents the total time all customers have spent in the system (measured in units of customer-seconds) during the interval $(0, t)$; let us denote this cumulative area by yet). Moreover, let λ_t be defined as the average arrival rate (customers per second) during the interval $(0, t)$; that is,

$$\lambda_t \triangleq \frac{\alpha(t)}{t} \quad (2.24)$$

We may define T_t as the system time per customer averaged over all customers in the interval $(0, t)$; since yet) represents the accumulated customer-seconds up to time t , we may divide by the number of arrivals up to that point to obtain

$$T_t = \frac{yet}{\alpha(t)}$$

Lastly, let us define $\bar{N}_t$ as the average number of customers in the queueing system during the interval $(0, t)$; this may be obtained by dividing the accumulated number of customer-seconds by the total interval length t

thusly

$$\bar{N}_t = \frac{y(t)}{t}$$

From these last three equations we see

$$\bar{N}_t = \lambda_t T_t$$

Let us now assume that our queueing system is such that the following limits exist as $t \rightarrow \infty$:

$$\lambda = \lim_{t \rightarrow \infty} \lambda_t,$$

$$T = \lim_{t \rightarrow \infty} T_t,$$

Note that we are using our former definitions for λ and T representing the average customer arrival rate and the average system time, respectively. If these last two limits exist, then so will the limit for $\bar{N}_t$ which we denote by $\bar{N}$ now representing the average number of customers in the system; that is,

$$\bar{N} = \lambda T \quad - \quad (2.25)$$

This last is the result we were seeking and is known as *Little's result*. It states that *the average number of customers in a queueing system is equal to the average arrival rate of customers to that system, times the average time spent in that system*.^{*} The above proof does not depend upon any specific assumptions regarding the arrival distribution $A(t)$ or the service time distribution $B(x)$; nor does it depend upon the number of servers in the system or upon the particular queueing discipline within the system. This result existed as a "folk theorem" for many years; the first to establish its validity in a formal way was J. D. C. Little [LITT 61] with some later simplifications by W. S. Jewell [JEWE 67], S. Eilon [EILO 69] and S. Stidham [STID 74]. It is important to note that we have not precisely defined the boundary around our queueing system. For example, the box in Figure 2.1 could apply to the entire system composed of queue and server, in which case $\bar{N}$ and T as defined refer to quantities for the entire system; on the other hand, we could have considered the boundary of the queueing system to contain only the queue itself, in which case the relationship would have been

$$\bar{N}_q = \lambda W \quad - \quad (2.26)$$

where $\bar{N}_q$ represents the average number of customers in the queue and, as defined earlier, W refers to the average time spent waiting in the queue. As a third possible alternative the queueing system defined could have surrounded

* An intuitive proof of Little's result depends on the observation that an arriving customer should *find* the same average number, $\bar{N}$, in the system as he *leaves behind* upon his departure. This latter quantity is simply the arrival rate λ times his average time in system, T .

only the server (or servers) itself; in this case our equation would have reduced to

$$\bar{N}_s = \lambda \bar{x} \quad (2.27)$$

where $\bar{N}_s$ refers to the average number of customers in the service facility (or facilities) and $\bar{x}$, of course, refers to the average time spent in the service box. Note that it is always true that

$$T = \bar{x} + W \quad (2.28)$$

The queueing system could refer to a specific class of customers, perhaps based on priority or some other attribute of this class, in which case the same relationship would apply. In other words, the average arrival rate of customers to a "queueing system" times the average time spent by customers in that "system" is equal to the average number of customers in the "system," regardless of how we define that "system."

We now discuss a basic parameter ρ , which is commonly referred to as the *utilization factor*. The utilization factor is in a fundamental sense really the ratio R/C , which we introduced in Chapter I. It is the ratio of the rate at which "work" enters the system to the maximum rate (capacity) at which the system can perform this work; the work an arriving customer brings into the system equals the number of seconds of service he requires. So, in the case of a single-server system, the definition for ρ becomes

$$\begin{aligned} \rho &\triangleq (\text{average arrival rate of customers}) \times (\text{average service time}) \\ &= \lambda \bar{x} \end{aligned} \quad (2.29)$$

This last is true since a single-server system has a maximum capacity for doing work, which equals 1 sec/sec and each arriving customer brings an amount of work equal to $\bar{x}$ sec; since, on the average, λ customers arrive per second, then $\lambda \bar{x}$ sec of work are brought in by customers each second that passes, on the average. In the case of multiple servers (say, m servers) the definition remains the same when one considers the ratio R/C , where now the work capacity of the system is m sec/sec; expressed in terms of system parameters we then have

$$\rho \triangleq \frac{\lambda \bar{x}}{m} \quad (2.30)$$

Equations (2.29) and (2.30) apply in the case when the maximum service rate is independent of the system state; if this is not the case, then a more careful definition must be provided. The rate at which work enters the system is sometimes referred to as the *traffic intensity* of the system and is usually expressed in *Erlangs*; in single-server systems, the utilization factor is equal to the traffic intensity whereas for (m) multiple servers, the traffic intensity equals $m\rho$. So long as $0 \leq \rho < 1$, then ρ may be interpreted as

$$\rho = E[\text{fraction of busy servers}] \quad (2.31)$$

[In the case of an infinite number of servers, the utilization factor ρ plays no important part, and instead we are interested in the *number* of busy servers (and its expectation).]

Indeed, for the system $G/G/1$ to be stable, it must be that $R < C$, that is, $0 \leq \rho < 1$. Occasionally, we permit the case $\rho = 1$ within the range of stability (in particular for the system $O/O/1$). Stability here once again refers to the fact that limiting distributions for all random variables of interest exist, and that all customers are eventually served. In such a case we may carry out the following simple calculation. We let τ be an arbitrarily long time interval; during this interval we expect (by the law of large numbers) with probability 1 that the number of arrivals will be very nearly equal to $\lambda\tau$. Moreover, let us define P_0 as the probability that the server is idle at some randomly selected time. We may, therefore, say that during the interval τ , the server is busy for $\tau - \tau P_0$ sec, and so with probability 1, the number of customers served during the interval τ is very nearly $(\tau - \tau P_0)/\bar{x}$. We may now equate the number of arrivals to the number served during this interval, which gives, for large τ ,

$$\lambda\tau \cong \frac{(\tau - \tau P_0)}{\bar{x}}$$

Thus, as $\tau \rightarrow \infty$ we have $\lambda\bar{x} = 1 - P_0$; using Definition (2.29) we finally have the important conclusion for $G/G/1$

$$\rho = 1 - P_0 \quad (2.32)$$

The interpretation here is that ρ is merely the fraction of time the server is busy; this supports the conclusion in Eq. (2.27) in which $\lambda\bar{x} = \rho$ was shown equal to the average number of customers in the service facility.

This, then, is a rapid look at an overall queueing system in which we have exposed some of the basic stochastic processes, as well as some of the important definitions and notation we will encounter. Moreover, we have established Little's result, which permits us to calculate the average number in the system once we have calculated the average time in the system (or vice versa). Now let us move on to a more careful study of the important stochastic processes in our queueing systems.

2.2*. DEFINITION AND CLASSIFICATION OF STOCHASTIC PROCESSES

At the end of Appendix II a definition is given for a stochastic process, which in essence states that it is a family of random variables $X(t)$ where the

• The reader may choose to skip Sections 2.2, 2.3, and 2.4 at this point and move directly to Section 2.5. He may then refer to this material only as he feels he needs to in the balance of the text.

random variables are "indexed" by the time parameter t . For example, the number of people sitting in a movie theater as a function of time is a stochastic process, as is also the atmospheric pressure in that movie theater as a function of time (at least those functions may be *modeled* as stochastic processes). Often we refer to a stochastic process as a random process. A random process may be thought of as describing the motion of a particle in some space. The classification of a random process depends upon three quantities: the *state space*; the *index (time) parameter*; and the *statistical dependencies* among the random variables $X(t)$ for different values of the index parameter t . Let us discuss each of these in order to provide the general framework for random processes.

First we consider the *state space*. The set of possible values (or states) that $X(t)$ may take on is called its state space. Referring to our analogy with regard to the motion of a particle, if the positions that particle may occupy are finite or countable, then we say we have a *discrete-state* process, often referred to as a *chain*. The state space for a chain is usually the set of integers $\{0, 1, 2, \dots\}$. On the other hand, if the permitted positions of the particle are over a finite or infinite continuous interval (or set of such intervals), then we say that we have a *continuous-state* process.

Now for the *index (time) parameter*. If the permitted times at which changes in position may take place are finite or countable, then we say we have a *discrete-(time) parameter* process; if these changes in position may occur anywhere within (a set of) finite or infinite intervals on the time axis, then we say we have a *continuous-parameter* process. In the former case we often write X_n rather than $X(t)$. X_n is often referred to as a random or stochastic *sequence* whereas $X(t)$ is often referred to as a random or stochastic *process*.

The truly distinguishing feature of a stochastic process is the relationship of the random variables $X(t)$ or X_n to other members of the same family. As defined in Appendix II, one must specify the complete joint distribution function among the random variables (which we may think of as vectors denoted by the use of boldface) $\mathbf{X} = [X(t_1), X(t_2), \dots, X(t_n)]$, namely,

$$F_{\mathbf{X}}(\mathbf{x}; \mathbf{t}) \triangleq P[X(t_1) \leq x_1, \dots, X(t_n) \leq x_n] \quad (2.33)$$

for all $\mathbf{x} = (x_1, x_2, \dots, x_n)$, $\mathbf{t} = (t_1, t_2, \dots, t_n)$, and $n \geq 1$. As mentioned there, this is a formidable task; fortunately, many interesting stochastic processes permit a simpler description. In any case, it is the function $F_{\mathbf{X}}(\mathbf{x}; \mathbf{t})$ that really describes the dependencies among the random variables of the stochastic process. Below we describe some of the usual types of stochastic processes that are characterized by different kinds of dependency relations among their random variables. We provide this classification in order to give the reader a global view of this field so that he may better understand in which particular

regions he is operating as we proceed with our study of queueing theory and its related stochastic processes.

(a) Stationary Processes. As we discuss at the very end of Appendix II, a stochastic process $X(l)$ is said to be stationary if $F_X(x; t)$ is invariant to shifts in time for all values of its arguments; that is, given any constant τ the following must hold :

$$F_X(x; t + \tau) = F_X(x; t) \quad (2.34)$$

where the notation $t + \tau$ is defined as the vector $(t_1 + \tau, t_2 + \tau, \dots, t_n + \tau)$.

An associated notion, that of *wide-sense stationarity*, is identified with the random process $X(l)$ if merely both the first and second moments are independent of the location on the time axis, that is, if $E[X(l)]$ is independent of l and if $E[X(l)X(l + T)]$ depends only upon T and not upon l . Observe that all stationary processes are wide-sense stationary, but not conversely. The theory of stationary random processes is, as one might expect, simpler than that for nonstationary processes.

(b) Independent Processes. The simplest and most trivial stochastic process to consider is the random sequence in which $\{X_n\}$ forms a set of independent random variables, that is, the joint pdf defined for our stochastic process in Appendix II must factor into the product, thusly

$$\begin{aligned} f_X(\mathbf{x}; \mathbf{t}) &\triangleq f_{X_1 \dots X_n}(x_1, \dots, x_n; t_1, \dots, t_n) \\ &= f_{X_1}(x_1; t_1) \cdots f_{X_n}(x_n; t_n) \end{aligned} \quad (2.35)$$

In this case we are stretching things somewhat by calling such a sequence a random process since there is no structure or dependence among the random variables. In the case of a continuous random process, such an independent process may be defined, and it is commonly referred to as "white noise" (an example is the time derivative of Brownian motion).

(c) Markov Processes. In 1907 A. A. Markov published a paper [MARK 07] in which he defined and investigated the properties of what are now known as Markov processes. In fact, what he created was a simple and highly useful form of dependency among the random variables forming a stochastic process, which we now describe.

A Markov process with a discrete state space is referred to as a Markov chain. The discrete-time Markov chain is the easiest to conceptualize and understand. A set of random variables $\{X_n\}$ forms a Markov chain if the probability that the next value (state) is X_{n+1} depends only upon the current value (state) X_n and not upon any previous values. Thus we have a random sequence in which the dependency extends backwards one unit in time. That

is, the way in which the entire past history affects the future of the process is completely summarized in the current value of the process.

In the case of a discrete-time Markov chain the instants when state changes may occur are preordained to be at the integers $0, 1, 2, \dots, n, \dots$. In the case of the continuous-time Markov chain, however, the transitions between states may take place at any instant in time. Thus we are led to consider the random variable that describes how long the process remains in its current (discrete) state before making a transition to some other state. Because the Markov property insists that the past history be completely summarized in the specification of the current state, then we are not free to require that a specification also be given as to how long the process has been in its current state! This imposes a heavy constraint on the distribution of time that the process may remain in a given state. In fact, as we shall see in Eq. (2.85), this state time must be *exponentially* distributed. In a real sense, then, the exponential distribution is a continuous distribution which is "memoryless" (we will discuss this notion at considerable length later in this chapter). Similarly, in the discrete-time Markov chain, the process may remain in the given state for a time that must be *geometrically* distributed; this is the only *discrete* probability mass function that is memoryless. This memoryless property is required of all Markov chains and restricts the generality of the processes one would like to consider.

Expressed analytically the *Markov property* may be written as

$$\begin{aligned} P[X(t_{n+1}) = x_{n+1} | X(t_n) = x_n, X(t_{n-1}) = x_{n-1}, \dots, X(t_1) = x_1] \\ = P[X(t_{n+1}) = x_{n+1} | X(t_n) = x_n] \end{aligned} \quad (2.36)$$

where $t_1 < t_2 < \dots < t_n < t_{n+1}$ and x_i is included in some discrete state space.

The consideration of Markov processes is central to the study of queueing theory and much of this text is devoted to that study. Therefore, a good portion of this chapter deals with discrete- and continuous-time Markov chains.

(d) Birth-death Processes. A very important special class of Markov chains has come to be known as the birth-death process. These may be either discrete- or continuous-time processes in which the defining condition is that state transitions take place between *neighboring* states only. That is, one may choose the set of integers as the discrete state space (with no loss of generality) and then the birth-death process requires that if $X_n = i$, then $X_{n+1} = i - 1$, i , or $i + 1$ and no other. As we shall see, birth-death processes have played a significant role in the development of queueing theory. For the moment, however, let us proceed with our general view of stochastic processes to see how each fits into the general scheme of things.

(e) *Semi-Markov Processes*. We begin by discussing discrete-time semi-Markov processes. The discrete-time Markov chain had the property that at every unit interval on the time axis the process was required to make a transition from the current state to some other state (possibly back to the same state). The transition probabilities were completely arbitrary; however, the requirement that a transition be made at every unit time (which really came about because of the Markov property) leads to the fact that the time spent in a state is geometrically distributed [as we shall see in Eq. (2.66)]. As mentioned earlier, this imposes a strong restriction on the kinds of processes we may consider. If we wish to relax that restriction, namely, to permit an arbitrary distribution of time the process may remain in a state, then we are led directly into the notion of a discrete-time *semi-Markov process*; specifically, we now permit the times between state transitions to obey an *arbitrary* probability distribution. Note, however, that at the instants of state transitions, the process behaves just like an ordinary Markov chain and, in fact, at those instants we say we have an *imbedded* Markov chain.

Now the definition of a continuous-time semi-Markov process follows directly. Here we permit state transitions at any instant in time. However, as opposed to the Markov process which required an exponentially distributed time in state, we now permit an arbitrary distribution. This then affords us much greater generality, which we are happy to employ in our study of queueing systems. Here, again, the imbedded Markov process is defined at those instants of state transition. Certainly, the class of Markov processes is contained within the class of semi-Markov processes.

(f) *Random Walks*. In the study of random processes one often encounters a process referred to as a *random walk*. A random walk may be thought of as a particle moving among states in some (say, discrete) state space. What is of interest is to identify the *location* of the particle in that state space. The salient feature of a random walk is that the next position the process occupies is equal to the previous position plus a random variable whose value is drawn independently from an arbitrary distribution; this distribution, however, does not change with the state of the process.* That is, a sequence of random variables $\{S_n\}$ is referred to as a random walk (starting at the origin) if

$$S_n = X_1 + X_2 + \dots + X_n \quad n = 1, 2, \dots \quad (2.37)$$

where $S_0 = 0$ and $X_1, X_2, \dots$ is a sequence of independent random variables with a common distribution. The index n merely counts the number of state transitions the process goes through; of course, if the instants of these transitions are taken from a discrete set, then we have a discrete-time random

* Except perhaps at some boundary states.

walk, whereas if they are taken from a continuum, then we have a continuous-time random walk. In any case, we assume that the interval between these transitions is distributed in an arbitrary way and so a random walk is a special case of a semi-Markov process.* In the case when the common distribution for X_n is a discrete distribution, then we have a discrete-state random walk; in this case the transition probability P_{ij} of going from state i to state j will depend only upon the difference in indices $j - i$ (which we denote by q_{j-i}).

An example of a continuous-time random walk is that of Brownian motion; in the discrete-time case an example is the total number of heads observed in a sequence of independent coin tosses.

A random walk is occasionally referred to as a process with "independent increments."

(g) Renewal Processes. A renewal process is related to a random walk. However, the interest is not in following a particle among many states but rather in *counting transitions* that take place as a function of time. That is, we consider the real time axis on which is laid out a sequence of points; the distribution of time between adjacent points is an arbitrary *common* distribution and each point corresponds to an instant of a state transition. We assume that the process begins in state 0 [i.e., $X(0) = 0$] and increases by unity at each transition epoch; that is, $X(t)$ equals the *number* of state transitions that have taken place by t . In this sense it is a special case of a random walk in which $q_0 = 1$ and $q_i = 0$ for $i \neq 1$. We may think of Eq. (2.37) as describing a renewal process in which S_n is the random variable denoting the *time* at which the n th transition takes place. As earlier, the sequence $\{X_n\}$ is a set of independent identically distributed random variables where X_n now represents the time between the $(n - 1)$ th and n th transition. One should be careful to distinguish the interpretation of Eq. (2.37) when it applies to renewal processes as here and when it applies to a random walk as earlier. The difference is that here in the renewal process the equation describes the *time* of the n th renewal or transition, whereas in the random walk it describes the *state* of the process and the time between state transitions is some other random variable.

An important example of a renewal process is the set of arrival instants to the $G/G/m$ queue. In this case, X_n is identified with the interarrival time.

• Usually, the distribution of time between intervals is of little concern in a random walk; emphasis is placed on the value (position) S_n after n transitions. Often, it is assumed that this distribution of interval time is memoryless, thereby making the random walk a special case of Markov processes; we are more generous in our definition here and permit an arbitrary distribution.

† It may be considered to be a special case of the random walk as defined in (f) above. A renewal process is occasionally referred to as a *recurrent* process.

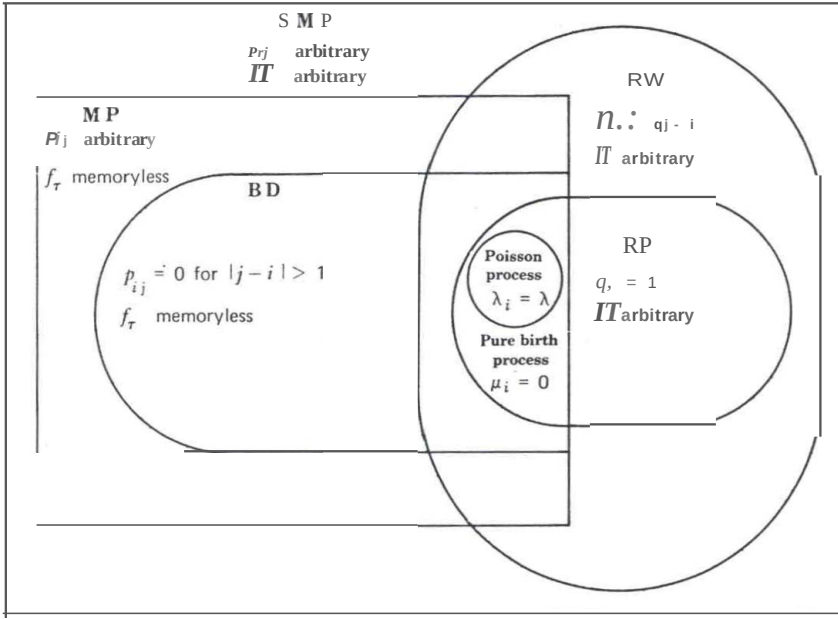


Figure 2.4 Relationships among the interesting random processes. SMP: Semi-Markov process; MP: Markov process; RW: Random walk; RP: Renewal process; BD: Birth-Death Process.

So there we have it—a self-consistent classification of some interesting stochastic processes. In order to aid the reader in understanding the relationship among Markov processes, semi-Markov processes, and their special cases, we have prepared the diagram of Figure 2.4, which shows this relationship for discrete-state systems. The figure is in the form of a Venn diagram. Moreover, the symbol P_{ij} denotes the probability of making a transition next to state j given that the process is currently in state i . Also, f_τ denotes the distribution of time between transitions; to say that “ f_τ is memoryless” implies that if it is a discrete-time process, then f_τ is a geometric distribution, whereas if it is a continuous-time process, then f_τ is an exponential distribution. Furthermore, it is implied that f_τ may be a function both of the current and the next state for the process.

The figure shows that birth-death processes form a subset of Markov processes, which themselves form a subset of the class of semi-Markov processes. Similarly, renewal processes form a subset of random walk processes which also are a subset of semi-Markov processes. Moreover, there are some renewal processes that may also be classified as birth-death

processes. Similarly, those Markov processes for which $p_{ij} = q_{j-i}$ (that is, where the transition probabilities depend only upon the difference of the indices) overlap those random walks where q_j is memoryless. A random walk for which q_j is memoryless and for which $q_{j-i} = 0$ when $|j-i| > 1$ overlaps the class of birth-death processes. If in addition to this last requirement our random walk has $q_0 = 1$, then we have a process that lies at the intersection of all five of the processes shown in the figure. This is referred to as a "pure birth" process; although f_i must be memoryless, it may be a distribution which depends upon the state itself, i ; is independent of the state (thus giving a constant "birth rate") then we have a process that is figuratively and literally at the "center" of the study of stochastic processes and enjoys the nice properties of each! This very special case is referred to as the *Poisson process* and plays a major role in queueing theory. We shall develop its properties later in this chapter.

So much for the classification of stochastic processes at this point. Let us now elaborate upon the definition and properties of discrete-state Markov processes. This will lead us naturally into some of the elementary queueing systems. Some of the required theory behind the more sophisticated continuous-state Markov processes will be developed later in this work as the need arises. We begin with the simpler discrete-state, discrete-time Markov chains in the next section and follow that with a section on discrete-state, continuous-time Markov chains.

2.3. DISCRETE-TIME MARKOV CHAINS'

As we have said, Markov processes may be used to describe the motion of a particle in some space. We now consider discrete-time Markov chains, which permit the particle to occupy discrete positions and permit transitions between these positions to take place only at discrete times. We present the elements of the theory by carrying along the following contemporary example.

Consider the hippie who hitchhikes from city to city across the country. Let X_n denote the city in which we find our hippie at noon on day n . When he is in some particular city i , he will accept the first ride leaving in the evening from that city. We assume that the travel time between any two cities is negligible. Of course, it is possible that no ride comes along, in which case he will remain in city i until the next evening. Since vehicles heading for various neighboring cities come along in some unpredictable fashion, the hippie's position at some time in the future is clearly a random variable. It turns out that this random variable may properly be described through the use of a Markov chain.

• See footnote on p. 19.

We have the following definition

DEFINITION: The sequence of random variables $X_1, X_2, \dots$ forms a discrete-time Markov chain if for all n ($n = 1, 2, \dots$) and all possible values of the random variables we have (for $i_1 < i_2 < \dots < i_n$) that

$$\begin{aligned} P[X_n = j \mid X_1 = i_1, X_2 = i_2, \dots, X_{n-1} = i_{n-1}] \\ = P[X_n = j \mid X_{n-1} = i_{n-1}] \end{aligned} \quad - (2.38)$$

In terms of our example, this definition merely states that the city next to be visited by the hippie depends only upon the city in which he is currently located and not upon all the previous cities he has visited. In this sense the memory of the random process, or Markov chain, goes back only to the most recent position of the particle (hippie). When $X_n = j$ (the hippie is in city j on day n), then the system is said to be in state E_j at time n (or at the n th step). To get our hippie started on day 0 we begin with some initial probability distribution $P[X_0 = j]$. The expression on the right side of Eq. (2.38) is referred to as the (one-step) *transition probability* and gives the conditional probability of making a transition from state $E_{i_{n-1}}$ at step $n - 1$ to state E_j at the n th step in the process. It is clear that if we are given the initial state probability distribution and the transition probabilities, then we can uniquely find the probability of being in various states at time n [see Eqs. (2.55) and (2.56) below].

If it turns out that the transition probabilities are independent of n , then we have what is referred to as a *homogeneous* Markov chain and in that case we make the further definition

$$P_{ij} \triangleq P[X_n = j \mid X_{n-1} = i] \quad (2.39)$$

which gives the probability of going to state E_j on the next step, given that we are currently at state E_i . What follows refers to homogeneous Markov chains only. These chains are such that their transition probabilities are stationary with time": therefore, given the current city or state (pun) the probability of various states m steps into the future depends only upon m and not upon the current time; it is expedient to define the *m-step* transition probabilities as

$$p_{ij}^{(m)} \triangleq P[X_{n+m} = j \mid X_n = i] \quad - (2.40)$$

From the Markov property given in Eq. (2.38) it is easy to establish the following recursive formula for calculating $p_{ij}^{(m)}$:

$$p_{ij}^{(m)} = \sum_k p_{ik}^{(m-1)} p_{kj} \quad m = 2, 3, \dots \quad (2.41)$$

This equation merely says that if we are to travel from E_i to E_j in m steps,

• Note that although this is a Markov process with stationary transitions, it need not be a stationary random process.

then we must do so by first traveling from E_i to *some* state E_k in $m - 1$ steps and then from E_k to E_j in one more step; the probability of these last two independent events (remember this is a Markov chain) is the product of the probability of each and if we sum this product over all possible intermediate states E_k , we arrive at $p_{ij}^{(m)}$.

We say that a Markov chain is *irreducible** if every state can be reached from every other state; that is, for each pair of states (E_i and E_j), there exists an integer m_0 (which may depend upon i and j) such that

$$p_{ij}^{(m_0)} > 0$$

Further, let A be the set of all states in a Markov chain. Then a subset of states A_1 is said to be *closed* if no one-step transition is possible from any state in A_1 to any state in A_1^c (the complement of the set A_1). If A_1 consists of a single state, say E_i , then it is called an *absorbing* state; a necessary and sufficient condition for E_i to be an absorbing state is $P_{ii} = 1$. If A is closed and does not contain any proper subset which is closed, then we have an *irreducible* Markov chain as defined above. On the other hand, if A contains proper subsets that are closed, then the chain is said to be *reducible*. If a closed subset of a reducible Markov chain contains no closed subsets of itself, then it is referred to as an *irreducible sub-Markov chain*; these subchains may be studied independently of the other states.

It may be that our hippie prefers not to return to a previously visited city. However, due to his mode of travel this may well happen, and it is important for us to define this quantity. Accordingly, let

$$f_j^{(n)} \triangleq P[\text{first return to } E_j \text{ occurs } n \text{ steps after leaving } E_j]$$

It is then clear that the probability of our hippie *ever* returning to city j is given by

$$f_j = \sum_{n=1}^{\infty} f_j^{(n)} = P[\text{ever returning to } E_j]$$

It is now possible to classify states of a Markov chain according to the value obtained for f_j . In particular, if $f_j = 1$ then state E_j is said to be *recurrent*; if on the other hand, $f_j < 1$, then state E_j is said to be *transient*. Furthermore, if the only possible steps at which our hippie can return to state E_j are $y, 2y, 3y, \dots$ (where $y > 1$ and is the largest such integer), then state E_j is said to be *periodic* with period y ; if $y = 1$, then E_j is *aperiodic*.

Considering states for which $f_j = 1$, we may then define the *mean recurrence time* of E_j as

$$M_j \triangleq \sum_{n=1}^{\infty} n f_j^{(n)} \quad (2.42)$$

• Many of the interesting Markov chains which one encounters in queueing theory are irreducible.

This is merely the average time to return to E ; With this we may then classify states even further. In particular, if $M_j = \infty$, then E_j is said to be *recurrent null*, whereas if $M_j < \infty$, then E_j is said to be *recurrent nonnull*. Let us define $\pi_j^{(n)}$ to be the probability of finding the system in state E_j at the n th step, that is,

$$\pi_j^{(n)} \triangleq P[X_n = j] \quad (2.43)$$

We may now state (without proof) two important theorems. The first comments on the set of states for an irreducible Markov chain.

Theorem 1 The states of an irreducible Markov chain are either all transient or all recurrent nonnull or all recurrent null. If periodic, then all states have the same period y .

Assuming that our hippie wanders forever, he will pass through the various cities of the nation many times, and we inquire as to whether or not there exists a *stationary* probability distribution $\{\pi_j\}$ describing his probability of being in city j at some time arbitrarily far into the future. [A probability distribution P_j is said to be a *stationary distribution* if when we choose it for our initial state distribution (that is, $\pi_j^{(0)} = P_j$) then for all n we will have $\pi_j^{(n)} = P_j$.] Solving for $\{\pi_j\}$ is a most important part of the analysis of Markov chains. Our second theorem addresses itself to this question.

Theorem 2 In an irreducible and aperiodic homogeneous Markov chain the limiting probabilities

$$\pi_j = \lim_{n \rightarrow \infty} \pi_j^{(n)} \quad (2.44)$$

always exist and are independent of the initial state probability distribution. Moreover, either

- (a) *all states are transient or all states are recurrent null in which cases $\pi_j = 0$ for all j and there exists no stationary distribution, or*
- (b) *all states are recurrent nonnull and then $\pi_j > 0$ for all j , in which case the set $\{\pi_j\}$ is a stationary probability distribution and*

$$\pi_j = \frac{1}{M_j} \quad (2.45)$$

In this case the quantities π_j are uniquely determined through the following equations

$$1 = \sum \pi_i \quad (2.46)$$

$$\pi_j = \sum \pi_i p_{ij} \quad (2.47)$$

We now introduce the notion of *ergodicity*. A state E_i is said to be ergodic if it is aperiodic, recurrent, and nonnull; that is, if $f_j = 1$, $M_j < \infty$, and $\gamma = 1$. If all states of a Markov chain are ergodic, then the Markov chain itself is said to be ergodic. Moreover, a Markov chain is said to be ergodic if the probability distribution $\{\pi_j^{(n)}\}$ as a function of n always converges to a limiting stationary distribution $\{\pi_j\}$, which is independent of the initial state distribution. It is easy to show that all states of an aperiodic irreducible Markov chain are ergodic. Moreover, among Foster's criteria [FELL 66] it can be shown that an irreducible and aperiodic Markov chain is ergodic if the set of linear equations given in Eq. (2.47) has a nonnull solution for which $\sum_j \pi_j < \infty$. The limiting probabilities $\{\pi_j\}$, of an ergodic Markov chain are often referred to as the *equilibrium* probabilities in the sense that the effect of the initial state distribution π_{j0} has disappeared.

By way of example, let's place the hippie in our fictitious land of Hatafla, and let us consider the network given in Figure 1.1 of Chapter I. In order to simplify this example we will assume that the cities of Nonabel, Cadabra, and Oriac have been bombed out and that the resultant road network is as given in Figure 2.5. In this figure the ordered links represent permissible directions of road travel; the numbers on these links represent the probability (P_{ij}) that the hippie will be picked up by a car traveling over that road, given that he is hitchhiking from the city where the arrow emanates. Note that from the city of Sucsamad our hippie has probability $1/2$ of remaining in that city until the next day. Such a diagram is referred to as a *state-transition* diagram. The parenthetical numbers following the cities will henceforth be used instead of the city names.

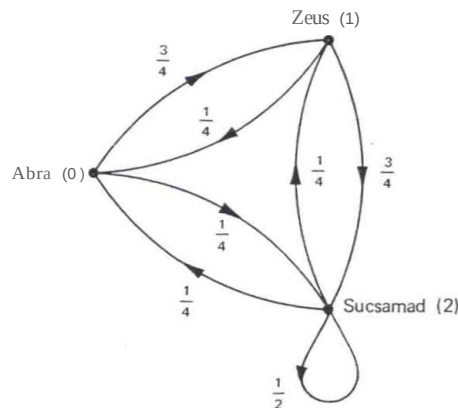


Figure 2.5 A Markov chain.

• A finite Markov chain is one with a finite number of states. If an irreducible Markov chain is of type (a) in Theorem 2 (i.e., recurrent null or transient) then it cannot be finite.

In order to continue our example we now define, in general, the *transition probability matrix* $\mathbf{P}$ as consisting of elements P_{ij} , that is,

$$\mathbf{P} = [p_{ij}] \quad - \quad (2.48)$$

If we further define the probability vector $\boldsymbol{\pi}$ as

$$\boldsymbol{\pi} = [\pi_0, \pi_1, \pi_2, \dots] \quad (2.49)$$

then we may rewrite the set of relations in Eq. (2.47) as

$$\mathbf{re} = \boldsymbol{\pi} \mathbf{P} \quad - \quad (2.50)$$

For our example shown in Figure 2.5 we have

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

and so we may solve Eq. (2.50) by considering the three equations derivable from it, that is,

$$\begin{aligned} \pi_0 &= 0\pi_0 + \frac{1}{4}\pi_1 + \frac{1}{4}\pi_2 \\ \pi_1 &= \frac{3}{4}\pi_0 + 0\pi_1 + \frac{1}{4}\pi_2 \\ \pi_2 &= \frac{1}{4}\pi_0 + \frac{3}{4}\pi_1 + \frac{1}{2}\pi_2 \end{aligned} \quad (2.51)$$

Note from Eq. (2.51) that the first of these three equations equals the negative sum of the second and third, indicating that there is a linear dependence among them. It *always* will be the case that one of the equations will be linearly dependent on the others, and it is therefore necessary to introduce the additional conservation relationship as given in Eq. (2.46) in order to solve the system. In our example we then require

$$1 = \pi_0 + \pi_1 + \pi_2 \quad (2.52)$$

Thus the solution is obtained by simultaneously solving any two of the

equations given by Eq. (2.51) along with Eq. (2.52). Solving we obtain

$$\begin{aligned} \pi_0 &= \frac{1}{5} = 0.20 \\ \pi_1 &= \frac{7}{25} = 0.28 \\ \pi_2 &= \frac{13}{25} = 0.52 \end{aligned} \quad (2.53)$$

This gives us the equilibrium (stationary) state probabilities. It is clear that this is an ergodic Markov chain (it is finite and irreducible).

Often we are interested in the transient behavior of the system. The transient behavior involves solving for $\pi_j^{(n)}$, the probability of finding our hippie in city j at time n . We also define the probability vector at time n as

$$\pi^{(n)} \triangleq [\pi_0^{(n)}, \pi_1^{(n)}, \pi_2^{(n)}, \dots] \quad - \quad (2.54)$$

Now using the definition of transition probability and making use of Definition (2.48) we have a method for calculating $\pi_j^{(n)}$ expressible in terms of P and the initial state distribution $\pi^{(0)}$. That is,

$$\pi^{(1)} = \pi^{(0)}P$$

Similarly, we may calculate the state probabilities at the second step by

$$\begin{aligned} \pi^{(2)} &= \pi^{(1)}P \\ &= [\pi^{(0)}P]P \\ &= \pi^{(0)}P^2 \end{aligned}$$

From this last we can then generalize to the result

$$\pi^{(n)} = \pi^{(0)}P^n \quad n = 1, 2, \dots \quad - \quad (2.55)$$

which may be solved recursively to obtain

$$\pi^{(n)} = \pi^{(0)}P^n \quad n = 1, 2, \dots \quad - \quad (2.56)$$

Equation (2.55) gives the general method for calculating the state probabilities n steps into a process, given a transition probability matrix P and an initial state vector $\pi^{(0)}$. From our earlier definitions, we have the stationary probability vector

$$\pi = \lim_{n \rightarrow \infty} \pi^{(n)}$$

assuming the limit exists. (From Theorem 2, we know that *this* will be the case if we have an irreducible aperiodic homogeneous Markov chain.)

Then, from Eq. (2.55) we find

$$\lim_{n \rightarrow \infty} \pi^{(n)} = \lim_{n \rightarrow \infty} \pi^{(n-1)} \mathbf{P}$$

and so

$$\tau_t = \tau_t \mathbf{P}$$

which is Eq. (2.50) again. Note that the solution for τ_t is independent of the initial state vector. Applying this to our example, let us assume that our hippie begins in the city of Abra at time 0 with probability 1, that is

$$\tau_{t(0)} = [1, 0, 0] \quad (2.57)$$

From this we may calculate the sequence of values $\tau_{t(n)}$ and these are given in the chart below. The limiting value τ_t as given in Eq. (2.53) is also entered in this chart.

n	0		2	3	4	∞
$\pi_0^{(n)}$	1	0	0.250	0.187	0.203	0.20
$\pi_1^{(n)}$	0	0.75	0.062	0.359	0.254	0.28
$\pi_2^{(n)}$	0	0.25	0.688	0.454	0.543	0.52

We may alternatively have chosen to assume that the hippie begins in the city of Zeus with probability 1, which would give rise to the initial state vector

$$\tau_{t(0)} = [0, 1, 0] \quad (2.58)$$

and which results in the following table:

n	0		2	3	4	
$\pi_0^{(n)}$	0	0.25	0.187	0.203	0.199	0.20
$\pi_1^{(n)}$	1	0	0.375	0.250	0.289	0.28
$\pi_2^{(n)}$	0	0.75	0.438	0.547	0.512	0.52

Similarly, beginning in the city of Sucsamad we find

$$\tau_{t(0)} = [0, 0, 1] \quad (2.59)$$

n	0		2	3	4	
$\pi_0^{(n)}$	0	0.25	0.187	0.203	0.199	0.20
$\pi_1^{(n)}$	0	0.25	0.313	0.266	0.285	0.28
$\pi_2^{(n)}$	1	0.50	0.500	0.531	0.516	0.52

From these calculations we may make a number of observations. First, we

see that after only four steps the quantities $\pi_i^{(n)}$ for a given value of i are almost identical regardless of the city in which we began. The rapidity with which these quantities converge, as we shall soon see, depends upon the eigenvalues of P . In all cases, however, we observe that the limiting values at infinity are rapidly approached and, as stated earlier, are independent of the initial position of the particle.

In order to get a better physical feel for what is occurring, it is instructive to follow the probabilities for the various states of the Markov chain as time evolves. To this end we introduce the notion of *baricentric coordinates*, which are extremely useful in portraying probability vectors. Consider a probability vector with N components (i.e., a Markov process with N states in our case) and a tetrahedron in $N - 1$ dimensions. In our example $N = 3$ and so our tetrahedron becomes an equilateral triangle in two dimensions. In general, we let the height of this tetrahedron be unity. Any probability vector $\pi^{(n)}$ may be represented as a point in this $N - 1$ space by identifying each component of that probability vector with a distance from one face of the tetrahedron. That is, we measure from face j a distance equal to the probability associated with that component $\pi_j^{(n)}$; if we do this for each face and therefore for each component, we will specify one point within the tetrahedron and that point correctly identifies our probability vector. Each unique probability vector will map into a unique point in this space, and it is easy to determine the probability measure from its location in that space. In our example we may plot the three initial state vectors as given in Eqs. (2.57)-(2.59) as shown in Figure 2.6. The numbers in parentheses represent which probability components are to be measured from the face associated with those numbers. The initial state vector corresponding to Eq. (2.59), for

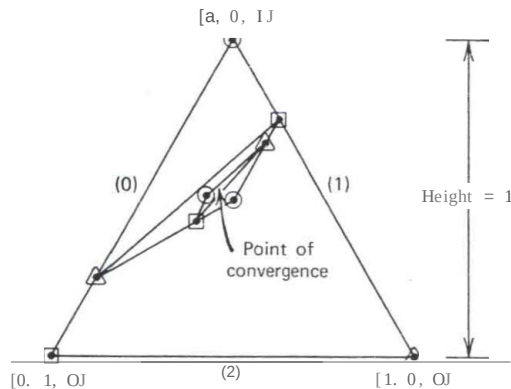


Figure 2.6 Representation of the convergence of a Markov chain.

example, will appear at the apex of the triangle and is indicated as such. In our earlier calculations we followed the progress of our probability vectors beginning with three initial state probability vectors. Let us now follow these paths simultaneously and observe, for example, that the vector $[0, 0, 1]$, following Eq. (2.59), moves to the position $[0.25, 0.25, 0.5]$; the vector $[0, 1, 0]$ moves to the position $[0.25, 0, 0.75]$, and the vector $[1, 0, 0]$ moves to the position $[0, 0.75, 0.25]$. Now it is clear that had we started with an initial state vector anywhere within the original equilateral triangle, that point would have been mapped into the *interior* of the smaller triangle, which now joins the three points just referred to and which represent possible positions of the original state vectors. We note from the figure that this new triangle is a shrunken version of the original triangle. If we now continue to map these three points into the second step of the process as given by the three charts above, we find an even smaller triangle interior to both the first and the second triangles, and this region represents the possible locations of any original state vector after *two* steps into the process. Clearly, this shrinking will continue until we reach a convergent point. This convergent point will in the limit be exactly that given by Eq. (2.53)! Thus we can see the way in which the possible positions of our probability vectors move around in our space.

The calculation of the transient response $1t^{(n)}$ from Eqs. (2.55) or (2.56) is extremely tedious if we desire more than just the first few terms. In order to obtain the general solution, we often resort to transform methods. Below we demonstrate this method in general and then apply it to our hippie hitchhiking example. This will give us an opportunity to apply the z-transform calculations that we have introduced in Appendix I.* Our point of departure is Eq. (2.55). That equation is a difference equation among vectors. The fact that it is a difference equation suggests the use of c-transforms as in Appendix I, and so we naturally define the following *vector* transform (the vectors in no way interfere with our transform approach except that we must be careful when taking inverses):

$$\Pi(z) \triangleq \sum_{n=0}^{\infty} 1t^{(n)} z^n \quad (2.60)$$

This transform will certainly exist in the unit disk, that is, $|z| \leq 1$. We now apply the transform method to Eq. (2.55) over its range of application ($11 = 1, 2, \dots$); this we do by first multiplying that equation by z^n and then summing from 1 to infinity, thus

$$\sum_{n=1}^{\infty} 1t^{(n)} z^n = \sum_{n=1}^{\infty} 1t^{(n-1)} P_{11} z^n$$

*The steps involved in applying this method are summarized on pp. 74-5 of this chapter.

We have now reduced our infinite set of difference equations to a single algebraic equation. Following through with our method we must now try to identify our vector transform $D(z)$. Our left-hand side contains all but the initial term of this transform and so we have

$$\Pi(z) - \pi^{(0)} = z \left(\sum_{n=1}^{\infty} \pi^{(n-1)} z^{n-1} \right) P$$

The parenthetical term on the right-hand side of this last equation is recognized as $D(z)$ simply by changing the index of summation. Thus we find

$$\Pi(z) - \pi^{(0)} = zD(z)P$$

z is merely a scalar in this vector equation and may be moved freely across vectors and matrices. Solving this matrix equation we immediately come up with a general solution for our vector transform:

$$\Pi(z) = \pi^{(0)}[I - zP]^{-1} \quad (2.61)$$

where I is the identity matrix and the $(-I)$ notation implies the matrix inverse. If we can invert this equation, we will have, by the uniqueness of transforms, the transient solution; that is, using the double-headed, double-barred arrow notation as in Appendix I to denote transform pairs, we have

$$D(z) \Leftrightarrow \tau_{t(n)} = \pi^{(0)}P^n \quad (2.62)$$

In this last we have taken advantage of Eq. (2.56). Comparing Eqs. (2.61) and (2.62) we have the obvious transform pair

$$[I - zP]^{-1} \Leftrightarrow P^n \quad (2.63)$$

Of course P^n is precisely what we are looking for in order to obtain our transient solution since this will directly give us $\tau_{t(n)}$ from Eq. (2.56). All that is required, therefore, is that we form the matrix inverse indicated in Eq. (2.63). In general this becomes a rather complex task when the number of states in our Markov chain is at all large. Nevertheless, this is one formal procedure for carrying out the transient analysis.

Let us apply these techniques to our hippie hitchhiking example. Recall that the transition probability matrix P was given by

$$P = \begin{vmatrix} 0 & 3 & 1 \\ 4 & 4 & 4 \\ 1 & 0 & 3 \\ 4 & 1 & 1 \\ 4 & 4 & 2 \end{vmatrix}$$

First we must form

$$1 - zP = \begin{vmatrix} 1 & -\frac{3}{4}z & \frac{1}{4}z \\ -\frac{1}{4}z & & -\frac{3}{4}z \\ -\frac{1}{4}z & -\frac{1}{4}z & 1 - \frac{1}{2}z \end{vmatrix}$$

Next, in order to find the inverse of this matrix we must form its determinant thus:

$$\det(\mathbf{I} - zP) = 1 - \frac{1}{2}z - \frac{7}{16}z^2 - \frac{1}{16}z^3$$

which factors nicely into

$$\det(\mathbf{I} - zP) = (1 - z)\left(1 + \frac{1}{4}z\right)^2$$

It is easy to show that $z = 1$ is always a root of the determinant for an irreducible Markov chain (and, as we shall see, gives rise to our equilibrium solution). We now proceed with the calculation of the matrix inverse using the usual methods to arrive at

$$[\mathbf{I} - zP]^{-1} = \frac{1}{(1 - z)\left(1 + \frac{1}{4}z\right)^2} \times \begin{vmatrix} 1 - \frac{1}{2}z - \frac{3}{16}z^2 & \frac{3}{4}z - \frac{5}{16}z^2 & \frac{1}{4}z + \frac{9}{16}z^2 \\ \frac{1}{4} - \frac{1}{16}z & 1 - \frac{1}{2}z - \frac{1}{16}z^2 & \frac{3}{4}z + \frac{1}{16}z^2 \\ \frac{1}{4}z + \frac{1}{16}z^2 & \frac{1}{4}z + \frac{3}{16}z^2 & 1 - \frac{3}{16}z^2 \end{vmatrix}$$

Having found the matrix inverse, we are now faced with finding the inverse transform of this matrix which will yield P^n . This we do as usual by carrying out a partial fraction expansion (see Appendix I). The fact that we have a matrix presents no problem; we merely note that each element in the matrix is itself a rational function of z which must be expanded in partial fractions term by term. (This task is simplified if the matrix is written as the sum of three matrices: a constant matrix; a constant matrix times z ; and a constant matrix times z^2 .) Since we have three roots in the denominator of our rational functions we expect three terms in our partial fraction expansion. Carrying

out this expansion and separating the three terms we find

$$\begin{aligned}
 [I - zP]^{-1} = & \frac{1/25}{1-z} \begin{bmatrix} 5 & 7 & 13 \\ 5 & 7 & 13 \\ 5 & 7 & 13 \end{bmatrix} + (1 + \frac{1/5}{1-z/4}) \begin{bmatrix} 0 & -8 & 8 \\ 0 & 2 & -2 \\ 0 & 2 & -2 \end{bmatrix} \\
 & + \frac{1/25}{1+z/4} \begin{bmatrix} 20 & 33 & -53 \\ -5 & 8 & -3 \\ -5 & -17 & 22 \end{bmatrix} \quad (2.64)
 \end{aligned}$$

We observe immediately from this expansion that the matrix associated with the root $(1 - e)$ gives precisely the equilibrium solution we found by direct methods [see Eq. (2.53)]; the fact that each row of this matrix is identical reflects the fact that the equilibrium solution is independent of the initial state. The other matrices associated with roots greater than unity in absolute value will always be what are known as differential matrices (each of whose rows must sum to zero). Inverting on z we finally obtain (by our tables in Appendix I)

$$\begin{aligned}
 P^n = & \frac{1}{25} \begin{bmatrix} 5 & 7 & 13 \\ 5 & 7 & 13 \\ 5 & 7 & 13 \end{bmatrix} + \frac{1}{5} (n+1) \left(-\frac{1}{4}\right)^n \begin{bmatrix} 0 & -8 & 8 \\ 0 & 2 & -2 \\ 0 & 2 & -2 \end{bmatrix} \\
 & + \frac{1}{25} \left(-\frac{1}{4}\right)^n \begin{bmatrix} 20 & 33 & -53 \\ -5 & 8 & -3 \\ -5 & -17 & 22 \end{bmatrix} \quad n = 0, 1, 2, \dots \quad (2.65)
 \end{aligned}$$

This is then the complete solution since application of Eq. (2.56) directly gives $\pi_{t(n)}$, which is the transient solution we were seeking. Note that for $n = 0$ we obtain the identity matrix whereas for $n = 1$ we must, of course, obtain the transition probability matrix P . Furthermore, we see that in this case we have two transient matrices, which decay in the limit leaving only the constant matrix representing our equilibrium solution. When we think about the decay of the transient, we are reminded of the shrinking triangles in Figure 2.6. Since the transients decay at a rate related to the characteristic values (one over the zeros of the determinant) we therefore expect the permitted positions in Figure 2.6 to decay with n in a similar fashion. In fact, it can be shown that these triangles shrink by a constant factor each time n increases by 1. This shrinkage factor for any Markov process can be shown to be equal to the absolute value of the product of the characteristic values of its transition probability matrix; in our example we have characteristic values equal to 1, $1/4$, $1/4$. Their product is $1/16$ and this indeed is the factor by which the area of our triangles decreases each time n is increased.

This method of transform analysis is extended in two excellent volumes by Howard [HOWA 71] in which he treats such problems and discusses additional approaches such as the flow-graph method of analysis.

Throughout this discussion of discrete-time Markov chains we have not explicitly addressed ourselves to the memoryless property* of the time that the system spends in a given state. Let us now prove that the number of time units that the system spends in the same state is *geometrically* distributed; the geometric distribution is the unique discrete memoryless distribution. Let us assume the system has just entered state E_i ; It will remain in this state at the next step with probability P_{ii} ; similarly, it will leave this state at the next step with probability $1 - P_{ii}$. If indeed it does remain in this state at the next step, then the probability of its remaining for an additional step is again P_{ii} and similarly the conditional probability of its leaving at this second step is given by $1 - P_{ii}$. And so it goes. Furthermore, due to the Markov property the fact that it has remained in a given state for a known number of steps in no way affects the probability that it leaves at the next step. Since these probabilities are independent, we may then write

$$P[\text{system remains in } E_i \text{ for exactly } m \text{ additional steps given that it has just entered } E_i] = (1 - P_{ii})P_{ii}^m \quad (2.66)$$

This, of course, is the geometric distribution as we claimed. A similar argument will be given later for the continuous-time Markov chain.

So far we have concerned ourselves principally with homogeneous Markov processes. Recall that a homogeneous Markov chain is one for which the transition probabilities are independent of time. Among the quantities we were able to calculate was the m -step transition probability $p_{ij}^{(m)}$, which gave the probability of passing from state E_i to state E_j in m steps; the recursive formula for this calculation was given in Eq. (2.41). We now wish to take a more general point of view and permit the transition probabilities to depend upon time. We intend to derive a relationship not unlike Eq. (2.41), which will form our point of departure for many further developments in the application of Markov processes to queueing problems. For the time being we continue to restrict ourselves to discrete-time, discrete-state Markov chains.

Generalizing the homogeneous definition for the multistep transition probabilities given in Eq. (2.40) we now define

$$p_{ij}(m, n) \triangleq P[X_n = j \mid X_m = i] \quad (2.67)$$

which gives the probability that the system will be in state E_j at step n , given

* The memoryless property is discussed in some detail later.

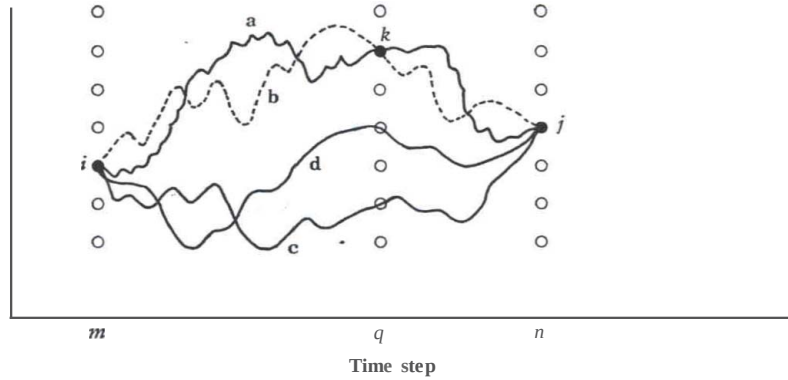


Figure 2.7 Sample paths of a stochastic process.

that it was in state E_i at step m , where $n \geq m$. As discussed in the homogeneous case, it certainly must be true that if our process goes from state E_i at time m to state E_j at time n , then at some intermediate time q it must have passed through some state E_k . This is depicted in Figure 2.7. In this figure we have shown four sample paths of a stochastic process as it moves from state E_i at time m to state E_j at time n . We have plotted the state of the process vertically and the discrete time steps horizontally. (We take the liberty of drawing continuous curves rather than a sequence of points for convenience.) Note that sample paths a and b both pass through state E_k at time q , whereas sample paths c and d pass through other intermediate states at time q . We are certain of one thing only, namely, that we must pass through *some* intermediate state at time q . We may then express $p_{ij}(m, n)$ as the sum of probabilities for all of these (mutually exclusive) intermediate states; that is,

$$p_{ij}(m, n) = \sum_k P[X_n = j, X_q = k | X_m = i] \quad (2.68)$$

for $m \leq q \leq n$. This last equation must hold for any stochastic process (not necessarily Markovian) since we are considering all mutually exclusive and exhaustive possibilities. From the definition of conditional probability we may rewrite this last equation as

$$p_{ij}(m, n) = \sum_k P[X_q = k | X_m = i] P[X_n = j | X_m = i, X_q = k] \quad (2.69)$$

Now we invoke the Markov property and observe that

$$P[X_n = j | X_m = i, X_q = k] = P[X_n = j | X_q = k]$$

Applying this to Eq. (2.69) and making use of our definition in Eq. (2.67) we finally arrive at

$$p_{ij}(m, n) = \sum_k p_{ik}(m, q) p_{kj}(q, n) \quad (2.70)$$

for $m \leq q \leq n$. Equation (2.70) is known as the *Chapman-Kolmogorov* equation for discrete-time Markov processes. Were this a homogeneous Markov chain then from the definition in Eq. (2.40) we would have the relationship $p_{ij}(m, n) = p_{ij}^{(n-m)}$ and in the casewhenn $= q + 1$ our Chapman-Kolmogorov equation would reduce to our earlier Eq. (2.41). The Chapman-Kolmogorov equation states that we can partition any $n - m$ step transition probability into the sum of products of a $q - m$ and an $n - q$ step transition probability to and from the intermediate states that might have been occupied at some time q within the interval. Indeed we are permitted to choose any partitioning we wish, and we will take advantage of this shortly.

It is convenient at this point to write the Chapman-Kolmogorov equation in matrix form. We have in the past defined P as the matrix containing the elements p_{ij} in the case of a homogeneous Markov chain. Since these quantities may now depend upon time, we define $p(n)$ to be the one-step transition probability matrix at time n , that is,

$$p(n) \triangleq [p_{ij}(n, n + 1)] \quad - \quad (2.71)$$

Of course, $p(n) = P$ if the chain is homogeneous. Also, for the homogeneous case we found that the n -step transition probability matrix was equal to P^n . In the nonhomogeneous case we must make a new definition and for this purpose we use the symbol $H(m, n)$ to denote the following multistep transition probabilitymatrix:

$$H(m, n) \triangleq [P_{ij}(m, n)] \quad - \quad (2.72)$$

Note that $H(n, n + 1) = p(n)$ and that in the homogeneous case $H(m, m + n) = P^n$. With these definitions we may then rewrite the Chapman-Kolmogorov equation in matrix form as

$$H(m, n) = H(m, q)H(q, n) \quad - \quad (2.73)$$

for $m \leq q \leq n$. To complete the definition we require that $H(l, l) = I$, where I is the identity matrix. All of the matrices we are considering are square matrices with dimensionality equal to the number of states of the Markov chain. A solution to Eq. (2.73) will consist of expressing $H(m, n)$ in terms of the given matrices $p(n)$.

As mentioned above, we are free to choose q to lie anywhere in the interval between m and n . Let us begin by choosing $q = n - 1$. In this case Eq. (2.70) becomes

$$p_{ij}(m, n) = \sum_k p_{ik}(m, n - 1)p_{kj}(n - 1, n) \quad (2.74)$$

which in matrix form may be written as

$$H(m, n) = H(m, n - 1)P(n - 1) \quad - \quad (2.75)$$

Equations (2.74) and (2.75) are known as the *forward* Chapman-Kolmogorov equations for discrete-time Markov chains since they are written at the forward (most recent time) end of the interval. On the other hand, we could have chosen $q = m + 1$, in which case we obtain

$$p_{ij}(m, n) = \sum_k P_{ik}(m, m+1)P_{kj}(m+1, n) \quad (2.76)$$

whose matrix form is

$$H(m, n) = P(m)H(m+1, n) \quad (2.77)$$

These last two are referred to as the *backward* Chapman-Kolmogorov equations since they occur at the backward (oldest time) end of the interval.

Since the forward and backward equations both describe the same discrete-time Markov chain, we would expect their solutions to be the same, and indeed this is the case. The general form of the solution is

$$H(m, n) = P(m)P(m+1) \dots P(n-1) \quad m \leq n-1 \quad (2.78)$$

That this solves Eqs. (2.75) and (2.77) may be established by direct substitution. We observe in the homogeneous case that this yields $H(m, n) = p_{nm}$ as we have seen earlier. By similar arguments we find that the time-dependent probabilities $\{\pi_j^{(n)}\}$ defined earlier may now be obtained through the following equation:

$$\pi^{(n+1)} = \pi^{(n)}P(n) \quad \text{---}$$

whose solution is

$$\pi^{(n+1)} = \pi^{(0)}P(0)P(1) \dots P(n) \quad \text{---} \quad (2.79)$$

These last two equations correspond to Eqs. (2.55) and (2.56), respectively, for the homogeneous case. The Chapman-Kolmogorov equations give us a means for describing the time-dependent probabilities of many interesting queueing systems that we develop in later chapters.*

Before leaving discrete-time Markov chains, we wish to introduce the special case of discrete time *birth-death processes*. A birth-death process is an example of a Markov process that may be thought of as modeling changes in the size of a population. In what follows we say that the system is in state E_k when the population consists of k members. We further assume that changes in population size occur by at most one; that is, a "birth" will change the population's size to one greater, whereas a "death" will lower the population size to one less. In considering birth-death processes we do not permit multiple births or bulk disasters; such possibilities will be considered

* It is clear from this development that all Markov processes must satisfy the Chapman-Kolmogorov equations. Let us note, however, that all processes that satisfy the Chapman-Kolmogorov equation are not necessarily Markov processes; see, for example, p. 203 of [PARZ 62].

later in the text and correspond to random walks. We will consider the Markov chain to be homogeneous in that the transition probabilities P_{ij} do not change with time; however, certainly they will be a function of the state of the system. Thus we have that for our discrete-time birth-death process

$$P_{ij} = \begin{cases} 1 - b_i - d_i, & j = i - 1 \\ b_i, & j = i \\ 0, & j = i + 1 \\ 0, & \text{otherwise} \end{cases} \quad (2.80)$$

Here d_i is the probability that at the next time step a single death will occur, driving the population size down to $i - 1$, given that the population size now is i . Similarly, b_i is the probability that a single birth will occur, given that the current size is i , thereby driving the population size to $i + 1$ at the next time step. $1 - b_i - d_i$ is the probability that neither of these events will occur and that at the next time step the population size will not change. Only these three possibilities are permitted. Clearly $d_0 = 0$, since we can have no deaths when there is no one in the population to die. However, contrary to intuition we do permit $b_0 > 0$; this corresponds to a birth when there are no members in the population. Whereas this may seem to be spontaneous generation, or perhaps divine creation, it does provide a meaningful model in terms of queueing theory. The model is as follows: The population corresponds to the customers in the queueing system; a death corresponds to a customer departure from that system; and a birth corresponds to a customer arrival to that system. Thus we see it is perfectly feasible to have an arrival (a birth) to an empty system! The stationary probability transition matrix for the general birth-death process then appears as follows:

$$P = \begin{pmatrix} 1 - b_0 & b_0 & 0 & 0 & 0 & 0 & 0 & 0 \\ d_1 & 1 - b_1 - d_1 & b_1 & 0 & 0 & 0 & 0 & 0 \\ 0 & d_2 & 1 - b_2 - d_2 & b_2 & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & d_j & 1 - b_j - d_j & b_j & 0 & 0 & \cdots & \cdots \end{pmatrix}$$

If we are dealing with a finite chain, then the last row of this matrix would be $[0 \dots 0 \ d_N \ 1 - d_N]$, which illustrates the fact that no births are permitted when the population has reached its maximum size N . We see that the P

matrix has nonzero terms only along the main diagonal and along the diagonals directly above and below it. This is a highly specialized form for the transition probability matrix, and as such we might expect that it can be solved. To solve the birth-death process means to find the solution for the state probabilities $\pi_n(t)$. As we have seen, the general form of solution for these probabilities is given in Eqs. (2.55) and (2.56) and the equation that describes the limiting solution (as $n \rightarrow \infty$) is given in Eq. (2.50). We also demonstrated earlier the z-transform method for finding the solution. Of course, due to this special structure of the birth-death transition matrix, we might expect a more explicit solution. We defer discussion of the solution to the material on continuous-time Markov chains, which we now investigate.

2.4. CONTINUOUS-TIME MARKOV CHAINS*

If we allow our particle in motion to occupy positions (take-on values) from a discrete set, but permit it to change positions or states at *any* point in time, then we say we have a continuous-time Markov chain. We may continue to use our example of the hippie hitchhiking from city to city, where now his transitions between cities may occur at any time of day or night. We let $X(t)$ denote the city in which we find our hippie at time t . $X(t)$ will take on values from a discrete set, which we will choose to be the ordered integers and which will be in one-to-one correspondence with the cities which our hippie may visit.

In the case of a continuous-time Markov chain, we have the following definition :

DEFINITION: The random process $X(t)$ forms a continuous-time Markov chain if for all integers n and for any sequence $t_1, t_2, \dots, t_{n+1}$ such that $t_1 < t_2 < \dots < t_{n+1}$ we have

$$\begin{aligned} P[X(t_{n+1}) = j \mid X(t_1) = i_1, X(t_2) = i_2, \dots, X(t_n) = i_n] \\ = P[X(t_{n+1}) = j \mid X(t_n) = i_n] \end{aligned} \quad (2.81)$$

This definition is the continuous-time version of that given in Eq. (2.38). The interpretation here is also the same, namely, that the future of our hippie's travels depends upon the past only through the current city in which we find him. The development of the theory for continuous time parallels that for discrete time quite directly as one might expect and, therefore, our explanations will be a bit more concise. Moreover, we will not overly concern

* See footnote on p. 19.

† An alternate definition for a discrete-state continuous-time Markov process is that the following relation must hold :

$$P[X(t) = j \mid X(\tau) = i \text{ for } \tau_1 \leq \tau \leq \tau_2 < t] = P[X(\tau_2) = j \mid X(\tau_1) = i]$$

ourselves with some of the deeper questions of convergence of limits in passing from discrete to continuous time; for a careful treatment the reader is referred to [PARZ 62, FELL 66].

Earlier we stated for any Markov process that the time which the process spends in any state must be "memoryless"; this implies that the discrete-time Markov chains must have geometrically distributed state times [which we have already proved in Eq. (2.66)] and that continuous-time Markov chains must have exponentially distributed state times. Let us now prove this last statement. For this purpose let τ_i be a random variable that represents the time which the process spends in state E_i . Recall the Markov property which states that the way in which the past trajectory of the process influences the future development is completely specified by giving the current state of the process. In particular, we need not specify how long the process has been in its current state. This means that the remaining time in E_i must have a distribution that depends only upon i and not upon how long the process has been in E_i . We may write this in the following form:

$$P[T_i > s + 1 | T_i > s] = h(1)$$

where $h(1)$ is a function only of the additional time 1 (and not of the expended time s)*. We may rewrite this conditional probability as follows:

$$P[T_i > s + 1 | T_i > s] = \frac{P[\tau_i > s + 1, \tau_i > s]}{P[T_i > s]}$$

$$\frac{P[T_i > s + 1]}{P[T_i > s]}$$

This last step follows since the event $T_i > s + 1$ implies the event $\tau_i > s$. Rewriting this last equation and introducing $h(1)$ once again we find

$$P[T_i > s + 1] = P[T_i > s]h(1) \quad (2.82)$$

Setting $s = 0$ and observing that $P[T_i > 0] = 1$ we have immediately that

$$P[T_i > 1] = h(1)$$

Using this last equation in Eq. (2.82) we then obtain

$$P[T_i > s + 1] = P[T_i > s]P[T_i > 1] \quad (2.83)$$

for $s, 1 \geq 0$. (Setting $s = 1 = 0$ we again require $P[T_i > 0] = 1$.) We now show that the only continuous distribution satisfying Eq. (2.83) is the

* The symbol s is used as a time variable in this section only and should not be confused with its use as a transform variable elsewhere.

exponential distribution. First we have, by definition, the following general relationship:

$$\begin{aligned} \frac{d}{dt} (P[T_i > t]) &= \frac{d}{dt} (1 - P[\tau_i \leq t]) \\ &= -JT_i(t) \end{aligned} \quad (2.84)$$

where we use the notation $f_{\tau_i}(t)$ to denote the pdf for τ_i . Now let us differentiate Eq. (2.83) with respect to s , yielding

$$\frac{dP[\tau_i > s + t]}{ds} = -JT_i(s)P[T_i > t]$$

where we have taken advantage of Eq. (2.84). Dividing both sides by $P[T_i > t]$ and setting $s = 0$ we have

$$\frac{dP[T_i > t]}{P[T_i > t]} = -f_{\tau_i}(0) ds$$

If we integrate this last from 0 to t we obtain

$$\log_e P[\tau_i > t] = -f_{\tau_i}(0)t$$

or

$$P[T_i > t] = e^{-f_{\tau_i}(0)t}$$

Now we use Eq. (2.84) again to obtain the pdf for τ_i as

$$JT_i(t) = f_{\tau_i}(0)e^{-f_{\tau_i}(0)t} \quad (2.85)$$

which holds for $t \geq 0$. There we have it: the pdf for the time the process spends in state E , is exponentially distributed with the parameter $f_{\tau_i}(0)$, which may depend upon the state E . We will have much more to say about this exponential distribution and its importance in Markov processes shortly.

In the case of a discrete-time homogeneous Markov chain we defined the transition probabilities as $P_{ij} = P[X_n = j | X_{n-1} = i]$ and also the m -step transition probabilities as $p_{ij}^{(m)} = P[X_{n+m} = j | X_n = i]$; these quantities were independent of n due to the homogeneity of the Markov chain. In the case of the nonhomogeneous Markov chain we found it necessary to identify points along the time axis in an absolute fashion and were led to the important transition probability definition $p_{ij}(m, n) = P[X_n = j | X_m = i]$. In a completely analogous way we must now define for our continuous-time Markov chains the following time-dependent transition probability:

$$p_{ij}(s, t) \triangleq P[X(t) = j | X(s) = i] \quad (2.86)$$

where $X(t)$ is the position of the particle at time $t \geq s$. Just as we considered three successive time instants $m \leq q \leq n$ for the discrete case, we may

consider the following three successive time instants for our continuous time chain $s \leq u \leq t$. We may then refer back to Figure 2.7 and identify some sample paths for what we will now consider to be a continuous-time Markov chain; the critical observation once again is that in passing from state E , at time s to state E , at time t , the process must pass through some intermediate state E , at the intermediate time u . We then proceed exactly as we did in deriving Eq. (2.70) and arrive at the following Chapman-Kolmogorov equation for continuous-time Markov chains:

$$p_{ij}(s, t) = \sum_k p_{ik}(s, u) p_{kj}(u, t) \quad (2.87)$$

where $i, j = 0, 1, 2, \dots$. We may put this equation into matrix form if we first define the matrix consisting of elements $P_{ij}(S, t)$ as

$$H(s, t) \triangleq [P_{ij}(S, t)] \quad (2.88)$$

Then the Chapman-Kolmogorov equation becomes.

$$H(s, t) = H(s, u)H(u, t) \quad s \leq u \leq t \quad (2.89)$$

[We define $H(/, t) = I$, the identity matrix.]

In the case of a homogeneous discrete-time Markov chain we found that the matrix equation $\mathbf{1}t = \mathbf{1}t\mathbf{P}$ had to be investigated in order to determine if the chain was ergodic, and so on; also, the transient solution in the non-homogeneous case could be determined from $\mathbf{1}t^{n+1} = \boldsymbol{\pi}^{(0)}\mathbf{P}(0)\mathbf{P}(1) \dots \mathbf{P}(n)$, which was given in terms of the time-dependent transition probabilities $p_{ij}(m, n)$. For the continuous-time Markov chain the one-step transition probabilities are replaced by the infinitesimal rates to be defined below; as we shall see they are given in terms of the time derivative of $P_{ij}(S, t)$ as $t \rightarrow s$.

What we wish now to do is to form the continuous-time analog of the forward and backward equations. So far we have reached Eq. (2.89), which is analogous to Eq. (2.73) in the discrete-time case. We wish to extract the analog for Eqs. (2.74)-(2.77), which show both the term-by-term and matrix form of the forward and backward equations, respectively. We choose to do this in the case of the forward equation, for example, by starting with Eq. (2.75), namely, $H(m, n) = H(m, n-1)P(n-1)$, and allowing the unit time interval to shrink toward zero. To this end we use this last equation and form the following difference:

$$\begin{aligned} H(m, n) - H(m, n-1) &= H(m, n-1)P(n-1) - H(m, n-1) \\ &= H(m, n-1)[P(n-1) - \mathbf{I}] \end{aligned} \quad (2.90)$$

We must now consider some limits. Just as in the discrete case we defined $p_{en} = H(n, n+1)$, we find it convenient in this continuous-time case to

define the following matrix:

$$P(t) \triangleq [p_{ij}(t, t + \Delta t)] \quad - \quad (2.91)$$

Furthermore we identify the matrix $H(s, t)$ as the limit of $H(m, n)$ as our time interval shrinks; similarly we see that the limit of p_{en} will be p_{et} . Returning to Eq. (2.90) we now divide both sides by the time step, which we denote by Δt , and take the limit as $\Delta t \rightarrow 0$. Clearly then the left-hand side limits to the derivative, resulting in

$$\frac{\partial H(s, t)}{\partial t} = H(s, t)Q(t) \quad s \leq t \quad - \quad (2.92)$$

where we have defined the matrix $Q(t)$ as the following limit:

$$Q(t) = \lim_{\Delta t \rightarrow 0} \frac{p_{et} - I}{\Delta t} \quad - \quad (2.93)$$

This matrix $Q(t)$ is known as the *infinitesimal generator* of the transition matrix function $H(s, t)$. Another more descriptive name for $Q(t)$ is the *transition rate* matrix; we will use both names interchangeably. The elements of $Q(t)$, which we denote by $q_{ij}(t)$, are the rates that we referred to earlier. They are defined as follows:

$$q_{ii}(t) = \lim_{\Delta t \rightarrow 0} \frac{p_{ii}(t, t + \Delta t) - 1}{\Delta t} \quad (2.94)$$

$$q_{ij}(t) = \lim_{\Delta t \rightarrow 0} \frac{p_{ij}(t, t + \Delta t)}{\Delta t} \quad i \neq j \quad (2.95)$$

These limits have the following interpretation. If the system at time t is in state E_i ; then the probability that a transition occurs (to any state other than E_i) during the interval $(t, t + \Delta t)$ is given by $-q_{ii}(t)\Delta t + o(\Delta t)$.^{*} Thus we may say that $-q_{ii}(t)$ is the *rate* at which the process departs from state E_i when it is in that state. Similarly, given that the system is in state E_i at time t , the conditional probability that it will make a transition from this state to state E_j in the time interval $(t, t + \Delta t)$ is given by $q_{ij}(t)\Delta t + o(\Delta t)$. Thus

• As usual, the notation $o(\Delta t)$ denotes a function that goes to zero with Δt faster than Δt itself, that is,

$$\lim_{\Delta t \rightarrow 0} \frac{o(\Delta t)}{\Delta t} = 0$$

More generally, one states that the function $g(t)$ is $o(y(t))$ as $t \rightarrow t_1$ if

$$\lim_{t \rightarrow t_1} \frac{|g(t)|}{|y(t)|} = 0$$

See also Chapter 8, P: 284 for a definition of $o(\cdot)$.

$q_{ij}(t)$ is the *rate* at which the process moves from E_j to E_i , given that the system is currently in the state E_j . Since it is always true that $\sum_j p_{ij}(s, t) = 1$ then we see that Eqs. (2.94) and (2.95) imply that

$$\sum_j q_{ij}(t) = 0 \quad \text{for all } i \quad (2.96)$$

Thus we have interpreted the terms in Eq. (2.92); this is nothing more than the *forward* Chapman-Kolmogorov equation for the continuous-time Markov chain.

In a similar fashion, beginning with Eq. (2.77) we may derive the *backward* Chapman-Kolmogorov equation

$$\frac{\partial \mathbf{H}(s, t)}{\partial s} = -\mathbf{Q}(s)\mathbf{H}(s, t) \quad s \leq t \quad (2.97)$$

The forward and backward matrix equations just derived may be expressed through their individual terms as follows. The forward equation gives us [with the additional condition that the passage to the limit in Eq. (2.95) is uniform in i for fixed j]

$$\frac{\partial p_{ij}(s, t)}{\partial t} = q_{ij}(t)p_{ij}(s, t) + \sum_{k \neq j} q_{kj}(t)p_{ik}(s, t) \quad (2.98)$$

The initial state E_j at the initial time s affects the solution of this set of differential equations only through the initial conditions

$$p_{ij}(s, s) = \begin{cases} 1 & \text{if } j = i \\ 0 & \text{if } j \neq i \end{cases}$$

From the backward matrix equation we obtain

$$\frac{\partial p_{ij}(s, t)}{\partial s} = -q_{ii}(s)p_{ij}(s, t) - \sum_{k \neq i} q_{ik}(s)p_{kj}(s, t) \quad (2.99)$$

The "initial" conditions for this equation are

$$p_{ij}(t, t) = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases}$$

These equations [(2.98) and (2.99)] uniquely determine the transition probabilities $p_{ij}(s, t)$ and must, of course, also satisfy Eq. (2.87) as well as the initial conditions.

In matrix notation we may exhibit the solution to the forward and backward Eqs. (2.92) and (2.97), respectively, in a straightforward manner; the

result is'

$$H(s, I) = \exp \left[\int_s^t Q(u) du \right] J \quad - \quad (2.100)$$

We observe that this solution also satisfies Eq. (2.89) and is a continuous-time analog to the discrete-time solution given in Eq. (2.78).

Now for the state probabilities themselves: In analogy with $\pi_j^{(n)}$ we now define

$$\pi_j(t) \triangleq P[X(t) = j] \quad - \quad (2.101)$$

as well as the vector of these probabilities

$$n(t) \triangleq [\pi_1(t), \pi_2(t), \dots] \quad (2.102)$$

If we are given the initial state distribution $n(0)$ then we can solve for the time-dependent state probabilities from

$$n(t) = n(0)H(0, t) \quad (2.103)$$

where a general solution may be seen from Eq. (2.100) to be

$$n(t) = n(0) \exp \left[\int_0^t Q(u) du \right] \quad - \quad (2.104)$$

This corresponds to the discrete-time solution given in Eq. (2.79). The matrix differential equation corresponding to Eq. (2.103) is easily seen to be

$$\frac{dn(t)}{dt} = n(t)Q(t)$$

This last is similar in form to Eq. (2.92) and may be expressed in terms of its elements as

$$\frac{d\pi_j(t)}{dt} = q_{jj}(t)\pi_j(t) + \sum_{k \neq j} q_{kj}(t)\pi_k(t) \quad (2.105)$$

The similarity between Eqs. (2.105) and (2.98) is not accidental. The latter describes the probability that the process is in state E_i at time t given that it was in state E_i at time s . The former merely gives the probability that the system is in state E_i at time t ; information as to where the process began is given in the initial state probability vector $n(0)$. If indeed $\pi_k(0) = 1$ for $k = i$ and $\pi_k(0) = 0$ for $k \neq i$, then we are stating for sure that the system was in state E_i at time 0. In this case $\pi_j(t)$ will be identically equal to $P_{ij}(0, t)$. Both forms for this probability are often used; the form $P_{ij}(s, t)$ is used when

• The expression e^{Pt} , where P is a square matrix is defined as the following matrix power series:

$$e^{Pt} = I + Pt + \frac{P^2 t^2}{2!} + \frac{P^3 t^3}{3!} + \dots$$

we want to specifically show the initial state; the form $\mathcal{T}(I)$ is used when we choose to neglect or imply the initial state.

We now consider the case where our continuous-time Markov chain is *homogeneous*. In this case we drop the dependence upon time and adopt the following notation:

$$P_i(I) \triangleq p_0(s, s + I) \quad (2.106)$$

$$q_{ij} \triangleq q_{ij}(I) \quad i, j = 1, 2, \dots \quad (2.107)$$

$$H(I) \triangleq H(s, s + I) = [p_{ij}(t)] \quad (2.108)$$

$$Q \triangleq Q(I) = [q_{ij}] \quad (2.109)$$

In this case we may list in rapid order the corresponding results. First, the Chapman-Kolmogorov equations become

$$P_{ij}(s + I) = \sum_k p_{ik}(s) p_{kj}(t) \quad (2.110)$$

and in matrix form*

$$H(s + I) = H(s)H(I) \quad (2.111)$$

The forward and backward equations become, respectively,

$$\frac{dp_{ij}(t)}{dt} = q_{ji}p_{ij}(t) + \sum_{k \neq j} q_{kj}p_{ik}(t) \quad (2.112)$$

and

$$-\frac{dp_{ij}(t)}{dt} = -q_{ii}p_{ij}(t) - \sum_{k \neq i} q_{ik}p_{kj}(t) \quad (2.113)$$

and in matrix form these become, respectively,

$$\frac{dH(I)}{dI} = H(I)Q \quad (2.114)$$

and

$$-\frac{dH(I)}{dI} = QH(I) \quad (2.115)$$

with the common initial condition $H(0) = I$. The solution for this matrix is given by

$$H(I) = e^{QI} \quad (2.116)$$

Now for the state probabilities themselves we have the differential equation

$$\frac{d\pi_j(t)}{dt} = q_{jj}\pi_j(t) + \sum_{k \neq j} q_{kj}\pi_k(t) \quad (2.117)$$

which in matrix form is

$$\frac{d\mathcal{T}(I)}{dI} = \mathcal{T}(I)Q \quad (2.118)$$

- The corresponding discrete-time result is simply $p^{m+n} = p^m p^n$.

For an irreducible homogeneous Markov chain it can be shown that the following limits always exist and are independent of the initial state of the chain, namely,

$$\lim_{t \rightarrow \infty} p_{ij}(t) = \pi_j$$

This set $\{\pi_j\}$ will form the limiting state probability distribution. For an ergodic Markov chain we will have the further limit, which will be independent of the initial distribution, namely,

$$\lim_{t \rightarrow \infty} \pi_j(t) = \pi_j$$

This limiting distribution is given uniquely as the solution of the following system of linear equations:

$$\langle\langle_i \rangle\rangle_j + \sum_{k \neq j} q_{kj} \pi_k = 0 \quad (2.115)$$

In matrix form this last equation may be expressed as

$$\mathbf{1} \mathbf{t} \mathbf{Q} = 0 \quad (2.116)$$

where we have used the obvious notation $\mathbf{1} \mathbf{t} = [\pi_1, \pi_2, \dots]$. This last equation coupled with the probability conservation relation, namely,

$$\sum \pi_j = 1 \quad (2.117)$$

uniquely gives us our limiting state probabilities. We compare the Eq. (2.116) with our earlier equation for discrete-time Markov chains, namely, $\mathbf{1} \mathbf{t} = \mathbf{1} \mathbf{t} \mathbf{P}$: here $\mathbf{P}$ was the matrix of transition *probabilities*, whereas the infinitesimal generator $\mathbf{Q}$ is a matrix of transition *rates*.

This completes our discussion of discrete-state Markov chains. In the table on pp. 402-403, we summarize the major results for the four cases considered here. For a further discussion, the reader is referred to [BHAR 60].

Having discussed discrete-state Markov chains (both in discrete and continuous time) it would seem natural that we next consider continuous-state Markov processes. This we will not do, but rather we postpone consideration of such material until we require it [viz. in Chapter 5 we consider Takacs' integrodifferential equation for M/G/I, and in Chapter 2 (Volume II) we develop the Fokker-Planck equation for use in the diffusion approximation for queues]. One would further expect that following the study of Markov processes, we would then investigate renewal processes, random walks, and finally, semi-Markov processes. Here too, we choose to postpone such discussions until they are needed later in the text (e.g., the discussion in Chapter 5 of Markov chains imbedded in semi-Markov processes).

Indeed it is fair to say that much of the balance of this textbook depends upon additional material from the theory of stochastic processes and will be developed as needed. For the time being we choose to specialize the results we have obtained from the continuous-time Markov chains to the class of birth-death processes, which, as we have forewarned, play a major role in queueing systems analysis. This will lead us directly to the important Poisson process.

2.5. BIRTH-DEATH PROCESSES

Earlier in this chapter we said that a birth-death process is the special case of a Markov process in which transitions from state E_k are permitted only to neighboring states E_{k+1} , E_k , and E_{k-1} . This restriction permits us to carry the solution much further in many cases. These processes turn out to be excellent models for all of the material we will study under elementary queueing theory in Chapter 3, and as such forms our point of departure for the study of queueing systems. The discrete-time birth-death process is of less interest to us than the continuous-time case, and, therefore, discrete-time birth-death processes are not considered explicitly in the following development; needless to say, an almost parallel treatment exists for that case. Moreover, transitions of the form from state E_k back to E_k are of direct interest only in the discrete-time Markov chains; in the continuous-time Markov chains, the rate at which the process returns to the state that it currently occupies is infinite, and the astute reader should have observed that we very-carefully subtracted this term out of our definition for $qu(t)$ in Eq. (2.94). Therefore, our main interest will focus on continuous-time birth-death processes with discrete state space in which transitions only to neighboring states E_{k+1} or E_{k-1} from state E_k are permitted.*

Earlier we described a birth-death process as one that is appropriate for modeling changes in the size of a population. Indeed, when the process is said to be in state E_k we will let this denote the fact that the population at that time is of size k . Moreover, a transition from E_k to E_{k+1} will signify a "birth" within the population, whereas a transition from E_k to E_{k-1} will denote a "death" in the population.

Thus we consider changes in size of a population where transitions from state E_k take place to *nearest neighbors* only. Regarding the nature of births and deaths, we introduce the notion of a *birth rate* λ_k , which describes the

* This is true in the one-dimensional case. Later, in Chapter 4, we consider multidimensional systems for which the states are described by discrete vectors, and then each state has two neighbors in *each* dimension. For example, in the two-dimensional case, the state descriptor is a couplet (k_1, k_2) denoted by E_{k_1, k_2} , whose four neighbors are E_{k_1-1, k_2} , E_{k_1+1, k_2} , E_{k_1, k_2-1} , and E_{k_1, k_2+1} .

rate at which births occur when the population is of size k . Similarly, we define a **death rate** μ_k , which is the rate at which deaths occur when the population is of size k . Note that these birth and death rates are independent of time and depend only on E_k ; thus we have a continuous-time homogeneous Markov chain of the birth-death type. We adopt this special notation since it leads us directly into the queueing system notation; note that, in terms of our earlier definitions, we have

$$\lambda_k = q_{k,k+1}$$

and

$$\mu_k = q_{k,k-1}$$

The nearest-neighbor condition requires that $q_{kj} = 0$ for $|k - j| > 1$. Moreover, since we have previously shown in Eq. (2.96) that $\sum_j q_{kj} = 0$, then we require

$$q_{kk} = -(\mu_k + \lambda_k) \quad (2.118)$$

Thus our infinitesimal generator for the general homogeneous birth-death process takes the form

$$Q = \begin{vmatrix} -\lambda_0 & \lambda_0 & 0 & 0 & 0 \\ \mu_1 & -(\lambda_1 + \mu_1) & \lambda_1 & 0 & 0 \\ 0 & \mu_2 & -(\lambda_2 + \mu_2) & \lambda_2 & 0 \\ 0 & 0 & \mu_3 & -(\lambda_3 + \mu_3) & \lambda_3 \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{vmatrix}$$

Note that except for the main, upper, and lower diagonals, all terms are zero.

To be more explicit, the assumptions we need for the birth-death process are that it is a homogeneous Markov chain $X(t)$ on the states $0, 1, 2, \dots$, that births and deaths are independent (this follows directly from the Markov property), and

$$\begin{aligned} B_1: \quad & P[\text{exactly 1 birth in } (t, t + \Delta t) \mid k \text{ in population}] \\ & = \lambda_k \Delta t + o(\Delta t) \end{aligned}$$

$$\begin{aligned} D_1: \quad & P[\text{exactly 1 death in } (t, t + \Delta t) \mid k \text{ in population}] \\ & = \mu_k \Delta t + o(\Delta t) \end{aligned}$$

$$\begin{aligned} B_2: \quad & P[\text{exactly 0 births in } (t, t + \Delta t) \mid k \text{ in population}] \\ & = 1 - \lambda_k \Delta t + o(\Delta t) \end{aligned}$$

$$\begin{aligned} D_2: \quad & P[\text{exactly 0 deaths in } (t, t + \Delta t) \mid k \text{ in population}] \\ & = 1 - \mu_k \Delta t + o(\Delta t) \end{aligned}$$

From these assumptions we see that multiple births, multiple deaths, or in fact, both a birth and a death in a small time interval are prohibited in the sense that each such multiple event is of order $o(\Delta t)$.

What we wish to solve for is the probability that the population size is k at some time t ; this we denote by'

$$Pk(t) \triangleq P[X(t) = k] \quad (2.119)$$

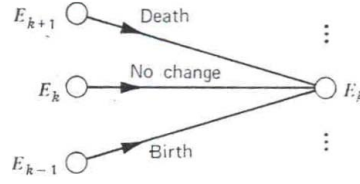
This calculation could be carried out *directly* by using our result in Eq. (2.114) for $\pi_j(t)$ and our specific values for q_{ij} . However, since the derivation of these equations for the birth-death process is so straightforward and follows from first principles, we choose *not* to use the heavy machinery we developed in the previous section, which tends to camouflage the simplicity of the basic approach, but rather to rederive them below. The reader is encouraged to identify the parallel steps in this development and compare them to the more general steps taken earlier. Note in terms of our previous definition that $Pk(t) = \pi_k(t)$. Moreover, we are "suppressing" the initial conditions temporarily, and will introduce them only when required.

We begin by expressing the Chapman-Kolmogorov dynamics, which are quite trivial in this case. In particular, we focus on the possible motions of our particle (that is, the number of members in our population) during an interval $(t, t + \Delta t)$. We will find ourselves in state E_k at time $t + \Delta t$ if one of the three following (mutually exclusive and exhaustive) eventualities occurred:

1. that we had k in the population at time t and no state changes occurred;
2. that we had $k - 1$ in the population at time t and we had a birth during the interval $(t, t + \Delta t)$;
3. that we had $k + 1$ members in the population at time t and we had one death during the interval $(t, t + \Delta t)$.

These three cases are portrayed in Figure (2.8). The probability for the first of these possibilities is merely the probability $Pk(t)$ that we were in state E_k at time t times the probability $p_{kk}(\Delta t)$ that we moved from state E_k to state E_k (i.e., had neither a birth nor a death) during the next Δt seconds; this is represented by the first term on the right-hand side of Eq. (2.120) below. The second and third terms on the right-hand side of that equation correspond, respectively, to the second and third cases listed above. We need not concern ourselves specifically with transitions from states other than nearest neighbors to state E_k since we have assumed that such transitions in an interval of

* We use $X(t)$ here to denote the number in system at time t to be consistent with the use of $X(t)$ for our general stochastic process. Certainly we could have used $N(t)$ as defined earlier; we use $N(t)$ outside of this chapter.



Time

Figure 2.8 Possible transitions into E_k .

duration Δt are of order $o(\Delta t)$. Thus we may write

$$\begin{aligned}
 P_k(t + \Delta t) &= P_k(t)p_{k,k}(\Delta t) \\
 &\quad + P_{k-1}(t)p_{k-1,k}(\Delta t) \\
 &\quad + P_{k+1}(t)p_{k+1,k}(\Delta t) \\
 &\quad + o(\Delta t) \quad k \geq 1
 \end{aligned} \tag{2.120}$$

We may add the three probabilities above since these events are clearly mutually exclusive. Of course, Eq. (2.120) only makes sense in the case for $k \geq 1$, since clearly we could not have had -1 members in the population. For the case $k = 0$ we need the special boundary equation given by

$$\begin{aligned}
 P_0(t + \Delta t) &= P_0(t)p_{00}(\Delta t) \\
 &\quad + P_1(t)p_{10}(\Delta t) \\
 &\quad + o(\Delta t) \quad k = 0
 \end{aligned} \tag{2.121}$$

Furthermore, it is also clear for all values of t that we must conserve our probability, and this is expressed in the following equation:

$$\sum_{k=0}^{\infty} P_k(t) = 1 \tag{2.122}$$

To solve the system represented by Eqs. (2.120)-(2.122) we must make use of our assumptions B_1 , D_1 , B_2 , and D_2 , in order to evaluate the coefficients

in these equations. Carrying out this operation our equations convert to

$$\begin{aligned}
 P_k(t + \Delta t) = & P_k(t)[1 - \lambda_k \Delta t + o(\Delta t)][1 - \mu_k \Delta t + o(\Delta t)] \\
 & + P_{k-1}(t)[\lambda_{k-1} \Delta t + o(\Delta t)] \\
 & + P_{k+1}(t)[\mu_{k+1} \Delta t + o(\Delta t)] \\
 & + o(\Delta t) \quad k \geq 1 \quad (2.123)
 \end{aligned}$$

$$\begin{aligned}
 P_0(t + \Delta t) = & P_0(t)[1 - \lambda_0 \Delta t + o(\Delta t)] \\
 & + P_1(t)[\mu_1 \Delta t + o(\Delta t)] \\
 & + o(\Delta t) \quad k = 0 \quad (2.124)
 \end{aligned}$$

In Eq. (2.124) we have used the assumption that it is impossible to have a death when the population is of size 0 (i.e., $\mu_0 = 0$) and the assumption that one indeed can have a birth when the population size is 0 ($\lambda_0 \geq 0$). Expanding the right-hand side of Eqs. (2.123) and (2.124) we have

$$\begin{aligned}
 P_k(t + \Delta t) = & P_k(t) - (\lambda_k + \mu_k) \Delta t P_k(t) + \lambda_{k-1} \Delta t P_{k-1}(t) \\
 & + \mu_{k+1} \Delta t P_{k+1}(t) + o(\Delta t) \quad k \geq 1 \\
 P_0(t + \Delta t) = & P_0(t) - \lambda_0 \Delta t P_0(t) + \mu_1 \Delta t P_1(t) + o(\Delta t) \quad k = 0
 \end{aligned}$$

If we now subtract $P_k(t)$ from both sides of each equation and divide by Δt , we have the following:

$$\begin{aligned}
 \frac{P_k(t + \Delta t) - P_k(t)}{\Delta t} = & -(\lambda_k + \mu_k)P_k(t) + \lambda_{k-1}P_{k-1}(t) \\
 & + \mu_{k+1}P_{k+1}(t) + \frac{o(\Delta t)}{\Delta t} \quad k \geq 1 \quad (2.125)
 \end{aligned}$$

$$\frac{P_0(t + \Delta t) - P_0(t)}{\Delta t} = -\lambda_0 P_0(t) + \mu_1 P_1(t) + \frac{o(\Delta t)}{\Delta t} \quad k = 0 \quad (2.126)$$

Taking the limit as Δt approaches 0 we see that the left-hand sides of Eqs. (2.125) and (2.126) represent the formal derivative of $P_k(t)$ with respect to t and also that the term $o(\Delta t)/\Delta t$ goes to 0. Consequently, we have the resulting equations:

$$\begin{aligned}
 \frac{dP_k(t)}{dt} = & -(\lambda_k + \mu_k)P_k(t) + \lambda_{k-1}P_{k-1}(t) + \mu_{k+1}P_{k+1}(t) \quad k \geq 1 \\
 \frac{dP_0(t)}{dt} = & -\lambda_0 P_0(t) + \mu_1 P_1(t) \quad k = 0
 \end{aligned} \quad (2.127)$$

The set of equations given by (2.127) is clearly a set of *differential-difference* equations and represents the dynamics of our probability system; we

recognize them as Eq. (2.114) and their solution will give the behavior of $P_k(t)$. It remains for us to solve them. (Note that this set was obtained by essentially using the Chapman-Kolmogorov equations.)

In order to solve Eqs. (2.127) for the time-dependent behavior $P_k(t)$ we now require our initial conditions: that is, we must specify $P_k(0)$ for $k = 0, 1, 2, \dots$. In addition, we further require that Eq. (2.122) be satisfied.

Let us pause temporarily to describe a simple *inspection* technique for finding the differential-difference equations given above. We begin by observing that an alternate way for displaying the information contained in the Q matrix is by means of the *state-transition-rate diagram*. In such a diagram the state E_k is represented by an oval surrounding the number k . Each nonzero infinitesimal rate $q_{jj'}$ (the elements of the Q matrix) is represented in the state-transition-rate diagram by a directed branch pointing from E_j to $E_{j'}$ and labeled with the value $q_{jj'}$. Furthermore, since it is clear that the terms along the main diagonal of Q contain no new information [see Eqs. (2.96) and (2.118)] we do not include the "self"-loop from E_j back to E_j . Thus the state-transition-rate diagram for the general birth-death process is as shown in Figure 2.9.

In viewing this figure we may truly think of a particle in motion moving among these states; the branches identify the permitted transitions and the branch labels give the infinitesimal rates at which these transitions take place. We emphasize that the labels on the ordered links refer to birth and death *rates* and *not to probabilities*. If one wishes to convert these labels to probabilities, one must multiply each by the quantity dt to obtain the probability of such a transition occurring in the next interval of time whose duration is dt . In that case it is also necessary to put self-loops on each state indicating the probability that in the next interval of time dt the system remains in the given state. Note that the state-transition-rate diagram contains exactly the same information as does the transition-rate matrix Q .

Concentrating on state E_k we observe that one may enter it only from state E_{k-1} or from state E_{k+1} and similarly one leaves state E_k only by entering state E_{k-1} or state E_{k+1} . From this picture we see why such processes are referred to as "nearest-neighbor" birth-death processes.

Since we are considering a dynamic situation it is clear that the difference between the rate at which the system enters E_k and the rate at which the system leaves E_k must be equal to the rate of change of "flow" into that state. This

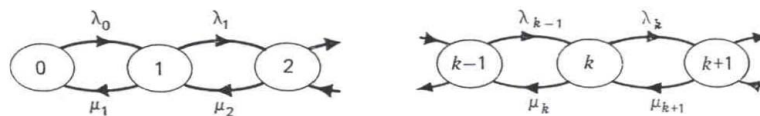


Figure 2.9 State-transition-rate diagram for the birth-death process.

notion is *crucial* and provides for us a simple intuitive means for writing down the equations of motion for the probabilities $P_k(t)$. Specifically, if we focus upon state E_k we observe that the rate at which *probability* "flows" into this state at time t is given by

$$\text{Flow rate into } E_k = \lambda_{k-1}P_{k-1}(t) + \mu_{k+1}P_{k+1}(t)$$

whereas the flow rate out of that state at time t is given by

$$\text{Flow rate out of } E_k = (\lambda_k + \mu_k)P_k(t)$$

Clearly the difference between these two is the effective probability flow rate into this state, that is,

$$\frac{dP_k(t)}{dt} = \lambda_{k-1}P_{k-1}(t) + \mu_{k+1}P_{k+1}(t) - (\lambda_k + \mu_k)P_k(t) \quad - \quad (2.128)$$

But this is exactly Eq. (2.127)! Of course, we have not attended to the details for the boundary state E_0 but it is easy to see that the rate argument just given leads to the correct equation for $k = 0$. Observe that each term in Eq. (2.128) is of the form: probability of being in a particular state at time t multiplied by the infinitesimal rate of leaving that state. It is clear that what we have done is to draw an imaginary boundary surrounding state E_k and have calculated the probability flow rates crossing that boundary, where we place opposite signs on flows entering as opposed to leaving; this total computation is then set equal to the time derivative of the probability flow rate into that state.

Actually there is no reason for selecting a single state as the "system" for which the flow equations must hold. In fact one may enclose any number of states within a contour and then write a flow equation for all flow crossing that boundary. The only danger in dealing with such a conglomerate set is that one may write down a *dependent* set of equations rather than an independent set; on the other hand, if one systematically encloses each state singly and writes down a conservation law for each, then one is guaranteed to have an independent set of equations for the system with the qualification that the conservation of probability given by Eq. (2.122) must also be applied.* Thus we have a simple inspection technique for arriving at the equations of motion for the birth-death process. As we shall see later this approach is perfectly suitable for other Markov processes (including semi-Markov processes) and will be used extensively; these observations also lead us to the notion of global and local balance equations (see Chapter 4).

At this point it is important for the reader to recognize and accept the fact that the birth-death process described above is capable of providing the

* When the number of states is finite (say, K states) then any set of $K - 1$ single-node state equations will be independent. The additional equation needed is Eq. (2.122).

framework for discussing a large number of important and interesting problems in queueing theory. The direct solution for appropriate special cases of Eq. (2.127) provides for us the transient behavior of these queueing systems and is of less interest to this book than the equilibrium or steady-state behavior of queues.* However, for purposes of illustration and to elaborate further upon these equations, we now consider some important examples.

The simplest system to consider is a *pure birth* system in which we assume $\mu_k = 0$ for all k (note that we have now entered the next-to-innermost circle in Figure 2.4!). Moreover, to simplify the problem we will assume that $\lambda_k = \lambda$ for all $k = 0, 1, 2, \dots$ (Now we have entered the innermost circle! We therefore expect some marvelous properties to emerge.) Substituting this into our Eqs. (2.127) we have

$$\begin{aligned} \frac{dP_k(t)}{dt} &= -\lambda P_k(t) + \lambda P_{k-1}(t) & k \geq 1 \\ \frac{dP_0(t)}{dt} &= -\lambda P_0(t) & k = 0 \end{aligned} \quad (2.129)$$

For simplicity we assume that the system begins at time 0 with 0 members, that is,

$$P_k(0) = \begin{cases} 1 & k = 0 \\ 0 & k \neq 0 \end{cases} \quad (2.130)$$

Solving for $P_0(t)$ we have immediately

$$P_0(t) = e^{-\lambda t}$$

Inserting this last into Eq. (2.129) for $k = 1$ results in

$$\frac{dP_1(t)}{dt} = -\lambda P_1(t) + \lambda P_0(t)$$

The solution to this differential equation is clearly

$$P_1(t) = \lambda t e^{-\lambda t}$$

Continuing by induction, then, we finally have as a solution to Eq. (2.129)

$$P_k(t) = \frac{(\lambda t)^k}{k!} e^{-\lambda t} \quad k \geq 0, t \geq 0 \quad (2.131)$$

This is the celebrated *Poisson* distribution. It is a pure birth process with constant birth rate λ and gives rise to a sequence of birth epochs which are

* Transient behavior is discussed elsewhere in this text, notably in Chapter 2 (Vol. II). For an excellent treatment the reader is referred to [COHEN 69].

said to constitute a *Poisson process*. Let us study the Poisson process more carefully and show its relationship to the exponential distribution.

The Poisson process is central to much of elementary and intermediate queueing theory and is widely used in their development. The special position of this process comes about for two reasons. First, as we have seen, it is the "innermost circle" in Figure 2.4 and, therefore, enjoys a number of marvelous and simplifying analytical and probabilistic properties; this will become undeniably apparent in our subsequent development. The second reason for its great importance is that, in fact, numerous natural physical and organic processes exhibit behavior that is probably meaningfully modeled by Poisson processes. For example, as Fry [FRY 28] so graphically points out, one of the first observations of the Poisson process was that it properly represented the number of army soldiers killed due to being kicked (in the head?) by their horses. Other examples include the sequence of gamma rays emitting from a radioactive particle, and the sequence of times at which telephone calls are originated in the telephone network. In fact, it was shown by Palm [PALM 43] and Khinchin [KHIN 60] that in many cases the sum of a large number of independent stationary renewal processes (each with an arbitrary distribution of renewal time) will tend to a Poisson process. This is an important limit theorem and explains why Poisson processes appear so often in nature where the aggregate effect of a large number of individuals or particles is under observation.

Since this development is intended for our use in the study of queueing systems, let us immediately adopt queueing notation and also condition ourselves to discussing a Poisson process as the *arrival of customers* to some queueing facility rather than as the birth of new members in a population. Thus λ is the average rate at which these customers arrive. With the initial condition in Eq. (2.130), $P_k(t)$ gives the probability that k arrivals occur during the time interval $(0, t)$. It is intuitively clear, since the average arrival rate is λ per second, that the average number of arrivals in an interval of length t must be λt . Let us carry out the calculation of this last intuitive statement. Defining K as the number of arrivals in this interval of length t [previously we used $x(t)$] we have

$$\begin{aligned} E[K] &= \sum_{k=0}^{\infty} k P_k(t) \\ &= e^{-\lambda t} \sum_{k=0}^{\infty} k \frac{(\lambda t)^k}{k!} \\ &= e^{-\lambda t} \sum_{k=1}^{\infty} \frac{(\lambda t)^k}{(k-1)!} \\ &= e^{-\lambda t} \lambda t \sum_{k=0}^{\infty} \frac{(\lambda t)^k}{k!} \end{aligned}$$

By definition, we know that $e^x = 1 + x + \frac{x^2}{2!} + \dots$ and so we get

$$E[K] = \lambda t \quad - (2.132)$$

Thus clearly the expected number of arrivals in $(0, t)$ is equal to λt .

We now proceed to calculate the variance of the number of arrivals. In order to do this we find it convenient to first calculate the following moment

$$\begin{aligned} E[K(K-1)] &= \sum_{k=0}^{\infty} k(k-1)P_k(t) \\ &= e^{-\lambda t} \sum_{k=0}^{\infty} k(k-1) \frac{(\lambda t)^k}{k!} \\ &= e^{-\lambda t} (\lambda t)^2 \sum_{k=2}^{\infty} \frac{(\lambda t)^{k-2}}{(k-2)!} \\ &= e^{-\lambda t} (\lambda t)^2 \sum_{k=0}^{\infty} \frac{(\lambda t)^k}{k!} \\ &= (\lambda t)^2 \end{aligned}$$

Now forming the variance in terms of this last quantity and in terms of $E[K]$, we have

$$\begin{aligned} \sigma_K^2 &= E[K(K-1)] + E[K] - (E[K])^2 \\ &= (\lambda t)^2 + \lambda t - (\lambda t)^2 \end{aligned}$$

and so

$$\sigma_K^2 = \lambda t \quad - (2.133)$$

Thus we see that the mean and variance of the Poisson process are identical and each equal to λt .

In Figure 2.10 we plot the family of curves $P_k(t)$ as a function of k and as a function of λt (a convenient normalizing form for t),

Recollect from Eq. (11.27) in Appendix II that the z-transform (probability generating function) for the probability mass distribution of a discrete random variable K where

$$g_k = P[K = k]$$

is given by

$$\begin{aligned} G(z) &= E[z^K] \\ &= \sum_k z^k g_k \end{aligned}$$

for $|z| \leq 1$. Applying this to the Poisson distribution derived above we have

$$\begin{aligned} E[z^K] &= \sum_{k=0}^{\infty} z^k P_k(t) \\ &= \sum_{k=0}^{\infty} e^{-\lambda t} \frac{(\lambda t z)^k}{k!} \\ &= e^{-\lambda t + \lambda t z} \end{aligned}$$

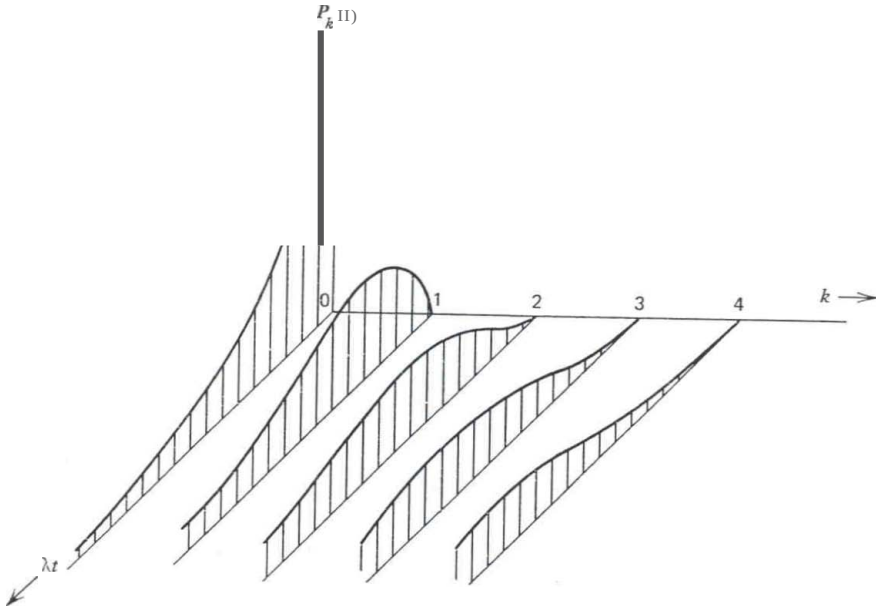


Figure 2.10 The Poisson distribution.

and so

$$G(z) = E[z^K] = e^{\lambda t(z-1)} \quad - (2.134)$$

We shall make considerable use of this result for the z-transform of a Poisson distribution. For example, we may now easily calculate the mean and variance as given in Eqs. (2.132) and (2.133) by taking advantage of the special properties of the z-transform (see Appendix II) as follows*:

$$G^{(1)}(1) = \frac{\partial}{\partial z} E[z^K] \Big|_{z=1} = E[K]$$

Applying this to the Poisson distribution, we get

$$\begin{aligned} E[K] &= \lambda t e^{\lambda t(z-1)} \Big|_{z=1} \\ &= \lambda t \end{aligned}$$

Also

$$\sigma_K^2 = G^{(2)}(1) + G^{(1)}(1) - [G^{(1)}(1)]^2$$

Thus, for the Poisson distribution,

$$\begin{aligned} \sigma_K^2 &= (\lambda t)^2 e^{\lambda t(z-1)} \Big|_{z=1} + \lambda t - (\lambda t)^2 \\ &= \lambda t \end{aligned}$$

This confirms our earlier calculations.

- The shorthand notation for derivatives given in Eq. (11.25) should be reviewed.

We have introduced the Poisson process here as a pure birth process and we have found an expression for $P_k(t)$, the probability distribution for the number of arrivals during a given interval of length t . Now let us consider the joint distribution of the arrival instants when it is known beforehand that exactly k arrivals have occurred during that interval. We break the interval $(0, t)$ into $2k + 1$ intervals as shown in Figure 2.11. We are interested in A_k , which is defined to be the event that exactly one arrival occurs in each of the intervals $\{\beta_i\}$ and that no arrival occurs in any of the intervals $\{\alpha_i\}$. We wish to calculate the probability that the event A_k occurs given that exactly k arrivals have occurred in the interval $(0, t)$: from the definition of conditional probability we thus have

$$P[A_k | \text{exactly } k \text{ arrivals in } (0, t)] = \frac{P[A_k \text{ and exactly } k \text{ arrivals in } (0, t)]}{P[\text{exactly } k \text{ arrivals in } (0, t)]} \quad (2.135)$$

When we consider Poisson arrivals in nonoverlapping intervals, we are considering independent events whose joint probability may be calculated as the product of the individual probabilities (i.e., the Poisson process has independent increments). We note from Eq. (2.131), therefore, that

$$P[\text{one arrival in interval of length } \beta_i] = \lambda \beta_i e^{-\lambda \beta_i}$$

and

$$P[\text{no arrival in interval of length } \alpha_i] = e^{-\lambda \alpha_i}$$

Using this in Eq. (2.135) we have directly

$$\begin{aligned} P[A_k | \text{exactly } k \text{ arrivals in } (0, t)] &= (\lambda \beta_1 \lambda \beta_2 \dots \lambda \beta_k e^{-\lambda \beta_1} e^{-\lambda \beta_2} \dots e^{-\lambda \beta_k}) (e^{-\lambda \alpha_1} e^{-\lambda \alpha_2} \dots e^{-\lambda \alpha_{k+1}}) \\ &= \frac{(\lambda t)^k / k!}{t^k} e^{-\lambda t} \end{aligned} \quad (2.136)$$

On the other hand, let us consider a new process that selects k points in the interval $(0, t)$ independently where each point is uniformly distributed over this interval. Let us now make the same calculation that we did for the Poisson process, namely,

$$p[A_k | \text{exactly } k \text{ arrivals in } (0, t)] = \left(\frac{\beta_1}{t}\right) \left(\frac{\beta_2}{t}\right) \dots \left(\frac{\beta_k}{t}\right) k! \quad (2.137)$$

where the term $k!$ comes about since we do not distinguish among the permutations of the k points among the k chosen intervals. We observe that the two conditional probabilities given in Eqs. (2.136) and (2.137) are the same and, therefore, conclude that if an interval of length t contains exactly k arrivals from a Poisson process, then the joint distribution of the instants when these arrivals occurred is the same as the distribution of k points uniformly distributed over the same interval.

Furthermore, it is easy to show from the properties of our birth process that the Poisson process is one with independent increments; that is, defining $X(s, s+t)$ as the number of arrivals in the interval $(s, s+t)$ then the following is true:

$$P[X(s, s+t) = k] = \frac{(\lambda t)^k e^{-\lambda t}}{k!}$$

regardless of the location of this interval.

We would now like to investigate the intimate relationship between the Poisson process and the exponential distribution. This distribution also plays a central role in queueing theory. We consider the random variable l , which we recall is the *time between adjacent arrivals* in a queueing system, and whose PDF and pdf are given by $A(t)$ and $a(t)$, respectively, as already agreed for the interarrival times. From its definition, then, $a(t)\Delta t + o(\Delta t)$ is the probability that the next arrival occurs at least t sec and at most $(t + \Delta t)$ sec from the time of the last arrival.

Since the definition of $A(t)$ is merely the probability that the time between arrivals is $\leq t$, it must clearly be given by

$$A(t) = 1 - P[l > t]$$

But $P[l > t]$ is just the probability that no arrivals occur in $(0, t)$, that is, $P(0, t)$. Therefore, we have

$$A(t) = 1 - P(0, t)$$

and so from Eq. (2.131), we obtain the PDF (in the Poisson case)

$$A(t) = 1 - e^{-\lambda t} \quad t \geq 0 \quad (2.138)$$

Differentiating, we obtain the pdf

$$a(t) = \lambda e^{-\lambda t} \quad t \geq 0 \quad (2.139)$$

This is the well-known *exponential* distribution; its pdf and PDF are shown in Figure 2.12.

What we have shown by Eqs. (2.138) and (2.139) is that for a Poisson arrival process, the time between arrivals is exponentially distributed; thus we say that the Poisson arrival process has exponential interarrival times.

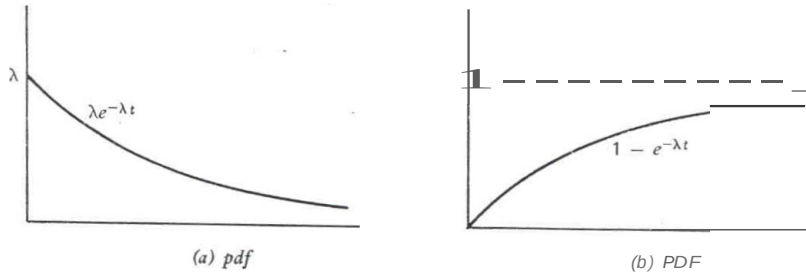


Figure 2.12 Exponential distribution.

The most amazing characteristic of the exponential distribution is that it has the remarkable *memoryless* property, which we introduced in our discussion of Markov processes. As the name indicates, the past history of a random variable that is distributed exponentially plays no role in predicting its future; precisely, we mean the following. Consider that an arrival has just occurred at time 0. If we inquire as to what our feeling is regarding the distribution of time until the next arrival, we clearly respond with the pdf given in Eq. (2.139). Now let some time pass, say, t_0 sec, during which no arrival occurs. We may at this point in time again ask, "What is the probability that the next arrival occurs t sec from *now*?" This question is the same question we asked at time 0 except we now know that the time between arrivals is at least t_0 sec. To answer the second question, we carry out the following calculations:

$$\begin{aligned}
 P[\tilde{t} \leq t + t_0 | i > t_0] &= \frac{P[t_0 \leq \tilde{t} \leq t + t_0]}{P[\tilde{t} > t_0]} \\
 &= \frac{P[i \leq t + t_0] - P[\tilde{t} < t_0]}{P[i > t_0]}
 \end{aligned}$$

Due to Eq. (2.138) we then have

$$P[\tilde{t} \leq t + t_0 | \tilde{t} > t_0] = \frac{1 - e^{-\lambda(t+t_0)} - (1 - e^{-\lambda t_0})}{1 - (1 - e^{-\lambda t_0})}$$

and so

$$P[i \leq t + t_0 | i > t_0] = 1 - e^{-\lambda t} \quad (2.140)$$

This result shows that the distribution of remaining time until the next arrival, given that t_0 sec has elapsed since the last arrival, is identically equal to the unconditional distribution of interarrival time. The impact of this statement is that our probabilistic feeling regarding the time until a future arrival occurs is independent of how long it has been since the last arrival occurred. That is, the future of an exponentially distributed random variable

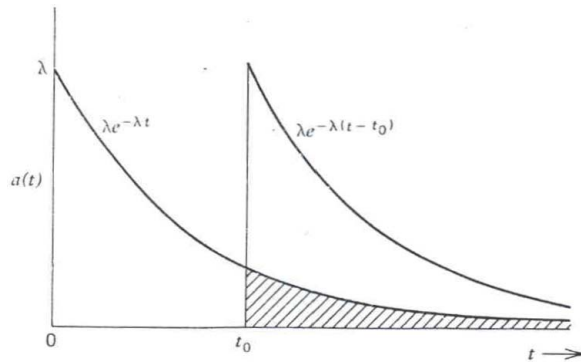


Figure 2.13 The memoryless property of the exponential distribution.

is independent of the past history of that variable and this distribution remains constant in time. The exponential distribution is the *Dilly* continuous distribution with this property. (In the case of a discrete random variable we have seen that the geometric distribution is the only discrete distribution with that same property.) We may further appreciate the nature of this memoryless property by considering Figure 2.13. In this figure we show the exponential density $\lambda e^{-\lambda t}$. Now given that t_0 sec has elapsed, in order to calculate the density function for the time until the next arrival, what one must do is to take that portion of the density function lying to the right of the point t_0 (shown shaded) and recognize that this region represents our probabilistic feeling regarding the future; the portion of the density function in the interval from 0 to t_0 is past history and involves no more uncertainty. In order to make the shaded region into a bona fide density function, we must magnify it in order to increase its total area to unity; the appropriate magnification takes place by dividing the function representing the tail of this distribution by the area of the shaded region (which must, of course, be $P[J > t_0]$). This operation is identical to the operation of creating a conditional distribution by dividing a joint distribution by the probability of the condition. Thus the shaded region magnifies into the second indicated curve in Figure 2.13. This new function is an *exact replica* of the original density function as shown from time 0, except that it is shifted t_0 sec to the right. No other density function has the property that its tail everywhere possesses the exact same shape as the entire density function.

We now use the memoryless property of the exponential distribution in order to close the circle regarding the relationship between the Poisson and exponential distributions. Equation (2.140) gives an expression for the PDF of the interarrival time conditioned on the fact that it is at least as large as t_0 . Let us position ourselves at time t_0 and ask for the probability that the next

arrival occurs within the next Δt sec. From Eq. (2.140) we have

$$\begin{aligned} P[i \leq t_0 + \Delta t | i > t_0] &= 1 - e^{-\lambda \Delta t} \\ &= 1 - \left[1 - \lambda \Delta t + \frac{(\lambda \Delta t)^2}{2!} - \dots \right] \\ &= \lambda \Delta t + o(\Delta t) \end{aligned} \quad (2.141)$$

Equation (2.141) tells us, given that an arrival has not yet occurred, that the probability of it occurring in the next interval of length Δt sec is $\lambda \Delta t + o(\Delta t)$. But this is exactly assumption B[from the opening paragraphs of this section. Furthermore, the probability of no arrival in the interval $(t_0, t_0 + \Delta t)$ is calculated as

$$\begin{aligned} P[i > t_0 + \Delta t | i > t_0] &= 1 - P[i \leq t_0 + \Delta t | i > t_0] \\ &= 1 - (1 - e^{-\lambda \Delta t}) \\ &= e^{-\lambda \Delta t} \\ &= 1 - \lambda \Delta t + \frac{(\lambda \Delta t)^2}{2!} - \dots \\ &= 1 - \lambda \Delta t + o(\Delta t) \end{aligned}$$

This corroborates assumption B₂. Furthermore,

$$\begin{aligned} P[2 \text{ or more arrivals in } (t_0, t_0 + \Delta t)] \\ &= 1 - P[\text{none in } (t_0, t_0 + \Delta t)] - P[\text{one in } (t_0, t_0 + \Delta t)] \\ &= 1 - [1 - \lambda \Delta t + o(\Delta t)] - [\lambda \Delta t + o(\Delta t)] \\ &= o(\Delta t) \end{aligned}$$

This corroborates the "multiple-birth" assumption. Our conclusion, then, is that the assumption of exponentially distributed interarrival times (which are independent one from the other) implies that we have a Poisson process, which implies we have a constant birth rate. The converse implications are also true. This relationship is shown graphically in Figure 2.14 in which the symbol $\leftrightarrow$ here denotes implication in both directions.

Let us now calculate the mean and variance for the exponential distribution as we did for the Poisson process. We shall proceed using two methods (the direct method and the transform method). We have

$$\begin{aligned} E[i] &\triangleq \int_0^{\infty} t a(t) dt \\ &= \int_0^{\infty} t \lambda e^{-\lambda t} dt \end{aligned}$$

We use a trick here to evaluate the (simple) integral by recognizing that the integrand is no more than the partial derivative of the following integral,

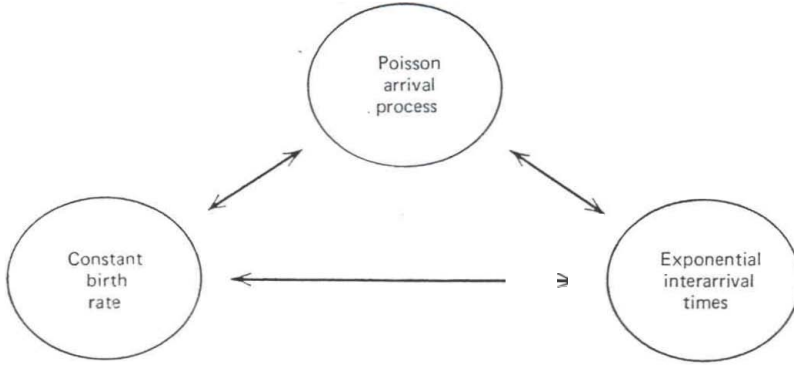


Figure 2.14 The memoryless triangle.

which may be evaluated by inspection:

$$\begin{aligned} \int_0^{\infty} t \lambda e^{-\lambda t} dt &= -\lambda \frac{\partial}{\partial \lambda} \int_0^{\infty} e^{-\lambda t} dt \\ &= -\lambda \frac{\partial}{\partial \lambda} \left(\frac{1}{\lambda} \right) = -\lambda \left(-\frac{1}{\lambda^2} \right) \end{aligned}$$

and so

$$\bar{t} = \frac{1}{\lambda} \quad (2.142)$$

Thus we have that the average interarrival time for an exponential distribution is given by $1/\lambda$. This result is intuitively pleasing if we examine Eq. (2.141) and observe that the probability of an arrival in an interval of length Δt is given by $\lambda \Delta t + o(\Delta t)$ and thus λ itself must be the average rate of arrivals; thus the average time between arrivals must be $1/\lambda$. In order to evaluate the variance, we first calculate the second moment for the interarrival time as follows:

$$\begin{aligned} E[(\bar{t})^2] &= \int_0^{\infty} t^2 a(t) dt \\ &= \int_0^{\infty} t^2 \lambda e^{-\lambda t} dt \\ &= \lambda \frac{\partial^2}{\partial \lambda^2} \int_0^{\infty} e^{-\lambda t} dt \\ &= \lambda \frac{\partial^2}{\partial \lambda^2} \left(\frac{1}{\lambda} \right) \end{aligned}$$

Thus the variance is given by

$$\begin{aligned}\sigma_i^2 &= E[(i)^2] - V^2 \\ &= \frac{2}{\lambda^2} - \left(\frac{1}{\lambda}\right)^2\end{aligned}$$

and so

$$\sigma_i^2 = \frac{1}{\lambda^2} \quad (2.143)$$

As usual, these two moments could more easily have been calculated by first considering the Laplace transform of the probability density function for this random variable. The notation for the Laplace transform of the interarrival pdf is $A^*(s)$. In this special case of the exponential distribution we then have the following:

$$\begin{aligned}A^*(s) &\triangleq \int_0^\infty e^{-st} a(t) dt \\ &= \int_0^\infty e^{-st} \lambda e^{-\lambda t} dt\end{aligned}$$

and so

$$A^*(s) = \frac{\lambda}{s + \lambda} \quad (2.144)$$

Equation (2.144) thus gives the Laplace transform for the exponential density function. From Appendix II we recognize that the mean of this density function is given by

$$\begin{aligned}f &= - \left. \frac{dA^*(s)}{ds} \right|_{s=0} \\ &= \left. \frac{\lambda}{(s + \lambda)^2} \right|_{s=0} \\ &= \frac{1}{\lambda}\end{aligned}$$

The second moment is also calculated in a similar fashion:

$$\begin{aligned}E[(i)^2] &= \left. \frac{d^2 A^*(s)}{ds^2} \right|_{s=0} \\ &= \left. \frac{2\lambda}{(s + \lambda)^3} \right|_{s=0} \\ &= \frac{2}{\lambda}\end{aligned}$$

and so

$$\sigma_i^2 = E[(i)^2] - (i)^2$$

Thus we see the ease with which moments can be calculated by making use of transforms.

Note also, that the coefficient of variation [see Eq. (II.23)] for the exponential is

$$C_a \triangleq \frac{\sigma_i}{\bar{i}} = 1 \quad (2.145)$$

It will be of further interest to us later in the text to be able to calculate the pdf for the time interval X required in order to collect k arrivals from a Poisson process. Let us define this random variable in terms of the random variables I_n where I_n = time between n th and $(n - 1)$ th arrival (where the "zeroth" arrival is assumed to occur at time 0). Thus

$$X = \sum_{n=1}^k t_n$$

We define $f_X(x)$ to be the pdf for this random variable. From Appendix II we should immediately recognize that the density of X is given by the convolution of the densities on each of the I_n 's, since they are independently distributed. Of course, this convolution operation is a bit lengthy to carry out, so let us use our further result in Appendix II, which tells us that the Laplace transform of the pdf for the sum of independent random variables is equal to the product of the Laplace transforms of the density for each. In our case each I_n has a common exponential distribution and therefore the Laplace transform for the pdf of X will merely be the k th power of $A^*(s)$ where $A^*(s)$ is given by Eq. (2.144); that is, defining

$$X^*(s) = \int_0^\infty e^{-sx} f_X(x) dx$$

for the Laplace transform of the pdf of our sum, we have

$$X^*(s) = [A^*(s)]^k$$

thus

$$X^*(s) = \left(\frac{\lambda}{s + \lambda} \right)^k \quad (2.146)$$

We must now invert this transform. Fortunately, we identify the needed transform pair as entry 10 in Table 1.4 of Appendix I. Thus the density function we are looking for, which describes the time required to observe k arrivals, is given by

$$f_X(x) = \frac{\lambda(\lambda x)^{k-1}}{(k-1)!} e^{-\lambda x} \quad x \geq 0 \quad (2.147)$$

This family of density functions (one for each value of k) is referred to as the family of *Erlang* distributions. We will have considerable use for this family later when we discuss the method of stages, in Chapter 4.

So much for the Poisson arrival process and its relation to the exponential distribution. Let us now return to the birth-death equations and consider a more general pure birth process in which we permit state-dependent birth rates λ_k (for the Poisson process, we had $\lambda_k = \lambda$). We once again insist that the death rates $\mu_k = 0$. From Eq. (2.127) this yields the set of equations

$$\begin{aligned} \frac{dP_k(t)}{dt} &= -\lambda_k P_k(t) + \lambda_{k-1} P_{k-1}(t) & k \geq 1 \\ \frac{dP_0(t)}{dt} &= -\lambda_0 P_0(t) & k = 0 \end{aligned} \quad (2.148)$$

Again, let us assume the initial distribution as given in Eq. (2.130), which states that (with probability one) the population begins with 0 members at time 0. Solving for p_{0et} we have

$$p_{0et} = e^{-\lambda_0 t}$$

The general solution* for $P_k(t)$ is given below with an explicit expression for the first two values of k :

$$\begin{aligned} P_k(t) &= e^{-\lambda_k t} \left[\lambda_{k-1} \int_0^t P_{k-1}(x) e^{\lambda_k x} dx + P_k(0) \right] & k = 0, 1, 2, \dots \quad (2.149) \\ P_1(t) &= \frac{\lambda_0 (e^{-\lambda_0 t} - e^{-\lambda_1 t})}{\lambda_1 - \lambda_0} \\ P_2(t) &= \frac{\lambda_0 \lambda_1}{\lambda_1 - \lambda_0} \left[\frac{e^{-\lambda_2 t}}{\lambda_2 - \lambda_1} - \frac{e^{-\lambda_1 t}}{\lambda_1 - \lambda_0} - \frac{e^{-\lambda_2 t}}{\lambda_2 - \lambda_0} + \frac{e^{-\lambda_0 t}}{\lambda_0} \right] \end{aligned}$$

As a third example of the time-dependent solution to the birth-death equations let us consider a *pure death* process in which a population is initiated with, say, N members and all that can happen to this population is that members die; none are born. Thus $\lambda_k = 0$ for all k , and $\mu_k = \mu \geq 0$

* The validity of this solution is easily verified by substituting Eq. (2.149) into Eq. (2.148).

for $k = 1, 2, \dots, N$. For this constant death rate process we have

$$\frac{dP_k(t)}{dt} = -\mu P_k(t) + \mu P_{k+1}(t) \quad 0 < k < N$$

$$\frac{dP_N(t)}{dt} = -\mu P_N(t) \quad k = N$$

$$\frac{dP_0(t)}{dt} = \mu P_1(t) \quad k = 0$$

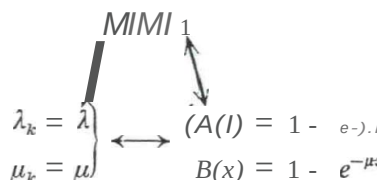
Proceeding as earlier and using induction we obtain the solution

$$P_k(t) = \frac{(\mu t)^{N-k}}{(N-k)!} e^{-\mu t} \quad 0 < k \leq N \quad (2.150)$$

$$\frac{dP_0(t)}{dt} = \frac{\mu(\mu t)^{N-1}}{(N-1)!} e^{-\mu t} \quad k = 0$$

Note the similarity of this last result to the Erlang distribution.

The last case we consider is a birth-death process in which all birth coefficients are equal to λ for $k \geq 0$ and all death coefficients are equal to μ for $k \geq 1$. This birth-death process with constant coefficients is of primary importance and forms perhaps the simplest interesting model of a queueing system. It is the celebrated *M/M/1* queue; recall that the notation denotes a single-server queue with a Poisson arrival process and an exponential distribution for service time (from our earlier discussion we recognize that this is the memoryless system). Thus we may say



$$\left. \begin{array}{l} \lambda_k = \lambda \\ \mu_k = \mu \end{array} \right\} \longleftrightarrow \begin{array}{l} A(l) = 1 - e^{-\mu l} \\ B(x) = 1 - e^{-\mu x} \end{array} \quad (2.151)$$

It should be clear why $A(l)$ is of exponential form from our earlier discussion relating the exponential interarrival distribution with the Poisson arrival process. In a similar fashion, since the death rate is constant ($\mu_k = \mu$, $k = 1, 2, \dots$) then the same reasoning leads to the observation that the time between deaths is also exponentially distributed (in this case with a parameter μ). However, deaths correspond in the queueing system to **service**

completions and, therefore, the service-time distribution $B(x)$ must be of exponential form. The interpretation of the condition $\mu_0 = 0$, which says that the death rate is zero when the population size is zero, corresponds in our queueing system to the condition that no service may take place when no customers are present. The behavior of the system $M/M/1$ will be studied throughout this text as we introduce new methods and new measures of performance; we will constantly check our sophisticated advanced techniques against this example since it affords one of the simplest applications of many of these advanced methods. Moreover, much of the behavior manifested in this system is characteristic of more complex queueing system behavior, and so a careful study here will serve to familiarize the reader with some important queueing phenomena.

Now for our first exposure to the $M/M/1$ system behavior. From the general equation for $P_k(t)$ given in Eq. (2.127) we find for this case that the corresponding differential-difference equations are

$$\begin{aligned} \frac{dP_k(t)}{dt} &= -(\lambda + \mu)P_k(t) + \lambda P_{k-1}(t) + \mu P_{k+1}(t) & k \geq 1 \\ \frac{dP_0(t)}{dt} &= -\lambda P_0(t) + \mu P_1(t) & k = 0 \end{aligned} \quad (2.152)$$

Many methods are available for solving this set of equations. Here, we choose to use the method of z-transforms developed in Appendix I. We have already seen one application of this method earlier in this chapter [when we defined the transform in Eq. (2.60) and applied it to the system of equations (2.55) to obtain the algebraic equation (2.61)]. Recall that the steps involved in applying the method of z-transforms to the solution of a set of difference equations may be summarized as follows:

1. Multiply the k th equation by z^k .
2. Sum all those equations that have the same form (typically true for $k = K, K+1, \dots$).
3. In this single equation, attempt to identify the z-transform for the unknown function. If all but a finite set of terms for the transform are present, then add the missing terms to get the function and then explicitly subtract them out in the equation.
4. Make use of the K "boundary" equations (namely, those that were omitted in step 2 above for $k = 0, 1, \dots, K-1$) to eliminate unknowns in the transformed equation.
5. Solve for the desired transform in the resulting algebraic, matrix or

differential equation. Use the conservation relationship, Eq. (2.122), to eliminate the last unknown term.†

6. Invert the solution to get an explicit solution in terms of k .
7. If step 6 cannot be carried out, then moments may be obtained by differentiating with respect to z and setting $z = 1$.

Let us apply this method to Eq. (2.152). First we define the time-dependent transform

$$P(z, t) \triangleq \sum_{k=0}^{\infty} P_k(t) z^k \quad (2.153)$$

Next we multiply the k th differential equation by z^k (step 1) and then sum over all permitted k ($k = 1, 2, \dots$) (step 2) to yield a single differential equation for the z -transform of $P_k(t)$:

$$\sum_{k=1}^{\infty} \frac{dP_k(t)}{dt} z^k = -(\lambda + \mu) \sum_{k=1}^{\infty} P_k(t) z^k + \lambda \sum_{k=1}^{\infty} P_{k-1}(t) z^k + \mu \sum_{k=1}^{\infty} P_{k+1}(t) z^k$$

Property 14 from Table I.1 in Appendix I permits us to move the differentiation operator outside the summation sign in this last equation. This summation then appears very much like $P(z, t)$ as defined above, except that it is missing the term for $k = 0$; the same is true of the first summation on the right-hand side of this last equation. In these two cases we need merely add and subtract the term $P_0(t)z^0$ which permits us to form the transform we are seeking. The second summation on the right-hand side is clearly $zP(z, t)$ since it contains an extra factor of z , but no missing terms. The last summation is missing a factor of z as well as the first two terms of this sum. We have now

We sometimes obtain a differential equation at this stage if our original set of difference equations was, in fact, a set of differential-difference equations. When this occurs, we arc tively back to step 1 of this procedure as far as the differential variable (usually time) concerned. We then proceed through steps 1-5 a second time using the Laplace transform **if this new variable; our transform multiplier becomes e^{-st} , our sums become integrals.** id our "tricks" become the properties associated with Laplace transforms (see Appendix . Similar "returns to step 1" occur whenever a function of more than one variable is msformed; for each discrete variable, we require a z -transform and, for each continuous riable, we require a Laplace transform.

When additional unknowns remain, we must appeal to the analyticity of the transform rd observe that in its region of analyticity the transform must have a zero to cancel each ile (singularity) if the transform is to remain bounded. These additional conditions mpletely remove any remaining unknowns. This procedure will often be used and plained in the next few chapters.

carried out step 3, which yields

$$\begin{aligned} \frac{\partial}{\partial t} [P(z, t) - P_0(t)] \\ = -(\lambda + \mu)[P(z, t) - P_0(t)] + \lambda z P(z, t) + \frac{\mu}{z} [P(z, t) - P_0(t)] \end{aligned} \quad (2.154)$$

The equation for $k = 0$ has so far not been used and we now apply it as described in step 4 ($K = I$), which permits us to eliminate certain terms in Eq. (2.154):

$$\frac{\partial}{\partial t} P(z, t) = -\lambda P(z, t) - \mu [P(z, t) - P_0(t)] + \lambda z P(z, t) + \frac{\mu}{z} [P(z, t) - P_0(t)]$$

Rearranging this last equation we obtain the following linear, first-order (partial) differential equation for $P(z, t)$:

$$z \frac{\partial}{\partial t} P(z, t) = (1 - z)[(\mu - \lambda z)P(z, t) - \mu P_0(t)] \quad (2.155)$$

This differential equation requires further transforming, as mentioned in the first footnote to step 5. We must therefore define the Laplace transform for our function $P(z, t)$ as follows*:

$$P^*(z, s) \triangleq \int_0^\infty e^{-st} P(z, t) dt \quad (2.156)$$

Returning to step I, applying this transform to Eq. (2.155), and taking advantage of property II in Table 1.3 in Appendix I, we obtain

$$z[sP^*(z, s) - P(z, 0)] = (1 - z)[(\mu - \lambda z)P^*(z, s) - \mu P_0^*(s)] \quad (2.157)$$

where we have defined $P_0^*(s)$ to be the Laplace transform of $P_0(t)$, that is,

$$P_0^*(s) \triangleq \int_0^\infty e^{-st} P_0(t) dt \quad (2.158)$$

We have now transformed the set of differential-difference equations for $P_k(t)$ both on the discrete variable k and on the continuous variable t . This has led us to Eq. (2.157), which is a simple algebraic equation in our twice-transformed function $P^*(z, s)$, and this we may write as

$$P^*(z, s) = \frac{zP(z, 0) - \mu(1 - z)P_0^*(s)}{sz - (1 - z)(\mu - \lambda z)} \quad (2.159)$$

* For convenience we take the lower limit of integration to be 0^+ rather than our usual convention of using 0^- with the nonnegative random variables we often deal with. As a consequence, we must include the initial condition $P(=, 0^+)$ in Eq. (2.157).

Let us carry this argument just a bit further. From the definition in Eq. (2.153) we see that

$$P(\cdot, 0+) = \sum_{k=0}^{\infty} P_k(0+) Z^k \quad (2.160)$$

Of course, $P_k(0+)$ is just our initial condition; whereas earlier we took the simple point of view that the system was empty at time 0 [that is, $P_0(0+) = 1$ and all other terms $P_k(0+) = 0$ for $k \neq 0$], we now generalize and permit i customers to be present at time 0, that is,

$$P_k(0+) = \begin{cases} 1 & k = i \\ 0 & k \neq i \end{cases} \quad (2.161)$$

When $i = 0$ we have our original initial condition. Substituting Eq. (2.161) into Eq. (2.160) we see immediately that

$$P(\cdot, 0+) = Z^i$$

which we may place into Eq. (2.159) to obtain

$$P^*(Z, s) = \frac{Z^{i+1} - \mu(1 - Z)P_0^*(s)}{sZ - (1 - Z)(\mu - \lambda Z)} \quad (2.162)$$

We are almost finished with step 5 except for the fact that the unknown function $P_0^*(s)$ appears in our equation. The second footnote to step 5 tells us how to proceed. From here on the analysis becomes a bit complex and it is beyond our desire at this point to continue the calculation; instead we relegate the excruciating details to the exercises below (see Exercise 2.20). It suffices to say that $P_0^*(s)$ is determined through the denominator roots of Eq. (2.162), which then leaves us with an explicit expression for our double transform. We are now at step 6 and must attempt to invert on both the transform variables; the exercises require the reader to show that the result of this inversion yields the final solution for our transient analysis, namely,

$$P_k(t) = e^{-(\lambda + \mu)t} \left[\rho^{(k-i)/2} I_{k-i}(at) + \rho^{k-i-1/2} I_{k+i+1}(at) + (1 - \rho) \rho^k \sum_{j=k+i+2}^{\infty} \rho^{-j/2} I_j(at) \right] \quad (2.163)$$

where

$$\rho = \frac{\lambda}{\mu} \quad (2.164)$$

$$a = 2\mu\rho^{1/2} \quad (2.165)$$

and

$$I_k(x) \triangleq \sum_{m=0}^{\infty} \frac{(x/2)^{k+2m}}{(k+m)! m!} \quad k \geq -1 \quad (2.166)$$

EXERCISES

- 2.1. Consider K independent sources of customers where the interarrival time between customers for each source is exponentially distributed with parameter λ_k (i.e., each source is a Poisson process). Now consider the arrival stream, which is formed by merging the input from each of the K sources defined above. Prove that this merged stream is also Poisson with parameter $\lambda = \lambda_1 + \lambda_2 + \dots + \lambda_K$.
- 2.2. Referring back to the previous problem, consider this merged Poisson stream and now assume that we wish to break it up into several branches. Let P_i be the probability that a customer from the merged stream is assigned to the substream i . If the overall rate is λ customers per second, and if the substream probabilities P_i are chosen for each customer independently, then show that each of these substreams is a Poisson process with rate λp_i .
- 2.3. Let $\{X_j\}$ be a sequence of identically distributed mutually independent Bernoulli random variables (with $P[X_j = 1] = p$, and $P[X_j = 0] = 1 - p$). Let $S_N = X_1 + \dots + X_N$ be the sum of a random number N of Bernoulli random variables. If N has a Poisson distribution with

where $I_k(x)$ is the modified Bessel function of the first kind of order k . This last expression is most disheartening. What it has to say is that an appropriate model for the *simplest interesting* queueing system (discussed further in the next chapter) leads to an ugly expression for the time-dependent behavior of its state probabilities. As a consequence, we can only hope for greater complexity and obscurity in attempting to find time-dependent behavior of more general queueing systems.

More will be said about time-dependent results later in the text. Our main purpose now is to focus upon the *equilibrium* behavior of queueing systems rather than upon their transient behavior (which is far more difficult). In the next chapter the equilibrium behavior for birth-death queueing systems will be studied and in Chapter 4 more general Markovian queues in equilibrium will be considered. Only when we reach Chapter 5, Chapter 8, and then Chapter 2 (Volume II) will the time-dependent behavior be considered again. Let us now proceed to the simplest equilibrium behavior.

REFERENCES

- BHAR 60 Bharucha-Reid, A. T., *Elements of the Theory of Markov Processes and Their Applications*, McGraw-Hill (New York) 1960.
- COHE 69 Cohen, J., *The Single Server Queue*, North Holland (Amsterdam), 1969.
- EILO 69 Eilon, S., "A Simpler Proof of $L = \lambda W$," *Operations Research*, 17, 915-916 (1969).
- FELL 66 Feller, W., *An Introduction to Probability Theory and Its Applications*. Vol. II, Wiley (New York), 1966.
- FRY 28 Fry, T. C., *Probability and Its Engineering Uses*, Van Nostrand, (New York), 1928.
- HOWA 71 Howard, R. A., *Dynamic Probabilistic Systems*, Vol. I (Markov Models) and Vol. II (Semi-Markov and Decision Processes), Wiley

- (c) Solve for the equilibrium probability vector π .
- (d) What is the mean recurrence time for state E_2 ?
- (e) For which values of α and p will we have $\pi_1 = \pi_2 = \pi_3$? (Give a physical interpretation of this case.)

- 2.6. Consider the discrete-state, discrete-time Markov chain whose transition probability matrix is given by

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

- (a) Find the stationary state probability vector π .
 - (b) Find $\|I - ZP\|_1$.
 - (c) Find the general form for P^n .
- 2.7. Consider a Markov chain with states $E_0, E_1, E_2, \dots$ and with transition probabilities

$$p_{ij} = e^{-\lambda} \sum_{n=0}^j \binom{j}{n} P^n q^{i-n} \left(\frac{\lambda^{j-n}}{j-n}\right)'$$

where $p + q = 1$ ($0 < p < 1$).

- (a) Is this chain irreducible? Periodic? Explain.
- (b) We wish to find

$$\pi_i = \text{equilibrium probability of } E_i$$

Write π_i in terms of p_{ij} and π_j for $j = 0, 1, 2, \dots$.

- (c) From (b) find an expression relating $P(z)$ to $P[1 + p(z - 1)]$, where

$$P(z) = \sum_{i=0}^{\infty} \pi_i z^i$$

- (d) Recursively (i.e., repeatedly) apply the result in (c) to itself and show that the n th recursion gives

$$P(z) = e^{\lambda(z-1)(1+p+p^2+\dots+p^{n-1})} [1 + p^n(z-1)]$$

- (e) From (d) find $P(z)$ and then recognize π_i .

- 2.8. Show that any point in or on the equilateral triangle of unit height shown in Figure 2.6 represents a three-component probability vector in the sense that the sum of the distances from any such point to each of the three sides must always equal unity.

- 2.9. Consider a pure birth process with constant birth rate λ . Let us consider an interval of length T , which we divide up into m segments each of length T/m . Define $\Delta t = T/m$.
- For Δt small, find the probability that a single arrival occurs in each of exactly k of the m intervals and that no arrivals occur in the remaining $m - k$ intervals.
 - Consider the limit as $\Delta t \rightarrow 0$, that is, as $m \rightarrow \infty$ for fixed T , and evaluate the probability $P_k(T)$ that exactly k arrivals occur in the interval of length T .
- 2.10. Consider a population of bacteria of size $N(t)$ at time t for which $N(0) = 1$. We consider this to be a pure birth process in which any member of the population will split into two new members in the interval $(t, t + \Delta t)$ with probability $\lambda \Delta t + o(\Delta t)$ or will remain unchanged in this interval with probability $1 - \lambda \Delta t + o(\Delta t)$ as $\Delta t \rightarrow 0$.
- Let $P_k(t) = P[N(t) = k]$ and write down the set of differential-difference equations that must be satisfied by these probabilities.
 - From part (a) show that the a-transform $P(z, t)$ for $N(t)$ must satisfy

$$P(z, t) = \frac{ze^{-\lambda t}}{1 - z + ze^{\lambda t}}$$

- Find $E[N(t)]$.
 - Solve for $P_k(t)$.
 - Solve for $P(z, t)$, $E[N(t)]$ and $P_k(t)$ that satisfy the initial condition $N(0) = 1$.
 - Consider the corresponding deterministic problem in which each bacterium splits into two every $1/\lambda$ sec and compare with the answer in part (c).
- 2.11. Consider a birth-death process with coefficients

$$\lambda_k = \begin{cases} \lambda & k=0 \\ 0 & k \neq 0 \end{cases} \quad \mu_k = \begin{cases} \mu & k=1 \\ 0 & k \neq 1 \end{cases}$$

which corresponds to an *MIMI* queueing system where there is no room for waiting customers.

- Give the differential-difference equations for $P_k(t)$ ($k = 0, 1$).
- Solve these equations and express the answers in terms of $P_0(0)$ and $P_1(0)$.

2.12. Consider a birth-death queueing system in which

$$\begin{aligned}\lambda_k &= \lambda & k \geq 0 \\ \mu_k &= k\mu & k \geq 0\end{aligned}$$

(a) For all k , find the differential-difference equations for

$$Pk(t) = P[k \text{ in system at time } t]$$

(b) Define the z -transform

$$P(z, t) = \sum_{k=0}^{\infty} P_k(t) z^k$$

and find the *partial* differential equation that $P(z, t)$ must satisfy.

(c) Show that the solution to this equation is

$$P(z, t) = \exp\left(\frac{\lambda}{\mu}(1 - e^{-\mu t})(z - 1)\right)$$

with the initial condition $P(z, 0) = 1$.

(d) Comparing the solution in part (c) with Eq. (2.134), give the expression for $P_k(t)$ by inspection.

(e) Find the limiting values for these probabilities as $t \rightarrow \infty$.

2.13. Consider a system in which the birth rate decreases and the death rate increases as the number in the system k increases, that is,

$$\lambda_k = \begin{cases} (K - k)\lambda & k \leq K \\ 0 & k > K \end{cases} \quad \mu_k = \begin{cases} k\mu & k \leq K \\ 0 & k > K \end{cases}$$

Write down the differential-difference equations for

$$Pk(t) = P[k \text{ in system at time } t].$$

2.14. Consider the case of a linear birth-death process in which $\lambda_k = k\lambda$ and $\mu_k = k\mu$:

(a) Find the partial-differential equation that must be satisfied by $P(z, t)$ as defined in Eq. (2.153).

(b) Assuming that the population size is one at time zero, show that the function that satisfies the equation in part (a) is

$$P(z, t) = \frac{\mu(1 - e^{(\lambda-\mu)t}) - (\lambda - \mu e^{(\lambda-\mu)t})z}{\mu - \lambda e^{(\lambda-\mu)t} - \lambda(1 - e^{(\lambda-\mu)t})z}$$

- (c) Expanding $P(z, t)$ in a power series show that

$$\begin{aligned} P_k(t) &= [1 - \alpha(t)][1 - \beta(t)][\beta(t)]^{k-1} \quad k = 1, 2, \dots \\ p_{\text{out}}(t) &= \alpha(t) \end{aligned}$$

and find $\alpha(t)$ and $\beta(t)$.

- (d) Find the mean and variance for the number in system at time t .
 (e) Find the limiting probability that the population dies out by time t for $t \rightarrow \infty$.

- 2.15. Consider a linear birth-death process for which $\lambda_k = k\lambda + \alpha$ and $\mu_k = ku$,

- (a) Find the differential-difference equations that must be satisfied by $P_k(t)$.
 (b) From (a) find the partial-differential equation that must be satisfied by the time-dependent transform defined as

$$P(z, t) = \sum_{k=0}^{\infty} P_k(t) z^k$$

- (c) What is the value of $P(1, t)$? Give a verbal interpretation for the expression

$$\bar{N}(t) = \lim_{z \rightarrow 1} \frac{\partial}{\partial z} P(z, t)$$

- (d) Assuming that the population size begins with i members at time 0, find an ordinary differential equation for $\bar{N}(t)$ and then solve for $\bar{N}(t)$. Consider the case $\lambda = \mu$ as well as $\lambda \neq \mu$.
 (e) Find the limiting value for $\bar{N}(t)$ in the case $\lambda < \mu$ (as $t \rightarrow \infty$).

- 2.16. Consider the equations of motion in Eq. (2.148) and define the Laplace transform

$$P_k^*(s) = \int_0^{\infty} P_k(t) e^{-st} dt$$

For our initial condition we will assume $p_{\text{out}}(t) = 1$ for $t = 0$. Transform Eq. (2.148) to obtain a set of linear difference equations in $\{P_k^*(s)\}$.

- (a) Show that the solution to the set of equations is

$$P_k^*(s) = \frac{\prod_{i=0}^{k-1} \lambda_i}{\prod_{i=0}^k (s + \lambda_i)}$$

- (b) From (a) find $p_{\text{out}}(t)$ for the case $\lambda_i = \lambda$ ($i = 0, 1, 2, \dots$).

- 2.17. Consider a time interval $(0, t)$ during which a Poisson process generates arrivals at an average rate λ . Derive Eq. (2.147) by considering the two events: exactly $k - 1$ arrivals occur in the interval $(0, t - \Delta t)$ and the event that exactly one arrival occurs in the interval $(t - \Delta t, t)$. Considering the limit as $\Delta t \rightarrow 0$ we immediately arrive at our desired result.
- 2.18. A barber opens up for business at $t = 0$. Customers arrive at random in a Poisson fashion; that is, the pdf of interarrival time is $a(t) = \lambda e^{-\lambda t}$. Each haircut takes X sec (where X is some random variable). Find the probability P that the second arriving customer will not have to wait and also find W , the average value of his waiting time for the two following cases:
- $X = c = \text{constant}$.
 - X is exponentially distributed with pdf:

$$b(x) = \mu e^{-\mu x}$$

- 2.19. At $t = 0$ customer A places a request for service and finds all m servers busy and n other customers waiting for service in an $M/M/m$ queueing system. All customers wait as long as necessary for service, waiting customers are served in order of arrival, and no new requests for service are permitted after $t = 0$. Service times are assumed to be mutually independent, identical, exponentially distributed random variables, each with mean duration $1/\mu$.
- Find the expected length of time customer A spends waiting for service in the queue.
 - Find the expected length of time from the arrival of customer A at $t = 0$ until the system becomes completely empty (all customers complete service).
 - Let X be the order of completion of service of customer A; that is, $X = k$ if A is the k th customer to complete service after $t = 0$. Find $P[X = k]$ ($k = 1, 2, \dots, m + n + 1$).
 - Find the probability that customer A completes service before the customer immediately ahead of him in the queue.
 - Let $\tilde{w}$ be the amount of time customer A waits for service. Find $P[\tilde{w} > x]$.
- 2.20. In this problem we wish to proceed from Eq. (2.162) to the transient solution in Eq. (2.163). Since $P^*(z, s)$ must converge in the region $|z| \leq 1$ for $\text{Re}(s) > 0$, then, in this region, the zeros of the denominator in Eq. (2.162) must also be zeros of the numerator.
- Find those two values of z that give the denominator zeros, and denote them by $\alpha_1(s)$, $\alpha_2(s)$ where $|\alpha_2(s)| < |\alpha_1(s)|$.

- (b) Using Rouché's theorem (see Appendix I) show that the denominator of $P^*(z, s)$ has a single zero within the unit disk $|z| \leq 1$.
- (c) Requiring that the numerator of $P^*(z, s)$ vanish at $z = \alpha_2(s)$ from our earlier considerations, find an explicit expression for $P_0^*(s)$.
- (d) Write $P^*(z, s)$ in terms of $\alpha_1(s) = \alpha_1$ and $\alpha_2(s) = \alpha_2$. Then show that this equation may be reduced to

$$P^*(z, s) = \frac{(z' + \alpha_2 z^{i-1} + \dots + \alpha_2^i) + \alpha_2^{i+1}/(1 - \alpha_2)}{\lambda \alpha_1 (1 - z/\alpha_1)}$$

- (e) Using the fact that $|\alpha_2| < 1$ and that $\alpha_1 \alpha_2 = \mu/\lambda$ show that the inversion on z yields the following expression for $P_k^*(s)$, which is the Laplace transform for our transient probabilities $P_k(t)$:

$$P_k^*(s) = \frac{1}{\lambda} \left[\alpha_1^{i-n-1} + \left(\frac{\mu}{\lambda}\right) \alpha_1^{i-n-3} + \left(\frac{\mu}{\lambda}\right)^2 \alpha_1^{i-n-5} + \dots \right. \\ \left. + \left(\frac{\mu}{\lambda}\right)^i \alpha_1^{i-n-1} + \left(\frac{\lambda}{\mu}\right)^{n+1} \sum_{k=n+i+2}^{\infty} \left(\frac{\mu}{\lambda \alpha_1}\right)^k \right]$$

- (f) In what follows we take advantage of property 4 in Table 1.3 and also we make use of the following transform pair:

$$kp k^a t^{-1} f_k(at) \Leftrightarrow \left[\frac{S + \sqrt{s^2 - \frac{4\lambda\mu}{2\lambda}}}{2\lambda} \right]^{-k}$$

where p and a are as defined in Eqs. (2.164), (2.165) and where $f_k(x)$ is the modified Bessel function of the first kind of order k as defined in Eq. (2.166). Using these facts and the simple relations among Bessel functions, namely,

$$\frac{2k}{x} I_k(x) = I_{k-1}(x) - I_{k+1}(x) \quad \text{and} \quad I_k(x) = I_{-k}(x)$$

show that Eq. (2.163) is the inverse transform for the expression shown in part (e).

- 2.21. The random variables $X_1, X_2, \dots, X_i, \dots$ are independent, identically distributed random variables each with density $f_X(x)$ and characteristic function $\phi_X(u) = E[e^{juX}]$. Consider a Poisson process $N(t)$ with parameter λ which is independent of the random variables X_i . Consider now a second random process of the form

$$X(t) = \sum_{i=1}^{N(t)} X_i$$

This second random process is clearly a family of staircase functions where the jumps occur at the discontinuities of the random process $N(t)$; the magnitudes of such jumps are given by the random variables X_i . Show that the characteristic function of this second random process is given by

$$\phi_{X(t)}(u) = e^{i u [\phi_X(u) - 1] t}$$

- 2.22. Passengers and taxis arrive at a service point from independent Poisson processes at rates λ , μ , respectively. Let the queue size at time t be q , a negative value denoting a line of taxis, a positive value denoting a queue of passengers. Show that, starting with $q_0 = 0$, the distribution of q , is given by the difference between independent Poisson variables of means λt , μt . Show by using the normal approximation that if $\lambda = \mu$, the probability that $-k \leq q \leq k$ is, for large t , $(2k + 1)(4\pi\lambda t)^{-1/2}$.

PART II

ELEMENTARY QUEUEING THEORY

Elementary here means that all the systems we consider are pure Markovian and, therefore, our state description is convenient and manageable. In Part I we developed the time-dependent equations for the behavior of birth-death processes; here in Chapter 3 we address the equilibrium solution for these systems. The key equation in this chapter is Eq. (3.11), and the balance of the material is the simple application of that formula. It, in fact, is no more than the solution to the equation $\boldsymbol{\pi} = \boldsymbol{\pi}\mathbf{P}$ derived in Chapter 2. The key tool used here is again that which we find throughout the text, namely, the calculation of flow rates across the boundaries of a closed system. In the case of equilibrium we merely ask that the rate of flow into be equal to the rate of flow out of a system. The application of these basic results is more than just an exercise for it is here that we first obtain some equations of use in engineering and designing queueing systems. The classical *M/M/1* queue is studied and some of its important performance measures are evaluated. More complex models involving finite storage, multiple servers, finite customer population, and the like, are developed in the balance of this chapter. In Chapter 4 we leave the birth-death systems and allow more general Markovian queues, once again to be studied in equilibrium. We find that the techniques here are similar to our earlier ones, but find that no general solution such as Eq. (3.11) is available; each system is a case unto itself and so we are rapidly led into the solutions of difference equations, which force us to look carefully at the method of z-transforms for these solutions. The ingenious method of stages introduced by Erlang is considered here and its generality discussed. At the end of the chapter we introduce (for later use in Volume II) networks of Markovian queues in which we take exquisite advantage of the memoryless properties that Markovian queues provide even in a network environment. At this point, however, we have essentially exhausted the use of the memoryless distribution and we must depart from that crutch in the following parts.

Birth-Death Queueing Systems in Equilibrium

In the previous chapter we studied a variety of stochastic processes. We indicated that Markov processes play a fundamental role in the study of queueing systems, and after presenting the main results from that theory, we then considered a special form of Markov process known as the birth-death process. We also showed that birth-death processes enjoy a most convenient property, namely, that the time between births and the time between deaths (when the system is nonempty) are each exponentially distributed.* We then developed Eq. (2.127), which gives the basic equations of motion for the general birth-death process with stationary birth and death rates.] The solution of this set of equations gives the transient behavior of the queueing process and some important special cases were discussed earlier. In this chapter we study the limiting form of these equations to obtain the equilibrium behavior of birth-death queueing systems.

The importance of elementary queueing theory comes from its historical influence as well as its ability to describe behavior that is to be found in more complex queueing systems. The methods of analysis to be used in this chapter in large part do *not* carryover to the more involved queueing situations; nevertheless, the obtained results *do* provide insight into the basic behavior of many of these other queueing systems.

It is necessary to keep in mind how the birth-death process describes queueing systems. As an example, consider a doctor's office made up of a waiting room (in which a queue is allowed to form, unfortunately) and a service facility consisting of the doctor's examination room. Each time a patient enters the waiting room from outside the office we consider this to be an *arrival* to the queueing system; on the other hand, this arrival may well be considered to be a *birth* of a new member of a population, where the population consists of all patients present. In a similar fashion, when a patient leaves

* This comes directly from the fact that they are Markov processes.

† In addition to these equations, one requires the conservation relation given in Eq. (2.122) and a set of initial conditions $\{P_k(0)\}$.

the office after being treated, he is considered to be a *departure* from the queueing system; in terms of a birth-death process this is considered to be a *death* of a member of the population.

We have considerable freedom in constructing a large number of queueing systems through the choice of the birth coefficients λ_k and death coefficients μ_k , as we shall see shortly. First, let us establish the general solution for the equilibrium behavior.

3.1. GENERAL EQUILIBRIUM SOLUTION

As we saw in Chapter 2 the time-dependent solution of the birth-death system quickly becomes unmanageable when we consider any sophisticated set of birth-death coefficients. Furthermore, were we always capable of solving for $P_k(t)$ it is not clear how useful that set of functions would be in aiding our understanding of the behavior of these queueing systems (too much information is sometimes a curse!). Consequently, it is natural for us to ask whether the probabilities $P_k(t)$ eventually settle down as t gets large and display no more "transient" behavior. This inquiry on our part is analogous to the questions we asked regarding the existence of π_k in the limit of $\pi_k(t)$ as $t \rightarrow \infty$. For our queueing studies here we choose to denote the limiting probability as P_k rather than π_k , purely for convenience. Accordingly, let

$$P_k \triangleq \lim_{t \rightarrow \infty} P_k(t) \quad (3.1)$$

where P_k is interpreted as the limiting probability that the system contains k members (or equivalently is in state E_k) at some arbitrary time in the distant future. The question regarding the existence of these limiting probabilities is of concern to us, but will be deferred at this point until we obtain the general steady-state or limiting solution. It is important to understand that whereas P_k (assuming it exists) is no longer a function of t , we are not claiming that the process does not move from state to state in this limiting case; certainly, the number of members in the population will change with time, but the *long-run* probability of finding the system with k members will be properly described by P_k .

Accepting the existence of the limit in Eq. (3.1), we may then set $\lim_{t \rightarrow \infty} dP_k(t)/dt$ as $t \rightarrow \infty$ equal to zero in the Kolmogorov forward equations (of motion) for the birth-death system [given in Eqs. (2.127)] and immediately obtain the result

$$0 = -(\lambda_k + \mu_k)P_k + \lambda_{k-1}P_{k-1} + \mu_{k+1}P_{k+1} \quad k \geq 1 \quad (3.2)$$

$$0 = -\lambda_0 P_0 + \mu_1 P_1 \quad k = 0 \quad (3.3)$$

The annoying task of providing a separate equation for $k = 0$ may be overcome by agreeing once and for all that the following birth and death

coefficients are identically equal to 0:

$$\begin{aligned}\lambda_{-1} &= \lambda_{-2} = \lambda_{-3} = &= 0 \\ \mu_0 &= \mu_{-1} = \mu_{-2} = &= 0\end{aligned}$$

Furthermore, since it is perfectly clear that we cannot have a negative number of members in our population, we will, in most cases, adopt the convention that

$$P_{-1} = p_{-2} = P_{-3} = \dots = 0$$

Thus, for all values of k , we may reformulate Eqs. (3.2) and (3.3) into the following set of difference equations for $k = \dots, -2, -1, 0, 1, 2, \dots$

$$0 = -(\lambda_k + \mu_k)p_k + \lambda_{k-1}p_{k-1} + \mu_{k+1}p_{k+1} \quad (3.4)$$

We also require the conservation relation

$$\sum_{k=0}^{\infty} p_k = 1 \quad (3.5)$$

Recall from the previous chapter that the limit given in the Eq. (3.1) is independent of the initial conditions.

Just as we used the state-transition-rate diagram as an inspection technique for writing down the equations of motion in Chapter 2, so may we use the same concept in writing down the *equilibrium* equations [Eqs. (3.2) and (3.3)] directly from that diagram. In this equilibrium case it is clear that flow must be *conserved* in the sense that the input flow must equal the output flow from a given state. For example, if we look at Figure 2.9 once again and concentrate on state E_k in equilibrium, we observe that

$$\text{Flow rate into } E_k = \lambda_{k-1}p_{k-1} + \mu_{k+1}p_{k+1}$$

and

$$\text{Flow rate out of } E_k = (\lambda_k + \mu_k)p_k$$

In equilibrium these two must be the same and so we have immediately

$$\lambda_{k-1}p_{k-1} + \mu_{k+1}p_{k+1} = (\lambda_k + \mu_k)p_k \quad (3.6)$$

But this last is just Eq. (3.4) again! *By inspection we have established the equilibrium difference equations for our system.* The same comments apply here as applied earlier regarding the conservation of flow across *any* closed boundary; for example, rather than surrounding each state and writing down its equation we could choose a sequence of boundaries the first of which surrounds E_0 , the second of which surrounds E_0 and E_1 and so on, each time adding the next higher-numbered state to get a new "boundary". In such an example the k th boundary (which surrounds states $E_0, E_1, \dots, E_{k-1}$) would

lead to the following simple conservation of flow relationship:

$$\lambda_k P_k = \mu_k P_{k-1} \quad (3.7)$$

This last set of equations is equivalent to drawing a vertical line separating adjacent states and equating flows across this boundary; this set of difference equations is equivalent to our earlier set.

The solution for P_k in Eq. (3.4) may be obtained by at least two methods. One way is first to solve for P_1 in terms of P_0 by considering the case $k = 1$. that is.

$$P_1 = \frac{\lambda_0}{\mu_1} P_0 \quad (3.8)$$

We may then consider Eq. (3.4) for the case $k = 2$ and using Eq. (3.8) obtain

$$0 = -(\lambda_1 + \mu_1)P_1 + \lambda_0 P_0 + \mu_2 P_2$$

$$0 = -(\lambda_1 + \mu_1) \frac{\lambda_0}{\mu_1} P_0 + \lambda_0 P_0 + \mu_2 P_2$$

$$0 = -\frac{\lambda_1 \lambda_0}{\mu_1} P_0 - \lambda_0 P_0 + \lambda_0 P_0 + \mu_2 P_2$$

and so

$$P_2 = \frac{\lambda_0 \lambda_1}{\mu_1 \mu_2} P_0 \quad (3.9)$$

If we examine Eqs. (3.8) and (3.9) we may justifiably guess that the general solution to Eq. (3.4) must be

$$\frac{\lambda_0 \lambda_1 \cdots \lambda_{k-1}}{\mu_1 \mu_2 \cdots \mu_k} P_0 \quad (3.10)$$

To validate this assertion we need merely use the inductive argument and apply Eq. (3.10) to Eq. (3.4) solving for P_k . Carrying out this operation we do, in fact, find that (3.10) is the solution to the general birth-death process in this steady-state or limiting case. We have thus expressed all equilibrium probabilities P_k in terms of a single unknown constant P_0 :

$$P_k = P_0 \prod_{i=0}^{k-1} \frac{\lambda_i}{\mu_{i+1}} \quad k=0, 1, 2, \dots \quad (3.11)$$

(Recall the usual convention that an empty product is unity by definition.) Equation (3.5) provides the additional condition that allows us to determine P_0 ; thus, summing over all k , we obtain

$$1 + \sum_{k=1}^{\infty} \prod_{i=0}^{k-1} \frac{\lambda_i}{\mu_{i+1}} = 1 \quad (3.12)$$

This "product" solution for $P_{\gg}(k = 0, 1, 2, \dots)$ simply obtained, is a *principal* equation in elementary queueing theory and, in fact, is the point of departure for all of our further solutions in this chapter.

A second easy way to obtain the solution to Eq. (3.4) is to rewrite that equation as follows:

$$\lambda_{k-1}p_{k-1} - \mu_k p_k = \lambda_k p_k - \mu_{k+1} p_{k+1} \quad (3.13)$$

Defining

$$g_k = \lambda_k p_k - \mu_{k+1} p_{k+1} \quad (3.14)$$

we have from Eq. (3.13) that

$$g_{k-1} = g_k \quad (3.15)$$

Clearly Eq. (3.15) implies that

$$gk = \text{constant with respect to } k \quad (3.16)$$

However, since $\lambda_{-1} = \mu_0 = 0$, Eq. (3.14) gives

$$g_{-1} = 0$$

and so the constant in Eq. (3.16) must be 0. Setting gk equal to 0, we immediately obtain from Eq. (3.14)

$$P_{HI} = \frac{\lambda_k}{\mu_{k+1}} P_k \quad (3.17)$$

Solving Eq. (3.17) successively beginning with $k = 0$ we obtain the earlier solution, namely, Eqs. (3.11) and (3.12).

We now address ourselves to the *existence* of the steady-state probabilities $P_{\gg}$ given by Eqs. (3.11) and (3.12). Simply stated, in order for those expressions to represent a probability distribution, we usually require that $P_{\gg} > 0$. This clearly places a condition upon the birth and death coefficients in those equations. Essentially, what we are requiring is that the system occasionally empties; that this is a condition for stability seems quite reasonable when one interprets it in terms of real life situations.* More precisely, we may classify the possibilities by first defining the two sums

$$S_1 \triangleq \sum_{k=0}^{\infty} \prod_{i=0}^{k-1} \frac{\lambda_i}{\mu_{i+1}} \quad (3.18)$$

$$S_2 \triangleq \sum_{k=0}^{\infty} \left(1 / \left(\lambda_k \prod_{i=0}^{k-1} \frac{\lambda_i}{\mu_{i+1}} \right) \right) \quad (3.19)$$

* It is easy to construct counterexamples to this case, and so we require the precise arguments which follow.

All states E_k of our birth-death process will be *ergodic* if and only if

$$\begin{aligned} \text{Ergodic:} \quad 5, &< \infty \\ 5_2 &= \infty \end{aligned}$$

On the other hand, all states will be *recurrent null* if and only if

$$\begin{aligned} \text{Recurrent null:} \quad 5, &= \infty \\ 5_2 &= \infty \end{aligned}$$

Also, all states will be *transient* if and only if

$$\begin{aligned} \text{Transient:} \quad 5, &= \infty \\ 5_2 &< \infty \end{aligned}$$

It is the ergodic case that gives rise to the equilibrium probabilities $\{p_k\}$ and that is of most interest to our studies. We note that the condition for ergodicity is met whenever the sequence $\{\lambda_k/\mu_k\}$ remains below unity from some k onwards, that is, if there exists some k_0 such that for all $k \geq k_0$ we have

$$\frac{\lambda_k}{\mu_k} < 1 \quad (3.20)$$

We will find this to be true in most of the queueing systems we study.

We are now **ready** to apply our general solution as given in Eqs. (3.11) and (3.12) to some very important special cases. Before we launch headlong into that discussion, let us put at ease those readers who feel that the birth-death constraints of permitting only nearest-neighbor transitions are too confining. It is true that the solution given in Eqs. (3.11) and (3.12) applies only to nearest-neighbor birth-death processes. However, rest assured that the equilibrium methods we have described can be extended to more general than nearest-neighbor systems; these generalizations are considered in Chapter 4.

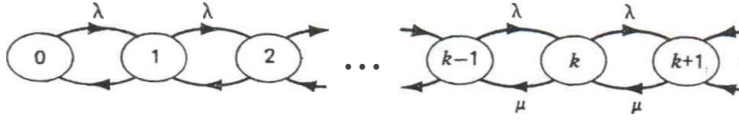
3.2. M/M/1: THE CLASSICAL QUEUEING SYSTEM

As mentioned in Chapter 2, the celebrated *MMI* queue is the simplest nontrivial interesting system and may be described by selecting the birth-death coefficients as follows:

$$\begin{aligned} \lambda_k &= \lambda & k = 0, 1, 2, \dots \\ \mu_k &= \mu & k = 1, 2, 3, \dots \end{aligned}$$

That is, we set all birth* coefficients equal to a constant λ and all death*

* In this case, the average interarrival time is $\bar{t} = 1/\lambda$ and the average service time is $\bar{x} = 1/\mu$; this follows since $\bar{t}$ and $\bar{x}$ are both exponentially distributed.

Figure 3.1 State-transition-rate diagram for $M/M/1$.

coefficients equal to a constant μ . We further assume that infinite queueing space is provided and that customers are served in a first-come-first-served fashion (although this last is not necessary for many of our results). For this important example the state-transition-rate diagram is as given in Figure 3.1.

Applying these coefficients to Eq. (3.11) we have

$$p_k = p_0 \prod_{i=0}^{k-1} \frac{\lambda}{\mu}$$

or

$$p_k = p_0 \left(\frac{\lambda}{\mu} \right)^k \quad (3.21)$$

The result is immediate. The conditions for our system to be ergodic (and, therefore, to have an equilibrium solution $P \gg 0$) are that $S_1 < \infty$ and $S_2 = \infty$; in this case the first condition becomes

$$S_1 = \sum_{k=0}^{\infty} \frac{p_k}{p_0} = \sum_{k=0}^{\infty} \left(\frac{\lambda}{\mu} \right)^k < \infty$$

The series on the left-hand side of the inequality will converge if and only if $\lambda/\mu < 1$. The second condition for ergodicity becomes

$$S_2 = \sum_{k=0}^{\infty} \frac{1}{\lambda(p_k/p_0)} = \sum_{k=0}^{\infty} \frac{1}{\lambda} \left(\frac{\mu}{\lambda} \right)^k = \infty$$

This last condition will be satisfied if $\lambda/\mu \leq 1$; thus the necessary and sufficient condition for ergodicity in the $M/M/1$ queue is simply $\lambda < \mu$. In order to solve for p_0 we use Eq. (3.12) [or Eq. (3.5) as suits the reader] and obtain

$$p_0 = 1 / \left[1 + \sum_{k=1}^{\infty} \left(\frac{\lambda}{\mu} \right)^k \right]$$

The sum converges since $\lambda < \mu$ and so

$$1 + \frac{\lambda/\mu}{1 - \lambda/\mu}$$

Thus

$$p_0 = 1 - \frac{\lambda}{\mu} \quad (3.22)$$

From Eq. (2.29) we have $p = \lambda/\mu$. From our stability conditions, we therefore require that $0 \leq p < 1$; note that this insures that $p_0 > 0$. From Eq. (3.21) we have, finally,

$$p_k = (1 - p)p^k \quad k = 0, 1, 2, \dots \quad (3.23)$$

Equation (3.23) is indeed the solution for the steady-state probability of finding k customers in the system.* We make the important observation that p_k depends upon λ and μ only through their ratio p .

The solution given by Eq. (3.23) for this fundamental system is graphed in Figure 3.2 for the case of $p = 1/2$. Clearly, this is the geometric distribution (which shares the fundamental memoryless property with the exponential distribution). As we develop the behavior of the MIMII queue, we shall continue to see that almost all of its important probability distributions are of the memoryless type.

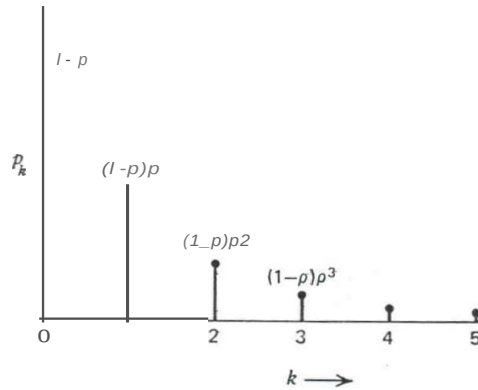
An important measure of a queueing system is the average number of customers in the system $\bar{N}$. This is clearly given by

$$\begin{aligned} \bar{N} &= \sum_{k=0}^{\infty} k p_k \\ &= (1 - p) \sum_{k=0}^{\infty} k p^k \end{aligned}$$

Using the trick similar to the one used in deriving Eq. (2.142) we have

$$\begin{aligned} \bar{N} &= (1 - p) p \frac{\partial}{\partial p} \sum_{k=0}^{\infty} p^k \\ &= (1 - p) p \frac{\partial}{\partial p} \left(\frac{1}{1 - p} \right) \\ \bar{N} &= \frac{p}{1 - p} \quad (3.24) \end{aligned}$$

* If we inspect the transient solution for M/M/1 given in Eq. (2.163), we see the term $(1 - p)p^k$; the reader may verify that, for $p < 1$, the limit of the transient solution agrees with our solution here.

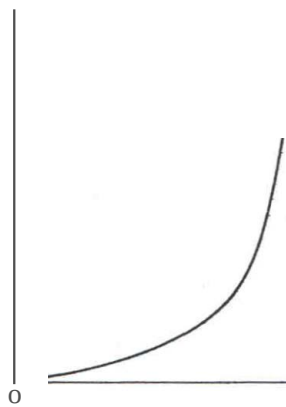
Figure 3.2 The solution for P_k in the system *MIM/1*.

The behavior of the expected number in the system is plotted in Figure 3.3. By similar methods we find that the variance of the number in the system is given by

$$\sigma_N^2 = \sum_{k=0}^{\infty} (k - \bar{N})^2 p_k$$

$$\sigma_N^2 = \frac{\rho}{(1 - \rho)^2} \quad - (3.25)$$

We may now apply Little's result directly from Eq. (2.25) in order to obtain

Figure 3.3 The average number in the system *MIM/1*.

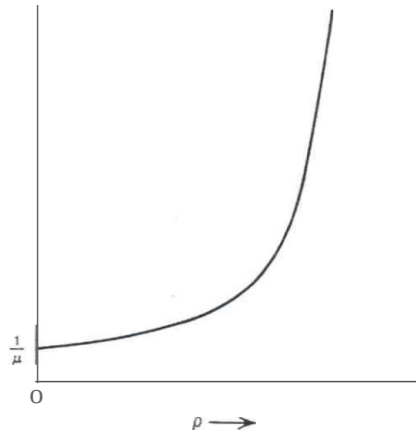


Figure 3.4 Average delay as a function of ρ for $MIM/1$.

T , the average *time* spent in the system as follows:

$$\begin{aligned}
 T &= \frac{\bar{N}}{\lambda} \\
 T &= \left(\frac{\rho}{1-\rho} \right) \left(\frac{1}{\lambda} \right) \\
 T &= \frac{1/\mu}{1-\rho} \quad - (3.26)
 \end{aligned}$$

This dependence of average time on the utilization factor ρ is shown in Figure 3.4. The value obtained by T when $\rho = 0$ is exactly the average service time expected by a customer; that is, he spends no time in queue and $1/\mu$ sec in service on the average.

The behavior given by Eqs. (3.24) and (3.26) is rather dramatic. As ρ approaches unity, both the average number in the system and the average time in the system grow in an unbounded fashion.* Both these quantities have a

* We observe at $\rho = 1$ that the system behavior is unstable; this is not surprising if one recalls that $\rho < 1$ was our condition for ergodicity. What is perhaps surprising is that the behavior of the average number $\bar{N}$ and of the average system time T deteriorates so badly as $\rho \rightarrow 1$ from below; we had seen for steady flow systems in Chapter I that so long as $R < C$ (which corresponds to the case $\rho < 1$) no queue formed and smooth, rapid flow proceeded through the system. Here in the $MIM/1$ queue we find this is no longer true and that we pay an extreme penalty when we attempt to run the system near (but below) its capacity. The

simple pole at $p = 1$. This type of behavior with respect to p as p approaches 1 is characteristic of almost every queueing system one can encounter. We will see it again in *M/G/1* in Chapter 5 as well as in the heavy traffic behavior of *G/G/1* (and also in the tight bounds on *G/G/1* behavior) in Volume II, Chapter 2.

Another interesting quantity to calculate is the probability of finding at least k customers in the system:

$$\begin{aligned} P[\geq k \text{ in system}] &= \sum_{i=k}^{\infty} P_i \\ &= \sum_{i=k}^{\infty} (1-p)^i \\ P[\geq k \text{ in system}] &= p^k \end{aligned} \quad - (3.27)$$

Thus we see that the probability of exceeding some limit on the number of customers in the system is a geometrically decreasing function of that number and decays very rapidly.

With the tools at hand we are now in a position to develop the probability density function for the time spent in the system. However, we defer that development until we treat the more general case of *M/G/1* in Chapter 5 [see Eq. (5.118)]. Meanwhile, we proceed to discuss numerous other birth-death queues in equilibrium.

3.3. DISCOURAGED ARRIVALS

This next example considers a case where arrivals tend to get discouraged when more and more people are present in the system. One possible way to model this effect is to choose the birth and death coefficients as follows:

$$\begin{aligned} \lambda_k &= \frac{\alpha}{k+1} & k &= 0, 1, 2, \dots \\ \mu_k &= \mu & k &= 1, 2, 3, \dots \end{aligned}$$

We are here assuming an harmonic discouragement of arrivals with respect to the number present in the system. The state-transition-rate diagram in this

intuitive explanation here is that with random flow (e.g., *M/M/1*) we get occasional bursts of traffic which temporarily overwhelm the server; while it is still true that the server will be idle on the average $1-p = P_0$ of the time this average idle time will not be distributed uniformly within small time intervals but will only be true in the long run. On the other hand, in the steady flow case (which corresponds to our system *D/D/1*) the system idle time will be distributed quite uniformly in the sense that after every service time (of exactly $1/\mu$ sec) there will be an idle time of exactly $(1/\lambda) - (1/\mu)$ sec. Thus it is the variability in both the interarrival time and in the service time which gives rise to the disastrous behavior near $p = 1$; any reduction in the variation of either random variable will lead to a reduction in the average waiting time, as we shall see again and again.

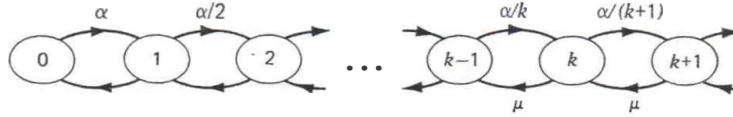


Figure 3.5 State-transition-rate diagram for discouraged arrivals.

case is as shown in Figure 3.5. We apply Eq. (3.11) immediately to obtain ρ ;

$$P_k = P_0 \prod_{i=0}^{k-1} \frac{\alpha/(i+1)}{\mu} \quad (3.28)$$

$$P_k = P_0 \left(\frac{\alpha}{\mu} \right)^k \frac{1}{k!} \quad (3.29)$$

Solving for P_0 from Eq. (3.12) we have

$$P_0 = \frac{1}{\left[1 + \sum_{k=1}^{\infty} \left(\frac{\alpha}{\mu} \right)^k \frac{1}{k!} \right]}$$

$$P_0 = e^{-\alpha/\mu}$$

From Eq. (2.32) we have therefore,

$$\rho = 1 - e^{-\alpha/\mu} \quad (3.30)$$

Note that the ergodic condition here is merely $\alpha/\mu < \infty$. Going back to Eq. (3.29) we have the final solution

$$P_k = \frac{(\alpha/\mu)^k}{k!} e^{-\alpha/\mu} \quad k = 0, 1, 2, \dots \quad (3.31)$$

We thus have a Poisson distribution for the number of customers in the system of discouraged arrivals! From Eqs. (2.131) and (2.132) we have that the expected number in the system is

$$\bar{N} = \frac{\alpha}{\mu}$$

In order to calculate T , the average time spent in the system, we may use Little's result again. For this we require λ , which is directly calculated from $\rho = \lambda \bar{x} = \lambda/\mu$; thus from Eq. (3.30)

$$\rho = \mu \rho = \mu(1 - e^{-\alpha/\mu})$$

Using this* and Little's result we then obtain

$$T = \frac{\alpha}{\mu^2(1 - e^{-\alpha/\mu})} \quad (3.32)$$

* Note that this result could have been obtained from $\lambda = \sum_k \lambda_k P_k$. The reader should verify this last calculation.

3.4. $M/M/\infty$: RESPONSIVE SERVERS (INFINITE NUMBER OF SERVERS)

Here we consider the case that may be interpreted either as that of a responsive **server** who accelerates her service rate linearly when more customers are waiting or may be interpreted as the case where there is always a new clerk or server available for each arriving customer. In particular, we set

$$\begin{aligned}\lambda_k &= \lambda & k &= 0, 1, 2, \dots \\ \mu_k &= k\mu & k &= 1, 2, 3, \dots\end{aligned}$$

Here the state-transition-rate diagram is that shown in Figure 3.6. Going directly to Eq. (3.11) for the solution we obtain

$$P_k = P_0 \prod_{i=0}^{k-1} \left(\frac{\lambda}{i\mu} + 1 \right) \quad (3.33)$$

Need we go any further? The reader should compare Eq. (3.33) with Eq. (3.28). These two are in fact equivalent for $\alpha = i$, and so we immediately have the solutions for P_k and $\bar{N}$,

$$P_k = \frac{(\lambda/\mu)^k}{k!} e^{-\lambda/\mu} \quad k = 0, 1, 2, \dots \quad (3.34)$$

$$\bar{N} = \frac{\lambda}{\mu}$$

Here, too, the ergodic condition is simply $\lambda/\mu < \infty$. It appears then that a system of discouraged arrivals behaves exactly the same as a system that includes a responsive server. However, Little's result provides a different (and simpler) form for T here than that given in Eq. (3.32); thus

$$T = \frac{1}{\mu}$$

This answer is, of course, obvious since if we use the interpretation where each arriving customer is granted his own server, then his time in system will be merely his service time which clearly equals $1/\mu$ on the average.

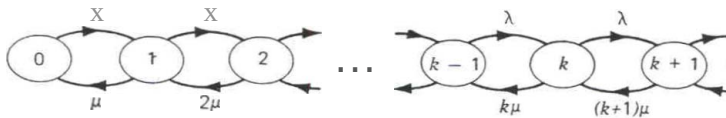


Figure 3.6 State-transition-rate diagram for the infinite-server case $M/M/\infty$.

3.5. M/M/m: THE m-SERVER CASE

Here again we consider a system with an unlimited waiting room and with a constant arrival rate λ . The system provides for a *maximum* of m servers. This is within the reach of our birth-death formulation and leads to

$$\begin{aligned}\lambda_k &= \lambda \quad k = 0, 1, 2, \dots \\ \mu_k &= \min [k\mu, m\mu] \\ &= \begin{cases} k\mu & 0 \leq k \leq m \\ m\mu & m \leq k \end{cases}\end{aligned}$$

From Eq. (3.20) it is easily seen that the condition for ergodicity is $\lambda/m\mu < 1$. The state-transition-rate diagram is shown in Figure 3.7. When we go to solve for P_k from Eq. (3.11) we find that we must separate the solution into two parts, since the dependence of μ_k upon k is also in two parts. Accordingly, for $k \leq m$,

$$\begin{aligned}P_k &= P_0 \prod_{i=0}^{k-1} \left(\frac{\lambda}{i+1} \frac{1}{\mu} \right) \\ &= \frac{P_0 \left(\frac{\lambda}{\mu} \right)^k}{k!}\end{aligned}\quad (3.35)$$

Similarly, for $k \geq m$,

$$\begin{aligned}P_k &= P_0 \prod_{i=0}^{m-1} \left(\frac{\lambda}{i+1} \frac{1}{\mu} \right) \prod_{j=m}^{k-1} \frac{\lambda}{m\mu} \\ &= P_0 \left(\frac{\lambda}{\mu} \right)^k \frac{1}{m! m^{k-m}}\end{aligned}\quad (3.36)$$

Collecting together the results from Eqs. (3.35) and (3.36) we have

$$P_k = \begin{cases} \frac{P_0 (mp)^k}{k!} \\ \frac{P_0 (p)^{kmm}}{m!} \end{cases} \quad k \geq m \quad (3.37)$$

where

$$p = \frac{\lambda}{m\mu} < 1 \quad (3.38)$$

This expression for p follows that in Eq. (2.30) and is consistent with our

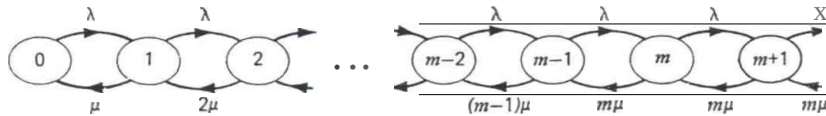


Figure 3.7 State-transition-rate diagram for M/M/m.

definition in terms of the expected fraction of busy servers. We may now solve for P_0 from Eq. (3.12), which gives us

$$P_0 = \left[1 + \sum_{k=1}^{m-1} \frac{(mp)^k}{k!} + \sum_{k=m}^{\infty} \frac{(mp)^k}{m!} \right]^{-1}$$

and so

$$P_0 = \left[\sum_{k=0}^{m-1} \frac{(mp)^k}{k!} + \frac{(mp)^m}{m!} \left(\frac{1}{1-p} \right) \right]^{-1} \quad (3.39)$$

The probability that an arriving customer is forced to join the queue is given by

$$\begin{aligned} P[\text{queueing}] &= \sum_{k=m}^{\infty} P_k \\ &= \sum_{k=m}^{\infty} P_0 \frac{(mp)^k}{m!} \frac{1}{m^{k-m}} \end{aligned}$$

Thus

$$P[\text{queueing}] = \left[\sum_{k=0}^{m-1} \frac{(mp)^k}{k!} + \frac{(mp)^m}{m!} \left(\frac{1}{1-p} \right) \right]^{-1} \left(\frac{(mp)^m}{m!} \right) \left(\frac{1}{1-p} \right) \quad (3.40)$$

This probability is of wide use in telephony and gives the probability that no trunk (i.e., server) is available for an arriving call (customer) in a system of m trunks; it is referred to as *Erlang's C formula* and is often denoted* by $C(m, \lambda/\mu)$.

3.6. M/M/1/K: FINITE STORAGE

We now consider for the first time the case of a queueing system in which there is a maximum number of customers that may be stored; in particular, we assume the system can hold at most a total of K customers (including the customer in service) and that any further arriving customers will in fact be refused entry to the system and will depart immediately without service. Newly arriving customers will continue to be generated according to a Poisson process but only those who find the system with strictly less than K customers will be allowed entry. In telephony the refused customers are considered to be "lost"; for the system in which $K = 1$ (i.e., no waiting room at all) this is referred to as a "blocked calls cleared" system with a single server.

* Europeans use the symbol $E_{2,m}(\lambda/\mu)$.

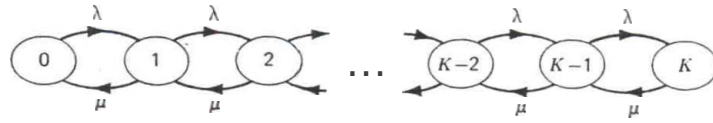


Figure 3.8 State-transition-rate diagram for the case of finite storage room M/M/1/K.

It is interesting that we are capable of accommodating this seemingly complex system description with our birth-death model. In particular, we accomplish this by effectively "turning off" the Poisson input as soon as the systems fills up, as follows:

$$\lambda_k = \begin{cases} \lambda & k < K \\ 0 & k \geq K \end{cases}$$

$$\mu_k = \mu \quad k = 1, 2, \dots, K$$

From Eq. (3.20), we see that this system is always ergodic. The state-transition-rate diagram for this finite Markov chain is shown in Figure 3.8. Proceeding directly with Eq. (3.11) we obtain

$$P_k = P_0 \prod_{i=0}^{k-1} \frac{\lambda}{\mu} \quad k \leq K$$

or

$$P_k = P_0 \left(\frac{\lambda}{\mu} \right)^k \quad (3.41)$$

Of course, we also have

$$P_k = 0 \quad k > K \quad (3.42)$$

In order to solve for P_0 we use Eqs. (3.41) and (3.42) in Eq. (3.12) to obtain

$$P_0 = \left[1 + \sum_{k=1}^K \left(\frac{\lambda}{\mu} \right)^k \right]^{-1}$$

$$= \left[1 + \frac{(\lambda/\mu)(1 - (\lambda/\mu)^K)}{1 - \lambda/\mu} \right]^{-1}$$

and so

$$P_0 = \frac{1 - \lambda/\mu}{1 - (\lambda/\mu)^{K+1}}$$

Thus, finally,

$$P_k = \begin{cases} \frac{1 - \lambda/\mu}{1 - (\lambda/\mu)^{K+1}} \left(\frac{\lambda}{\mu} \right)^k & 0 \leq k \leq K \\ 0 & \text{otherwise} \end{cases} \quad (3.43)$$

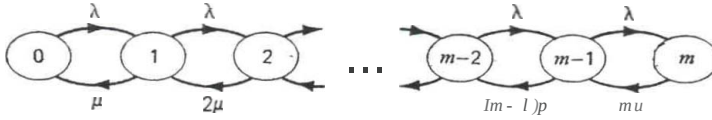


Figure 3.9 State-transition-rate diagram for m-server loss system M/M/m/m.

For the case of blocked calls cleared ($K = 1$) we have

$$P_k = \begin{cases} \frac{1}{1 + \lambda/\mu} & k=0 \\ \frac{\lambda/\mu}{1 + \lambda/\mu} & k = 1 = K \\ 0 & \text{otherwise} \end{cases} \quad (3.44)$$

3.7. $MjM/m/m$: m-SERVER LOSS SYSTEMS

Here we have again a blocked calls cleared situation in which there are available m servers. Each newly arriving customer is given his private server; however, if a customer arrives when all servers are occupied, that customer is lost. We create this artifact as above by choosing the following birth and death coefficients:

$$\lambda_k = \begin{cases} \lambda & k < m \\ 0 & k \geq m \end{cases}$$

$$\mu_k = k\mu \quad k = 1, 2, \dots, m$$

Here again, ergodicity is always assured. This finite state-transition-rate diagram is shown in Figure 3.9.

Applying Eq. (3.11) we obtain

$$P_k = P_0 \prod_{i=0}^{k-1} \left(\frac{\lambda}{i\mu} + 1 \right) \mu \quad k \leq m$$

or

$$P_k = \frac{P_0 \left(\frac{\lambda}{\mu} \right)^k}{k!} \quad k \leq m \quad (3.45)$$

Solving for P_0 we have

$$P_0 = \frac{1}{\sum_{k=0}^m \left(\frac{\lambda}{\mu} \right)^k \frac{1}{k!}}$$

This particular system is of great interest to those in telephony [so much so that a special case of Eq. (3.45) has been tabulated and graphed in many books

on telephony]. Specifically, P_m describes the fraction of time that all m servers are busy. The name given to this probability expression is *Erlang's loss formula* and it is given by

$$P_m = \frac{(\lambda/\mu)^m / m!}{\sum_{k=0}^m (\lambda/\mu)^k / k!} \quad - (3.46)$$

This equation is also referred to as *Erlang's B formula* and is commonly denoted* by $B(m, \lambda/\mu)$. Formula (3.46) was first derived by Erlang in 1917!

3.8. M/M/I//Mt : FINITE CUSTOMER POPULATION- SINGLE SERVER

Here we consider the case where we no longer have a Poisson input process with an infinite user population, but rather have a *finite* population of possible users. The system structure is such that we have a total of M users; a customer is either in the system (consisting of a queue and a single server) or outside the system and in some sense "arriving." In particular, when a customer is in the "arriving" condition then the time it takes him to arrive is a random variable with an exponential distribution whose mean is $1/\lambda$ sec. All customers act independently of each other. As a result, when there are k customers in the system (queue plus service) then there are $M - k$ customers in the arriving state and, therefore, the total average arrival rate in this state is $\lambda(M - k)$. We see that this system is in a strong sense self-regulating. By this we mean that when the system gets BUSY, with many of these customers in the queue, then the rate at which additional customers arrive is in fact reduced, thus lowering the further congestion of the system. We model this quite appropriately with our birth-death process choosing for parameters

$$\lambda_k = \begin{cases} \lambda(M - k) & \text{OS } k \leq M \\ 0 & \text{otherwise} \end{cases}$$

$$\mu_k = \mu \quad k = 1, 2, \dots$$

The system is ergodic. We assume that we have sufficient room to contain M customers in the system. The finite state-transition-rate diagram is shown in Figure 3.10. Using Eq. (3.11) we solve for p_k as follows:

$$p_k = p_0 \prod_{i=0}^{k-1} \frac{\lambda(M - i)}{\mu} \quad 0 \leq k \leq M$$

* Europeans use the notation $E1_m(\lambda/\mu)$.

† Recall that a blank entry in either of the last two optional positions in this notation means an entry of ∞ ; thus here we have the system M/M/I/oo/M.

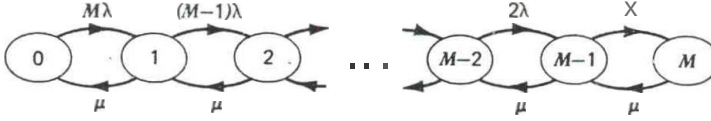


Figure 3.10 State-transition-rate diagram for single-server finite population system $M/M/I/M$.

Thus

$$p_k = \begin{cases} p_0 \left(\frac{\lambda}{\mu}\right)^k \frac{M!}{(M-k)!} & 0 \leq k \leq M \\ 0 & k > M \end{cases} \quad - (3.47)$$

In addition, we obtain for p_0

$$p_0 = \left[\sum_{k=0}^M \left(\frac{\lambda}{\mu}\right)^k \frac{M!}{(M-k)!} \right]^{-1} \quad - (3.48)$$

3.9. $M/M/\infty/M$: FINITE CUSTOMER POPULATION- "INFINITE" NUMBER OF SERVERS

We again consider the finite population case, but now provide a separate server for each customer in the system. We model this as follows:

$$\lambda_k = \begin{cases} \lambda(M-k) & 0 \leq k \leq M \\ 0 & \text{otherwise} \end{cases}$$

$$\mu_k = k\mu \quad k = 1, 2, \dots$$

Clearly, this too is an ergodic system. The finite state-transition-rate diagram is shown in Figure 3.11. Solving this system, we have from Eq. (3.11)

$$p_k = \frac{p_0 \prod_{i=0}^{k-1} \lambda(M-i)}{(k+1)\mu}$$

$$= p_0 \left(\frac{\lambda}{\mu}\right)^k \binom{M}{k} \quad 0 \leq k \leq M \quad (3.49)$$

where the binomial coefficient is defined in the usual way,

$$\binom{M}{k} \triangleq \frac{M!}{k!(M-k)!}$$

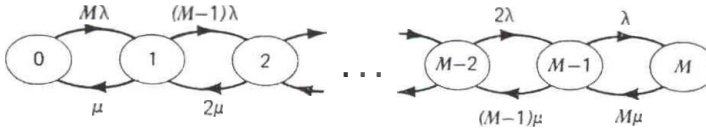


Figure 3.11 State-transition-rate diagram for "infinite"-server finite population system $M/M/\infty/M$.

Solving for P_0 we have

$$p_0 = \left[\sum_{k=0}^M \left(\frac{\lambda}{\mu} \right)^k \binom{M}{k} \right]^{-1}$$

and so

$$p_0 = \frac{1}{(1 + \lambda/\mu)^M}$$

Thus

$$p_k = \begin{cases} \frac{\left(\frac{\lambda}{\mu} \right)^k \binom{M}{k}}{(1 + \lambda/\mu)^M} & 0 \leq k \leq M \\ 0 & \text{otherwise} \end{cases} \quad (3.50)$$

We may easily calculate the expected number of people in the system from

$$\begin{aligned} \bar{N} &= \sum_{k=0}^M k p_k \\ &= \frac{\sum_{k=0}^M k \left(\frac{\lambda}{\mu} \right)^k \binom{M}{k}}{(1 + \lambda/\mu)^M} \end{aligned}$$

Using the partial-differentiation trick such as for obtaining Eq. (3.24) we then have

$$\bar{N} = \frac{M\lambda/\mu}{1 + \lambda/\mu}$$

3.10. M/M/m/K/M: FINITE POPULATION, m-SERVER CASE, FINITE STORAGE

This rather general system is the most complicated we have so far considered and will reduce to all of the previous cases (except the example of discouraged arrivals) as we permit the parameters of this system to vary. We assume we have a finite population of M customers, each with an "arriving" parameter λ . In addition, the system has m servers, each with parameter μ . The system also has finite storage room such that the total number of customers in the system (queueing plus those in service) is no more than K . We assume $M \geq K \geq m$; customers arriving to find K already in the system are "lost" and return immediately to the arriving state as if they had just completed service. This leads to the following set of birth-death coefficients:

$$\begin{aligned} \lambda'_k &= \begin{cases} \lambda(M - k) & 0 \leq k \leq K - 1 \\ 0 & \text{otherwise} \end{cases} \\ \mu_k &= \begin{cases} k\mu & 0 \leq k \leq m \\ m\mu & k \geq m \end{cases} \end{aligned}$$

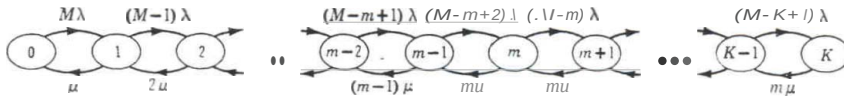


Figure 3.12 State-transition-rate diagram for m-server, finite storage, finite population system M/M/m/K/M.

In Figure 3.12 we see the most complicated of our finite state-transition-rate diagrams. In order to apply Eq. (3.11) we must consider two regions. First, for the range $0 \leq k \leq m-1$ we have

$$p_k = p_0 \prod_{i=0}^{k-1} \frac{\lambda(M-i)}{(i+1)\mu} = p_0 \left(\frac{\lambda}{\mu}\right)^k \binom{M}{k} \quad 0 \leq k \leq m-1 \quad (3.51)$$

For the region $m \leq k \leq K$ we have

$$p_k = p_0 \prod_{i=0}^{m-1} \frac{\lambda(M-i)}{(i+1)\mu} \prod_{i=m}^{k-1} \frac{\lambda(M-K+i)}{m\mu} = p_0 \left(\frac{\lambda}{\mu}\right)^k \binom{M}{k} \frac{1}{m!} \prod_{i=m}^{k-1} (M-K+i) \quad \ll S e \gg \quad (3.52)$$

The expression for p_0 is rather complex and will not be given here, although it may be computed in a straightforward manner. In the case of a pure loss system (i.e., $M \geq K = m$), the stationary state probabilities are given by

$$p_k = \frac{\binom{M}{k} \left(\frac{\lambda}{\mu}\right)^k}{\sum_{i=0}^m \binom{M}{i} \left(\frac{\lambda}{\mu}\right)^i} \quad k = 0, 1, \dots, m \quad (3.53)$$

This is known as the *Engset* distribution.

We could continue these examples ad nauseam but we will instead take a benevolent approach and terminate the set of examples here. Additional examples are given in the exercises. It should be clear to the reader by now that a large number of interesting queueing structures can be modeled with the birth-death process. In particular, we have demonstrated the ability to model the multiple-server case, the finite-population case, the finite-storage case and combinations thereof. The common element in all of these is that the solution for the equilibrium probabilities $\{p_k\}$ is given in Eqs. (3.11) and (3.12). Only systems whose solutions are given by these equations have been considered in this chapter. However, there are many other Markovian systems that lend themselves to simple solution and which are important in queueing

theory. In the next chapter (4) we consider the equilibrium solution for Markovian queues; in Chapter 5 we will generalize to semi-Markov processes in which the service time distribution $B(x)$ is permitted to be general, and in Chapter 6 we revert back to the exponential service time case, but permit the interarrival time distribution $A(I)$ to be general; in both of these cases an imbedded Markov chain will be identified and solved. Only when both $A(I)$ and $B(x)$ are nonexponential do we require the methods of advanced queueing theory discussed in Chapter 8. (There are some special nonexponential distributions that may be described with the theory of Markov processes and these too are discussed in Chapter 4.)

EXERCISES

3.1. Consider a pure Markovian queueing system in which



$$\lambda_k = \begin{cases} \lambda & 0 \leq k \leq K \\ 2\lambda & K < k \end{cases}$$

$$\mu_k = \mu \quad k = 1, 2, \dots$$

- (a) Find the equilibrium probabilities P_k for the number in the system.
- (b) What relationship must exist among the parameters of the problem in order that the system be stable and, therefore, that this equilibrium solution in fact be reached? Interpret this answer in terms of the possible dynamics of the system.

3.2. Consider a Markovian queueing system in which

$$\lambda_k = \alpha^k \lambda \quad k \geq 0, 0 \leq \alpha < 1$$

$$\mu_k = \mu \quad k \geq 1$$

- (a) Find the equilibrium probability p_k of having k customers in the system. Express your answer in terms of P_0 .
- (b) Give an expression for P_{∞}
- 3.3.** Consider an $M/M/2$ queueing system where the average arrival rate is λ customers per second and the average service time is $1/\mu$ sec, where $\lambda < 2\mu$.
- (a) Find the differential equations that govern the time-dependent probabilities $P_k(t)$.
- (b) Find the equilibrium probabilities

$$P_k = \lim_{t \rightarrow \infty} P_k(t)$$

- 3.4. Consider an $M/M/1$ system with parameters λ, μ in which customers are impatient. Specifically, upon arrival, customers estimate their queueing time w and then join the queue with probability $e^{-\alpha w}$ (or leave with probability $1 - e^{-\alpha w}$). The estimate is $w = k/\mu$ when the new arrival finds k in the system. Assume $0 \leq \alpha$.
- In terms of P_0 , find the equilibrium probabilities P_k of finding k in the system. Give an expression for P_0 in terms of the system parameters.
 - For $0 < \alpha, 0 < \mu$ under what conditions will the equilibrium solution hold?
 - For $\alpha \rightarrow \infty$, find P_k explicitly and find the average number in the system.
- 3.5. Consider a birth-death system with the following birth and death coefficients:

$$\begin{aligned}\lambda_k &= (k+2)\lambda & k &= 0, 1, 2, \dots \\ \mu_k &= k\rho & k &= 1, 2, 3, \dots\end{aligned}$$

All other coefficients are zero.

- Solve for p_k . Be sure to express your answer explicitly in terms of λ, k , and μ only.
 - Find the average number of customers in the system.
- 3.6. Consider a birth-death process with the following coefficients:

$$\begin{aligned}\lambda_k &= \alpha k(K_2 - k) & k &= K_1, K_1 + 1, \dots, K_2 \\ \mu_k &= \beta k(k - K_1) & k &= K_1, K_1 + 1, \dots, K_2\end{aligned}$$

where $K_1 \leq K_2$ and where these coefficients are zero outside the range $K_1 \leq k \leq K_2$. Solve for P_k (assuming that the system initially contains $K_1 \leq k \leq K_2$ customers).

- 3.7. Consider an $M/M/m$ system that is to serve the pooled sum of two Poisson arrival streams; the i th stream has an average arrival rate given by λ_i and exponentially distributed service times with mean $1/\mu_i$ ($i = 1, 2$). The first stream is an ordinary stream whereby each arrival requires exactly one of the m servers; if all m servers are busy then any newly arriving customer of type 1 is lost. Customers from the second class each require the simultaneous use of m_0 servers (and will occupy them all simultaneously for the same exponentially distributed amount of time whose mean is $1/\mu_2$ sec); if a customer from this class finds less than m_0 idle servers then he too is lost to the system. Find the fraction of type 1 customers and the fraction of type 2 customers that are lost.

- 3.8. Consider a finite customer population system with a single server such as that considered in Section 3.8; let the parameters M, λ be replaced by M, λ' . It can be shown that if $M \rightarrow \infty$ and $\lambda' \rightarrow \lambda$ such that $\lim M\lambda' = \lambda$ then the finite population system becomes an infinite population system with exponential interarrival times (at a mean rate of λ customers per second). Now consider the case of Section 3.10; the parameters of that case are now to be denoted M, λ', m, μ in the obvious way. Show what value these parameters must take on if they are to represent the earlier cases described in Sections 3.2, 3.4, 3.5, 3.6, 3.7, 3.8, or 3.9.
- 3.9. Using the definition for $B(m, \lambda/\mu)$ in Section 3.7 and the definition of $C(m, \lambda/\mu)$ given in Section 3.5 establish the following for $\lambda/\mu > 0$, $m = 1, 2, \dots$

$$(a) \quad S\left(m, \frac{\lambda}{\mu}\right) < \sum_{k=m}^{\infty} \frac{(\lambda/\mu)^k}{k!} < C\left(m, \frac{\lambda}{\mu}\right)$$

$$(b) \quad C\left(m, \frac{\lambda}{\mu}\right) = \frac{B\left(m, \frac{\lambda}{\mu}\right)}{1 - \frac{\lambda}{\mu} \left[1 - B\left(m, \frac{\lambda}{\mu}\right)\right]}$$

$$(c) \quad B\left(m+1, \frac{\lambda}{\mu}\right) = \frac{\frac{\mu}{\lambda} B\left(m, \frac{\lambda}{\mu}\right)}{m+1 + \frac{\lambda}{\mu} B\left(m, \frac{\lambda}{\mu}\right)}$$

- 3.10. Here we consider an M/M/1 queue in discrete time where time is segmented into intervals of length q sec each. We assume that events can only occur at the ends of these discrete time intervals. In particular the probability of a single arrival at the end of such an interval is given by λq and the probability of no arrival at that point is $1 - \lambda q$ (thus at most one arrival may occur). Similarly the departure process is such that if a customer is in service during an interval he will complete service at the end of that interval with probability $1 - \sigma$ or will require at least one more interval with probability σ .
- (a) Derive the form for $a(l)$ and $b(x)$, the interarrival time and service time pdf's, respectively.
- (b) Assuming FCFS, write down the equilibrium equations that govern the behavior of $P_k = P[k \text{ customers in system at the end of a discrete time interval}]$ where k includes any arrivals who

have occurred at the end of this interval as well as any customers who are about to leave at this point.

- (c) Solve for the expected value of the number of customers at these points.

3.11. Consider an M/M/1 system with "feedback"; by this we mean that when a customer departs from service he has probability σ of rejoining the tail of the queue after a random feedback time, which is exponentially distributed (with mean $1/\gamma$ sec); on the other hand, with probability $1 - \sigma$ he will depart forever after completing service. It is clear that a customer may return many times to the tail of the queue before making his eventual final departure. Let p_{kj} be the equilibrium probability that there are k customers in the "system" (that is, in the queue and the service facility) and that there are j customers in the process of returning to the system.

- (a) Write down the set of difference equations for the equilibrium probabilities p_{kj} .

- (b) Defining the double z-transform

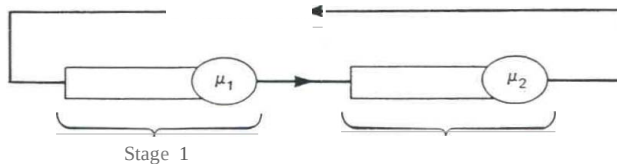
$$P(z_1, z_2) = \sum_{k=0}^{\infty} \sum_{j=0}^{\infty} p_{kj} z_1^k z_2^j$$

show that

$$\begin{aligned} \gamma z_2 - z_1 \frac{\partial P(z_1, z_2)}{\partial z_2} + \{\lambda(1 - \sigma) - z_1\} \\ + \mu \left[1 - \frac{1}{z_1} - \frac{\sigma}{z_1} \frac{z_2}{z_1} \right] P(z_1, z_2) \\ = \mu \left[1 - \frac{1}{z_1} - \frac{\sigma}{z_1} \frac{z_2}{z_1} \right] P(0, z_2) \end{aligned}$$

- (c) By taking advantage of the moment-generating properties of our z-transforms, show that the mean number in the "system" (queue plus server) is given by $\rho/(1 - \rho)$ and that the mean number returning to the tail of the queue is given by $\mu\sigma\rho/\gamma$, where $\rho = \lambda/(1 - \sigma)\mu$.

3.12. Consider a "cyclic queue" in which M customers circulate around through two queueing facilities as shown below.



Both servers are of the exponential type with rates μ_1 and μ_2 , respectively. Let

$$P_k = P[k \text{ customers in stage I and } M-k \text{ in stage 2}]$$

- (a) Draw the state-transition-rate diagram.
- (b) Write down the relationship among $\{p_k\}$.
- (c) Find

$$P(z) = \sum_{k=0}^M p_k z^k$$

- (d) Find p_k .

- 3.13. Consider an $M\{M\}$ queue with parameters λ and μ . A customer in the queue will defect (depart without service) with probability $\alpha \Delta t + o(\Delta t)$ in any interval of duration Δt .

- (a) Draw the state-transition-rate diagram.
- (b) Express p_{k+1} in terms of p_k .
- (c) For $\alpha = \mu$, solve for p_k ($k = 0, 1, 2, \dots$).

- 3.14. Let us elaborate on the $M\{M\}$ system of Section 3.6.

- (a) Evaluate P_0 when $\lambda = \mu$.
- (b) Find $\bar{N}$ for $\lambda \neq \mu$ and for $\lambda = \mu$.
- (c) Find T by carefully solving for the average arrival rate to the system:

Markovian Queues in Equilibrium

The previous chapter was devoted to the study of the birth-death product solution given in Eq. (3.11). The beauty of that solution lies not only in its simplicity but also in its broad range of application to queueing systems, as we have discussed. When we venture beyond the birth-death process into the more general Markov process, then the product solution mentioned above no longer applies; however, one seeks and often finds some other form of product solution for the pure Markovian systems. In this chapter we intend to investigate some of these Markov processes that are of direct interest to queueing systems. Most of what we say will apply to random walks of the Markovian type; we may think of these as somewhat more general birth-death processes where steps beyond nearest neighbors are permitted, but which nevertheless contain sufficient structure so as to permit explicit solutions. All of the underlying distributions are, of course, exponential.

Our concern here again is with equilibrium results. We begin by outlining a general method for finding the equilibrium equations by inspection. Then we consider the special Erlangian distribution E_r , which is applied to the queueing systems $M/ET/1$ and $ETM/1$. We find that the system $M/ET/1$ has an interpretation as a bulk arrival process whose general form we study further; similarly the system $ET/M/1$ may be interpreted as a bulk service system, which we also investigate separately. We then consider the more general systems $E_{r_a}/E_{r_b}/1$ and step beyond that to mention a broad class of $\bar{M}/G/1$ systems that are derivable from the Erlangian by "series-parallel" combinations. Finally, we consider the case of queueing networks in which all the underlying distributions once again are of the memoryless type. As we shall see in most of these cases we obtain a product form of solution.

1. THE EQUILIBRIUM EQUATIONS

Our point of departure is Eq. (2.116), namely, $1tQ = 0$, which expresses the equilibrium conditions for a general ergodic discrete-state continuous-time Markov process; recall that $1t$ is the row vector of equilibrium state probabilities and that Q is the infinitesimal generator whose elements are the

infinitesimal transition rates of our Markov process. As discussed in the previous chapter, we adopt the more standard queueing-theory notation and replace the vector τ_i with the row vector p whose k th element is the equilibrium probability p_k of finding the system in state E_k . Our task then is to solve

$$pQ = 0$$

with the additional conservation relation given in Eq. (2.117), namely,

$$\sum_k p_k = 1$$

This vector equation describes the "equations of motion" in equilibrium.

In Chapter 3 we presented a graphical inspection method for writing down equations of motion making use of the state-transition-rate diagram. For the equilibrium case that method was based on the observation that the probabilistic flow rate into a state must equal the probabilistic flow rate out of that state. It is clear that this notion of flow conservation applies more generally than only to the birth-death process, but in fact to any Markov chain. Thus we may construct "non-nearest-neighbor" systems and still expect that our flow conservation technique should work; this in fact is the case. Our approach then is to describe our Markov chain in terms of a state diagram and then apply conservation of flow to each state in turn. This graphical representation is often easier for this purpose than, in fact, is the verbal, mathematical, or matrix description of the system. Once we have this graphical representation we can, by inspection, write down the equations that govern the system dynamics. As an example, let us consider the very simple three-state Markov chain (which clearly is not a birth-death process since the transition $E_0 \rightarrow E_2$ is permitted), as shown in Figure 4.1. Writing down the flow conservation law for each state yields

$$\frac{3}{2}\lambda p_0 = \mu p_1 \quad (4.1)$$

$$(\lambda + \mu)p_1 = \lambda p_0 + \mu p_2 \quad (4.2)$$

$$\mu p_2 = \frac{\lambda}{2} p_0 + \lambda p_1 \quad (4.3)$$

where Eqs. (4.1), (4.2), and (4.3) correspond to the flow conservation for states E_0 , E_1 and E_2 , respectively. Observe also that the last equation is exactly the sum of the first two; we always have exactly one redundant equation in these finite Markov chains. We know that the additional equation required is

$$p_0 + p_1 + p_2 = 1$$

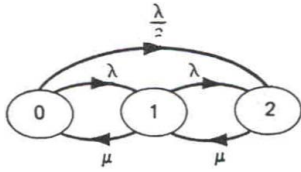


Figure 4.1 Example of non-nearest-neighbor system.

The solution to this system of equations gives

$$\begin{aligned}
 p_0 &= \left[1 + 2 \frac{\lambda}{\mu} + \frac{3}{2} \left(\frac{\lambda}{\mu} \right)^2 \right]^{-1} \\
 p_1 &= \frac{3}{2} \frac{\lambda}{\mu} p_0 \\
 p_2 &= \left[\frac{1}{2} \frac{\lambda}{\mu} + \frac{3}{2} \left(\frac{\lambda}{\mu} \right)^2 \right] p_0
 \end{aligned} \tag{4.4}$$

Voila! Simple as pie. In fact, it is as "simple" as inverting a set of simultaneous linear equations.

We take advantage of this inspection technique in solving a number of Markov chains in equilibrium in the balance of this chapter.*

As in the previous chapter we are here concerned with the limiting probability defined as $P_k = \lim_{t \rightarrow \infty} P[N(t) = k]$ as $t \rightarrow \infty$, assuming it exists. This probability may be interpreted as giving the proportion of *time* that the system spends in state E_k . One could, in fact, estimate this probability by measuring how often the system contained k customers as compared to the total measurement time. Another quantity of interest (perhaps of greater interest) in queueing systems is the probability that an arriving customer finds the system in state E_k ; that is, we consider the equilibrium probability

$$p'_k = P[\text{arriving customer finds the system in state } E_k]$$

in the case of an ergodic system. One might intuitively feel that in all cases $p'_k = p_k$, but it is easy to show that this is not generally true. For example, let us consider the (non-Markovian) system 0/O/1 in which arrivals are uniformly spaced in time such that we get one arrival every i sec exactly; the service-time requirements are identical for all customers and equal, say

* It should also be clear that this inspection technique permits us to write down the time-dependent state probabilities $P_k(t)$ directly as we have already seen for the case of birth-death processes; these time-dependent equations will in fact be exactly Eq. (2.114).

to $\bar{x}$ sec. We recognize this single-server system as an instance of steady flow through a single channel (remember the pineapple factory). For stability we require that $\bar{x} < \bar{t}$. Now it is clear that no arrival will ever have to wait once equilibrium is reached and, therefore, $'0 = 1$ and $'k = 0$ for $k = 1, 2, \dots$. Moreover, it is clear that the fraction of time that the system contains one customer (in service) is exactly equal to $\rho = \bar{x}/\bar{t}$, and the remainder of the time the system will be empty; therefore, we have $P_0 = 1 - \rho$, $p_1 = \rho$, $P_k = 0$ for $k = 2, 3, 4, \dots$. So we have a trivial example in which $p_k \neq 'k$. However, as is often the case, one's intuition has a basis in fact, and we find that there is a large class of queueing systems for which $p_k = 'k$ for all k . This, in fact, is the class of stable queueing systems with Poisson arrivals!

Actually, we can prove more, and as we show below for any queueing system with Poisson arrivals we must have

$$P_k(t) = R_k(t)$$

where $P_k(t)$ is, as before, the probability that the system is in state E_k at time t and where $R_k(t)$ is the probability that a customer arriving at time t finds the system in state E_k . Specifically, for our system with Poisson arrivals we define $A(t, t + \Delta t)$ to be the *event* that an arrival occurs in the interval $(t, t + \Delta t)$; then we have

$$R_k(t) \triangleq \lim_{\Delta t \rightarrow 0} P[N(t) = k \mid A(t, t + \Delta t)] \quad (4.5)$$

[where $N(t)$ gives the number in system at time t]. Using our definition of conditional probability we may rewrite $R_k(t)$ as

$$\begin{aligned} R_k(t) &= \lim_{\Delta t \rightarrow 0} \frac{P[N(t) = k, A(t, t + \Delta t)]}{P[A(t, t + \Delta t)]} \\ &= \lim_{\Delta t \rightarrow 0} \frac{P[A(t, t + \Delta t) \mid N(t) = k] P[N(t) = k]}{P[A(t, t + \Delta t)]} \end{aligned}$$

Now for the case of Poisson arrivals we know (due to the memoryless property) that the event $A(t, t + \Delta t)$ must be independent of the number in the system at time t (and also of the time t itself); consequently $P[A(t, t + \Delta t) \mid N(t) = k] = P[A(t, t + \Delta t)]$, and so we have

$$R_k(t) = \lim_{\Delta t \rightarrow 0} P[N(t) = k]$$

or

$$R_k(t) = P_k(t) \quad (4.6)$$

This is what we set out to prove, namely, that the time-dependent probability of an arrival finding the system in state E_k is exactly equal to the time-dependent probability of the system being in state E_k . Clearly this also applies to the equilibrium probability P_k that an arrival finds k customers in the system and the proportion of time p_k that the system finds itself with k customers. This equivalence does not surprise us in view of the memoryless property of the Poisson process, which as we have just shown generates a sequence of arrivals that take a really "random look" at the system.

4.2. THE METHOD OF STAGES-ERLANGIAN DISTRIBUTION E_r

The "method of stages" permits one to study queueing systems that are more general than the birth--death systems. This ingenious method is a further testimonial to the brilliance of A. K. Erlang, who developed it early in this century long before our tools of modern probability theory were available. Erlang recognized the extreme simplicity of the exponential distribution and its great power in solving Markovian queueing systems. However, he also recognized that the exponential distribution was not always an appropriate candidate for representing the true situation with regard to service times (and interarrival times). He must also have observed that to allow a more general service distribution would have destroyed the Markovian property and then would have required some more complicated solution method.* The inherent beauty of the Markov chain was not to be given up so easily. What Erlang conceived was the notion of *decomposing* the service time distribution into a collection of structured exponential distributions.

The principle on which the method of stages is based is the memoryless property of the exponential distribution; again we repeat that this lack of memory is reflected by the fact that the distribution of time remaining for an exponentially distributed random variable is independent of the acquired "age" of that random variable.

Consider the diagram of Figure 4.2. In this figure we are defining a service facility with an exponentially distributed service time pdf given by

$$b(x) \triangleq \frac{dB(x)}{dx} = \mu e^{-\mu x} \quad x \geq 0 \quad (4.7)$$

The notation of the figure shows an oval which represents the service facility and is labeled with the symbol μ , which represents the service-rate parameter

* As we shall see in Chapter 5, a newer approach to this problem, the "method of imbedded Markov chains," was not available at the time of Erlang.

† Identical observations apply also to the interarrival time distribution.

Service
facility

Figure 4.2 The single-stage exponential server.

as in Eq. (4.7). The reader will recall from Chapter 2 that the exponential distribution has a mean and variance given by

$$E[\tilde{x}] = \frac{1}{\mu}$$

$$\sigma_{\tilde{x}}^2 = \frac{1}{\mu^2}$$

where the subscript $\tilde{x}$ identifies this as the service time variance.

Now consider the system shown in Figure 4.3. In this figure the large oval represents the service facility. The internal structure of this service facility is revealed as a series or tandem connection of two smaller ovals. Each of these small ovals represents a single exponential server such as that depicted in Figure 4.2; in Figure 4.3, however, the small ovals are labeled internally with the parameter 2μ indicating that they each have a pdf given by

$$h(y) = 2\mu e^{-2\mu y} \quad y \geq 0 \quad (4.8)$$

Thus the mean and variance for $h(y)$ are $E(\tilde{y}) = 1/2\mu$ and $\sigma_{\tilde{y}}^2 = (1/2\mu)^2$. The fashion in which this two-stage service facility functions is that upon departure of a customer from this facility a new customer is allowed to enter from the left. This new customer enters stage I and remains there for an amount of time randomly chosen from $h(y)$. Upon his departure from this first stage he then proceeds immediately into the second stage and spends an amount of time there equal to a random variable drawn independently once again from $h(y)$. After this second random interval expires he then departs from the service facility and at this point only may a new customer enter the facility from the left. We see then, that one, and only one, customer is

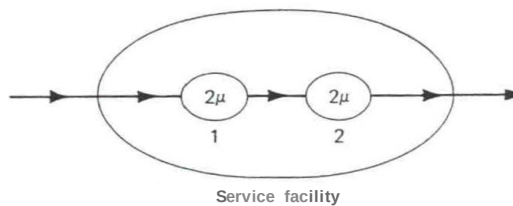


Figure 4.3 The two-stage Erlangian server.

allowed into the box entitled "service facility" at any time.* This implies that at least one of the two service stages must always be empty. We now inquire as to the specific distribution of total time spent in the service facility. Clearly this is a random variable, which is the sum of two independent and identically distributed random variables. Thus, as shown in Appendix II, we must form the convolution of the density function associated with each of the two summands. Alternatively, we may calculate the Laplace transform of the service time pdf as being equal to the product of the Laplace transform of the pdf's associated with each of the summands. Since both random variables are (independent and) identically distributed we must form the product of a function with itself. First, as always, we define the appropriate transforms as

$$B^*(s) \triangleq \int_0^{\infty} e^{-sx} b(x) dx \quad (4.9)$$

$$H^*(s) \triangleq \int_0^{\infty} e^{-sy} h(y) dy \quad (4.10)$$

From our earlier statements we have

$$B^*(s) = [H^*(s)]^2$$

But, we already know the transform of the exponential from Eq. (2.144) and so

$$H^*(s) = \frac{2\mu}{s + 2\mu}$$

Thus

$$B^*(s) = \left(\frac{2\mu}{s + 2\mu} \right)^2 \quad (4.11)$$

We must now invert Eq. (4.11). However, the reader may recall that we already have seen this form in Eq. (2.146) with its inverse in Eq. (2.147). Applying that result we have

$$b(x) = 2\mu(2\mu x)e^{-2\mu x} \quad x \geq 0 \quad (4.12)$$

We may now calculate the mean and variance of this two-stage system in one of three possible ways: by arguing on the basis of the structure in Figure 4.3; by using the moment generating properties of $B^*(s)$; or by direct calculation

* As an example of a two-stage service facility in which only one stage may be active at a time, consider a courtroom in a small town. A queue of defendants forms, waiting for trial. The judge tries a case (the first service stage) and then fines the defendant. The second stage consists of paying the fine to the court clerk. However, in this small town, the judge is also the clerk and so he moves over to the clerk's desk, collects the fine, releases the defendant, goes back to his bench, and then accepts the next defendant into "service."

from the density function given in Eq. (4.12). We choose the first of these three methods since it is most straightforward (the reader may verify the other two for his own satisfaction). Since the time spent in service is the sum of two random variables, then it is clear that the expected time in service is the sum of the expectations of each. Thus we have

$$E[\tilde{x}] = 2E[\tilde{y}] = \frac{1}{\mu}$$

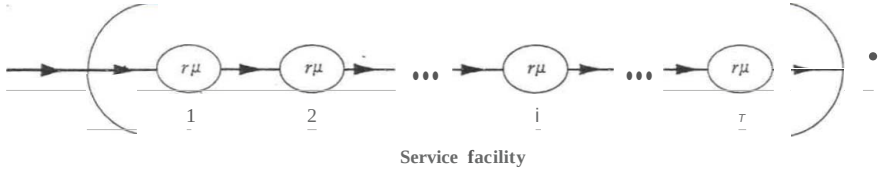
Similarly, since the two random variables being summed are independent, we may, therefore, sum their variances to find the variance of the sum:

$$\sigma_b^2 = \sigma_h^2 + \sigma_h^2 = \frac{1}{2\mu^2}$$

Note that we have arranged matters such that the mean time in service in the single-stage system of Figure 4.2 and the two-stage system of Figure 4.3 is the same. We accomplished this by speeding up each of the two-stage service stations by a factor of 2. Note further that the variance of the two-stage system is one-half the variance of the one-stage system.

The previous paragraph introduced the notion of a two-stage service facility but we have yet to discuss the crucial point. Let us consider the state variable for a queueing system with Poisson arrivals and a two-stage exponential server as given in Figure 4.3. As always, as part of our state description, we must record the number of customers waiting in the queue. In addition we must supply sufficient information about the service facility so as to summarize the relevant past history. Owing to the memoryless property of the exponential distribution it is enough to indicate which of the following three possible situations may be found within the service facility: either both stages are idle (indicating an empty service facility); or the first stage is busy and the second stage is idle; or the first stage is idle and the second stage is busy. This service-facility state information may be supplied by identifying the stage of service in which the customer may be found. Our state description then becomes a two-dimensional vector that specifies the number of customers in queue and the number of stages yet to be completed by our customer in service. The time this customer has already spent in his current stage of service is irrelevant in calculating the future behavior of the system. Once again we have a Markov process with a discrete (two-dimensional) state space!

The method generalizes and so now we consider the case in which we provide an r -stage service facility, as shown in Figure 4.4. In this system, of course, when a customer departs by exiting from the right side of the oval service facility a new customer may then enter from the left side and proceed one stage at a time through the sequence of r stages. Upon his departure from


 Figure 4.4 The r -stage Erlangian server E_r .

the r th stage a new customer again may then enter, and so on. The time that he spends in the i th stage is drawn from the density function

$$h(y) = r\mu e^{-r\mu y} \quad Y \geq 0 \quad (4.13)$$

The total time that a customer spends in this service facility is the sum of r independent identically distributed random variables, each chosen from the distribution given in Eq. (4.13). We have the following expectation and variance associated with each stage :

$$E[\tilde{y}] = \frac{1}{r\mu}$$

$$\sigma_h^2 = \left(\frac{1}{r\mu}\right)^2$$

It should be clear to the reader that we have chosen each stage in this system to have a service rate equal to $r\mu$ in order that the mean service time remain constant :

$$E[\tilde{x}] = r \left(\frac{1}{r\mu}\right) = \frac{1}{\mu}$$

Similarly, since the stage times are independent we may add the variances to obtain

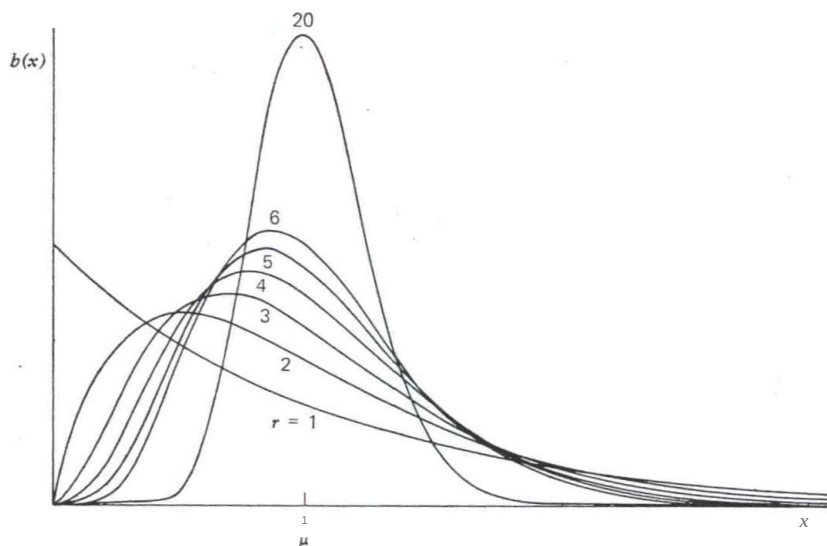
$$\sigma_b^2 = r \left(\frac{1}{r\mu}\right)^2 = \frac{1}{r\mu^2}$$

Also, we observe that the coefficient of variation [see Eq. (11.23)] is

$$C_s = \frac{1}{\sqrt{r}} \quad (4.14)$$

Once again we wish to solve for the pdf of the service time. This we do by generalizing the notions leading up to Eq. (4.11) to obtain

$$g^*(s) = \left(\frac{r\mu}{s + r\mu}\right)^r \quad (4.15)$$

Figure 4.5 The family of r -stage Erlangian distributions E_i :

Equation (4.15) is easily inverted as earlier to give

$$b(x) = \frac{r\mu(r\mu x)^{r-1}e^{-r\mu x}}{(r-1)!} \quad x \geq 0 \quad - (4.16)$$

This we recognize as the Erlangian distribution given in Eq. (2.147). We have carefully adjusted the mean of this density function to be independent of r . In order to obtain an indication of its width we must examine the standard deviation as given by

$$\sigma_r = \frac{1}{\sqrt{r}} \left(\frac{1}{\mu} \right)$$

Thus we see that the standard deviation for the r -stage Erlangian distribution is $1/\sqrt{r}$ times the standard deviation for the single stage. It should be clear to the sophisticated reader that as r increases, the density function given by Eq. (4.16) must approach that of the normal or Gaussian distribution due to the central limit theorem. This is indeed true but we give more in Eq. (4.16) by specifying the actual sequence of distributions as r increases to show the fashion in which the limit is approached. In Figure 4.5 we show the family of r -stage Erlangian distributions (compare with Figure 2.10). From this figure we observe that the mean holds constant as the width or standard deviation of the density shrinks by $1/\sqrt{r}$. Below, we show that the limit (as r goes to infinity) for this density function must, in fact, be a unit impulse function

(see Appendix I) at the point $x = 1/\mu$; this implies that the time spent in an infinite-stage Erlangian service facility approaches a constant with probability 1 (this constant, of course, equals the mean $1/\mu$). We see further that the peak of the family shown moves to the right in a regular fashion. To calculate the location of the peak, we differentiate the density function as given in Eq. (4.16) and set this derivative equal to zero to obtain

$$\frac{db(x)}{dx} = \frac{(r\mu)^2(r-1)(r\mu x)^{r-2}e^{-r\mu x}}{(r-1)!} - \frac{(r\mu x)^{r-1}e^{-r\mu x}(r\mu)^2}{(r-1)!} = 0$$

or

$$(r-1) = r\mu x$$

and so we have

$$x_{\text{peak}} = \left(\frac{r-1}{r}\right)\frac{1}{\mu} \quad (4.17)$$

Thus we see that the location of the peak moves rather quickly toward its final location at $1/\mu$.

We now show that the limiting distribution is, in fact, a unit impulse by considering the limit of the Laplace transform given in Eq. (4.15):

$$\begin{aligned} \lim_{r \rightarrow \infty} B^*(s) &= \lim_{r \rightarrow \infty} \left(\frac{r\mu}{s + r\mu} \right)^r \\ &= \lim_{r \rightarrow \infty} \left(1 + \frac{1}{s/r\mu} \right)^{-r} \\ \lim_{r \rightarrow \infty} B^*(s) &= e^{-s/\mu} \end{aligned} \quad (4.18)$$

We recognize the inverse transform of this limiting distribution from entry 3 in Table 1.4 of Appendix I; it is merely a unit impulse located at $x = 1/\mu$.

Thus the family of Erlangian distributions varies over a fairly broad range; as such, it is extremely useful for approximating empirical (and even theoretical) distributions. For example, if one had measured a service-time operation and had sufficient data to give acceptable estimates of its mean and variance only, then one could select one member of this two-parameter family such that $1/\mu$ matched the mean and $1/r\mu^2$ matched the variance; this would then be a method for approximating $B(x)$ in a way that permits solution of the queueing system (as we shall see below). If the measured coefficient of variation exceeds unity, we see from Eq. (4.14) that this procedure fails, and we must use the hyperexponential distribution described later or some other distribution.

It is clear for each member of this family of density functions that we may describe the state of the service facility by merely giving the number of stages yet to be completed by a customer in service. We denote the r -stage Erlangian

distribution by the symbol E , (not to be confused with the notation for the state of a random process). Since our state variable is discrete, we are in a position to analyze the queueing system* $M/Er/1$. This we do in the following section. Moreover, we will use the same technique in *Section 4.4* to decompose the interarrival time distribution $A(t)$ into an r -stage Erlangian distribution. Note in these next two sections that we neurotically require at least one of our distributions to be a pure exponential (this is also true for Chapters 5 and 6).

4.3. THE QUEUE $M/Er/1$

Here we consider the system for which

$$\begin{aligned} a(t) &= \lambda e^{-\lambda t} \quad t \geq 0 \\ b(x) &= \frac{r\mu(r\mu x)^{r-1} e^{-r\mu x}}{(r-1)!} \quad x \geq 0 \end{aligned}$$

Since in addition to specifying the number of customers in the system (as in Chapter 3), we must also specify the number of stages remaining in the service facility for the man in service, it behooves us to represent each customer in the queue as possessing r stages of service yet to be completed for him. Thus we agree to take the state variable as the total number of service stages yet to be completed by all customers in the system at the time the state is described.] In particular, if we consider the state at a time when the system contains k customers and when the i th stage of service contains the customer in service we then have that the number of stages contained in the total system is

$$\begin{aligned} j &\triangleq \text{number of stages left in total system} \\ &= (k-1)r + (r-i+1) \end{aligned}$$

Thus

$$j = rk - i + 1 \quad (4.19)$$

As usual, p_k is defined as the equilibrium probability for the number of customers in the system; we further define

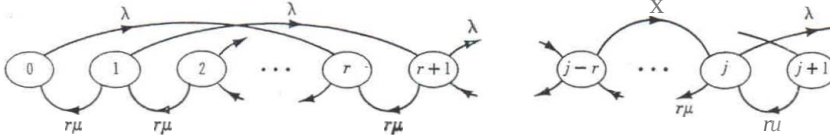
$$P_j \triangleq P[j \text{ stages in system}] \quad (4.20)$$

The relationship between customers and stages allows us to write

$$p_k = \sum_{j=(k-1)r+1}^{kr} P_j; \quad k = 1, 2, 3, \dots$$

* Clearly this is a special case of the system $M/G/H$ which we will analyze in Chapter 5 using the imbedded Markov chain approach.

† Note that this converts our proposed two-dimensional state vector into a one-dimensional description.

Figure 4.6 State-transition-rate diagram for number of stages: *MIEr1*.

And now for the beauty of Erlang's approach: We may represent the state-transition-rate diagram for stages in our system as shown in Figure 4.6. Focusing on state E_j ; we see that it is entered from below by a state which is r positions to its left and also entered from above by state E_{j+r} ; the former transition is due to the arrival of r new stages when a new customer enters, and the latter is due to the completion of one stage within the r -stage service facility. Furthermore we may leave state E_j ; at a rate λ due to an arrival and at a rate $r\mu$ due to a service completion. Of course, we have special boundary conditions for states $E_0, E_1, \dots, E_{r-1}$. In order to handle the boundary situation simply let us agree, as in Chapter 3, that state probabilities with negative subscripts are in fact zero. We thus define

$$P_j = 0 \quad j < 0 \quad (4.21)$$

We may now write down the system state equations immediately by using our flow conservation inspection method. (Note that we are writing the forward equations in equilibrium.) Thus we have

$$\lambda P_0 = r\mu P_1 \quad (4.22)$$

$$(\lambda + r\mu)P_j = \lambda P_{j-r} + r\mu P_{j+1} \quad j = 1, 2, \dots \quad (4.23)$$

Let us now use our "familiar" method of solving difference equations, namely the z -transform. Thus we define

$$P(z) = \sum_{j=0}^{\infty} P_j z^j;$$

As usual, we multiply the j th equation given in Eq. (4.23) by z^j ; and then sum over all applicable j . This yields

$$\sum_{j=1}^{\infty} (\lambda + r\mu) P_j z^j = \sum_{j=1}^{\infty} \lambda P_{j-r} z^j + \sum_{j=1}^{\infty} r\mu P_{j+1} z^j;$$

Rewriting we have

$$(\lambda + r\mu) \left[\sum_{j=0}^{\infty} P_j z^j - P_0 \right] = \lambda z^r \sum_{j=1}^{\infty} P_{j-r} z^{j-r} + \frac{r\mu}{z} \sum_{j=1}^{\infty} P_{j+1} z^{j+1}$$

Recognizing $P(z)$, we then have

$$(\lambda + r\mu)[P(z) - P_0] = \lambda z^r P(z) + \frac{r\mu}{z} [P(z) - P_0 - P_1 z]$$

The first term on the right-hand side of this last equation is obtained by taking special note of Eq. (4.21). Simplifying we have

$$P(z) = \frac{P_0[\lambda + r\mu - (r\mu/z)] - r\mu P_1}{\lambda + r\mu - \lambda z^r - (r\mu/z)}$$

We may now use Eq. (4.22) to simplify this last further:

$$P(z) = \frac{r\mu P_0[1 - (1/z)]}{\lambda + r\mu - \lambda z^r - (r\mu/z)}$$

yielding finally

$$P(z) = \frac{r\mu P_0(1 - z)}{r\mu + \lambda z^{r+1} - (\lambda + r\mu)z} \quad (4.24)$$

We may evaluate the constant P_0 by recognizing that $P(1) = 1$ and using L'Hospital's rule, thus

$$P(1) = 1 = \frac{r\mu P_0}{r\mu - \lambda}$$

giving (observe that $P_0 = P_0$)

$$P_0 = \frac{\lambda}{\mu}$$

In this system the arrival rate is λ and the average service time is held fixed at $1/\mu$ independent of r . Thus we recognize that our utilization factor is

$$\rho \triangleq \frac{\lambda}{\mu} \quad (4.25)$$

Substituting back into Eq. (4.24) we find

$$P(z) = \frac{r\mu(1 - \rho)(1 - z)}{r\mu + \lambda z^{r+1} - (\lambda + r\mu)z} \quad (4.26)$$

We must now invert this z -transform to find the distribution of the number of stages in the system.

The case $r = 1$, which is clearly the system $M/M/1$, presents no difficulties; this case yields

$$P(z) = \frac{\mu(1 - \rho)(1 - z)}{\mu + \lambda z^2 - (\lambda + \mu)z} = \frac{(1 - \rho)(1 - z)}{1 + \rho z^2 - (1 + \rho)z}$$

The denominator factors into $(1 - z)(1 - \rho z)$ and so canceling the common term $(1 - z)$ we obtain

$$P(z) = \frac{1 - \rho}{1 - \rho z}$$

We recognize this function as entry 6 in Table 1.2 of Appendix I, and so we have immediately

$$P_k = (1 - \rho)\rho^k \quad k = 0, 1, 2, \dots \quad (4.27)$$

Now in the case $r = 1$ it is clear that $p_k = P_k$ and so Eq. (4.27) gives us the distribution of the number of customers in the system $M/M/1$, as we had seen previously in Eq. (3.23).

For arbitrary values of r things are a bit more complex. The usual approach to inverting a z -transform such as that given in Eq. (4.26) is to make a partial fraction expansion and then to invert each term by inspection; let us follow this approach. Before we can carry out this expansion we must identify the $r + 1$ zeroes of the denominator polynomial. Unity is easily seen to be one such. The denominator may therefore be written as $(1 - z)[r\mu - \lambda(z + z_2 + \dots + z_r)]$, where the remaining r zeroes (which we choose to denote by $z_1, z_2, \dots, z_r$) are the roots of the bracketed expression. Once we have found these roots* (which are unique) we may then write the denominator as $r\mu(1 - z)(1 - z_1/z) \dots (1 - z_r/z)$. Substituting this back into Eq. (4.26) we find

$$P(z) = \frac{1 - \rho}{(1 - z_1/z)(1 - z_2/z) \dots (1 - z_r/z)}$$

Our partial fraction expansion now yields

$$P(z) = (1 - \rho) \sum_{i=1}^r \frac{A_i}{(1 - z/z_i)} \quad (4.28)$$

where

$$A_i = \prod_{\substack{n=1 \\ n \neq i}}^r \frac{1}{(1 - z_i/z_n)}$$

We may now invert Eq. (4.28) by inspection (from entry 6 in Table 1.2) to obtain the final solution for the distribution of the number of stages in the system, namely,

$$r_j = (1 - \rho) \sum_{i=1}^r A_i (z_i)^{-j} \quad j = 1, 2, \dots, r \quad - \quad (4.29)$$

* Many of the analytic problems in queueing theory reduce to the (difficult) task of locating the roots of a function.

and where as before $P_0 = 1 - \rho$. Thus we see for the system $M/E_r/1$ that the distribution of the number of stages in the system is a weighted sum of *geometric* distributions. The waiting-time distribution may be calculated using the methods developed later in Chapter 5.

4.4. THE QUEUE $Er/M/1$

Let us now consider the queueing system $Er/M/1$ for which

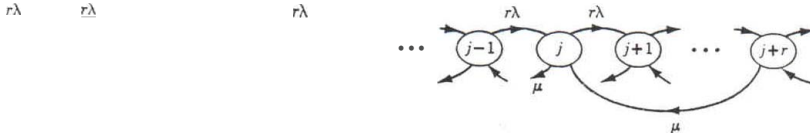
$$a(t) = \frac{r^k (r\lambda t)^{r-1} e^{-r\lambda t}}{(r-1)!} \quad t \geq 0 \quad (4.30)$$

$$b(x) = \mu e^{-\mu x} \quad x \geq 0 \quad (4.31)$$

Here the roles of interarrival time and service time are interchanged from those of the previous section; in many ways these two systems are duals of each other. The system operates as follows: Given that an arrival has just occurred, then one immediately introduces a new "arriving" customer into an r -stage Erlangian facility much like that in Figure 4.4; however, rather than consider this to be a service facility we consider it to be an "arriving" facility. When this arriving customer is inserted from the left side he must then pass through r exponential stages each with parameter $r\lambda$. It is clear that the pdf of the time spent in the arriving facility will be given by Eq. (4.30). When he exits from the right side of the arriving facility he is then said to "arrive" to the queueing system $ErM/1$. Immediately upon his arrival, a new customer (taken from an infinite pool of available customers) is inserted into the left side of the arriving box and the process is repeated. Once having arrived, the customer joins the queue, waits for service, and is then served according to the distribution given in Eq. (4.31). It is clear that an appropriate state description for this system is to specify not only the number of customers in the system, but also to identify which stage in the arriving facility the arriving customer now occupies. We will consider that each customer who has already arrived (but not yet departed) is contributing r stages of "arrival"; in addition we will count the number of stages so far completed by the arriving customer as a further contribution to the number of arrival stages in the system. Thus our state description will consist of the total number of stages of arrival currently in the system; when we find k customers in the system and when our arriving customer is in the i th stage of arrival ($1 \leq i \leq r$) then the total number of stages of arrival in the system is given by

$$j = rk + i - 1$$

Once again let us use the definition given in Eq. (4.20) so that P_i is defined to be the number of *arrival* stages in the system; as always P_k will be the

Figure 4.7 State-transition-rate diagram for number of stages: $ET/M/I$.

equilibrium probability for number of *customers* in the system, and clearly they are related through

$$P_k = \sum_{j=rk}^{r(k+1)-1} P_j;$$

The system we have defined is an irreducible ergodic Markov chain with its state-transition-rate diagram for stages given in Figure 4.7. Note that when a customer departs from service, he "removes" r stages of "arrival" from the system. Using our inspection method, we may write down the equilibrium equations as

$$r\lambda P_0 = \mu P_r \quad (4.32)$$

$$r\lambda P_j = r\lambda P_{j-1} + \mu P_{j+r} \quad 1 \leq j \leq r-1 \quad (4.33)$$

$$(r\lambda + \mu)P_j = r\lambda P_{j-1} + \mu P_{j+r} \quad r \leq j \quad (4.34)$$

Again we define the z -transform for these probabilities as

$$P(z) = \sum_{j=0}^{\infty} P_j z^j$$

Let us now apply our transform method to the equilibrium equations. Equations (4.33) and (4.34) are almost identical except that the former is missing the term μP_j ; consequently let us operate upon the equations in the range $j \geq 1$, adding and subtracting the missing terms as appropriate. Thus we obtain

$$\sum_{j=1}^{\infty} (\mu + r\lambda) P_j z^j - \sum_{j=1}^{r-1} \mu P_j z^j = \sum_{j=1}^{\infty} r\lambda P_{j-1} z^j + \sum_{j=1}^{\infty} \mu P_{j+r} z^j$$

Identifying the transform in this last equation we have

$$(\mu + r\lambda)[P(z) - P_0] - \sum_{j=1}^{r-1} \mu P_j z^j = r\lambda z P(z) + \frac{\mu}{z^r} [P(z) - \sum_{j=0}^r P_j z^j]$$

We may now use Eq. (4.32) to eliminate the term P_r and then finally solve for our transform to obtain

$$P(z) = \frac{(1 - zI) \sum_{j=0}^{r-1} P_j z^j}{rpz^{r+1} - (1 + rp)z^r + 1} \quad (4.35)$$

where as always we have defined $\rho = \lambda \bar{x} = \lambda/\mu$. We must now study the poles (zeroes of the denominator) for this function. The denominator polynomial has $r + 1$ zeroes of which unity is one such [the factor $(1 - z)$ is almost always present in the denominator]. Of the remaining r zeroes it can be shown (see Exercise 4.10) that exactly $r - 1$ of them lie in the range $|z| < 1$ and the last, which we shall denote by z_0 , is such that $|z_0| > 1$. We are still faced with the numerator summation that contains the unknown probabilities P_i ; we must now appeal to the second footnote in step 5 of our z -transform procedure (see Chapter 2, pp. 74-75), which takes advantage of the observation that the a -transform of a probability distribution must be analytic in the range $|z| < 1$ in the following way. Since $P(z)$ must be bounded in the range $|z| < 1$ [see Eq. (II.28)] and since the denominator has $r - 1$ zeroes in this range, then certainly the numerator must also have zeroes at the same $r - 1$ points. The numerator consists of two factors; the first of the form $(1 - z)^r$ all of whose zeroes have absolute value equal to unity; and the second in the form of a summation. Consequently, the "compensating" zeroes in the numerator must come from the summation itself (the summation is a polynomial of degree $r - 1$ and therefore has exactly $r - 1$ zeroes). These observations, therefore, permit us to equate the numerator sum to the denominator (after its two roots at $z = 1$ and $z = z_0$ are factored out) as follows:

$$\frac{\sum_{i=0}^{\infty} \rho^i P_i z^{r+i} - (1 - \rho)^r z^r + 1}{(1 - z)(1 - z/z_0)} = K \sum_{j=0}^{r-1} P_j z^j;$$

where K is a constant to be evaluated below. This computation permits us to rewrite Eq. (4.35) as

$$P(z) = \frac{(1 - z^r)}{K(1 - z)(1 - z/z_0)}$$

But since $P(1) = 1$ we find that

$$K = r/(1 - 1/z_0)$$

and so we have

$$P(z) = \frac{(1 - z^r)(1 - 1/z_0)}{r(1 - z)(1 - z/z_0)} \quad (4.36)$$

We now know all there is to know about the poles and zeroes of $P(z)$; we are, therefore, in a position to make a partial fraction expansion so that we may invert on z . Unfortunately $P(z)$ as expressed in Eq. (4.36) is not in the proper form for the partial fraction expansion, since the numerator degree is not less than the denominator degree. However, we will take advantage of property 8 in Table I.1 of Appendix I, which states that if $F(z) = \sum_{n=0}^{\infty} F_n z^n$ then

$z'F(z) \Leftrightarrow f_{n-r}$, where we recall that the notation $\Leftrightarrow$ indicates a transform pair. With this observation then, we carry out the following partial fraction expansion

$$P(z) = (1 - \mathcal{Z}[\frac{1}{z} + \frac{-1/rz_0}{1 - z/z_0}])$$

If we denote the inverse transform of the quantity in square brackets by f_j , then it is clear that the inverse transform for $P(z)$ must be

$$P_j = f_j - f_{j-r} \quad (4.37)$$

By inspection we see that

$$f_j = \begin{cases} (1 - z_0^{-j-1}) & j \geq 0 \\ 0 & j < 0 \end{cases} \quad (4.38)$$

First we solve for P_j in the range $j \geq r$; from Eqs. (4.37) and (4.38) we, therefore, have

$$P_j = \frac{1}{r} z_0^{-j-1} (1 - z_0^{-r}) \quad (4.39)$$

We may simplify this last expression by recognizing that the denominator of Eq. (4.35) must equal zero for $z = z_0$; this observation leads to the equality $rp(z_0 - 1) = 1 - z_0^{-r}$, and so Eq. (4.39) becomes

$$P_j = p(z_0 - 1) z_0^{r-j-1} \quad j \geq r \quad (4.40)$$

On the other hand, in the range $0 \leq j < r$ we have that $f_{j-r} = 0$, and so P_j is easily found for the rest of our range. Combining this and Eq. (4.40) we finally obtain the distribution for the number of arrival stages in our system:

$$P_j = \begin{cases} (1 - z_0^{-j-1}) & 0 \leq j < r \\ p(z_0 - 1) z_0^{r-j-1} & j \geq r \end{cases} \quad (4.41)$$

Using our earlier relationship between p_k and P_j we find (the reader should check this algebra for himself) that the distribution of the number of customers in the system is given by

$$P_k = \begin{cases} 1 - p & k = 0 \\ p(z_0^r - 1) z_0^{-rk} & k > 0 \end{cases} \quad (4.42)$$

We note that this distribution for number of customers is *geometric* with a slightly modified first term. We could at this point calculate the waiting time distribution, but we will postpone that until we study the system $G/M/1$ in Chapter 6.

4.5. BULK ARRIVAL SYSTEMS

In Section 4.3 we studied the system $M/ET\{1$ in which each customer had to pass through r stages of service to complete his total service. The key to the solution of that system was to count the number of service stages remaining in the system, each customer contributing r stages to that number upon his arrival into the system. We may look at the system from another point of view in **which** we consider each "customer" arrival to be in reality the arrival of r customers. Each of these r customers will require only a single stage of service (that is, the service time distribution is an exponential*). Clearly, these two points of view define identical systems: The former is the system $M/ET\{1$ and the latter is an $M/M/I$ system with "bulk" arrivals of size r . In fact, if we were to draw the state-transition-rate diagram for the number of *customers* in the system, then the bulk arrival system would lead to the diagram given in Figure 4.6; of course, that diagram was for the number of stages in the system $M/ET\{1$. As a consequence, we see that the generating function for the number of customers in the bulk arrival system must be given by Eq. (4.26) and that the distribution of number of customers in the system is given by Eq. (4.29) since we are equating stages in the original system to customers in the current system.

Since we are considering bulk arrival systems, we may as well be more generous and permit other than a fixed-size bulk to arrive at each (Poisson) arrival instant. What we have in mind is to permit a bulk (or group) at each arrival instant to be of random size where

$$g_i \triangleq P[\text{bulk size is } i] \quad (4.43)$$

(As an example, one may think of random-size families arriving at the doctor's office for individual vaccinations.) As usual, we will assume that the arrival rate (of bulks) is λ . Taking the number of customers in the system as our state variable, we have the state-transition-rate diagram of Figure 4.8. In this figure we have shown details only for state E_k for clarity. Thus we find that we can enter E_k from any state below it (since we permit bulks of any size to arrive); similarly, we can move from state E_k to any state above it, the net rate at which we leave E_k being $i.g. + i.g. + \dots = \lambda \sum_{i=1}^{\infty} g_i = \lambda$. If, as usual we define P_k to be the equilibrium probability for the number of customers in the system, then we may write down the following equilibrium

* To make the correspondence complete, the parameter for this exponential distribution should indeed be ru . However, in the following development, we will choose the parameter merely to be μ and recall this fact whenever we compare the bulk arrival system to the system $M/ET\{1$.

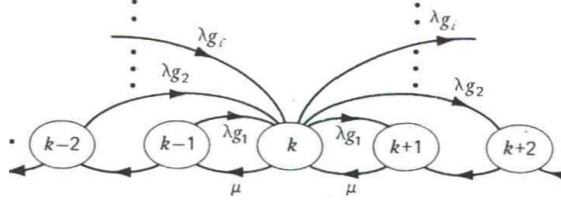


Figure 4.8 The bulk arrival state-transition-rate diagram.

equations using our inspection method:

$$(\lambda + \mu)p_k = \mu p_{k+1} + \sum_{i=0}^{k-1} p_i \lambda g_{k-i} \quad k \geq 1 \quad (4.44)$$

$$\lambda p_0 = \mu p_1 \quad (4.45)$$

Equation (4.44) has equated the rate out of state E_k (the left-hand side) to the rate into that state, where the first term refers to a service completion and the second term (the sum) refers to all possible ways that arrivals may occur and drive us into state E_k from below. Equation (4.45) is the single boundary equation for the state E_0 . As usual, we shall solve these equations using the method of z-transforms; thus we have

$$(\lambda + \mu) \sum_{k=1}^{\infty} p_k z^k = \frac{\mu}{z} \sum_{k=1}^{\infty} p_{k+1} z^{k+1} + \sum_{k=1}^{\infty} \sum_{i=0}^{k-1} p_i \lambda g_{k-i} z^k \quad (4.46)$$

We may interchange the order of summation for the double sum such that

$$\sum_{k=1}^{\infty} \sum_{i=0}^{k-1} = \sum_{i=0}^{\infty} \sum_{k=i+1}^{\infty}$$

and regrouping the terms, we have

$$\begin{aligned} \sum_{k=1}^{\infty} \sum_{i=0}^{k-1} p_i \lambda g_{k-i} z^k &= \lambda \sum_{i=0}^{\infty} p_i z^i \sum_{k=i+1}^{\infty} g_{k-i} z^{k-i} \\ &= \lambda \sum_{i=0}^{\infty} p_i z^i \sum_{j=1}^{\infty} g_j z^j \end{aligned} \quad (4.47)$$

The z-transform we are seeking is

$$P(z) \triangleq \sum_{k=0}^{\infty} p_k z^k$$

and we see from Eq. (4.47) that we should define the z-transform for the distribution of bulk size as*

$$G(z) \triangleq \sum_{k=1}^{\infty} g_k z^k \quad (4.48)$$

• We could just as well have permitted $g_0 > 0$, which would then have allowed zero-size bulks to arrive, and this would have put self-loops in our state-transition diagram corresponding to null arrivals. Had we done so, then the definition for $G(z)$ would have ranged from zero to infinity, and everything we say below applies for this case as well.

Extracting these transforms from Eq. (4.46) we have

$$(\lambda + \mu)[P(z) - P_0] = \frac{\mu}{z} [P(z) - P_0 - P_1 z] + \lambda P(z) G(z)$$

Note that the product $P(z)G(z)$ is a manifestation of property II in Table I. I of Appendix 1, since we have in effect formed the transform of the convolution of the sequence $\{P_k\}$ with that of $\{g_k\}$ in Eq. (4.44). Applying the boundary equation (4.45) and simplifying, we have

$$P(z) = \frac{\mu P_0 (1 - z)}{\mu(1 - z) - \lambda z(1 - G(z))}$$

To eliminate P_0 we use $P(1) = 1$; direct application yields the indeterminate form $\frac{0}{0}$ and so we must use L'Hospital's rule, which gives $P_0 = 1 - \rho$. We obtain

$$P(z) = \frac{\mu(1 - \rho)(1 - z)}{\mu(1 - z) - \lambda z[1 - G(z)]} \quad (4.49)$$

This is the final solution for the transform of number of customers in the bulk arrival M/M/1 system. Once the sequence $\{g_k\}$ is given, we may then face the problem of inverting this transform. One may calculate the mean and variance of the number of customers in the system in terms of the system parameters directly from $P(z)$ (see Exercise 4.8). Let us note that the appropriate definition for the utilization factor ρ must be carefully defined here. Recall that ρ is the average arrival rate of customers times the average service time. In our case, the average arrival rate of customers is the product of the average arrival rate of bulks and the average bulk size. From Eq. (11.29) we have immediately that the average bulk size must be $G'(1)$. Thus we naturally conclude that the appropriate definition for ρ in this system is

$$\rho = \frac{\lambda G'(1)}{\mu} \quad (4.50)$$

It is instructive to consider the special case where all bulk sizes are the same, namely,

$$g_k = \begin{cases} 1 & k = r \\ 0 & k \neq r \end{cases}$$

Clearly, this is the simplified bulk system discussed in the beginning of this section; it corresponds exactly to the system M/Er/1 (where we must make the minor modification as indicated in our earlier footnote that μ must now be replaced by $r\mu$). We find immediately that $G(z) = z^r$ and after substituting this into our solution Eq. (4.49) we find that it corresponds exactly to our earlier solution Eq. (4.26) as, of course, it must.

4.6. BULK SERVICE SYSTEMS

In Section 4.4 we studied the system $ErIM/I$ in which arrivals were considered to have passed through r stages of "arrival." We found it expedient in that case to take as our state variable the number of "arrival stages" that were in the system (where each fully arrived customer still in the system contributed r stages to that count). As we found an analogy between bulk *arrival* systems and the Erlangian service systems of Section 4.3, here also we find an analogy between bulk *service* systems and the Erlangian arrival systems studied in Section 4.4. Thus let us consider an $M/M/I$ system which provides service to groups of size r . That is, when the server becomes free he will accept a "bulk" of exactly r customers from the queue and administer service to them collectively; the service time for this group is drawn from an exponential distribution with parameter μ . If, upon becoming free, the server finds less than r customers in the queue, he then waits until a total of r accumulate and then accepts them for bulk service, and so on.* Customers arrive from a simple Poisson process, at a rate λ , one at a time. It should be clear to the reader that this bulk service system and the $ErIM/I$ are identical. Were we to draw the state-transition-rate diagram for the number of customers in the bulk service system, then we would find exactly the diagram of Figure 4.7 (with the parameter $r\lambda$ replaced by λ ; we must account for this parameter change, however, whenever we compare our bulk service system with the system $Er/M/I$). Since the two systems are equivalent, then the solution for the distribution of number of customers in the bulk service system must be given by Eq. (4.41) (since stages in the original system correspond to customers in the current system).

It certainly seems a waste for our server to remain idle when less than r customers are available for bulk service. Therefore let us now consider a system in which the server will, upon becoming free, accept r customers for bulk service if they are available, or if not will accept less than r if any are available. We take the number of customers in the system as our state variable and find Figure 4.9 to be the state-transition-rate diagram. In this figure we see that all states (except for state E_0) behave in the same way in that they are entered from their left-hand neighbor by an arrival, and from their neighbor r units to the right by a group departure, and they are exited by either an arrival or a group departure; on the other hand, state E_0 can be entered from anyone of the r states immediately to its right and can be exited only by an arrival. These considerations lead directly to the following set of equations for the equilibrium probability P_k of finding k customers in

• For example, the shared taxis in Israel do not (usually) depart until they have collected a full load of customers, all of whom receive service simultaneously.

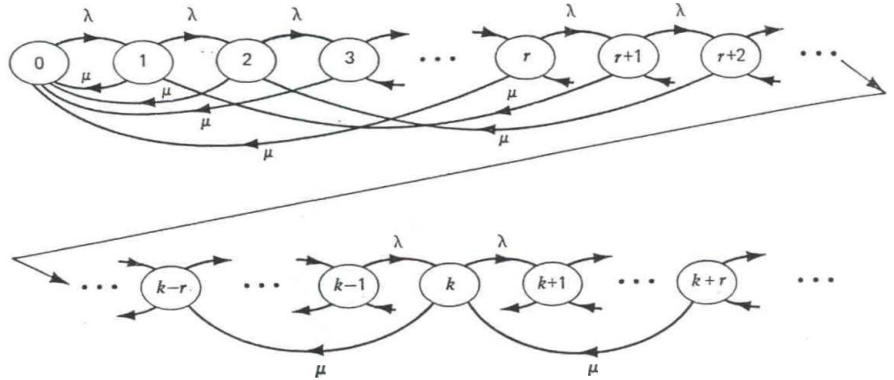


Figure 4.9 The bulk service state-transition-rate diagram.

the system:

$$\begin{aligned}(\lambda + \mu)p_k &= \mu p_{k+r} + \lambda p_{k-1} \quad k \geq 1 \\ \lambda p_0 &= \mu(p_1 + p_2 + \dots + p_r)\end{aligned}\quad (4.51)$$

Let us now apply our z-transform method; as usual we define

$$P(z) = \sum_{k=0}^{\infty} p_k z^k$$

We then multiply by z^k , sum, and then identify $P(z)$ to obtain in the usual way

$$(\lambda + \mu)[P(z) - p_0] = \frac{\mu}{z^r} \left[P(z) - \sum_{k=0}^r p_k z^k \right] + \lambda z P(z)$$

Solving for $P(z)$ we have

$$P(z) = \frac{\mu \sum_{k=0}^r p_k z^k - (\lambda + \mu)p_0 z^r}{\lambda z^{r+1} - (\lambda + \mu)z^r + \mu}$$

From our boundary Eq. (4.51) we see that the negative term in the numerator of this last equation may be written as

$$-z^r(\lambda p_0 + \mu p_0) = -\mu z^r \sum_{k=0}^r p_k$$

and so we have

$$P(z) = \frac{\sum_{k=0}^{r-1} p_k (z^k - z^r)}{r p z^{r+1} - (1 + r p) z^r + 1} \quad (4.52)$$

where we have defined $p = \lambda/\mu r$ since, for this system, up to r customers may be served simultaneously in an interval whose average length is $1/\mu$ sec. We

immediately observe that the denominator of this last equation is precisely the same as in Eq. (4.35) from our study of the system $ET/M/I$. Thus we may give the same arguments regarding the location of the denominator roots; in particular, of the $r + 1$ denominator zeroes, exactly one will occur at the point $z = 1$, exactly $r - 1$ will be such that $|z| < 1$, and only one will be found, which we will denote by z_0 , such that $|z_0| > 1$. Now let us study the numerator of Eq. (4.52). We note that this is a polynomial in z of degree r . Clearly one root occurs at $z = 1$. By arguments now familiar to us, $P(z)$ must remain bounded in the region $|z| < 1$, and so the $r - 1$ remaining zeroes of the numerator must exactly match the $r - 1$ zeroes of the denominator for which $|z| < 1$; as a consequence of this the two polynomials of degree $r - 1$ must be proportional, that is,

$$\frac{K \sum_{k=0}^{T-1} p_k(z^k - zT)}{1 - z} = \frac{rpz^{T+1} - (1 + rp)z^T + 1}{(1 - z)(1 - z/z_0)}$$

Taking advantage of this last equation we may then cancel common factors in the numerator and denominator of Eq. (4.52) to obtain

$$P(z) = \frac{1}{K(1 - z/z_0)}$$

The constant K may be evaluated in the usual way by requiring that $P(1) = 1$, which provides the following simple form for our generating function:

$$P(z) = \frac{1}{1 - z/z_0} \quad (4.53)$$

This last we may invert by inspection to obtain finally the distribution for the number of customers in our bulk service system

$$p_k = \left(1 - \frac{1}{z_0}\right) \left(\frac{1}{z_0}\right)^k \quad k = 0, 1, 2, \dots \quad -(4.54)$$

Once again we see the familiar geometric distribution appear in the solution of our Markovian queueing systems!

4.7. SERIES-PARALLEL STAGES: GENERALIZATIONS

How general is the method of stages studied in Section 4.3 for the system $M/Er/1$ and studied in Section 4.4 for the system ErM/I ? The Erlangian distribution is shown in Figure 4.5; recall that we may select its mean by appropriate choice of μ and may select a range of standard deviations by adjusting r . Note, however, that we are restricted to accept a coefficient of

variation that is less than **that** of the exponential distribution [from Eq. (4.14) we see that $C_b = 1/r$ whereas for $r = 1$ the exponential gives $C_b = 1$] and so in some sense Erlangian random variables are "more regular" than exponential variables. This situation is certainly less than completely general.

One direction for generalization would be to remove the restriction that one of our two basic queueing distributions must be exponential; that is, we certainly could consider the system *ErJErJI* in which we have an *ra*-stage Erlangian distribution for the interarrival times and an *rb*-stage Erlangian distribution for the service times.* On the other hand, we could attempt to generalize by broadening the class of distributions we consider beyond that of the Erlangian. This we do next.

We wish to find a stage-type arrangement that gives larger coefficients of variation than the exponential. One might consider a generalization of the *r*-stage Erlangian in which we permit each stage to have a *different* service rate (say, the *i*th stage has rate μ_i). Perhaps this will extend the range of C_b above unity. In this case we will have instead of Eq. (4.15) a Laplace transform for the service-time pdf given by

$$B^*(s) = \left(\frac{\mu_1}{s + \mu_1} \right) \left(\frac{\mu_2}{s + \mu_2} \right) \cdots \left(\frac{\mu_r}{s + \mu_r} \right) \quad (4.55)$$

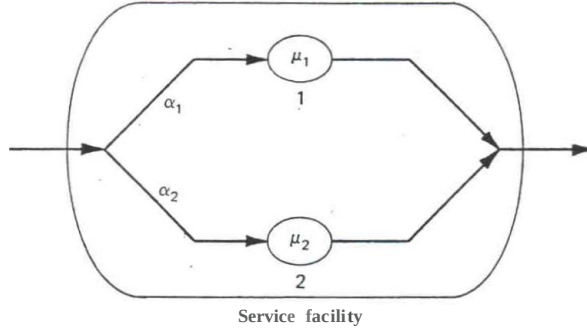
The service time density $h(x)$ will merely be the convolution of *r* exponential densities each with its own parameter μ_i . The squared coefficient of variation in this case is easily shown [see Eq. (11.26), Appendix II] to be

$$C_b^2 = \left(\sum_i \frac{1}{\mu_i^2} \right) / \left(\sum_i \frac{1}{\mu_i} \right)^2$$

But for real $a_i \geq 0$, it is always true that $\sum_i a_i^2 \leq (\sum_i a_i)^2$ since the right-hand side contains the left-hand side plus the sum of all the nonnegative cross terms. Choosing $a_i = 1/\mu_i$, we find that $C_b^2 \leq 1$. Thus, unfortunately, no generalization to larger coefficients of variation is obtained this way.

We previously found that sending a customer through an increasing sequence of faster exponential stages in *series* tended to reduce the variability of the service time, and so one might expect that sending him through a *parallel* arrangement would increase the variability. This in fact is true. Let us therefore consider the two-stage parallel service system shown in Figure 4.10. The situation may be contrasted to the service structure shown in Figure 4.3. In Figure 4.10 an entering customer approaches the large oval (which represents the service facility) from the left. Upon entry into the

* We consider this shortly.

Figure 4.10 A two-stage parallel server H_2 .

facility he will proceed to service stage 1 with probability α_1 or will proceed to service stage 2 with probability α_2 , where $\alpha_1 + \alpha_2 = 1$. He will then spend an exponentially distributed interval of time in the i th such stage whose mean is $1/\mu_i$ sec. After that interval the customer departs and only then is a new customer allowed into the service facility. It is clear from this description that the service time pdf will be given by

$$b(x) = \alpha_1 \mu_1 e^{-\mu_1 x} + \alpha_2 \mu_2 e^{-\mu_2 x} \quad x \geq 0$$

and also we have

$$B^*(s) = \alpha_1 \frac{\mu_1}{s + \mu_1} + \alpha_2 \frac{\mu_2}{s + \mu_2}$$

Of course the more general case with R parallel stages is shown in Figure 4.11. (Contrast this with Figure 4.4.) In this case, as always, at most one customer at any one time is permitted within the large oval representing the service facility. Here we assume that

$$\sum_{i=1}^R \alpha_i = 1 \quad (4.56)$$

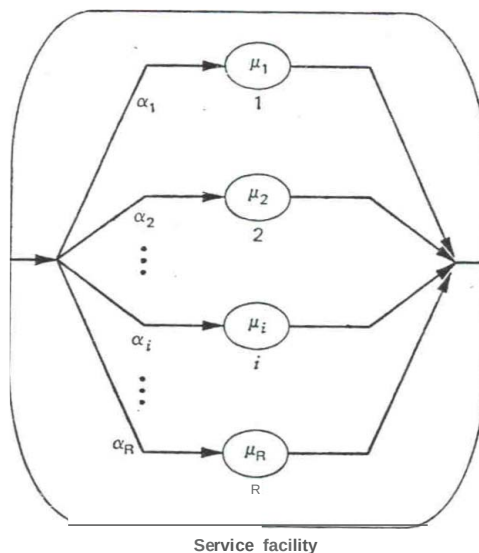
Clearly,

$$b(x) = \sum_{i=1}^R \alpha_i \mu_i e^{-\mu_i x} \quad x \geq 0 \quad - (4.57)$$

and

$$B^*(s) = \sum_{i=1}^R \alpha_i \frac{\mu_i}{s + \mu_i}$$

The pdf given in Eq. (4.57) is referred to as the *hyperexponential* distribution and is denoted by HR . Hopefully, the coefficient of variation ($C_b \triangleq \sigma_b/\bar{x}$) is now greater than unity and therefore represents a wider variation than

Figure 4.11 The R-stage parallel server HR .

that of the exponential. Let us prove this. From Eq. (II.26) we find immediately that

$$\bar{x} = \sum_{i=1}^R \frac{\alpha_i}{\mu_i}$$

$$\overline{x^2} = 2 \sum_{i=1}^R \frac{\alpha_i}{\mu_i^2}$$

Forming the square of the coefficient of variation we then have

$$C_b^2 \triangleq \frac{\sigma_b^2}{\bar{x}^2} = \frac{x^2 - \bar{x}^2}{\bar{x}^2}$$

$$= \frac{2 \sum_{i=1}^R \frac{\alpha_i}{\mu_i^2}}{\left(\sum_{i=1}^R \frac{\alpha_i}{\mu_i} \right)^2} - 1 \quad (4.58)$$

Now, Eq. (II.35), the Cauchy-Schwarz inequality, may also be expressed as follows (for a_i, b_i , real):

$$\left(\sum_i a_i b_i \right)^2 \leq \left(\sum_i a_i^2 \right) \left(\sum_i b_i^2 \right) \quad (4.59)$$

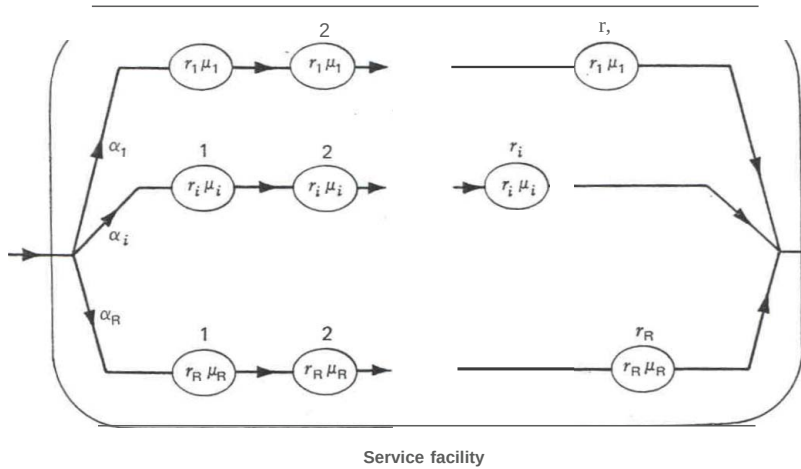


Figure 4.13 Series-parallel server.

stages into arbitrarily complex structures such as shown in Figure 4.13. This diagram shows R parallel "stages," the i th "stage" consisting of an r_i -stage series system ($i = 1, 2, \dots, R$); each stage in the i th series branch is an exponential service facility with parameter $r_i \mu_i$. It is clear that great generality can be built into such series-parallel systems. Within the service facility one and only one of the multitude of stages may be occupied by a customer and no new customer may enter the large oval (representing the service facility) until the previous customer departs. In all cases, however, we note that the state of the service facility is completely contained in the specification of the particular single stage of service in which the customer may currently be found. Clearly the pdf for the service time is calculable directly as above to give

$$b(x) = \sum_{i=1}^R \alpha_i \frac{r_i \mu_i (r_i \mu_i x)^{r_i-1}}{(r_i - 1)!} e^{-r_i \mu_i x} \quad x \geq 0 \quad (4.61)$$

and has a transform given by

$$B^*(s) = \sum_{i=1}^R \alpha_i \left(\frac{r_i \mu_i}{s + r_i \mu_i} \right)^{r_i} \quad (4.62)$$

One further way in which we may generalize our series-parallel server is to remove the restriction that each stage within the same series branch has the same service rate ($r_i \mu_i$); if indeed we permit the j th series stage in the i th

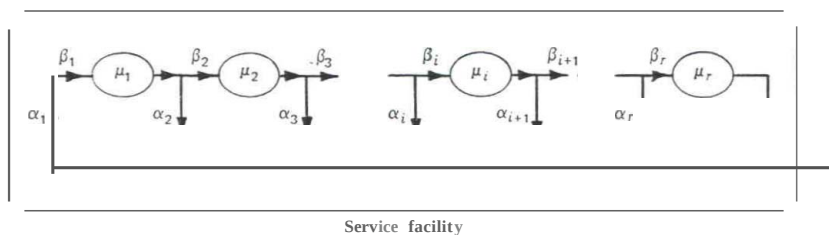


Figure 4.14 Another stage-type server.

parallel branch to have a service rate given by μ_{ij} , then we find that the Laplace transform of the service time density will be generalized to

$$B^*(s) = \sum_{j=1}^r \alpha_j \prod_{i=1}^r \left(\frac{\mu_{ij}}{s + \mu_{ij}} \right) \quad (4.63)$$

These generalities lead to rather complex system equations.

Another way to create the series-parallel effect is as follows. Consider the service facility shown in Figure 4.14. In this system there are r service stages only one of which may be occupied at a given time. Customers enter from the left and depart to the right. Before entering the i th stage an independent choice is made such that with probability β_i the customer will proceed into the i th exponential service stage and with probability α_i he will depart from the system immediately; clearly we require $\beta_i + \alpha_i = 1$ for $i = 1, 2, \dots, r$. After completing the r th stage he will depart from the system with probability 1. One may immediately write down the Laplace transform of the pdf for the system as follows:

$$B^*(s) = \alpha_1 + \sum_{i=1}^r \beta_1 \beta_2 \cdots \beta_i \alpha_{i+1} \prod_{j=1}^i \left(\frac{\mu_j}{s + \mu_j} \right) \quad (4.64)$$

where $\alpha_{r+1} = 1$. One is tempted to consider more general transitions among stages than that shown in this last figure; for example, rather than choosing only between immediate departure and entry into the next stage one might consider feedback or feedforward to other stages. Cox [COX 55] has shown that no further generality is introduced with this feedback and feedforward concept over that of the system shown in Figure 4.14.

It is clear that each of these last three expressions for $B^*(s)$ may be rewritten as a rational function of s , that is, as a ratio of polynomials in s . The positions of the poles (zeroes of the denominator polynomial) for $B^*(s)$ will of necessity be located on the negative real axis of the complex s -plane. This is not quite as general as we would like, since an arbitrary pdf for

service time may have poles located anywhere in the negative half s -plane [that is, for $\text{Re}(s) < 0$]. Cox [COX 55] has studied this problem and suggests that complex values for the exponential parameters $r_i \mu_i$ be permitted; the argument is that whereas this corresponds to no physically realizable exponential stage, so long as we provide poles in complex conjugate pairs then the entire service facility will have a real pdf, which corresponds to the feasible cases. If we permit complex-conjugate pairs of poles then we have *complete generality* in synthesizing any rational function of s for our service-time transform $B^*(s)$. In addition, we have in effect outlined a method of solving these systems by keeping track of the state of the service facility. Moreover, we can similarly construct an interarrival time distribution from series-parallel stages, and thereby we are capable of considering any $G/G/I$ system where the distributions have transforms that are rational functions of s . It is further true that any nonrational function of s may be approximated arbitrarily closely with rational functions.* Thus in principle we have solved a very general problem. Let us discuss this method of solution. The state description clearly will be the number of customers in the system, the stage in which the arriving customer finds himself within the (stage-type) arriving box and the stage in which the customer finds himself in service. From this we may draw a (horribly complicated) state-transition diagram. Once we have this diagram we may (by inspection) write down the equilibrium equations in a rather straightforward manner; this large set of equations will typically have many boundary conditions. However, these equations will all be linear in the unknowns and so the solution method is straightforward (albeit extremely tedious). What more natural setup for a computer solution could one ask for? Indeed, a digital computer is extremely adept at solving large sets of linear equations (such a task is much easier for a digital computer to handle than is a small set of nonlinear equations). In carrying out the digital solution of this (typically infinite) set of linear equations, we must reduce it to a finite set; this can only be done in an approximate way by first deciding at what point we are satisfied in truncating the sequence $P_0, p_1, p_2, \dots$. Then we may solve the finite set and perhaps extrapolate the

* In a real sense, then, we are faced with an approximation problem; how may we "best" approximate a given distribution by one that has a rational transform. If we are given a pdf in numerical form then Prony's method [WHIT 44] is one acceptable procedure. On the other hand, if the pdf is given analytically it is difficult to describe a general procedure for suitable approximation. Of course one would like to make these approximations with the fewest number of stages possible. We comment that if one wishes to fit the first and second moments of a given distribution by the method of stages then the number of stages cannot be significantly less than $1/C_b^2$; unfortunately, this implies that when the distribution tends to concentrate around a fixed value, then the number of stages required grows rather quickly.

solution to the infinite set; all this is in way of approximation and hopefully we are able to carry out the computation far enough so that the neglected terms are indeed negligible.

One must not overemphasize the usefulness of this procedure; this solution method is not as yet automated but does at least in principle provide a method of approach. Other analytic methods for handling the more complex queueing situations are discussed in the balance of this book.

4.8. NETWORKS OF MARKOVIAN QUEUES

We have so far considered Markovian systems in which each customer was demanding a single service operation from the system. We may refer to this as a "single-node" system. In this section we are concerned with multiple-node systems in which a customer requires service at more than one station (node). Thus we may think of a *network of nodes*, each of which is a service center (perhaps with multiple servers at some of the nodes) and each with storage room for queues to form. Customers enter the system at various points, queue for service, and upon departure from a given node then proceed to some other node, there to receive additional service. We are now describing the last category of flow system discussed in Chapter I, namely, stochastic flow in a network.

A number of new considerations emerge when one considers networks. For example, the topological structure of the network is important since it describes the permissible transitions between nodes. Also the paths taken by individual customers must somehow be described. Of great significance is the nature of the stochastic flow in terms of the basic stochastic processes describing that flow; for example, in the case of a tandem queue where customers departing from node i immediately enter node $i + 1$, we see that the interdeparture times from the former generate the interarrival times to the latter. Let us for the moment consider the simple two-node tandem network shown in Figure 4.15. Each oval in that figure describes a queueing system consisting of a queue and server(s); within each oval is given the node number. (It is important not to confuse these physical *network* diagrams with the abstract *state-transition-rate* diagrams we have seen earlier.) For the moment let us assume that a Poisson process generates the arrivals to the system at a rate λ , all of which enter node one; further assume that node one consists of a single exponential server at rate μ . Thus node one is exactly an $M/M/1$ queueing system. Also we will assume that node two has a single



Figure 4.15 A two-node tandem network.

exponential server also of rate μ . The basic question is to solve for the inter-arrival time distribution feeding node two; this certainly will be equivalent to the interdeparture time distribution from node one. Let $d(t)$ be the pdf describing the interdeparture process from node one and as usual let its Laplace transform be denoted by $D^*(s)$. Let us now calculate $D^*(s)$. When a customer departs from node one either a second customer is available in the queue and ready to be taken into service immediately or the queue is empty. In the first case, the time until this next customer departs from node one will be distributed exactly as a service time and in that case we will have

$$D^*(s) | \text{node one nonempty} = B^*(s)$$

On the other hand, if the node is empty upon this first customer's departure then we must wait for the sum of two intervals, the first being the time until the second customer arrives and the next being his service time; since these two intervals are independently distributed then the pdf of the sum must be the convolution of the pdf's for each. Certainly then the transform of the sum pdf will be the product of the transforms of the individual pdfs and so we have

$$D^*(s) | \text{node one empty} = \frac{\lambda}{s + \lambda} B^*(s)$$

where we have given the explicit expression for the transform of the inter-arrival time density. Since we have an exponential server we may also write $B^*(s) = \mu/(s + \mu)$; furthermore, as we shall discuss in Chapter 5 the probability of a departure leaving behind an empty system is the same as the probability of an arrival finding an empty system, namely, $1 - \rho$. This permits us to write down the unconditional transform for the interdeparture time density as

$$D^*(s) = (1 - \rho) D^*(s) | \text{node one empty} + \rho D^*(s) | \text{node one nonempty}$$

Using our above calculations we then have

$$D^*(s) = (1 - \rho) \left(\frac{\lambda}{s + \lambda} \right) \left(\frac{\mu}{s + \mu} \right) + \rho \left(\frac{\mu}{s + \mu} \right)$$

A little algebra gives

$$D^*(s) = \frac{\lambda}{s + \lambda} \quad (4.65)$$

and so the interdeparture time distribution is given by

$$D(t) = 1 - e^{-\lambda t} \quad t \geq 0 \quad \blacksquare$$

Thus we find the remarkable conclusion that the interdeparture times are exponentially distributed with the same parameter as the interarrival times! In other words (in the case of a stable stationary queueing system), a Poisson process driving an exponential server generates a Poisson process for departures. This startling result is usually referred to as *Burke's theorem* [BURK 56]; a number of others also studied the problem (see, for example, the discussion in [SAAT 65]). In fact, Burke's theorem says more, namely, that the steady-state output of a stable M/M/m queue with input parameter λ and service-time parameter μ for each of the m channels is in fact a Poisson process at the same rate λ . Burke also established that the output process was independent of the other processes in the system. It has also been shown that the M/M/m system is the only such FCFS system with this property. Returning now to Figure 4.15 we see therefore that node two is driven by an independent Poisson arrival process and therefore it too behaves like an M/M/1 system and so may be analyzed independently of node one. In fact Burke's theorem tells us that we may connect many multiple-server nodes (each server with exponential pdf) together in a feedforward* network fashion and still preserve this node-by-node decomposition.

Jackson [JACK 57] addressed himself to this question by considering an arbitrary network of queues. The system he studied consists of N nodes where the i th node consists of m_i exponential servers each with parameter μ_i ; further the i th node receives arrivals from outside the system in the form of a Poisson process at rate r_i . Thus if $N = 1$ then we have an M/M/1 system. Upon leaving the i th node a customer then proceeds to the j th node with probability r_{ji} ; this formulation permits the case where $r_{ii} \geq 0$. On the other hand, after completing service in the i th node the probability that the customer departs from the network (never to return again) is given by $1 - \sum_{j=1}^N r_{ji}$. We must calculate the total average arrival rate of customers to a given node. To do so, we must sum the (Poisson) arrivals from outside the system plus arrivals (not necessarily Poisson) from all internal nodes; that is, denoting the total average arrival rate to node i by λ_i we easily find that this set of parameters must satisfy the following equations:

$$\lambda_i = r_i + \sum_{j=1}^N \lambda_j r_{ji} \quad i=1, 2, \dots, N \quad - (4.66)$$

In order for all nodes in this system to represent ergodic Markov chains we require that $r_{ii} < m_i \mu_i$ for all i ; again we caution the reader not to confuse the nodes in this discussion with the system states of each node from our

* Specifically we do not permit feedback paths since this may destroy the Poisson nature of the feedback departure stream. In spite of this, the following discussion of Jackson's work points out that even networks with feedback are such that the individual nodes behave as if they were fed totally by Poisson arrivals, when in fact they are not.

previous discussions. What is amazing is that Jackson was able to show that each node (say the i th) in the network behaves as if it were an independent M/M/m system with a Poisson input rate λ_i . In general, the total input will *not* be a Poisson process. The state variable for this N-node system consists of the vector $(k_1, k_2, \dots, k_N)$, where k_i is the number of customers in the i th node [including the customer(s) in service]. Let the equilibrium probability associated with this state be denoted by $p(k_1, k_2, \dots, k_N)$. Similarly we denote the marginal distribution of finding k_i customers in the i th node by $p_i(k_i)$. Jackson was able to show that the joint distribution for all nodes factored into the product of each of the marginal distributions. that is,

$$p(k_1, k_2, \dots, k_N) = p_1(k_1)p_2(k_2) \cdots p_N(k_N) \quad - (4.67)$$

and $p_i(k_i)$ is given as the solution to the classical M/M/m system [see, for example, Eqs. (3.37)-(3.39) with the obvious change in notation]! This last result is commonly referred to as *Jackson's theorem*. Once again we see the "product" form of solution for Markovian queues in equilibrium.

A modification of Jackson's network of queues was considered by Gordon and Newell [GORD 67]. The modification they investigated was that of a *closed* Markovian network in the sense that a fixed and finite number of customers, say K , are considered to be in the system and are trapped in that system in the sense that no others may enter and none of these may leave: this corresponds to Jackson's case in which $\sum_{j=1}^N r_{ij} = 1$ and $Y_i = 0$ for all i . (An interesting example of this class of systems known as cyclic queues had been considered earlier by Koenigsberg [KOEN 58]; a cyclic queue is a tandem queue in which the last stage is connected back to the first.) In the general case considered by Gordon and Newell we do not quite expect a product solution since there is a dependency among the elements of the state vector $(k_1, k_2, \dots, k_N)$ as follows:

$$\sum_{i=1}^N k_i = K \quad (4.68)$$

As is the case for Jackson's model we assume that this discrete-state Markov process is irreducible and therefore a unique equilibrium probability distribution exists for $p(k_1, k_2, \dots, k_N)$. In this model, however, there is a finite number of states; in particular it is easy to see that the number of distinguishable states of the system is equal to the number of ways in which one can place K customers among the N nodes, and is equal to the binomial coefficient

$$\binom{N + K - 1}{N - 1}$$

The following equations describe the behavior of the equilibrium distribution of customers in this closed system and may be written by inspection as

$$\begin{aligned} p(k_1, k_2, \dots, k_N) \sum_{i=1}^N \delta_{k_i-1} \alpha_i(k_i) \mu_i \\ = \sum_{i=1}^N \sum_{j=1}^N \delta_{k_j-1} \alpha_i(k_i + 1) \mu_i r_{ij} p(k_1, k_2, \dots, k_i - 1, \dots, k_i + 1, \dots, k_N) \end{aligned} \quad (4.69)$$

where the discrete unit step-function defined in Appendix I takes the form

$$\delta_k \triangleq \begin{cases} 1 & k = 0, 1, 2, \dots \\ 0 & k < 0 \end{cases} \quad (4.70)$$

and is included in the equilibrium equations to indicate the fact that the service rate must be zero when a given node is empty; furthermore we define

$$\alpha_i(k_i) = \begin{cases} k_i! & k_i \leq m_i \\ m_i! & k_i \geq m_i \end{cases}$$

which merely gives the number of customers in service in the i th node when there are k_i customers at that node. As usual the left-hand side of Eq. (4.69) describes the flow of probability out of state $(k_1, k_2, \dots, k_N)$ whereas the right-hand side accounts for the flow of probability into that state from neighboring states. Let us proceed to write down the solution to these equations. We define the function $\beta_i(k_i)$ as follows:

$$\beta_i(k_i) = \begin{cases} k_i! & k_i \leq m_i \\ m_i! m_i^{k_i - m_i} & k_i \geq m_i \end{cases}$$

Consider a set of numbers $\{X_i\}$ which are solutions to the following set of linear equations:

$$\mu_i x_i = \sum_{j=1}^N \mu_j x_j r_{ji} \quad i = 1, 2, \dots, N \quad (4.71)$$

Note that this set of equations is in the same form as $\mathbf{1t} = \mathbf{1tP}$ where now the vector $\mathbf{1t}$ may be considered to be $(\mu_1 x_1, \dots, \mu_N x_N)$ and the elements of the matrix $\mathbf{P}$ are considered to be the elements r_{ij} .* Since we assume that the

* Again the reader is cautioned that, on the one hand, we have been considering Markov chains in which the quantities p_{ij} refer to the transition probabilities among the possible states that the system may take on, whereas, on the other hand, we have in this section in addition been considering a network of queueing systems in which the probabilities r_{ij} refer to transitions that customers make between nodes in that network.

matrix of transition probabilities (whose elements are β_{ij}) is irreducible, then by our previous studies we know that there must be a solution to Eqs. (4.71), all of whose components are positive; of course, they will only be determined to within a multiplicative constant since there are only $N - 1$ independent equations there. With these definitions the solution to Eq. (4.69) can be shown to equal

$$p(k_1, k_2, \dots, k_N) = \frac{1}{G(K)} \prod_{i=1}^N \frac{x_i^{k_i}}{\beta_i(k_i)} \quad (4.72)$$

where the normalization constant is given by

$$G(K) = \sum_{\mathbf{k} \in A} \prod_{i=1}^N \frac{x_i^{k_i}}{\beta_i(k_i)} \quad (4.73)$$

Here we imply that the summation is taken over all state vectors $\mathbf{k} \triangleq (k_1, \dots, k_N)$ that lie in the set A , and this is the set of all state vectors for which Eq. (4.68) holds. This then is the solution to the closed finite queueing network problem, and we observe once again that it has the product form.

We may expose the product formulation somewhat further by considering the case where $K \rightarrow \infty$. As it turns out, the quantities x_i/m_i are critical in this calculation; we will assume that there exists a unique such ratio that is largest and we will renumber the nodes such that $x_1/m_1 > x_i/m_i$ ($i \neq 1$). It can then be shown that $p(k_1, k_2, \dots, k_N) \rightarrow 0$ for any state in which $k_1 < \infty$. This implies that an infinite number of customers will form in node one, and this node is often referred to as the "bottleneck" for the given network. On the other hand, however, the marginal distribution $p(k_2, \dots, k_N)$ is well-defined in the limit and takes the form

$$p(k_2, k_3, \dots, k_N) = p_2(k_2)p_3(k_3) \cdots p_N(k_N) \quad (4.74)$$

Thus we see the product solution directly for this marginal distribution and, of course, it is similar to Jackson's theorem in Eq. (4.67); note that in one case we have an open system (one that permits external arrivals) and in the other case we have a closed system. As we shall see in Chapter 4, Volume II, this model has significant applications in time-shared and multi-access computer systems.

Jackson (JACK 63] earlier considered an even more general open queueing system, which includes the closed system just considered as a special case. The new wrinkles introduced by Jackson are, first, that the customer arrival process is permitted to depend upon the total number of customers in the system (using this, he easily creates closed networks) and, second, that the service rate at any node may be a function of the number of customers in that node. Thus defining

$$S(k) \triangleq k_1 + k_2 + \dots + k_N$$

we then permit the total arrival rate to be a function of $S(k)$ when the system state is given by the vector k . Similarly we define the exponential service rate at node i to be μ_{k_i} when there are k_i customers at that node (including those in service). As earlier, we have the node transition probabilities $'_{ij}$ ($i, j = 1, 2, \dots, N$) with the following additional definitions: $'_{0i}$ is the probability that the next externally generated arrival will enter the network at node i ; $'_{i, N+1}$ is the probability that a customer leaving node i departs from the system; and $'_{0, N+1}$ is the probability that the next arrival will require no service from the system and leave immediately upon arrival. Thus we see that in this case $y_i = \gamma_i(S(k))$, where $y(S(k))$ is the total external arrival rate to the system [conditioned on the number of customers $S(k)$ at the moment] from our external Poisson process. It can be seen that the probability of a customer arriving at node i_1 and then passing through the node sequence $i_2, i_3, \dots, i_n$ and then departing is given by $r_{0i_1} r_{i_1 i_2} r_{i_2 i_3} \dots r_{i_n, N+1}$. Rather than seek the solution of Eq. (4.66) for the traffic rates, since they are functions of the total number of customers in the system we rather seek the solution for the following equivalent set:

$$e_i = \gamma_i + \sum_{j=1}^N e_j '_{ji} \quad (4.75)$$

[In the case where the arrival rates are independent of the number in the system then Eqs. (4.66) and (4.75) differ by a multiplicative factor equal to the total arrival rate of customers to the system.] We assume that the solution to Eq. (4.75) exists, is unique, and is such that $e_i \geq 0$ for all i ; this is equivalent to assuming that with probability 1 a customer's journey through the network is of finite length. e_i is, in fact, the expected number of times a customer will visit node i in passing through the network.

Let us define the time-dependent state probabilities as

$$P_k(t) = P[\text{system (vector) state at time } t \text{ is } k] \quad (4.76)$$

By our usual methods we may write down the differential-difference equations governing these probabilities as follows:

$$\begin{aligned} \frac{d}{dt} P_k(t) = & - \left[\gamma(S(k)) + \sum_{i=1}^N \mu_{k_i} (1 - r_{ii}) \right] P_k(t) + \sum_{i=1}^N \gamma(S(k) - 1) r_{0i} P_{k(i^-)}(t) \\ & + \sum_{j=1}^N \mu_{k_j+1} r_{j, N+1} P_{k(i^+)}(t) + \sum_{\substack{i=1 \\ i \neq j}}^N \sum_{j=1}^N \mu_{k_i+1} r_{ji} P_{k(i, j)}(t) \end{aligned} \quad (4.77)$$

where terms are omitted when any component of the vector argument goes negative; $k(i^-) = k$ except for its i th component, which takes on the value

$k, -1$; $\mathbf{k}(i^+) = \mathbf{k}$ except for its i th component, which takes on the value $k, +1$; and $k(i, j) = \mathbf{k}$ except that its i th component is $k, -1$ and its j th component is $k, +1$ where $i \neq j$. Complex as this notation appears its interpretation should be rather straightforward for the reader. Jackson shows that the equilibrium distribution is unique (if it exists) and defines it in our earlier notation to be $\lim_{t \rightarrow \infty} P\mathbf{k}(t) \triangleq p_{\mathbf{k}} \triangleq p(k_1, k_2, \dots, k_N)$ as $t \rightarrow \infty$. In order to give the equilibrium solution for $p_{\mathbf{k}}$ we must unfortunately define the following further notation:

$$F(K) \triangleq \prod_{S(\mathbf{k})=0}^{K-1} y(S(\mathbf{k})) \quad K = 0, 1, 2, \dots \quad (4.78)$$

$$f(\mathbf{k}) \triangleq \prod_{i=1}^N \prod_{j=1}^{k_i} \frac{e_i}{\mu_{ji}} \quad (4.79)$$

$$H(K) \triangleq \sum_{\mathbf{k} \in A} f(\mathbf{k}) \quad (4.80)$$

$$G \triangleq \begin{cases} \sum_{K=0}^{\infty} F(K)H(K) & \text{if the sum converges} \\ \infty & \text{otherwise} \end{cases} \quad (4.81)$$

where the set A shown in Eq. (4.80) is the same as that defined for Eq. (4.73). In terms of these definitions then Jackson's more general theorem states that if $G < \infty$ then a unique equilibrium-state probability distribution exists for the general state-dependent networks and is given by

$$p_{\mathbf{k}} = \frac{1}{G} f(\mathbf{k}) F(S(\mathbf{k})) \quad (4.82)$$

Again we detect the product form of solution. It is also possible to show that in the case when arrivals are independent of the total number in the system [that is, $y \triangleq y(S(\mathbf{k}))$] then even in the case of state-dependent service rates Jackson's first theorem applies, namely, that the joint pdf factors into the product of the individual pdf's given in Eq. (4.67). In fact $PiCk_j$ turns out to be the same as the probability distribution for the number of customers in a single-node system where arrivals come from a Poisson process at rate $y e_j$; and with the state-dependent service rates μ_{k_i} such as we have derived for our general birth-death process in Chapter 3. Thus one impact of Jackson's second theorem is that for the constant-arrival-rate case, the equilibrium probability distributions of number of customers in the system at individual

centers are independent of other centers; in addition, each of these distributions is identical to the well-known single-node service center with the same parameters.* A remarkable result!

This last theorem is perhaps as far as one can get with simple Markovian networks, since it seems to extend Burke's theorem in its most general sense. When one relaxes the Markovian assumption on arrivals and/or service times, then extreme complexity in the interdeparture process arises not only from its marginal distribution, but also from its lack of independence on other state variables.

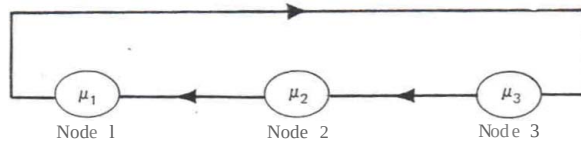
These Markovian queueing networks lead to rather depressing sets of (linear) system equations; this is due to the enormous (yet finite) state description. It is indeed remarkable that such systems do possess reasonably straightforward solutions. The key to solution lies in the observation that these systems may be represented as Markovian population processes, as neatly described by Kingman [KING 69] and as recently pursued by Chandy [CHAN 72]. In particular, a Markov population process is a continuous-time Markov chain over the set of finite-dimensional state vectors $k = (k_1, k_2, \dots, k_N)$ for which transitions are permitted only between states: k and $k(i+)$ (an external arrival at node i); k and $k(i-)$ (an external departure from node i); and k and $k(i, j)$ (an internal transfer from node i to node j). Kingman gives an elegant discussion of the interesting classes and properties of these processes (using the notion and properties of reversible Markov chains). Chandy discusses some of these issues by observing that the equilibrium probabilities for the system states obey not only the global-balance equations that we have so far seen (and typically which lead to product-form solutions) but also that this system of equations may be decomposed into many sets of smaller systems of equations, each of which is simpler to solve. This transformed set is referred to as the set of "local"-balance equations, which we now proceed to discuss.

The concept of local balance is most valuable when one deals with a network of queues. However, the concept does apply to single-node Markovian queues, and in fact we have already seen an example of local balance at play.

* This model also permits one to handle the closed queueing systems studied by Gordon and Newell. In order to create the constant total number of customers one need merely set $y(k) = 0$ for $k \geq K$ and $y(K - 1) = c_0$, where K is the fixed number one wishes to contain within the system. In order to keep the node transition probabilities identical in the open and closed systems, let us denote the former as earlier by r_{ij} ; and the latter now by r_{ij}' ; to make the limit of Jackson's general system equivalent to the closed system of Gordon and Newell we then require $r_{ij}' = r_{ij} + (r_{ij} \cdot N + 1)(rU_{ij})'$

† In Chapter 4, Volume II, we describe some recent results that do in fact extend the model to handle different customer classes and different service disciplines at each node (permitting, in some cases, more general service-time distributions).

‡ See the definitions following Eq. (4.77).

Figure 4.16 A simple cyclic network example: $N = 3$, $K = 2$.

Let us recall the global-balance equations (the flow-conservation equations) for the general birth-death process as exemplified in Eq. (3.6). This equation was obtained by balancing flow into and out of state E_k in Figure 2.9. We also commented at that time that a different boundary could be considered across which flow must be conserved, and this led to the set of equations (3.7). These latter equations are in fact local-balance equations and have the extremely interesting property that they *match* terms from the left-hand side of Eq. (3.6) with corresponding terms on the right-hand side; for example, the term $\lambda_{k-1}p_{k-1}$ on the left-hand side of Eq. (3.6) is seen to be equal to $\mu_k p_k$ on the right-hand side of that equation directly from Eq. (3.7), and by a second application of Eq. (3.7) we see that the two remaining terms in Eq. (3.6) must be equal. This is precisely the way in which local balance operates, namely, to observe that certain sets of terms in the global-balance equation must balance by themselves giving rise to a number of "local"-balance equations.

The significant observation is that, if we are dealing with an ergodic Markov process, then we know for sure that there is a unique solution for the equilibrium probabilities as defined by the generic equation $\pi = \pi P$. Second, if we decompose the global-balance equations for such a process by matching terms of the large global-balance equations into sets of smaller local-balance equations (and of course account for all the terms in the global balance), then any solution satisfied by this large set of local-balance equations must also satisfy the global-balance equations; the converse is not generally true. Thus any solution for the local-balance equations will yield the unique solution for our Markov process.

In the interesting case of a network of queues we define a *local-balance equation* (with respect to a given network state and a network node i) as one that equates the rate of flow out of that network state due to the departure of a customer from node i to the rate of flow into that network state due to the arrival of a customer to node i .^{*} This notion in the case of networks is best illustrated by the simple example shown in Figure 4.16. Here we show the case of a three-node network where the service rate in the i th node is given as

* When service is nonexponential but rather given in terms of a stage-type service distribution, then one equates arrivals to and departures from a given stage of service (rather than to and from the node itself).

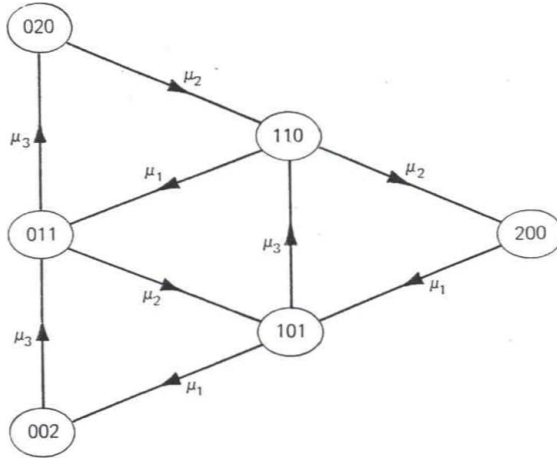


Figure 4.17 State-transition-rate diagram for example in Figure 4.16.

μ_i and is independent of the number of customers at that node; we assume there are exactly $K = 2$ customers circulating in this closed cyclic network. Clearly we have $r_{13} = r_{32} = r_{21} = 1$ and $r_{ij} = 0$ otherwise. Our state description is merely the triplet (k_1, k_2, k_3) , where as usual k_i gives the number of customers in node i and where we require, of course, that $k_1 + k_2 + k_3 = 2$. For this network we will therefore have exactly

$$\binom{N + K - 1}{N - 1} = 6$$

states with state-transition rates as shown in Figure 4.17.

For this system we have six global-balance equations (one of which will be redundant as usual; the extra condition comes from the conservation of probability); these are

$$\mu_1 p(2, 0, 0) = \mu_2 p(1, 1, 0) \quad (4.83)$$

$$\mu_2 p(0, 2, 0) = \mu_3 p(0, 1, 1) \quad (4.84)$$

$$\mu_3 p(0, 0, 2) = \mu_1 p(1, 0, 1) \quad (4.85)$$

$$\mu_1 p(1, 1, 0) + \mu_2 p(1, 1, 0) = \mu_2 p(0, 2, 0) + \mu_3 p(1, 0, 1) \quad (4.86)$$

$$\mu_2 p(0, 1, 1) + \mu_3 p(0, 1, 1) = \mu_3 p(0, 0, 2) + \mu_1 p(1, 1, 0) \quad (4.87)$$

$$\mu_1 p(1, 0, 1) + \mu_3 p(1, 0, 1) = \mu_2 p(0, 1, 1) + \mu_1 p(2, 0, 0) \quad (4.88)$$

Each of these global-balance equations is of the form whereby the left-hand side represents the flow out of a state and the right-hand side represents the flow into that state. Equations (4.83)-(4.85) are already local-balance equations as we shall see; Eqs. (4.86)-(4.88) have been written so that the first term on the left-hand side of each equation balances the first term on the right-hand side of the equation, and likewise for the second terms. Thus Eq. (4.86) gives rise to the following local-balance equations:

$$\mu_1 p(1, 1, 0) = \mu_2 p(0, 2, 0) \quad (4.89)$$

$$\mu_2 p(1, 1, 0) = \mu_3 p(1, 0, 1) \quad (4.90)$$

Note, for example, that Eq. (4.89) takes the rate out of state (1, 1, 0) due to a departure from node 1 and equates it to the rate into that state due to arrivals at node 1; similarly, Eq. (4.90) does likewise for departures and arrivals at node 2. This is the principle of local balance and we see therefore that Eqs. (4.83)-(4.85) are already of this form. Thus we generate nine local-balance equations* (four of which must therefore be redundant when we consider the conservation of probability), each of which is extremely simple and therefore permits a straightforward solution to be found. If this set of equations does indeed have a solution, then they certainly guarantee that the global equations are satisfied and therefore that the solution we have found is the unique solution to the original global equations. The reader may easily verify the following solution:

$$\begin{aligned} p(1, 0, 1) &= \frac{\mu_1}{\mu_3} p(2, 0, 0) \\ p(1, 1, 0) &= \frac{\mu_1}{\mu_2} p(2, 0, 0) \\ p(0, 1, 1) &= \frac{(\mu_1)^2}{\mu_2 \mu_3} p(2, 0, 0) \\ p(0, 0, 2) &= \left(\frac{\mu_1}{\mu_3} \right)^2 p(2, 0, 0) \\ p(0, 2, 0) &= \left(\frac{\mu_1}{\mu_2} \right)^2 p(2, 0, 0) \\ p(2, 0, 0) &= \left[1 + \frac{\mu_1}{\mu_3} + \frac{\mu_1}{\mu_2} + \frac{(\mu_1)^2}{\mu_2 \mu_3} + \left(\frac{\mu_1}{\mu_3} \right)^2 + \left(\frac{\mu_1}{\mu_2} \right)^2 \right]^{-1} \end{aligned} \quad (4.91)$$

Had we allowed all possible transitions among nodes (rather than the cyclic behavior in this example) then the state-transition-rate diagram would have

* The reader should write them out directly from Figure 4.17.

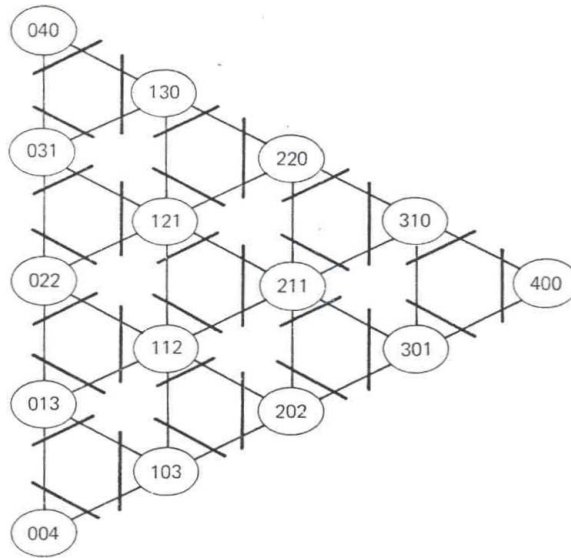


Figure 4.18 State-transition-rate diagram showing local balance ($N = 3, K = 4$).

permitted transitions in both directions where now only unidirectional transitions are permitted; however, it will always be true that only transitions to nearest-neighbor states (in this two-dimensional diagram) are permitted so that such a diagram can always be drawn in a planar fashion. For example, had we allowed four customers in an arbitrarily connected three-node network, then the state-transition-rate diagram would have been as shown in Figure 4.18. In this diagram we represent possible transitions between nodes by an undirected branch (representing two one-way branches in opposite directions). Also, we have collected together sets of branches by joining them with a heavy line, and these are meant to represent branches whose contributions appear in the same local-balance equation. These diagrams can be extended to higher dimensions when there are more than three nodes in the system. In particular, with four nodes we get a tetrahedron (that is, a three-dimensional simplex). In general, with N nodes we will get an $(N - 1)$ -dimensional simplex with $K + 1$ nodes along each edge (where $K =$ number of customers in the closed system). We note in these diagrams that all nodes lying in a given straight line (parallel to any base of the simplex) maintain one component of the state vector at a constant value and that this value increases or decreases by unity as one moves to a parallel set of nodes. The local-balance equations are identified as balancing flow in that set of branches that connects a given node on one of these constant lines to all other nodes on that constant line adjacent and parallel to this node, and that decreases by unity that component that had been held constant. In summary, then, the

local-balance equations are trivial to write down, and if one can succeed in finding a solution that satisfies them, then one has found the solution to the global-balance equations as well!

As we see, most of these Markovian networks lead to rather complex systems of linear equations. Wallace and Rosenberg [WALL 66] propose a numerical solution method for a large class of these equations which is computationally efficient. They discuss a computer program, which is designed to evaluate the equilibrium probability distributions of state variables in very large finite Markovian queueing networks. Specifically, it is designed to solve the equilibrium equations of the form given in Eqs. (2.50) and (2.116), namely, $\pi = \pi P$ and $\pi Q = 0$. The procedure is of the "power-iteration type" such that if $\pi^{(i)}$ is the i th iterate then $\pi^{(i+1)} = \pi^{(i)} R$ is the $(i+1)$ th iterate; the matrix R is either equal to the matrix $\alpha P + (1 - \alpha)I$ (where α is a scalar) or equal to the matrix $\beta Q + I$ (where β is a scalar and I is the identity matrix), depending upon which of the two above equations is to be solved. The scalars α and β are chosen carefully so as to give an efficient convergence to the solution of these equations. The speed of solution is quite remarkable and the reader is referred to [WALL 66] and its references for further details.

Thus ends our study of purely Markovian systems in equilibrium. The unifying feature throughout Chapters 3 and 4 has been that these systems give rise to product-type solutions; one is therefore urged to look for solutions of this form whenever Markovian queueing systems are encountered. In the next chapter we permit either $A(t)$ or $B(x)$ (but not both) to be of arbitrary form, requiring the other to remain in exponential form.

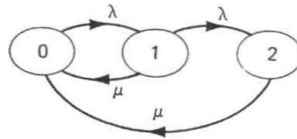
REFERENCES

- BURK 56 Burke, P. J., "The Output of a Queueing System," *Operations Research*, 4, 699-704 (1966).
- CHAN 72 Chandy, K. M., "The Analysis and Solutions for General Queueing Networks," Proc. Sixth Annual Princeton Conference on Information Sciences and Systems, Princeton University, March 1972.
- COX 55 Cox, D. R., "A Use of Complex Probabilities in the Theory of Stochastic Processes," *Proceedings Cambridge Philosophical Society*, 51, 313-319 (1955).
- GORD 67 Gordon, W. J. and G. F. Newell, "Closed Queueing Systems with Exponential Servers," *Operations Research*, 15, 254-265 (1967).
- JACK 57 Jackson, J. R., "Networks of Waiting Lines," *Operations Research*, S, 518-521 (1957).
- JACK 63 Jackson, J. R., "Jobshop-Like Queueing Systems," *Management Science*, 10, 131-142 (1963).
- KING 69 Kingman, J. F. C., "Markov Population Processes," *Journal of Applied Probability*, 6, 1-18 (1969).

- KOEN 58 Koenigsberg, E., "Cyclic Queues," *Operations Research Quarterly*, 9, 22-35 (1958).
- SAAT 65 Saaty, T. L., "Stochastic Network Flows: Advances in Networks of Queues," *Proc. Symp. Congestion Theory*, Univ. of North Carolina Press, (Chapel Hill), 86-107, (1965).
- WALL 66 Wallace, V. L. and R. S. Rosenberg, "Markovian Models and Numerical Analysis of Computer System Behavior," *AFIPS Spring Joint Computer Conference Proc.*, 141-148, (1966).
- WHIT 44 Whittaker, E. and G. Robinson, *The Calculus of Observations*, 4th ed., Blackie (London), (1944).

EXERCISES

- 4.1. Consider the Markovian queueing system shown below. Branch labels are birth and death rates. Node labels give the number of customers in the system.



- Solve for P_k
 - Find the average number in the system.
 - For $\lambda = \mu$, what values do we get for parts (a) and (b)? Try to interpret these results.
 - Write down the transition rate matrix Q for this problem and give the matrix equation relating Q to the probabilities found in part (a).
- 4.2. Consider an $E_k/E_n/1$ queueing system where no queue is permitted to form. A customer who arrives to find the service facility busy is "lost" (he departs with no service). Let E_{ij} be the system state in which the "arriving" customer is in the i th arrival stage and the customer in service is in the j th service stage (note that there is always some customer in the arrival mechanism and that if there is no customer in the service facility, then we let $j = 0$). Let $1/k\lambda$ be the average time spent in any arrival stage and $1/n\mu$ be the average time spent in any service stage.
- Draw the state transition diagram showing all the transition rates.
 - Write down the equilibrium equation for E_{ij} where $0 < i < k$, $0 < j < n$,

- 4.3. Consider an $M/E_r/1$ system in which *no* queue is allowed to form. Let j = the number of stages of service left in the system and let P_j be the equilibrium probability of being in state E_j .
- Find $P_j, j = 0, 1, \dots, r$.
 - Find the probability of a busy system.
- 4.4. Consider an $M/H/1$ system in which *no* queue is allowed to form. Service is of the hyperexponential type as shown in Figure 4.10 with $\mu_1 = 2\mu\alpha_1$ and $\mu_2 = 2\mu(1 - \alpha_1)$.
- Solve for the equilibrium probability of an empty system.
 - Find the probability that server 1 is occupied.
 - Find the probability of a busy system.
- 4.5. Consider an $M/M/1$ system with parameters λ and μ in which exactly two customers arrive at each arrival instant.
- Draw the state-transition-rate diagram.
 - By inspection, write down the equilibrium equations for p_k ($k = 0, 1, 2, \dots$).
 - Let $\rho = 2\lambda/\mu$. Express $P(z)$ in terms of ρ and z .
 - Find $P(z)$ by using the bulk arrival results from Section 4.5.
 - Find the mean and variance of the number of customers in the system from $P(z)$.
 - Repeat** parts (a)-(e) with exactly r customers arriving at each arrival instant (and $\rho = r\lambda/\mu$).
- 4.6. Consider an $M/M/1$ queueing system with parameters i and μ . At each of the arrival instants one new customer will enter the system with probability $1/2$ or two new customers will enter simultaneously with probability $1/2$.
- Draw the state-transition-rate diagram for this system.
 - Using the method of non-nearest-neighbor systems write down the equilibrium equations for P_s .
 - Find $P(z)$ and also evaluate any constants in this expression so that $P(z)$ is given in terms only of i and μ . If possible eliminate any common factors in the numerator and denominator of this expression [this makes life simpler for you in part (d)].
 - From part (c) find the expected number of customers in the system.
 - Repeat part (c) using the results obtained in Section 4.5 directly.
- 4.7. For the bulk arrival system of Section 4.5, assume (for $0 < \alpha < 1$) that
- $$g_i = (1 - \alpha)\alpha^i \quad i = 0, 1, 2, \dots$$
- Find p_k = equilibrium probability of finding k in the system.

For the bulk arrival system studied in Section 4.5, find the mean $\bar{N}$ and variance σ_N^2 for the number of customers in the system. Express your answers in terms of the moments of the bulk arrival distribution.

Consider an M/M/1 system with the following variation: Whenever the server becomes free, he accepts *two* customers (if at least two are available) from the queue into service simultaneously. Of these two customers, only one receives service; when the service for this one is completed, both customers depart (and so the other customer got a "free ride").

If only one customer is available in the queue when the server becomes free, then that customer is accepted alone and is serviced; if a new customer happens to arrive when this single customer is being served, then the new customer joins the old one in service and this new customer receives a "free ride."

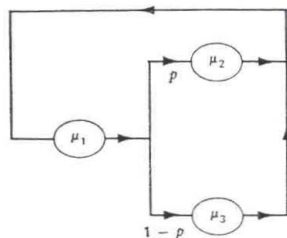
In all cases, the service time is exponentially distributed with mean $1/\mu$ sec and the average (Poisson) arrival rate is λ customers per second.

- Draw the appropriate state diagram.
- Write down the appropriate difference equations for P_k = equilibrium probability of finding k customers in the system.
- Solve for $P(z)$ in terms of P_0 and P_1 .
- Express P_i in terms of P_0 .

We consider the denominator polynomial in Eq. (4.35) for the system $E_r/M/1$. Of the $r + 1$ roots, we know that one occurs at $z = 1$. Use Rouché's theorem (see Appendix I) to show that exactly $r - 1$ of the remaining r roots lie in the unit disk $|z| \leq 1$ and therefore exactly one root, say z_0 , lies in the region $|z_0| > 1$.

Show that the solution to Eq. (4.71) gives a set of variables $\{X_i\}$ which guarantee that Eq. (4.72) is indeed the solution to Eq. (4.69).

- Draw the state-transition-rate diagram showing local balance for the case $(N = 3, K = 5)$ with the following structure:



- Solve for $p(k_1, k_2, k_3)$.

- 4.13.** Consider a two-node Markovian queueing network (of the more general type considered by Jackson) for which $N = 2$, $m_1 = m_2 = 1$, $\mu_{k_i} = \mu_i$ (constant service rate), and which has transition probabilities (r_{ij}) as described in the following matrix:

j	0	2	3
0	0	0	0
1	0	$1 - \alpha$	α
2	0	0	0

where $0 < \alpha < 1$ and nodes 0 and $N + 1$ are the "source" and "sink" nodes, respectively. We also have (for some integer K)

$$\gamma(S(k_1, k_2)) = \begin{cases} \infty & k_1 + k_2 \neq K \\ 0 & k_1 + k_2 = K \end{cases}$$

and assume the system initially contains K customers.

- Find e_i , ($i = 1, 2$) as given in Eq. (4.75).
- Since $N = 2$, let us denote $p(k_1, k_2) = p(k_1, K - k_1)$ by p_{k_1} . Find the balance equations for p_{k_1} .
- Solve these equations for p_{k_1} explicitly.
- By considering the fraction of time the first node is busy, find the time between customer departures from the network (via node 1, of course).

PART III

INTERMEDIATE QUEUEING THEORY

We are here concerned with those queueing systems for which we can still apply certain simplifications due to their Markovian nature. We encounter those systems that are representable as *imbedded* Markov chains, namely, the $M/G/1$ and the $G/M/m$ queues. In Chapter 5 we rapidly develop the basic equilibrium equations for $M/G/1$ giving the notorious *Pollaczek-Khinchin* equations for queue length and waiting time. We next discuss the busy period and, finally, introduce some moderately advanced techniques for studying these systems, even commenting a bit on the time-dependent solutions. Similarly for the queue $G/M/m$ in Chapter 6, we find that we can make some very specific statements about the equilibrium system behavior and, in fact, find that the conditional distribution of waiting time will always be exponential regardless of the interarrival time distribution! Similarly, the conditional queue-length distribution is shown to be geometric. We note in this part that the methods of solution are quite different from that studied in Part II, but that much of the underlying behavior is similar; in particular the mean queue size, the mean waiting time, and the mean busy period duration all are inversely proportional to $1 - \rho$ as earlier. In Chapter 7 we briefly investigate a rather pleasing interpretation of transforms in terms of probabilities.

The techniques we had used in Chapter 3 [the explicit product solution of Eq. (3.11)] and in Chapter 4 (flow conservation) are replaced by an indirect a -transform approach in Chapter 5. However, in Chapter 6, we return once again to the flow conservation inherent in the $\pi = \pi P$ solution.

S

The Queue MjGjl

That which makes elementary queueing theory so appealing is the simplicity of its state description.* In particular, all that is required in order to summarize the entire past history of the queueing system is a specification of the number of customers present. All other historical information is irrelevant to the future behavior of pure Markovian systems. Thus the state description is not only one dimensional but also countable (and in some cases finite). It is this latter property (the countability) that simplifies our calculations.

In this chapter and the next we study queueing systems that are driven by non-Markovian stochastic processes. As a consequence we are faced with new problems for which we must find new methods of solution.

In spite of the non-Markovian nature of these two systems there exists an abundance of techniques for handling them. Our approach in this chapter will be the method of the *imbedded Markov chain* due to Palm [PALM 43] and Kendall [KEND 51]. However, we have in reality already seen a second approach to this class of problems, namely, the *method of stages*, in which it was shown that so long as the interarrival time and service time pdf's have Laplace transforms that are rational, then the stage method can be applied (see Section 4.7); the disadvantage of that approach is that it merely gives a procedure for carrying out the solution but does not show the solution as an explicit expression, and therefore properties of the solution cannot be studied for a class of systems. The third approach, to be studied in Chapter 8, is to solve *Lindley's integral equation* [LIND 52]; this approach is suitable for the system G/G/1 and so obviously may be specialized to some of the systems we consider in this chapter. A fourth approach, the *method of supplementary*

* Usually a state description is given in terms of a vector which describes the system's state at time t . A vector $\mathbf{v}(t)$ is a *state vector* if, given $\mathbf{v}(t)$ and all inputs to this system during the interval (t, t_1) (where $t < t_1$), then we are capable of solving for the state vector $\mathbf{v}(t_1)$. Clearly it behooves us to choose a state vector containing that information that permits us to calculate quantities of importance for understanding system behavior.

† We saw in Chapter 4 that occasionally we record the number of stages in the system rather than the number of customers.

variables, is discussed in the exercises at the end of this chapter; more will be said about this method in the next section. We also discuss the *busy period analysis* [GAVE 59], which leads to the waiting-time distribution (see Section 5.10). Beyond these there exist other approaches to non-Markovian queueing systems, among which are the *random-walk* and *combinatorial* approaches [TAKA 67] and the *method of Green's function* [KEIL 65].

5.1. THE M/G/1 SYSTEM

The M/G/I queue is a single-server system with Poisson arrivals and arbitrary service-time distribution denoted by $B(x)$ [and a service time pdf denoted by $b(x)$]. That is, the interarrival time distribution is given by

$$A(t) = 1 - e^{-\lambda t} \quad t \geq 0$$

with an average arrival rate of λ customers per second, a mean interarrival time of $1/\lambda$ sec, and a variance $\sigma_a^2 = 1/\lambda^2$. As defined in Chapter 2 we denote the k th moment of service time by

$$x^k \triangleq \int_0^\infty x^k b(x) dx$$

and we sometimes express these service time moments by $b_k \triangleq x^k$.

Let us discuss the state description (vector) for the M/G/I system. If at some time t we hope to summarize the complete past history of this system, then it is clear that we must certainly specify $N(t)$, the number of customers present at time t . Moreover, we must specify $Xo(t)$, the service time already received by the customer in service at time t ; this is necessary since the service-time distribution is not necessarily of the memoryless type. (Clearly, we need not specify how long it has been since the last arrival entered the system, since the arrival process is of the memoryless type.) Thus we see that the random process $N(t)$ is a non-Markovian process. However, the vector $[N(t), Xo(t)]$ is a Markov process and is an appropriate state vector for the M/G/I system, since it completely summarizes all past history relevant to the future system development.

We have thus gone from a single-component description of state in elementary queueing theory to what appears to be a two-component description here in intermediate queueing theory. Let us examine the inherent difference between these two state descriptions. In elementary queueing theory, it is sufficient to provide $N(t)$, the number in the system at time t , and we then have a Markov process with a discrete-state space, where the states themselves are either finite or countable in number. When we proceed to the current situation where we need a two-dimensional state description, we find that the number in the system $N(t)$ is still denumerable, but now we must also

provide $X_0(t)$, the expended service time, which is *continuous*. We have thus evolved from a discrete-state description to a continuous-state description, and this essential difference complicates the analysis.

It is possible to proceed with a general theory based upon the couplet $[N(t), X_0(t)]$ as a state vector and such a method of solution is referred to as the *method of supplementary variables*. For a treatment of this sort the reader is referred to Cox [COX 55] and Kendall [KEND 53]; Henderson [HEND 72] also discusses this method, but chooses the remaining service time instead of the expended service time as **the** supplementary variable. In this text we choose to use the *method of the imbedded Markov chain* as discussed below. However, before we proceed with the method itself, it is clear that we should understand some properties of the expended service time; this we do in the following section.

5.2. THE PARADOX OF RESIDUAL LIFE: A **BIT OF** RENEWAL THEORY

We are concerned here with the case where an arriving customer finds a partially served customer in the service facility. Problems of this sort occur repeatedly in our studies, and so we wish to place this situation in a more general context. We begin with an apparent paradox illustrated through the following example. Assume that our hippie from Chapter 2 arrives at a roadside cafe at an arbitrary instant in time and begins hitchhiking. Assume further that automobiles arrive at this cafe according to a Poisson process at an average rate of λ cars per minute. How long must the hippie wait, on the average, until the next car comes along?

There are two apparently logical answers to this question. First, we might argue that since the average time between automobile arrivals is $1/\lambda$ min, and since the hippie arrives at a random point in time, then "obviously" the hippie will wait on the average $1/2\lambda$ min. On the other hand, we observe that since the Poisson process is memoryless, the time until the next arrival is independent of how long it has been since the previous arrival and therefore the hippie will wait on the average $1/\lambda$ min; this second argument can be extended to show that the average time from the last arrival until the hippie begins hitchhiking is also $1/\lambda$ min. The second solution therefore implies that the average time between the last car and the next car to arrive will be $2/\lambda$ min! It appears that this interval is twice as long as it should be for a Poisson process! Nevertheless, the second solution is the correct one, and so we are faced with an apparent paradox!

Let us discuss the solution to this problem in the case of an arbitrary interarrival time distribution. This study properly belongs to renewal theory, and we quote results freely from that field; most of these results can be found

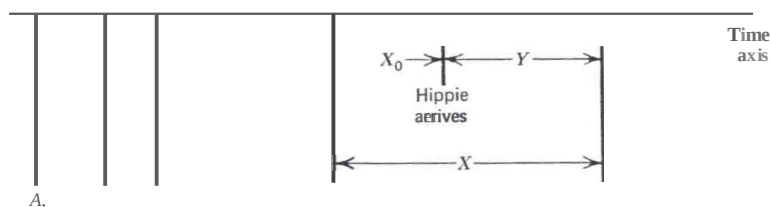


Figure 5.1 Life, age and residual life.

in the excellent monograph by Cox [COX 62] or in the fine expository article by Smith [SMIT 58]; the reader is also encouraged to see Feller [FELL 66]. The basic diagram is that given in Figure 5.1. In this figure we let A_k denote the k th automobile, which we assume arrives at time τ_k . We assume that the intervals $\tau_{k+1} - \tau_k$ are independent and identically distributed random variables with distribution given by

$$F(x) \triangleq P[\tau_{k+1} - \tau_k \leq x] \quad (5.1)$$

We further define the common pdf for these intervals as

$$f(x) \triangleq \frac{dF(x)}{dx} \quad (5.2)$$

Let us now choose a random point in time, say t , when our hippie arrives at the roadside cafe. In this figure, A_{n-1} is the last automobile to arrive prior to t and A_n will be the first automobile to arrive after t . We let X denote this "special" interarrival time and we let Y denote the time that our hippie must wait until the next arrival. Clearly, the sequence of arrival points $\{T_k\}$ forms a renewal process; renewal theory discusses the instantaneous replacement of components. In this case, $\{T_k\}$ forms the sequence of instants when the old component fails and is replaced by a new component. In the language of renewal theory X is said to be the *lifetime* of the component under consideration, Y is said to be the *residual life* of that component at time t , and $X_0 = X - Y$ is referred to as the *age* of that component at time t . Let us adopt that terminology and proceed to find the pdf for X and Y , the lifetime and residual life of our selected component. We assume that the renewal process has been operating for an arbitrarily long time since we are interested only in limiting distributions.

The amazing result we will find is that X is not distributed according to $F(x)$. In terms of our earlier example this means that the interval which the hippie happens to select by his arrival at the cafe is not a typical interval. In fact, herein lies the solution to our paradox: A long interval is more likely

to be "intercepted" by our hippie than a short one. In the case of a Poisson process we shall see that this bias causes the selected interval to be on the average twice as long as a typical interval.

Let the residual life have a distribution

$$\hat{F}(x) \triangleq \text{pr } Y \leq x \quad (5.3)$$

with density

$$j(x) = \frac{d\hat{F}(x)}{dx} \quad (5.4)$$

Similarly, let the selected lifetime X have a pdf $f(x)$ and PDF $F(x)$ where

$$F(x) \triangleq P[X \leq x] \quad (5.5)$$

In Exercise 5.2 we direct the reader through a rigorous derivation for the residual lifetime density $j(x)$. Rather than proceed through those details, let us give an intuitive derivation for the density that takes advantage of our physical intuition regarding this problem. Our basic observation is that long intervals between renewal points occupy larger segments of the time axis than do shorter intervals, and therefore it is more likely that our random point t will fall in a long interval. If we accept this, then we recognize that the probability that an interval of length x is chosen should be proportional to the length x as well as to the relative occurrence of such intervals [which is given by $I(x) dx$]. Thus, for the selected interval, we may write

$$J(x) dx = K x J(x) dx \quad (5.6)$$

where the left-hand side is $P[x < X \leq x + dx]$ and the right-hand side expresses the linear weighting with respect to interval length and includes a constant K , which must be evaluated so as to properly normalize this density. Integrating both sides of Eq. (5.6) we find that $K = 1/jm$ where

$$jm \triangleq E[T - \tau_{k-1}] \quad (5.7)$$

and is the common average time between renewals (between arrivals of automobiles). Thus we have shown that the density associated with the selected interval is given in terms of the density of typical intervals by

$$f_X(x) = \frac{xJ(x)}{jm} \quad (5.8)$$

This is our first result. Let us proceed now to find the density of residual life $j(x)$. If we are told that $X = x$, then the probability that the residual life Y does not exceed the value y is given by

$$\text{pr } Y \leq y \mid X = x = \frac{y}{x}$$

for $0 \leq y \leq x$; this last is true since we have randomly chosen a point within this selected interval, **and** therefore this point must be uniformly distributed within that interval. Thus we may write down the joint density of X and Y as

$$\begin{aligned} \text{pr } y < Y \leq y + dy, x < X \leq x + dx &= \left(\frac{dy}{x} \right) \left(\frac{xf(x)}{m_1} dx \right) \\ &= \frac{f(x) dy dx}{m_1} \end{aligned} \quad (5.9)$$

for $0 \leq y \leq x$. Integrating over x we obtain $f(y)$, which is the unconditional density for Y , namely,

$$f(y) dy = \int_{x=y}^{\infty} \frac{f(x) dy dx}{m_1}$$

This immediately gives the final result:

$$f(y) = \frac{1 - F(y)}{m_1} \quad (5.10)$$

This is our second result. It gives the density of residual life in terms of the common distribution of interval length and its mean.*

Let us express this last result in terms of transforms. Using our usual transform notation we have the following correspondences:

$$f(x) \leftrightarrow F^*(s)$$

$$\hat{f}(x) \leftrightarrow I^*(s)$$

Clearly, all the random variables we have been discussing in this section are nonnegative, and so the relationship in Eq. (5.10) may be transformed directly by use of entry 5 in Table 1.4 and entry 13 in Table 1.3 to give

$$\hat{F}^*(s) = \frac{1 - F^*(s)}{sm_1} \quad (5.11)$$

It is now a trivial matter to find the moments of residual life in terms of the moments of the lifetimes themselves. We denote the n th moment of the lifetime by m_n ; and the l th moment of the residual life by r_l ; that is,

$$m_n \triangleq E[(T_k - T_{k-1})^n] \quad (5.12)$$

$$r_l \triangleq E[Y^l] \quad (5.13)$$

Using our moment formula Eq. (11.26), we may differentiate Eq. (5.11) to obtain the moments of residual life. As $s \rightarrow 0$ we obtain indeterminate for m_s

• It may also be shown that the limiting pdf for age (0) is the same as for residual life (Y) given in Eq. (5.10).

which may be evaluated by means of L'Hospital's rule; this computation gives the moments of residual life as

$$r_n = \frac{\ln n + 1}{(n + 1)m_1} \quad -(5.14)$$

This important formula is most often used to evaluate '1' the mean residual life, which is found equal to

$$r_1 = \frac{m_2}{2m_1} \quad - (5.15)$$

and may also be expressed in terms of the lifetime variance (denoted by $\sigma^2 \triangleq m_2 - m_1^2$) to give

$$1 = \frac{m_1}{2} + \frac{\sigma^2}{2m_1} \quad (5.16)$$

This last form shows that the correct answer to the hippie paradox is $m_1/2$, half the mean interarrival time, *only* if the variance is zero (regularly spaced arrivals); however, for the Poisson arrivals, $m_1 = 1/\lambda$ and $\sigma^2 = 1/\lambda^2$, giving $1 = 1/\lambda = m_1$, which confirms our earlier solution to the hippie paradox of residual life. Note that $m_1/2 \leq 1$ and 1 will grow without bound as $\sigma^2 \rightarrow \infty$. The result for the mean residual life (1) is a rather counterintuitive result; we will see it appear again and again.

Before leaving renewal theory we take this opportunity to quote some other useful results. In the language of renewal theory the age-dependent failure rate rex) is defined as the instantaneous rate at which a component will fail given that it has already attained an age of x ; that is, $rex) dx \triangleq P[x < \text{lifetime of component} \leq x + dx | \text{lifetime} > x]$. From first principles, we see that this conditional density is

$$rex) = \frac{f(x)}{1 - F(x)} \quad - (5.17)$$

where once again $f(x)$ and $F(x)$ refer to the common distribution of component lifetime. The *renewal function* $H(x)$ is defined to be

$$H(x) \triangleq E[\text{number of renewals in an interval of length } x] \quad (5.18)$$

and the *renewal density* $hex)$ is merely the renewal rate at time x defined by

$$hex) \triangleq \frac{dH(x)}{dx} \quad (5.19)$$

Renewal theory seems to be obsessed with limit theorems, and one of the important results is the *renewal theorem*, which states that

$$\lim_{x \rightarrow \infty} hex) = \frac{1}{nH} \quad (5.20)$$

This merely says that in the limit one cannot identify when the renewal process began, and so the rate at which components are renewed is equal to the inverse of the average time between renewals (m). We note that $hex)$ is not a pdf; in fact, its integral diverges in the typical case. Nevertheless, it does possess a Laplace transform which we denote by $H^*(s)$. It is easy to show that the following relationship exists between this transform and the transform of the underlying pdf for renewals, namely:

$$H^*(s) = \frac{F^*(s)}{1 - F^*(s)} \quad (5.21)$$

This last is merely the transform expression of the *integral equation of renewal theory*, which may be written as

$$hex) = f(x) + \int_0^x h(x - t)f(t) dt \quad (5.22)$$

More will not be said about renewal theory at this point. Again the reader is urged to consult the references mentioned above.

5.3. THE IMBEDDED MARKOV CHAIN

We now consider the method of the imbedded Markov chain and apply it to the M/G/1 queue. The fundamental idea behind this method is that we wish to simplify the description of state from the two-dimensional description $[N(t), Xo(t)]$ into a one-dimensional description $N(t)$. If indeed we are to be successful in calculating future values for our state variable we must also implicitly give, along with this one-dimensional description of the number in system, the time expended on service for the customer in service. Furthermore (and here is the crucial point), we agree that we may gain this simplification by looking not at all points in time but rather at a select set of points in time. Clearly, these special epochs must have the property that, if we specify the number in the system at one such point and also provide future inputs to the system, then at the next suitable point in time we can again calculate the number in system; thus somehow we must implicitly be specifying the expended service for the man in service. How are we to identify a set of points with this property? There are many such sets. An extremely convenient set of points with this property is the set of *departure* instants from service. It is

clear if we specify the number of customers left behind by a departing customer that we can calculate this same quantity at some point in the future given only the additional inputs to the system. Certainly, we have specified the expended service time at these instants: it is in fact zero for the customer (if any) currently in service since he has just at that instant entered service!*

(There are other sets of points with this property, for example, the set of points that occur exactly 1 sec after customers enter service; if we specify the number in the system at these instants, then we are capable of solving for the number of customers in the system at such future instants of time. Such a set as just described is not as useful as the departure instants since we must worry about the case where a customer in service does not remain for a duration exceeding 1 sec.)

The reader should recognize that what we are describing is, in fact, a semi-Markov process in which the state transitions occur at customer departure instants. At these instants we define the imbedded Markov chain to be the number of customers present in the system immediately following the departure. The transitions take place only at the imbedded points and form a discrete-state space. The distribution of time between state transitions is equal to the service time distribution $R(x)$ whenever a departure leaves behind at least one customer, whereas it equals the convolution of the interarrival-time distribution (exponentially distributed) with $b(x)$ in the case that the departure leaves behind an empty system. In any case, the behavior of the chain at these imbedded points is completely describable as a Markov process, and the results we have discussed in Chapter 2 are applicable.

Our approach then is to focus attention upon departure instants from service and to specify as our state variable the *number of customers left behind* by such a departing customer. We will proceed to solve for the system behavior at these instants in time. Fortunately, the solution at these imbedded Markov points happens also to provide the solution for *all* points in time.†

In Exercise 5.7 the reader is asked to rederive some *M/G/1* results using the method of supplementary variables; this method is good at all points in time and (as it must) turns out to be identical to the results we get here by using the imbedded Markov chain approach. This proves once again that our solution

* Moreover, we assume that no service has been expended on any other customer in the queue.

† This happy circumstance is due to the fact that we have a Poisson input and therefore (as shown in Section 4.1) an arriving customer takes what amounts to a "random" look at the system. Furthermore, in Exercise 5.6 we assist the reader in **proving** that the limiting distribution for the number of customers left behind by a departure is the same as the limiting distribution of customers found by a new arrival for any system that changes state by unit step values (positive or negative); this result is true for arbitrary arrival- and arbitrary service-time distributions! Thus, for **M/G/1**, arrivals, departures, and random observers all see the *same* distribution of number in the system.

is good for all time. In the following pages we establish results for the queue-length distribution, the waiting-time distribution, and the busy-period distribution (all in terms of transforms); the waiting-time and busy-period duration results are in no way restricted by the imbedding we have described. So even if the other methods were not available, these results would still hold and would be unconstrained due to the imbedding process. As a final reassurance to the reader we now offer an intuitive justification for the equivalence between the limiting distributions seen by departures and arrivals. Taking the state of the system as "the number of customers therein, we may observe the changes in system state as time evolves; if we follow the system state in continuous time, then we observe that these changes are of the nearest-neighbor type. In particular, if we let E_k be the system state when k customers are in the system, then we see that the only transitions from this state are $E_k \rightarrow E_{k+1}$ and $E_k \rightarrow E_{k-1}$ (where this last can only occur if $k > 0$). This is denoted in Figure 5.2. We now make the observation that the number of transitions of the type $E_k \rightarrow E_{k+1}$ can differ by at most one from the number of transitions of the type $E_{k+1} \rightarrow E_k$. The former correspond to customer arrivals and occur at the arrival instants; the latter refer to customer departures and occur at the departure instants. After the system has been in operation for an arbitrarily long time, the number of such transitions upward must essentially equal the number of transitions downward. Since this up-and-down motion with respect to E_k occurs with essentially the same frequency, we may therefore conclude that the system states found by arrivals must have the same limiting distribution (r_k) as the system states left behind by departures (which we denote by d_k). Thus, if we let $N(t)$ be the number in the system at time t , we may summarize our two conclusions as follows:

1. For *Poisson* arrivals, it is always true that [see Eq. (4.6)]

$$P[N(t) = k] = P[\text{arrival at time } t \text{ finds } k \text{ in system}]$$
that is,

$$P_k(t) = R_k(t) \quad (5.23)$$
2. If in any (perhaps non-Markovian) system $N(t)$ makes only discontinuous changes of size (plus or minus) one, then if either one of the following limiting distributions exists, so does the other and they are equal (see Exercise 5.6):

$$\begin{aligned}
 r_k &\triangleq \lim_{t \rightarrow \infty} P[\text{arrival at } t \text{ finds } k \text{ customers in system}] \\
 d_k &\triangleq \lim_{t \rightarrow \infty} P[\text{departure at } t \text{ leaves } k \text{ customers behind}] \\
 r_k &= d_k
 \end{aligned} \quad (5.24)$$

Thus, for *M/G/1*,

$$r_k = p_k = d_k$$

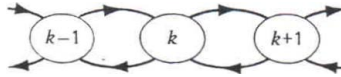


Figure 5.2 State transitions for unit step-change systems.

Our approach for the balance of this chapter is first to find the mean number in system, a result referred to as the Pollaczek-Khinchin mean-value formula.* Following that we obtain the generating function for the distribution of number of customers in the system and then the transform for both the waiting-time and total **system-time** distributions. These last transform results we shall refer to as Pollaczek-Khinchin transform equations.* Furthermore, we solve for the transform of the busy-period duration and for the number served in the busy period; we then show how to derive waiting-time results from the busy-period analysis. Lastly, we derive the **Takács** integrodifferential equation for the unfinished work in the system. We begin by defining some notation and identifying the transition probabilities associated with our *imbedded* Markov chain.

5.4. THE **TRANSITION** PROBABILITIES

We have already discussed the use of customer departure instants as a set of imbedded points in the time axis; at these instants we define the imbedded Markov chain as the number of customers left behind by these departures (this forms our imbedded Markov chain). It should be clear to the reader that this is a complete state description since we know for sure that zero service has so far been expended on the customer in service and that the time since the last arrival is irrelevant to the future development of the process, since the interarrival-time distribution is memoryless. Early in Chapter 2 we introduced some symbolical and graphical notation; we ask that the reader refresh his understanding of Figure 2.2 and that he recall the following definitions:

C_n represents the n th customer to enter the system

r ; = arrival time of C ;

t ; = $\tau_n - \tau_{n-1}$ = interarrival time between C_{n-1} and C ;

x_n = service time for C_n

In addition, we introduce two new random variables of considerable interest:

qn = number of customers left behind by departure of C_n from service

v_n = number of customers arriving during the service of C_n

• There is considerable disagreement within the queueing theory literature regarding the names for the mean-value and transform equations. Some authors refer to the mean-value expression as the Pollaczek-Khinchin formula, whereas others reserve that term for the transform equations. We attempt to relieve that confusion by adding the appropriate adjectives to these names.

We are interested in solving for the distribution of q'' , namely, $Piq; = kj$, which is, in fact, a time-dependent probability; its limiting distribution (as $n \rightarrow \infty$) corresponds to ek' which we know is equal to p_k , the basic distribution discussed in Chapters 3 and 4 previously. In carrying out that solution we will find that the number of arriving customers v_n plays a crucial role.

As in Chapter 2, we find that the transition probabilities describe our Markov chain; thus we define the one-step transition probabilities

$$p_{ij} \triangleq P[q_{n+1} = j | q_n = i] \quad (5.25)$$

Since these transitions are observed only at departures, it is clear that $q_{n+1} < q_n - 1$ is an impossible situation; on the other hand, $q_{n+1} \geq q_n - 1$ is possible for all values due to the arrivals $v_n + 1$. It is easy to see that the matrix of transition probabilities $P = [P_{ij}]$ ($i, j = 0, 1, 2, \dots$) takes the following form:

$$P = \begin{vmatrix} \alpha_0 & \alpha_1 & \alpha_2 & \alpha_3 \\ \alpha_0 & \alpha_1 & \alpha_2 & \alpha_3 \\ 0 & \alpha_0 & \alpha_1 & \alpha_2 \\ 0 & 0 & \alpha_0 & \alpha_1 \\ 0 & 0 & 0 & \alpha_0 \end{vmatrix}$$

where

$$\alpha_k \triangleq P[v_n + 1 = k] \quad (5.26)$$

For example, the j th component of the first row of this matrix gives the probability that the previous customer left behind an empty system and that during the service of C_{n+1} exactly j customers arrived (all of whom were left behind by the departure of $C_n + 1$); similarly, for other than the first row, the entry p_{ij} for $j \geq i - 1$ gives the probability that exactly $j - i + 1$ customers arrived during the service period for C_{n+1} , given that C_n left behind exactly i customers; of these i customers one was indeed $C_n + 1$ and this accounts for the $+1$ term in this last computation. The state-transition-probability diagram for this Markov chain is shown in Figure 5.3, in which we show only transitions out of E_i .

Let us now calculate α_k . We observe first of all that the arrival process (a Poisson process at a rate of λ customers per second) is independent of the state of the queueing system. Similarly, x''' the service time for C_n , is independent

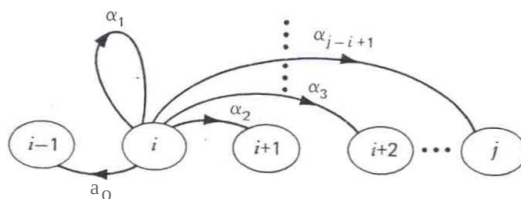


Figure 5.3 State-transition-probability diagram for the $M/G/I$ imbedded Markov Chain.

of u and is distributed according to $B(x)$. Therefore, v_n , the number of arrivals during the service time x_n depends only upon the duration of x_n and not upon u at all. We may therefore dispense with the subscripts on v_n and x_n , replacing them with the random variables $\tilde{v}$ and $\tilde{x}$ so that we may write $P[x_n \leq x] = P[\tilde{x} \leq x] = B(x)$ and $P[v_n = k] = P[u = k] = \alpha_k$. We may now proceed with the calculation of α_k . We have by the law of total probability

$$\alpha_k = P[u = k] = \int_0^\infty P[u = k, x < \tilde{x} \leq x + dx] dx$$

By conditional probabilities we further have

$$\alpha_k = \int_0^\infty P[u = k \mid \tilde{x} = x] b(x) dx \quad (5.27)$$

where again $b(x) = dB(x)/dx$ is the pdf for service time. Since we have a Poisson arrival process, we may replace the probability beneath the integral by the expression given in Eq. (2.131), that is,

$$\alpha_k = \int_0^\infty \frac{(\lambda x)^k}{k!} e^{-\lambda x} b(x) dx \quad (5.28)$$

This then completely specifies the transition probability matrix P .

We note that since $\alpha_k > 0$ for all $k \geq 0$ it is possible to reach all other states from any given state; thus our Markov chain is irreducible (and aperiodic). Moreover, let us make our usual definition:

$$\rho = \lambda \bar{x}$$

and point out that this Markov chain is ergodic if $\rho < 1$ (unless specified otherwise, we shall assume $\rho < 1$ below).

The stationary probabilities may be obtained from the vector equation $p = pP$ where $p = [p_0, p_1, p_2, \dots]$ whose k th component $p_k (= \dot{p}_k)$ is

merely the limiting probability that a departing customer will leave behind k customers, namely,

$$p_k = P[\tilde{q} = k] \quad (5.29)$$

In the following section we find the mean value $E[\tilde{q}]$ and in the section following that we find the z-transform for p_k .

5.5. THE MEAN QUEUE LENGTH

In this section we derive the Pollaczek-Khinchin formula for the mean value of the limiting queue length. In particular, we define

$$\tilde{q} = \lim q_n \quad (5.30)$$

which certainly will exist in the case where our imbedded chain is ergodic.

Our first step is to find an equation relating the random variable q_{n+1} to the random variable q_n by considering two cases. The first is shown in Figure 5.4 (using our time-diagram notation) and corresponds to the case where C_n leaves behind a nonempty system (i.e., $q_n > 0$). Note that we are assuming a first-come-first-served queueing discipline, although this assumption only affects waiting times and not queue lengths or busy periods. We see from Figure 5.4 that q_n is clearly greater than zero since c_{n+1} is already in the system when C_n departs. We purposely do not show when customer C_{n+2} arrives since that is unimportant to our developing argument. We wish now to find an expression for q_{n+1} , the number of customers left behind when c_{n+1} departs. This is clearly given as equal to q_n the number of customers present when C_n departed less 1 (since customer c_{n+1} departs himself) plus the number of customers that arrive during the service interval x_{n+1} . This last term is clearly equal to D_{n+1} by definition and is shown as a "set" of arrivals

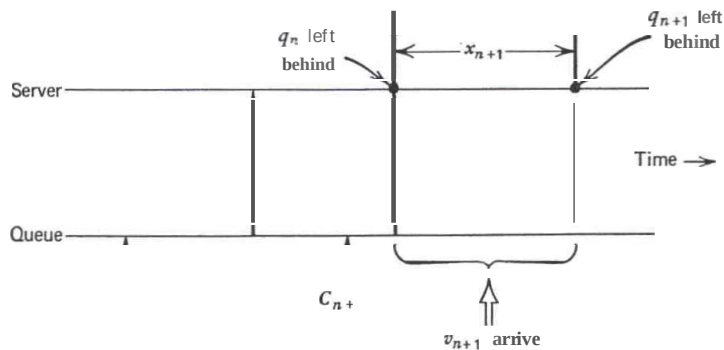
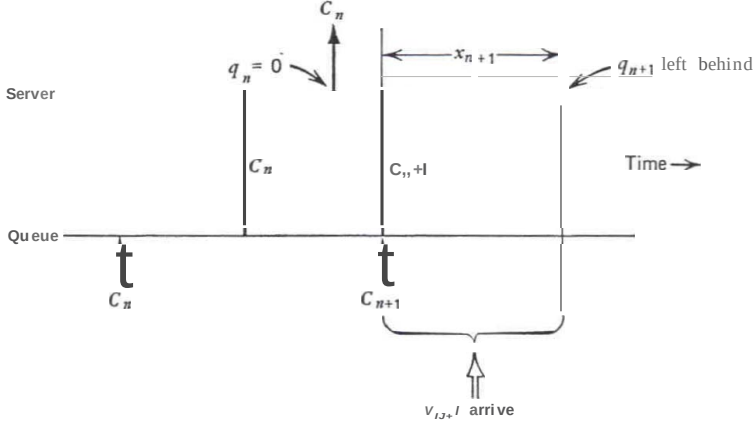


Figure 5.4 Case where $q_n > 0$.

Figure 5.5 Case where $q_n = 0$.

in the diagram. Thus we have

$$q_{n+1} = q_n - 1 + v_{n+1} \quad q_n > 0 \quad (5.31)$$

Now consider the second case where $q_n = 0$, that is, our departing customer leaves behind an empty system; this is illustrated in Figure 5.5. In this case we see that q_n is clearly zero since C_{n+1} has not yet arrived by the time C_n departs. Thus q_{n+1} , the number of customers left behind by the departure of C_{n+1} , is merely equal to the number of arrivals during his service time. Thus

$$q_{n+1} = v_{n+1} \quad q_n = 0 \quad (5.32)$$

Collecting together Eq. (5.31) and Eq. (5.32) we have

$$q_{n+1} = \begin{cases} q_n - 1 + v_{n+1} & q_n > 0 \\ v_{n+1} & q_n = 0 \end{cases} \quad (5.33)$$

It is convenient at this point to introduce Δ_k , the shifted discrete step function

$$\Delta_k = \begin{cases} 1 & k = 1, 2, \dots \\ 0 & k \leq 0 \end{cases} \quad (5.34)$$

which is related to the discrete step function δ_k [defined in Eq. (4.70)] through $\Delta_k = \delta_{k-1}$. Applying this definition to Eq. (5.33) we may now write the single defining equation for q_{n+1} as

$$q_{n+1} = q_n - \Delta_{q_n} + v_{n+1} \quad (5.35)$$

Equation (5.35) is the key equation for the study of M/G/I systems. It remains for us to extract from Eq. (5.35) the mean value* for qn . As usual, we concern ourselves not with the time-dependent behavior (which is inferred by the subscript n) but rather with the limiting distribution for the random variable qn , which we denote by $\tilde{q}$. Accordingly we assume that the j th moment of qn exists in the limit as n goes to infinity independent of n , namely,

$$\lim_{n \rightarrow \infty} E[q_n^j] = E[\tilde{q}^j] \quad (5.36)$$

(We are in fact requiring ergodicity here.)

As a first attempt let us hope that forming the expectation of both sides of Eq. (5.35) and then taking the limit as $n \rightarrow \infty$ will yield the average value we are seeking. Proceeding as described we have

$$E[q_{n+1}] = E[q_n] - E[\Delta_{q_n}] + E[v_{n+1}]$$

Using Eq. (5.36) we have, in the limit as $n \rightarrow \infty$,

$$E[\tilde{q}] = E[\tilde{q}] - E[\Delta_{\tilde{q}}] + E[\tilde{v}]$$

Alas, the expectation we were seeking drops out of this equation, which yields instead

$$E[\Delta_{\tilde{q}}] = E[\tilde{v}] \quad (5.37)$$

What insight does this last equation provide us? (Note that since $\tilde{v}$ is the number of arrivals during a customer's service time, which is independent of n , the index on u ; could have been dropped even before we went to the limit.) We have by definition that

$$E[\tilde{v}] = \text{average number of arrivals in a service time}$$

Let us now interpret the left-hand side of Eq. (5.37). By definition we may calculate this directly as

$$\begin{aligned} E[\Delta_{\tilde{q}}] &= \sum_{k=0}^{\infty} \Delta_k P[\tilde{q} = k] \\ &= \Delta_0 P[\tilde{q} = 0] + \Delta_1 P[\tilde{q} = 1] + \dots \end{aligned}$$

• We could at this point proceed to the next section to obtain the (z-transform of the) limiting distribution for number in system and from that expression evaluate the average number in system. Instead, let us calculate the average number in system directly from Eq. (5.35) following the method of Kendall [KENO 51]; we choose to carry out this extra work to demonstrate to the student the simplicity of the argument.

But, from the definition in Eq. (5.34) we may rewrite this as

$$E[\Delta_{\tilde{q}}] = 0\{P[\tilde{q} = 0]\} + 1\{P[\tilde{q} > 0]\}$$

or

$$E[\Delta_{\tilde{q}}] = P[\tilde{q} > 0] \quad (5.38)$$

Since we are dealing with a single-server system, Eq. (5.38) may also be written as

$$E[\Delta_{\tilde{q}}] = P[\text{busy system}] \quad (5.39)$$

And from our definition of the utilization factor we further have

$$P[\text{busy system}] = \rho \quad (5.40)$$

as we had observed* in Eq. (2.32). Thus from Eqs. (5.37), (5.39), and (5.40) we conclude that

$$E[\tilde{v}] = \rho \quad - \quad (5.41)$$

We thus have the perfectly reasonable conclusion that the expected number of arrivals per service interval is equal to $\rho (= \lambda \bar{x})$. For stability we of course require $\rho < 1$, and so Eq. (5.41) indicates that customers must arrive more slowly than they can be served (on the average).

We now return to the task of solving for the expected value of $\tilde{q}$. Forming the *first* moment of Eq. (5.35) yielded interesting results but failed to give the desired expectation. Let us now attempt to find this average value by *first squaring* Eq. (5.35) and *then* taking expectations as follows:

$$q_{n+1}^2 = q_n^2 + \Delta_{q_n}^2 + v_{n+1}^2 - 2q_n \Delta_{q_n} + 2q_n v_{n+1} - 2\Delta_{q_n} v_{n+1} \quad (5.42)$$

From our definition in Eq. (5.34) we have $(\Delta_{q_n})^2 = \Delta_{q_n}$ and also $q_n \Delta_{q_n} = q_n$. Applying this to Eq. (5.42) and taking expectations, we have

$$E[q_{n+1}^2] = E[q_n^2] + E[\Delta_{q_n}] + E[v_{n+1}^2] - 2E[q_n] + 2E[q_n v_{n+1}] - 2E[\Delta_{q_n} v_{n+1}]$$

In this equation, we have the expectation of the product of two random variables in the last two terms. However, we observe that v_{n+1} [the number of arrivals during the $(n+1)$ th service interval] is independent of q_n (the number of customers left behind by n). Consequently, the last two expectations may each be written as a product of the expectations. Taking the limit as n goes to infinity, and using our limit assumptions in Eq. (5.36), we have

$$0 = E[\Delta_{\tilde{q}}] + E[\tilde{v}^2] - 2E[\tilde{q}] + 2E[\tilde{q}]E[\tilde{v}] - 2E[\Delta_{\tilde{q}}]E[\tilde{v}]$$

* For any M/G/1 system, we see that $P[\tilde{q} = 0] = 1 - P[\tilde{q} > 0] = 1 - \rho$ and so $P[\text{new customer need 1101 queue}] = 1 - \rho$. This agrees with our earlier observation for G/G/1.

We now make use of Eqs. (5.37) and (5.41) to obtain, as an intermediate result for the expectation of $\tilde{q}$,

$$E[\tilde{q}] = \rho + \frac{E[\tilde{v}^2] - E[\tilde{v}]}{2(1 - \rho)} \quad (5.43)$$

The only unknown here is $E[\tilde{v}^2]$.

Let us solve not only for the second moment of $\tilde{v}$ but, in fact, let us describe a method for obtaining *all* the moments, Equation (5.28) gives an expression for $\alpha_k = P[\tilde{v} = k]$. From this expression we should be able to calculate the moments. However, we find it expedient first to define the z-transform for the random variable $\tilde{v}$ as

$$V(z) \triangleq E[z^{\tilde{v}}] \triangleq \sum_{k=0}^{\infty} P[\tilde{v} = k] z^k \quad (5.44)$$

Forming $V(z)$ from Eqs. (5.28) and (5.44) we have

$$V(z) = \sum_{k=0}^{\infty} \int_0^{\infty} \frac{(\lambda x)^k}{k!} e^{-\lambda x} b(x) dx z^k$$

Our summation and integral are well behaved, and we may interchange the order of these two operations to obtain

$$\begin{aligned} V(z) &= \int_0^{\infty} e^{-\lambda x} \left(\sum_{k=0}^{\infty} \frac{(\lambda x z)^k}{k!} \right) b(x) dx \\ &= \int_0^{\infty} e^{-\lambda x} e^{\lambda x z} b(x) dx \\ &= \int_0^{\infty} e^{-(\lambda - \lambda z)x} b(x) dx \end{aligned} \quad (5.45)$$

At this point we define (as usual) the Laplace transform $B^*(s)$ for the service time pdf as

$$B^*(s) \triangleq \int_0^{\infty} e^{-sx} b(x) dx$$

We note that Eq. (5.45) is of this form, with the complex variable s replaced by $\lambda - \lambda z$, and so we recognize the important result that

$$V(z) = B^*(\lambda - \lambda z) \quad (5.46)$$

This last equation is extremely useful and represents a relationship between the z-transform of the probability distribution of the random variable $\tilde{v}$ and the Laplace transform of the pdf of the random variable $\tilde{x}$ when the Laplace transform is evaluated at the critical point $\lambda - \lambda z$. These two random variables are such that $\tilde{v}$ represents the number of arrivals occurring during the

interval $\tilde{x}$ where the arrival process is Poisson at an average rate of λ arrivals per second. We will shortly have occasion to incorporate this interpretation of Eq. (5.46) in our further results.

From Appendix II we note that various derivatives of z-transforms evaluated for $z = 1$ give the various moments of the random variable under consideration. Similarly, the appropriate derivative of the Laplace transform evaluated at its argument $s = 0$ also gives rise to moments. In particular, from that appendix we recall that

$$B^{*(k)}(0) \triangleq \left. \frac{dB^*(s)}{ds^k} \right|_{s=0} = (-1)^k E[\tilde{x}^k] \quad (5.47)$$

$$V(l(1)) \triangleq \left. \frac{dV(z)}{dz} \right|_{z=1} = E[ii] \quad (5.48)$$

$$V^{(2)}(1) \triangleq \left. \frac{d^2V(z)}{dz^2} \right|_{z=1} = E[ii^2] - E[ii] \quad (5.49)$$

In order to simplify the notation for these limiting derivative operations, we have used the more usual superscript notation with the argument replaced by its limit. Furthermore, we now resort to the overbar notation to denote expected value of the random variable below that bar.† Thus Eqs. (5.47)-(5.49) become

$$B^*(0) = (-1)^k \bar{x}^k \quad (5.50)$$

$$V(l(1)) = \bar{v} \quad (5.51)$$

$$V(2l(1)) = \bar{v}^2 - \bar{v} \quad (5.52)$$

Of course, we must also have the conservation of probability given by

$$B^*(0) = V(1) = 1 \quad (5.53)$$

We now wish to exploit the relationship given in Eq. (5.46) so as to be able to obtain the moments of the random variable $\tilde{v}$ from the expressions given in Eqs. (5.50)-(5.53). Thus from Eq. (5.46) we have

$$\frac{dV(z)}{dz} = \frac{dB^*(z - \lambda z)}{dz} \quad (5.54)$$

† Recall from Eq. (2.19) that $E[x_n^k] \rightarrow \bar{x}^k = \bar{b}_k$ (rather than the more cumbersome notation $(\tilde{x})^k$ which one might expect). We take the same liberties with $\tilde{v}$ and $\tilde{q}$, namely, $(\tilde{v})^k = \bar{v}^k$ and $(\tilde{q})^k = \bar{q}^k$.

This last may be calculated as

$$\begin{aligned}\frac{dB^*(\lambda - \lambda z)}{dz} &= \left(\frac{dB^*(\lambda - \lambda z)}{d(\lambda - \lambda z)} \right) \left(\frac{d(\lambda - \lambda z)}{dz} \right) \\ &= -\lambda \frac{dB^*(y)}{dy}\end{aligned}\quad (5.55)$$

where

$$y = \lambda - \lambda z \quad (5.56)$$

Setting $z = 1$ in Eq. (5.54) we have

$$V'(1) = -\lambda \left. \frac{dB^*(y)}{dy} \right|_{z=1}$$

But from Eq. (5.56) the case $z = 1$ is the case $y = 0$, and so we have

$$V'(1) = -\lambda B^{*(1)}(0) \quad (5.57)$$

From Eqs. (5.50), (5.51), and (5.57), we finally have

$$\bar{v} = \lambda \bar{x} \quad (5.58)$$

But $\lambda \bar{x}$ is just ρ and we have once again established that which we knew from Eq. (5.41), namely, $\bar{v} = \rho$. (This certainly is encouraging.) We may continue to pick up higher moments by differentiating Eq. (5.54) once again to obtain

$$\frac{d^2 V(z)}{dz^2} = \frac{d^2 B^*(\lambda - \lambda z)}{dz^2} \quad (5.59)$$

Using the first derivative of $B^*(y)$ we now form its second derivative as follows:

$$\begin{aligned}\frac{d^2 B^*(\lambda - \lambda z)}{dz^2} &= \frac{d}{dz} \left[-\lambda \frac{dB^*(y)}{dy} \right] \\ &= -\lambda \left(\frac{d^2 B^*(y)}{dy^2} \right) \left(\frac{dy}{dz} \right)\end{aligned}$$

or

$$\frac{d^2 B^*(\lambda - \lambda z)}{dz^2} = \lambda^2 \frac{d^2 B^*(y)}{dy^2} \quad (5.60)$$

Setting z equal to 1 in Eq. (5.59) and using Eq. (5.60) we have

$$V^{(2)}(1) = \lambda^2 B^{*(2)}(0)$$

Thus, from earlier results in Eqs. (5.50) and (5.52), we obtain

$$v^2 - \bar{v} = \lambda^2 x^2 \quad - \quad (5.61)$$

We have thus finally solved for v^2 . This clearly is the quantity required in order to evaluate Eq. (5.43). If we so desired (and with suitable energy) we could continue this differentiation game and extract additional moments of $\bar{v}$ in terms of the moments of $\bar{x}$; we prefer not to yield to that temptation here.

Returning to Eq. (5.43) we apply Eq. (5.61) to obtain

$$ij = P + \frac{\lambda^2 x^2}{2(1 - \rho)} \quad (5.62)$$

This is the result we were after! It expresses the average queue size at customer departure instants in terms of known quantities, namely, the utilization factor ($\rho = \lambda \bar{x}$), λ , and x^2 (the second moment of the service-time distribution). Let us rewrite this result in terms of $C_b^2 = \sigma_b^2/(\bar{x})^2$, the squared coefficient of variation for service time :

$$\bar{q} = \frac{\rho}{1 - \rho} + \frac{\rho}{2(1 - \rho)} (1 + C_b^2) \quad - \quad (5.63)$$

This last is the extremely well-known formula for the average number of customers in an M/G/1 system and is commonly* referred to as the *Pollaczek-Khinchin (P-K) mean-value formula*. Note with emphasis that this average depends only upon the *first two* moments ($\bar{x}$ and x^2) of the service-time distribution. Moreover, observe that ij grows *linearly* with the variance of the service-time distribution (or, if you will, linearly with its squared coefficient of variation).

The P-K mean-value formula provides an expression for ij that represents the average number of customers in the system at departure instants; however, we already know that this also represents the average number at the arrival instants and, in fact, at all points in time. We already have a notation for the average number of customers in the system, namely $\bar{N}$, which we introduced in Chapter 2 and have used in previous chapters; we will continue to use the $\bar{N}$ notation outside of this chapter. Furthermore, we have defined $\bar{N}_q$ to be the average number of customers in the queue (not counting the customer in service). Let us take a moment to develop a relationship between these two quantities. By definition we have

$$\bar{N} \triangleq \sum_{k=0}^{\infty} k P[\tilde{q} = k] \quad (5.64)$$

* See footnote on p. 177.

Similarly we may calculate the average queue size by subtracting unity from this previous calculation so long as there is at least one customer in the system, that is (note the lower limit),

$$\bar{N}_q = \sum_{k=1}^{\infty} (k-1)P[\tilde{q} = k]$$

This easily gives us

$$\bar{N}_q = \sum_{k=0}^{\infty} kP[\tilde{q} = k] - \sum_{k=1}^{\infty} P[\tilde{q} = k]$$

But the second sum is merely p and so we have the result

$$\bar{N}_q = \bar{N} - p \quad - (5.65)$$

This simple formula gives the general relationship we were seeking.

As an example of the P-K mean-value formula, in the case of an *MIMfI* system, we have that the coefficient of variation for the exponential distribution is unity [see Eq. (2.145)]. Thus for this system we have

$$\frac{\bar{q}}{q - p} + \frac{p}{P} = \frac{2}{2(1 - p)} \quad (2)$$

or

$$\bar{q} = \frac{p}{1 - P} \quad \text{MIMfI} \quad (5.66)$$

Equation (5.66) gives the expected number of customers left behind by a departing customer. Compare this to the expression for the average number of customers in an *MIMfI* system as given in Eq. (3.24). They are identical and lend validity to our earlier statements that the method of the imbedded Markov chain in the *MIGfI* case gives rise to a solution that is good at all points in time. As a second example, let us consider the service-time distribution in which service time is a constant and equal to $\bar{x}$. Such systems are described by the notation *MIDfI*, as we mentioned earlier. In this case clearly $C_b^2 = 0$ and so we have

$$\frac{\bar{q}}{q - P} + \frac{p}{P} = \frac{1}{2(1 - p)} \quad - (5.67)$$

$$\bar{q} = \frac{p}{1 - P} - \frac{p}{2(1 - p)} \quad \text{MIDfI}$$

Thus the *MIDfI* system has $p^2/2(1 - p)$ fewer customers on the average than the *MIMfI* system, demonstrating the earlier statement that $\bar{q}$ increases with the variance of the service-time distribution.

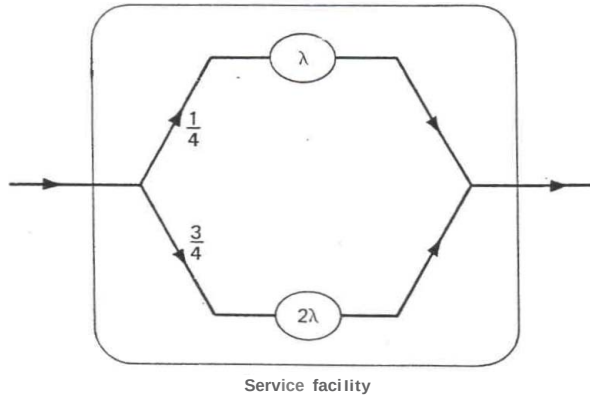


Figure 5.6 The $M/H_2/1$ example.

For a third example, we consider an $M/H_2/1$ system in which

$$b(x) = \frac{1}{4}\lambda e^{-\lambda x} + \frac{3}{4}(2\lambda)e^{-2\lambda x} \quad x \geq 0 \quad (5.6S)$$

That is, the service facility consists of two parallel service stages, as shown in Figure 5.6. Note that λ is also the arrival rate, as usual. We may immediately calculate $\bar{x} = 5/(S)$ and $\rho^2 = 31/(64\lambda^2)$, which yields $C_b^2 = 31/25$. Thus

$$\begin{aligned} \bar{q} &= \bar{p} + \frac{\rho^2(2.24)}{2(1 - \rho)} \\ &= \frac{\rho}{1 - \rho} + \frac{0.12\rho^2}{1 - \rho} \end{aligned}$$

Thus we see the (small) increase in $\bar{q}$ for the (small) increase in C_b^2 over the value of unity for $M/M/1$. We note in this example that ρ is fixed at $\rho = \lambda\bar{x} = 5/S$; therefore, $\bar{q} = 1.79$, whereas for $M/M/1$ at this value of ρ we get $\bar{q} = 1.66$. We have introduced this $M/H_2/1$ example here since we intend to carry it (and the $M/M/1$ example) through our $MIG/1$ discussion.

The main result of this section is the Pollaczek-Khinchin formula for the mean number in system, as given in Eq. (5.63). This result becomes a special case of our results in the next section, but we feel that its development has been useful as a pedagogical device. Moreover, in obtaining this result we established the basic equation for $MIG/1$ given in Eq. (5.35). We also obtained the general relationship between $V(z)$ and $B^*(z)$, as given in Eq. (5.46); from this we are able to obtain the moments for the number of arrivals during a service interval.

We have not as yet derived any results regarding *time* spent in the system; we are now in a position to do so. We recall Little's result:

$$\bar{N} = \lambda T$$

This result relates the expected number of customers $\bar{N}$ in a system to λ , the arrival rate of customers and to T , their average time in the system. For *MIG//* we have derived Eq. (5.63), which is the expected number in the system at customer departure instants. We may therefore apply Little's result to this expected number in order to obtain the average time spent in the system (queue + service). We know that $\bar{q}$ also represents the average number of customers found at random, and so we may equate $\bar{q} = \bar{N}$. Thus we have

$$\bar{N} = \rho + \rho \frac{(1 + C_b^2)}{2(1 - \rho)} = \lambda T$$

Solving for T we have

$$T = \bar{x} + \frac{\rho \bar{x}(1 + C_b^2)}{2(1 - \rho)} \quad (5.69)$$

This last is easily interpreted. The average total time spent in system is clearly the average time spent in service plus the average time spent in the queue. The first term above is merely the average service time and thus the second term must represent the average queueing time (which we denote by W). Thus we have that the average queueing time is

$$W = \frac{\rho \bar{x}(1 + C_b^2)}{2(1 - \rho)}$$

or

$$W = \frac{W_0}{1 - \rho} \quad - (5.70)$$

where $W_0 \triangleq \lambda \bar{x}^2 / 2$; W_0 is the average remaining service time for the customer (if any) found in service by a new arrival (work it out using the mean residual life formula). A particularly nice normalization factor is now apparent. Consider T , the average time spent in system. It is natural to compare this time to $\bar{x}$, the average service time required of the system by a customer. Thus the ratio $T/\bar{x}$ expresses the ratio of time spent in system to time required of the system and represents the *factor* by which the system inconveniences

customers due to the fact that they are sharing the system with other customers. If we use this normalization in Eqs. (5.69) and (5.70), we arrive at the following, where now time is expressed in units of average service intervals:

$$\frac{T}{\bar{x}} = \frac{1 + \rho(1 + C_b^2)}{2(1 - \rho)} \quad (5.71)$$

$$\frac{W}{\bar{x}} = \frac{\rho(1 + C_b^2)}{2(1 - \rho)} \quad (5.72)$$

Each of these last two equations is also referred to as the P-K mean-value formula [along with Eq. (5.63)]. Here we see the linear fashion in which the statistical fluctuations of the input processes create delays (i.e., $1 + C_b^2$ is the sum of the squared interarrival-time and service-time coefficients of variation). Further, we see the highly nonlinear dependence of delays upon the average load ρ .

Let us now compare the mean normalized queueing time for the systems" $M/M/1$ and $M/D/1$; these have a squared coefficient of variation C_b^2 equal to 1 and 0, respectively. Applying this to Eq. (5.72) we have

$$\frac{W}{\bar{x}} = \frac{\rho}{(1 - \rho)} \quad \text{MIM1} \quad (5.73)$$

$$\frac{W}{\bar{x}} = \frac{\rho}{2(1 - \rho)} \quad \text{MID1} \quad (5.74)$$

Note that the system with constant service time ($M/D/1$) has *half* the average waiting time of the system with exponentially distributed service time ($M/M/1$). Thus, as we commented earlier, the time in the system and the number in the system both grow in proportion to the variance of the service-time distribution.

Let us now proceed to find the *distribution* of the number in the system.

5.6. DISTRIBUTION OF NUMBER IN SYSTEM

In the previous sections we characterized the $M/G/1$ queueing system as an imbedded Markov chain and then established the fundamental equation (5.35) repeated here:

$$q_{n+1} = q_n - \Delta_{q_n} + v_{n+1} \quad (5.75)$$

By forming the average of this last equation we obtained a result regarding the utilization factor ρ [see Eq. (5.41)]. By first *squaring* Eq. (5.75) and then

• Of less interest is our highly specialized $M/H_2/1$ example for which we obtain $W/\bar{x} = 1.12\rho/(1 - \rho)$.

taking expectations we were able to obtain P-K formulas that gave the expected number in the system [Eq. (5.63)] and the normalized expected time in the system [Eq. (5.71)]. If we were now to seek the *second* moment of the number in the system we could obtain this quantity by first *cubing* Eq. (5.75) and then taking expectations. In this operation it is clear that the expectation $E[\tilde{q}^3]$ would cancel on both sides of the equation once the limit on n was taken; this would then leave an expression for the second moment of g . Similarly, all higher moments- can be obtained by raising Eq. (5.75) to successively higher powers and then forming expectations.* In this section, however, we choose to go after the *distribution* for qn itself (actually we consider the limiting random variable g). As it turns out, we will obtain a result which gives the z-transform for this distribution rather than the distribution itself. In principle, these last two are completely equivalent; in practice, we sometimes face great difficulty in inverting from the z-transform back to the distribution. Nevertheless, we can pick off the moments of the distribution of $\tilde{q}$ from the z-transform in extremely simple fashion by making use of the usual properties of transforms and their derivatives.

Let us now proceed to calculate the a-transform for the probability of finding k customers in the system immediately following the departure of a customer. We begin by defining the z-transform for the random variable qn as

$$Q_n(z) \triangleq \sum_{k=0}^{\infty} P[q_n = k] z^k \quad (5.76)$$

From Appendix II (and from the definition of expected value) we have that this z-transform (or probability generating function) is also given by

$$Q_n(z) \triangleq E[z^{qn}] \quad (5.77)$$

Of interest is the z-transform for our limiting random variable g :

$$Q(z) = \lim_{n \rightarrow \infty} Q_n(z) = \sum_{k=0}^{\infty} P[g = k] z^k = E[z^g] \quad (5.78)$$

As is usual in these definitions for transforms, the sum on the right-hand side of Eq. (5.76) converges to Eq. (5.77) only within some circle of convergence in the z-plane which defines a maximum value for $|z|$ (certainly $|z| \leq 1$ is allowed).

The system *MfG fl* is characterized by Eq. (5.75). We therefore use both sides of this equation as an exponent for z as follows:

$$z^{q_{n+1}} = z^{q_n - \Delta q_n + v_{n+1}}$$

* Specifically, the k th power leads to an expression for $E[\tilde{q}^{k-1}]$ that involves the first k moments of service time.

Let us now take expectations:

$$E[z^{q_{n+1}}] = E[z^{q_n - \Delta_{q_n} + v_{n+1}}]$$

Using Eq. (5.77) we recognize the left-hand side of this last as $Q_{n+1}(z)$. Similarly, we may write the right-hand side of this equation as the expectation of the product of two factors, giving us

$$Q_{n+1}(z) = E[z^{q_n - \Delta_{q_n}} z^{v_{n+1}}] \quad (5.79)$$

We now observe, as earlier, that the random variable v_{n+1} (which represents the number of arrivals during the service of C_{n+1}) is independent of the random variable q_n (which is the number of customers left behind upon the departure of C_n). Since this is true, then the two factors within the expectation on the right-hand side of Eq. (5.79) must themselves be independent (since functions of independent random variables are also independent). We may thus write the expectation of the product in that equation as the product of the expectations:

$$Q_{n+1}(z) = E[z^{q_n - \Delta_{q_n}}] E[z^{v_{n+1}}] \quad (5.80)$$

The second of these two expectations we again recognize as being independent of the subscript $n + 1$; we thus remove the subscript and consider the random variable $\tilde{v}$ again. From Eq. (5.44) we then recognize that the second expectation on the right-hand side of Eq. (5.80) is merely

$$E[z^{v_{n+1}}] = E[z^{\tilde{v}}] = V(z)$$

We thus have

$$Q_{n+1}(z) = V(z) E[z^{q_n - \Delta_{q_n}}] \quad (5.81)$$

The only complicating factor in this last equation is the expectation. Let us examine this term separately; from the definition of expectation we have

$$E[z^{q_n - \Delta_{q_n}}] = \sum_{k=0}^{\infty} P[q_n = k] z^{k - \Delta_k}$$

The difficult part of this summation is that the exponent on z contains Δ_k , which takes on one of two values according to the value of k . In order to simplify this special behavior we write the summation by exposing the first term separately:

$$E[z^{q_n - \Delta_{q_n}}] = P[q_n = 0] z^{0 - \Delta_0} + \sum_{k=1}^{\infty} P[q_n = k] z^{k - \Delta_k} \quad (5.82)$$

Regarding the sum in this last equation we see that it is almost of the form given in Eq. (5.76); the differences are that we have one fewer powers of z and also that we are missing the first term in the sum. Both these deficiencies may be corrected as follows:

$$\sum_{k=1}^{\infty} P[q_n = k] z^{k-t} = \frac{1}{z} \sum_{k=0}^{\infty} P[q_n = k] z^k - \frac{1}{z} P[q_n = 0] z^0 \quad (5.83)$$

Applying this to Eq. (5.82) and recognizing that the sum on the right-hand side of Eq. (5.83) is merely $Q_n(z)$, we have

$$E[z^{q_n - \Delta_{q_n}}] = P[q_n = 0] + Q_n(z) - P[q_n = 0]$$

We may now substitute this last in Eq. (5.81) to obtain

$$Q_{n+1}(z) = V(z) \left(P[q_n = 0] + \frac{Q_n(z) - P[q_n = 0]}{z} \right)$$

We now take the limit as n goes to infinity and recognize the limiting value expressed in Eq. (5.36). We thus have

$$Q(z) = V(z) \left(P[\tilde{q} = 0] + \frac{Q(z) - P[\tilde{q} = 0]}{z} \right) \quad (5.84)$$

Using $P[\tilde{q} = 0] = 1 - \rho$, and solving Eq. (5.84) for $Q(z)$ we find

$$Q(z) = V(z) \frac{(1 - \rho)(1 - 1/z)}{1 - V(z)/z} \quad (5.85)$$

Finally we multiply numerator and denominator of this last by $(-z)$ and use our result in Eq. (5.46) to arrive at the well-known equation that gives the z -transform for the number of customers in the system,

$$Q(z) = B^*(\lambda - \lambda z) \frac{(1 - \rho)(1 - 1/z)}{B^*(\lambda - \lambda z) - z} \quad (5.86)$$

We shall refer to this as one form of the *Pollaczek-Khinchin (P-K) transform equation*.^t

The P-K transform equation readily yields the moments for the distribution of the number of customers in the system. Using the moment-generating properties of our transform expressed in Eqs. (5.50)-(5.52) we see that certainly $Q(1) = 1$; when we attempt to set $z = 1$ in Eq. (5.86), we obtain an indeterminant form and so we are required to use L'Hospital's rule. In carrying out this operation we find that we must evaluate $\lim_{z \rightarrow 1} dB^*(\lambda - \lambda z)/dz$ as $z \rightarrow 1$, which was carried out in the previous section and shown to be equal to ρ . This computation verifies that $Q(1) = 1$. In Exercise 5.5, the reader is asked to show that $Q(1) = \bar{q}$.

^t This formula was found in 1932 by A. Y. Khinchin [KHIN 32]. Shortly we will derive two other equations (each of which follow from and imply this equation), which we also refer to as P-K transform equations; these were studied by F. Pollaczek [POLL 30] in 1930 and Khinchin in 1932. See also the footnote on p. 177.

[‡] We note that the denominator of the P-K transform equation must always contain the factor $(1 - \rho)$ since $B^*(0) = 1$.

Usually, the inversion of the P-K transform equation is difficult, and therefore one settles for moments. However, the system M/M/1 yields very nicely to inversion (and to almost everything else). Thus, by way of example, we shall find its distribution. We have

$$g^*(s) = \frac{\mu}{s + \mu} \quad \text{M/M/1} \quad (5.87)$$

Clearly, the region of convergence for this last form is $\text{Re}(s) > -\mu$. Applying this to the P-K transform equation we find

$$Q(z) = \left(\frac{\mu}{\lambda - \lambda z + \mu} \right) \frac{(1 - p)(1 - z)}{[\mu/(\lambda - \lambda z + \mu)] - z}$$

Noting that $\rho = \lambda/\mu$, we have

$$Q(z) = \frac{1 - \rho}{1 - \rho z} \quad (5.88)$$

Equation (5.88) is the solution for the z-transform of the distribution of the number of people in the system. We can reach a point such as this with many service-time distributions $B(x)$; for the exponential distribution we can evaluate the inverse transform (by inspection!). We find immediately that

$$P\{j = k\} = (1 - \rho)\rho^k \quad \text{M/M/1} \quad (5.89)$$

This then is the familiar solution for M/M/1. If the reader refers back to Eq. (3.23), he will find the same function for the probability of k customers in the M/M/1 system. However, Eq. (3.23) gives the solution for *all* points in time whereas Eq. (5.89) gives the solution only at the imbedded Markov points (namely, at the departure instants for customers). The fact that these two answers are identical is no surprise for two reasons: first, because we told you so (we said that the imbedded Markov points give solutions that are good at all points); and second, because we recognize that the M/M/1 system forms a continuous-time Markov chain.

As a second example, we consider the system M/H₂/1 whose pdf for service time was given in Eq. (5.68). By inspection we may find $B^*(s)$, which gives

$$g^*(s) = \left(\frac{1}{4} \right) \frac{\lambda}{s + \lambda} + \left(\frac{3}{4} \right) \frac{2\lambda}{s + 2\lambda} \\ \frac{7\lambda s + 8\lambda^2}{4(s + \lambda)(s + 2\lambda)} \quad (5.90)$$

where the plane of convergence is $\text{Re } (z) > -\lambda$. From the P-K transform equation we then have

$$Q(z) = \frac{(1-p)(1-z)[8+7(1-z)]}{8+7(1-z)-4z(2-z)(3-z)}$$

Factoring the denominator and canceling the common term $(1-z)$ we have

$$Q(z) = \frac{(1-p)(1-(7/15)z)}{[1-(2/5)z][1-(2/3)z]}$$

We now expand $Q(z)$ in partial fractions, which gives

$$Q(z) = (1-p)\left(\frac{1}{1-(2/5)z} + \frac{3(4)}{1-(2/3)z}\right)$$

This last may be inverted by inspection (by now the reader should recognize the sixth entry in Table 1.2) to give

$$P_k = P[ij = k] = (1-p)\left[\frac{1}{4}\left(\frac{2}{5}\right)^k + \frac{3}{4}\left(\frac{2}{3}\right)^k\right] \quad (5.91)$$

Lastly, we note that the value for p has already been calculated at 5/8, and so for a final solution we have

$$p_k = \frac{3}{32}\left(\frac{2}{5}\right)^k + \frac{9}{32}\left(\frac{2}{3}\right)^k \quad k = 0, 1, 2, \dots \quad (5.92)$$

It should not surprise us to find this sum of geometric terms for our solution.

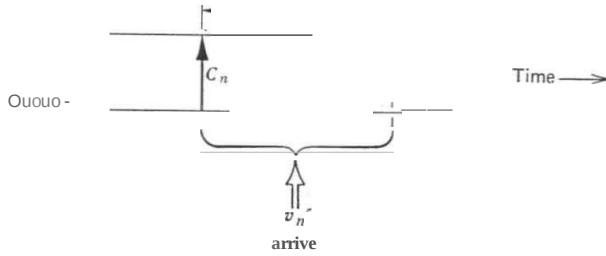
Further examples will be found in the exercises. For now we terminate the discussion of how *many* customers are in the system and proceed with the calculation of how *long* a customer spends in the system.

5.7. DISTRIBUTION OF WAITING TIME

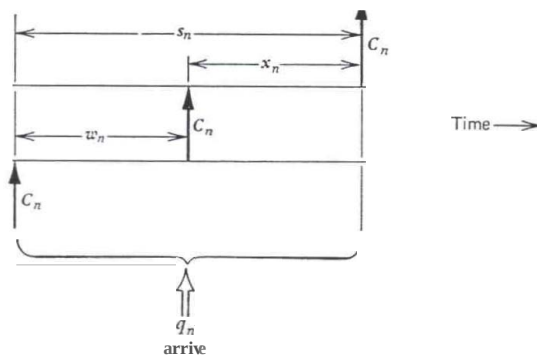
Let us now set out to find the distribution of time spent in the system and in the queue. These particular quantities are rather easy to obtain from our earlier principal result, namely, the P-K transform equation (and as we have said, lead to expressions which share that name). Note that the order in which customers receive service has so far not affected our results. Now, however, we must use our assumption that the order of service is first-come-first-served.

In order to proceed in the simplest possible fashion, let us re-examine the derivation of the following equation:

$$V(z) = B^*(i - \lambda z) \quad (5.93)$$

Figure 5.7 Derivation of $V(z) = B^*(\lambda - \lambda z)$.

In Figure 5.7, the reader is reminded of the structure from which we obtained this equation. Recall that $V(z)$ is the z -transform of the number of customer arrivals in a particular interval, where the arrival process is Poisson at a rate λ customers per second. The particular time interval involved happens to be the service interval for C_n ; this interval has distribution $B(x)$ with Laplace transform $B^*(s)$. The derived relation between $V(z)$ and $B^*(s)$ is given in Eq. (5.93). The important observation to make now is that a relationship of this form must exist between *any* two random variables where the one identifies the number of customer arrivals from a Poisson process and the other describes the time interval over which we are counting these customer arrivals. It clearly makes no difference what the interpretation of this time interval is, only that we give the distribution of its length; in Eq. (5.93) it just so happens that the interval involved is a service interval. Let us now direct our attention to Figure 5.8, which concentrates on the *time spent in the system* for C_n . In this figure we have traced the history of C_n . The interval labeled w_n identifies the time from when C_n enters the queue until that customer leaves the queue and enters service; it is clearly the *waiting time in queue* for C_n . We have also identified the service time x_n for C_n . We may thus

Figure 5.8 Derivation of $Q(z) = S^*(1 - \lambda z)$.

identify the total time spent in system s_n for C_n ,

$$s_n = w_n + x_n \quad (5.94)$$

We have earlier defined g_n as the number of customers left behind upon the departure of C_n . In considering a first-come-first-served system it is clear that all those customers present upon the arrival of C_n must depart before he does; consequently, those customers that C_n leaves behind him (a total of g_n) must be precisely those who arrive during his stay in the system. Thus, referring to Figure 5.8, we may identify those customers who arrive during the time interval s_n as being our previously defined random variable g_n . The reader is now asked to compare Figures 5.7 and 5.8. In both cases we have a Poisson arrival process at rate λ customers per second. In Figure 5.7 we inquire into the number of arrivals (u_n) during the interval whose duration is given by x_n ; in Figure 5.8 we inquire into the number of arrivals (g_n) during an interval whose duration is given by s_n . We now define the distribution for the total time spent in system for C_n as

$$S_n(y) \triangleq P[s_n \leq y] \quad (5.95)$$

Since we are assuming ergodicity, we recognize immediately that the limit of this distribution (as n goes to infinity) must be independent of n . We denote this limit by $S(y)$ and the limiting random variable by $\tilde{s}$ [i.e., $S_n(y) \rightarrow S(y)$ and $s_n \rightarrow \tilde{s}$]. Thus

$$S(y) \triangleq P[\tilde{s} \leq y] \quad (5.96)$$

Finally, we define the Laplace transform of the pdf for total time in system as •

$$S^*(s) \triangleq \int_0^\infty e^{-sy} dS(y) = E[e^{-s\tilde{s}}] \quad (5.97)$$

With these definitions we go back to the analogy between Figures 5.7 and 5.8. Clearly, since u_n is analogous to q_n , then $V(z)$ must be analogous to $Q(z)$, since each describes the generating function for the respective number distribution. Similarly, since x_n is analogous to s_n , then $B^*(s)$ must be analogous to $S^*(s)$. We have therefore by direct analogy from Eq. (5.93) that†

$$Q(z) = S^*(\lambda - \lambda z) \quad (5.98)$$

Since we already have an explicit expression for $Q(\cdot)$ as given in the P-K transform equation, we may therefore use that with Eq. (5.98) to give an explicit expression for $S^*(s)$ as

$$S^*(\lambda - \lambda z) = B^*(\lambda - \lambda z) \frac{(1 - p)(1 - z)}{B^*(\lambda - \lambda z) - z} \quad (5.99)$$

† This can be derived directly by the unconvinced reader in a fashion similar to that which led to Eqs. (5.28) and (5.46).

This last equation is just crying for the obvious change of variable

$$s = \lambda - \lambda z$$

which gives

$$z = 1 - \frac{s}{\lambda}$$

Making this change of variable in Eq. (5.99) we then have

$$S^*(s) = B^*(s) \frac{s(1-p)}{s - \lambda + \lambda B^*(s)} \quad (5.100)$$

Equation (5.100) is the desired explicit expression for the Laplace transform of the distribution of total time spent in the *MIGII* system. It is given in terms of known quantities derivable from the initial statement of the problem [namely, the specification of the service-time distribution $B(x)$ and the parameters λ and $\bar{x}$]. This is the second of the three equations that we refer to as the P-K transform equation.

From Eq. (5.100) it is trivial to derive the Laplace transform of the distribution of waiting time, which we shall denote by $W^*(s)$. We define the PDF for C_n 's waiting time (in queue) to be $Wn(y)$, that is,

$$Wn(y) \triangleq P[w_n \leq y] \quad (5.101)$$

Furthermore, we define the limiting quantities (as $n \rightarrow \infty$), $Wn(y) \rightarrow W(y)$ and $w_n \rightarrow \tilde{w}$, so that

$$W(y) \triangleq P[\tilde{w} \leq y] \quad (5.102)$$

The corresponding Laplace transform is

$$W^*(s) \triangleq \int_0^\infty e^{-sy} dW(y) = E[e^{-s\tilde{w}}] \quad (5.103)$$

From Eq. (5.94) we may derive the distribution of $\tilde{w}$ from the distribution of $\tilde{s}$ and $\tilde{x}$ (we drop subscript notation now since we are considering equilibrium behavior). Since a customer's service time is independent of his queueing time, we have that $\tilde{s}$, the time spent in system for some customer, is the sum of two independent random variables: $\tilde{w}$ (his queueing time) and $\tilde{x}$ (his service time). That is, Eq. (5.94) has the limiting form

$$\tilde{s} = \tilde{w} + \tilde{x} \quad (5.104)$$

As derived in Appendix II the Laplace transform of the pdf of a random variable that is itself the sum of two independent random variables is equal to the product of the Laplace transforms for the pdf of each. Consequently, we have

$$S^*(s) = W^*(s) B^*(s)$$

Thus from Eq. (5.100) we obtain immediately that

$$W^*(s) = \frac{s(1-\rho)}{s - \lambda + \lambda B^*(s)} \quad (5.105)$$

This is the desired expression for the Laplace transform of the queueing (waiting)-time distribution. Here we have the third equation that will be referred to as the P-K transform equation.

Let us rewrite the P-K transform equation for waiting time as follows:

$$W^*(s) = \frac{1-\rho}{1-\rho \frac{1-B^*(s)}{s\bar{x}}} \quad (5.106)$$

We recognize the bracketed term in the denominator of this equation to be exactly the Laplace transform associated with the density of residual service time from Eq. (5.11). Using our special notation for residual densities and their transforms, we define

$$\hat{B}^*(s) \triangleq \frac{1-B^*(s)}{s\bar{x}} \quad (5.107)$$

and are therefore permitted to write

$$W^*(s) = \frac{1-\rho}{1-\rho \hat{B}^*(s)} \quad (5.108)$$

This observation is truly amazing since we recognized at the outset that the problem with the M/G/1 analysis was to take account of the expended service time for the man in service. From that investigation we found that the residual service time remaining for the customer in service had a pdf given by $b(x)$, whose Laplace transform is given in Eq. (5.107). In a sense there is a poetic justice in its appearance at this point in the final solution. Let us follow Beneš [BENE 56] in inverting this transform in terms of these residual service time densities. Equation (5.108) may be expanded as the following power series:

$$W^*(s) = (1-\rho) \sum_{k=0}^{\infty} \rho^k [\hat{B}^*(s)]^k \quad (5.109)$$

From Appendix I we know that the k th power of a Laplace transform corresponds to the k -fold convolution of the inverse transform with itself. As in Appendix I the symbol $\otimes$ is used to denote the convolution operator, and we now choose to denote the k -fold convolution of a function $f(x)$ with itself by the use of a parenthetical subscript as follows:

$$f^{(k)}(x) \triangleq \underbrace{f(x) \otimes f(x) \otimes \dots \otimes f(x)}_{k\text{-fold convolution}} \quad (5.110)$$

Using this notation we may by inspection invert Eq. (5.109) to obtain the waiting-time pdf, which we denote by $w(y) \triangleq dW(y)/dy$; it is given by

$$w(y) = \sum_{k=0}^{\infty} (1-p)^k p k b \mathcal{L}(y) \quad (5.111)$$

This is a most intriguing result! It states that the waiting time pdf is given by a weighted sum of convolved residual service time pdf's. The interesting observation is that the weighting factor is simply $(1-p)^k p k$, which we now recognize to be the probability distribution for the number of customers in an M/M/1 system. Tempting as it is to try to give a physical explanation for the simplicity of this result and its relation to M/M/1, no satisfactory, intuitive explanation has been found to explain this dramatic form. We note that the contribution to the waiting-time density decreases geometrically with p in this series. Thus, for p not especially close to unity, we expect the high-order terms to be of less and less significance, and one practical application of this equation is to provide a rapidly converging approximation to the density of waiting time.

So far in this section we have established two principle results, namely, the P-K transform equations for time in system and time in queue given in Eqs. (5.100) and (5.105), respectively. In the previous section we have already given the first moment of these two random variables [see Eqs. (5.69) and (5.70)]. We wish now to give a recurrence formula for the moments of the waiting time. We denote the k th moment of the waiting time $E[w^k]$, as usual, by w^k . Takács [TAKA 62b] has shown that if $\bar{x}^{i+1}$ is finite, then so also are $\bar{w}, \bar{w}^2, \dots, \bar{w}^i$; we now adopt our slightly simplified notation for the i th moment of service time as follows: $h_i \triangleq \bar{x}^i$. The Takacs recurrence formula is

$$w^k = \frac{\lambda}{1-p} \sum_{i=1}^k \binom{k}{i} \frac{b}{i(i+1)} \bar{w}^{k-i} \quad (5.112)$$

where $\bar{w}_0 \triangleq 1$. From this formula we may write down the first couple of moments for waiting time (and note that the first moment of waiting time agrees with the P-K formula):

$$\bar{w} (= IV) = \frac{sb}{2(1-p)} \quad (5.113)$$

$$\bar{w}^2 = 2(\bar{w})^2 + \frac{\lambda b_s}{3(1-p)} \quad (5.114)$$

In order to obtain similar moments for the total time in system, that is, $E[S^k]$, which we denote by s^k , we need merely take advantage of Eq. (5.104); from this equation we find

$$s^k = (\bar{w} + \bar{x})^k \quad (5.115)$$

Using the binomial expansion and the independence between waiting time and service time for a given customer, we find

$$\overline{s^k} = \sum_{i=0}^k \binom{k}{i} \overline{w^{k-i}} \overline{b_i} \quad (5.116)$$

Thus calculating the moments of the waiting time from Eq. (5.112) also permits us to calculate the moments of time in system from this last equation. In Exercise 5.25, we derive a relationship between s_k and the moments of the number in system; the simplest of these is Little's result, and the others are useful generalizations.

At the end of Section 3.2, we promised the reader that we would develop the pdf for the time spent in the system for an $M/M/1$ queueing system. We are now in a position to fulfill that promise. Let us in fact find both the distribution of waiting time and distribution of system time for customers in $M/M/1$. Using Eq. (5.87) for the system $M/M/1$ we may calculate $S^*(s)$ from Eq. (5.100) as follows:

$$\begin{aligned} S^*(s) &= \frac{\mu}{(s + \mu)} \left[s - \frac{s(1-p)}{\lambda + \lambda\mu/(s + \mu)} \right] \\ S^*(s) &= \frac{\mu(1-p)}{s + \mu(1-p)} \quad \text{MIM/1} \end{aligned} \quad (5.117)$$

This equation gives the Laplace transform of the pdf for time in the system which we denote, as usual, by $s(y) \triangleq dS(y)/dy$. Fortunately (as is usual with the case $M/M/1$), we recognize the inverse of this transform by inspection. Thus we have immediately that

$$s(y) = \mu(1-p)e^{-\mu(1-p)y} \quad y \geq 0 \quad \text{MIM/1} \quad (5.118)$$

The corresponding PDF is given by

$$S(y) = 1 - e^{-\mu(1-p)y} \quad y \geq 0 \quad \text{MIM/1} \quad (5.119)$$

Similarly, from Eq. (5.105) we may obtain $W^*(s)$ as

$$\begin{aligned} W^*(s) &= \frac{s(1-p)}{s - \lambda + \lambda\mu/(s + \mu)} \\ &= \frac{(s + \mu)(1-p)}{s + (\mu - \lambda)} \end{aligned} \quad (5.120)$$

Before we can invert this we must place the right-hand side in proper form, namely, where the numerator polynomial is of lower degree than the denominator. We do this by dividing out the constant term and obtain

$$W^*(s) = (1-p) + \frac{\lambda(1-p)}{s + \mu(1-p)} \quad (5.121)$$

This
we
ApF
Imp

Fro

This
(
to
reF
Re
(5.
frc
di:
tir
th
(5

m
SI
th

•
p
a
t

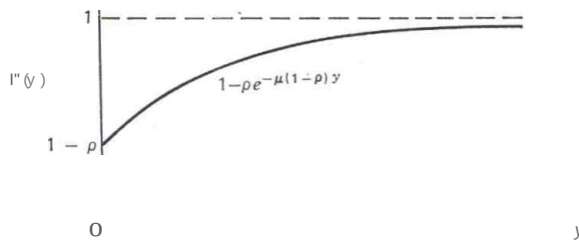


Figure 5.9 Waiting-time distribution for M/M/1.

This expression gives the Laplace transform for the pdf of waiting time which we denote, as usual, by $w(y) \triangleq dW(y)/dy$. From entry 2 in Table 1.4 of Appendix I, we recognize that the inverse transform of $(1 - p)$ must be an impulse at the origin; thus by inspection we have

$$w(y) = (1 - p)\delta(y) + \lambda(1 - p)e^{-\mu(1-p)y}, \quad y \geq 0 \quad \text{M/M/I} \quad (5.122)$$

From this we find the PDF of waiting time simply as

$$W(y) = 1 - pe^{-\mu(1-p)y}, \quad y \geq 0 \quad \text{M/M/I} \quad (5.123)$$

This distribution is shown in Figure 5.9.

Observe that the probability of not queueing is merely $1 - p$; compare this to Eq. (5.89) for the probability that $\tilde{q} = 0$. Clearly, they are the same; both represent the probability of not queueing. This also was found in Eq. (5.40). Recall further that the mean normalized queueing time was given in Eq. (5.73); we obtain the same answer, of course, if we calculate this mean value from (5.123). It is interesting to note for M/M/I that all of the interesting distributions are memoryless: this applies not only to the given interarrival time and service time processes, but also to the distribution of the number in the system given by Eq. (5.89), the pdf of time in the system given by Eq. (5.119), and the pdf of waiting time* given by Eq. (5.122).

It turns out that it is possible to find the density given in Eq. (5.118) by a more *direct* calculation, and we display this method here to indicate its simplicity. Our point of departure is our early result given in Eq. (3.23) for the probability of finding k customers in system upon arrival, namely,

$$p_k = (1 - p)p^k \quad (5.124)$$

* A simple exponential form for the tail of the waiting-time distribution (that is, the probabilities associated with long waits) can be derived for the system M/G/1. We postpone a discussion of this asymptotic result until Chapter 2, Volume II, in which we establish this result for the more general system G/G/1.

We repeat again that this is the same expression we found in Eq. (5.89) and we know by now that this result applies for all points in time. We wish to form the Laplace transform of the pdf of total time in the system by considering this Laplace transform *conditioned* on the number of customers found in the system upon arrival of a new customer. We begin as generally as possible and first consider the system $M/G/I$. In particular, we define the conditional distribution

$$S(y|k) = P[\text{customer's total time in system} \leq y \mid \text{he finds } k \text{ in system upon his arrival}]$$

We now define the Laplace transform of this conditional density

$$S^*(s|k) \triangleq \int_0^\infty e^{-sy} dS(y|k) \quad (5.125)$$

Now it is clear that if a customer finds no one in system upon his arrival, then he must spend an amount of time in the system exactly equal to his own service time, and so we have

$$S^*(s|0) = B^*(s)$$

On the other hand, if our arriving customer finds exactly one customer ahead of him, then he remains in the system for a time equal to the time to finish the man in service, plus his own service time; since these two intervals are independent, then the Laplace transform of the density of this sum must be the product of the Laplace transform of each density, giving

$$S^*(s|1) = B^*(s)B^*(s)$$

where $\hat{B}^*(s)$ is, again, the transform for the pdf for residual service time. Similarly, if our arriving customer finds k in front of him, then his total system time is the sum of the k service times associated with each of these customers plus his own service time. These $k + 1$ random variables are all independent, and k of them are drawn from the same distribution S_{ex} . Thus we have the k -fold product of $B^*(s)$ with $\hat{B}^*(s)$ giving

$$S^*(s|k) = [B^*(s)]^k \hat{B}^*(s) \quad (5.126)$$

Equation (5.126) holds for $M/G/I$. Now for our $M/M/I$ problem, we have that $B^*(s) = \mu/(s + \mu)$ and, similarly, for $\hat{B}^*(s)$ (memoryless); thus we have

$$S^*(s|k) = \left(\frac{\mu}{s + \mu} \right)^{k+1} \quad (5.127)$$

In order to obtain $S^*(s)$ we need merely weight the transform $S^*(s|k)$ with the probability P_k of our customer finding k in the system upon his arrival, namely,

$$S^*(s) = \sum_{k=0}^{\infty} S^*(s|k) P_k$$

Substituting Eqs. (5.127) and (5.124) into this last we have

$$S^*(s) = \sum_{k=0}^{\infty} \left(\frac{\mu}{s + \mu} \right)^k (1 - \rho) \rho^k$$

$$\frac{\mu(1 - \rho)}{s + \mu(1 - \rho)} \quad (5.128)$$

We recognize that Eq. (5.128) is identical to Eq. (5.117) and so the remaining steps leading to Eq. (5.118) follow immediately. This demonstration of a simpler method for calculating the distribution of system time in the M/M/1 queue demonstrates the following important fact: In the development of Eq. (5.128) we were required to consider a sum of random variables, each distributed by the same exponential distribution; the number of terms in that sum was itself a random variable distributed geometrically. What we found was that this geometrical weighting on a sum of identically distributed exponential random variables was itself exponential [see Eq. (5.118)]. This result is true in general, namely, that a geometric sum of exponential random variables is itself exponentially distributed.

Let us now carry out the calculations for our M/H₂/1 example. Using the expression for $B^*(s)$ given in Eq. (5.90), and applying this to the P-K transform equation for waiting-time density, we have

$$W^*(s) = \frac{4s(1 - \rho)(s + \lambda)(s + 2\lambda)}{4(s - \lambda)(s + \lambda)(s + 2\lambda) + 8\lambda^3 + 7\lambda^2 s}$$

This simplifies upon factoring the denominator, to give

$$W^*(s) = \frac{(1 - \rho)(s + \lambda)(s + 2\lambda)}{[s + (3/2)\lambda][s + (1/2)\lambda]}$$

Once again, we must divide numerator by denominator to reduce the degree of the numerator by one, giving

$$W^*(s) = (1 - \rho) + \frac{\lambda(1 - \rho)[s + (5/4)\lambda]}{[s + (3/2)\lambda][s + (1/2)\lambda]}$$

We may now carry out our partial-fraction expansion:

$$W^*(s) = (1 - \rho) \left[1 + \frac{\lambda/4}{s + (3/2)\lambda} + \frac{3\lambda/4}{s + (1/2)\lambda} \right]$$

This we may now invert by inspection to obtain the pdf for waiting time (and recalling that $\rho = 5/8$):

$$w(y) = \frac{3}{8} u(y) + \frac{3\lambda}{32} e^{-(3/2)\lambda y} + \frac{9\lambda}{32} e^{-(1/2)\lambda y} \quad y \geq 0 \quad (5.129)$$

This completes our discussion of the waiting-time and system-time distributions for M/G/1. We now introduce the busy period, an important stochastic process in queueing systems.

5.8. THE BUSY PERIOD AND ITS DURATION

We now choose to study queueing systems from a different point of view. We make the observation that the system passes through alternating cycles of busy period, idle period, busy period, idle period, and so on. Our purpose in this section is to derive the distribution for the length of the idle period and the length of the busy period for the M/G/1 queue.

As we already understand, the pertinent sequences of random variables that drive a queueing system are the instants of arrival and the sequence of service times. As usual let

C_n = the n th customer

τ_n = arrival time of C_n

$I_n = \tau_n - \tau_{n-1}$ = interarrival time between C_{n-1} and C_n

x_n = service time for C_n

We now recall the important stochastic process $V(t)$ as defined in Eq. (2.3):

$V(t) \triangleq$ the unfinished work in the system at time t
 $\triangleq$ the remaining time required to empty the system of all customers present at time t

This function $V(t)$ is appropriately referred to as the unfinished work at time t since it represents the interval of time that is required to empty the system completely if no new customers are allowed to enter after the instant t . This function is sometimes referred to as the "virtual" waiting time at time t since, for a first-come-first-served system it represents how long a (virtual) customer would wait in queue *if* he entered at time t ; however, this waiting-time interpretation is good only for first-come-first-served disciplines, whereas the unfinished work interpretation applies for all disciplines. Behavior of this function is extremely important in understanding queueing systems when one studies them from the point of view of the busy period.

Let us refer to Figure 5.10a, which shows the fashion in which busy periods alternate with idle periods. The busy-period durations are denoted by $Y_1, Y_2, Y_3, \dots$ and the idle period durations by $I_1, I_2, \dots$. Customer C_n ,

then terminates the busy period and initiates a new idle period. The idle period is terminated at time τ_4 when C_4 enters. This second busy period serves only one customer before the system goes idle again. The third busy period serves two customers. And so it continues. For reference we show in Figure 5.10b our usual double-time-axis representation for the same sequence of customer arrivals and service times drawn to the same scale as Figure 5.10a and under an assumed first-come-first-served discipline. Thus we can say that $Vet)$ is a function which has vertical jumps at the customer-arrival instants (these jumps equaling the service times for those customers) and decreases at a rate of 1 sec/sec so long as it is positive; when it reaches a value of zero, it remains there until the next customer arrival. This stochastic process is a continuous-state Markov process subject to discontinuous jumps; we have not seen such as this before.

Observe for Figure 5.10b that the departure instants may be obtained by extrapolating the linearly decreasing portion of $Vet)$ down to the horizontal axis; at these intercepts, a customer departure occurs and a new customer service begins. Again we emphasize that the last observation is good only for the first-come-first-served system. What is important, however, is to observe that the function $Vet)$ itself is *independent of the order of service!* The only requirement for this last statement to hold is that the server remain busy as long as some customer is in the system and that no customers depart before they are completely served; such a system is said to be "work conserving" (see Chapter 3, Volume II). The truth of this independence is evident when one considers the definition of $Vet)$.

Now for the idle-period and busy-period distributions. Recall

$$A(t) = P\{t. \leq t\} = 1 - e^{-\lambda t} \quad t \geq 0 \quad (5.130)$$

$$B(x) = P\{x. \leq x\}$$

where $A(t)$ and $B(x)$ are each independent of n . Our interest lies in the two following distributions:

$$F(y) \triangleq P\{J. \leq y\} \triangleq \text{idle-period distribution} \quad (5.131)$$

$$G(y) \triangleq P\{y. \leq y\} \triangleq \text{busy-period distribution} \quad (5.132)$$

The calculation of the idle-period distribution is trivial for the system M/G/1. Observe that when the system terminates a busy period, a new idle period must begin, and this idle period will terminate immediately upon the arrival of the next customer. Since we have a memoryless distribution, the time until the next customer arrival is distributed according to Eq. (5.130), and therefore we have

$$F(y) = 1 - e^{-\lambda y} \quad y \geq 0 \quad (5.133)$$

So much for the idle-time distribution in M/G/1.

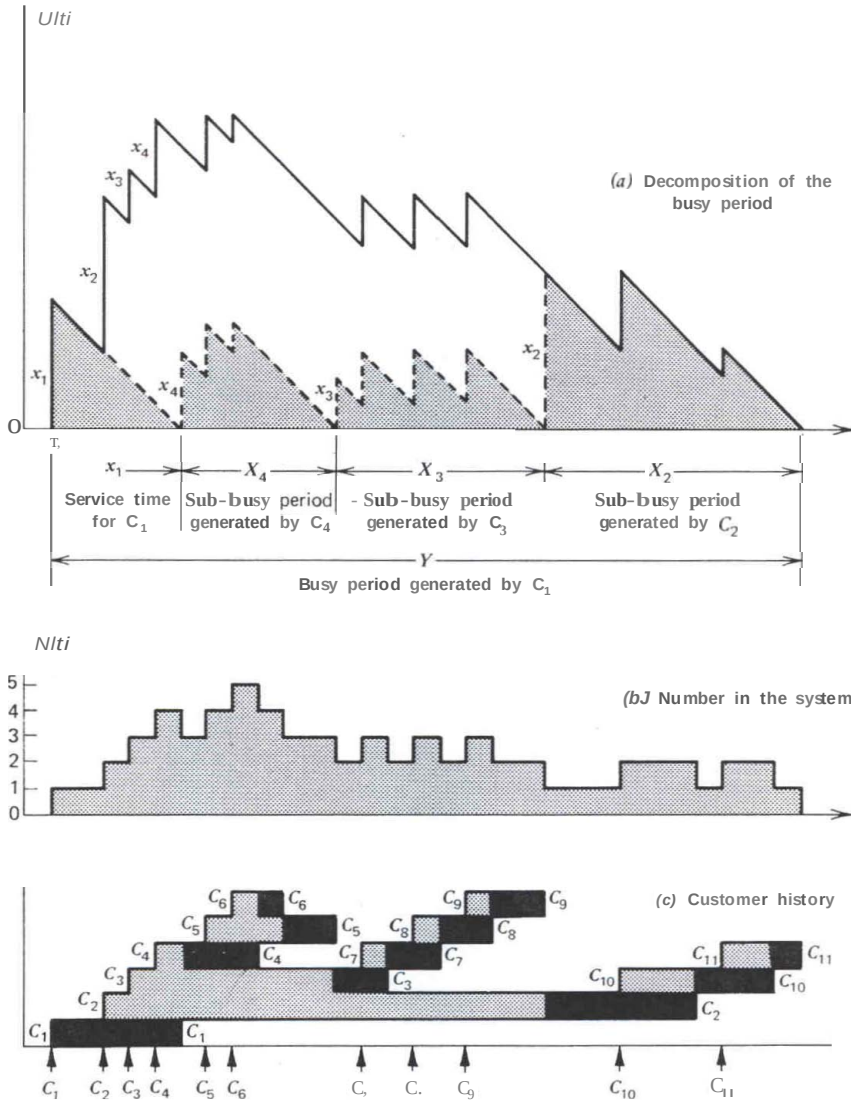


Figure 5.11 The busy period: last-come-first-served

Now for the busy-period distribution; this is not quite so simple. The reader is referred to Figure 5.11. In part (a) of this figure we once again observe the unfinished work $U(t)$. We assume that the system is empty just prior to the instant τ_1 , at which time customer C_1 initiates a busy period of duration Y . His service time is equal to x_1 . It is clear that this customer will depart from the system at a time $\tau_1 + x_1$. During his service other customers

may arrive to the system and it is they who will continue the busy period. For the function shown, three other customers (C_2 , C_3 , and C_4) arrive during the interval of C_1 's service. We now make use of a brilliant device due to Takács [TAKA 62a]. In particular, we choose to permute the order in which customers are served so as to create a last-come-first-served (LCFS) queueing discipline* (recall that the duration of a busy period is *independent* of the order in which customers are served). The motivation for the reordering of customers *will* soon be apparent. At the departure of C_1 we then take into service the *newest* customer, which in our example is C_4 . In addition, since all future arrivals during this busy period must be served before (LCFS!) any customers (besides C_4) who arrived during C_1 's service (in this case C_2 and C_3), then we may as well consider them to be (temporarily) out of the system. Thus, when C_4 enters service, it is as *if he initiated a new busy period*, which we will refer to as a "sub-busy period"; the sub-busy period generated by C_4 will have a duration X_4 exactly as long as it takes to service C_4 and all those who enter into the system to find it busy (remember that C_2 and C_3 are not considered to be in the system at this time). Thus in Figure 5.11a we show the sub-busy period generated by C_4 during which customers C_4 , C_3 , and C_2 get serviced in that order. At time $t_1 + X_1 + X_4$ this sub-busy period ends and we now continue the last-come-first-served order of service by bringing C_3 back into the system. It is clear that he may be considered as generating his own sub-busy period, of duration X_3 , during which all of his "descendants" receive service in the last-come-first-served order (namely, C_3 , C_7 , C_8 , and C_9). Finally, then, the system empties again, we reintroduce C_2 and permit his sub-busy period (of length X_2) to run its course (and complete the major busy period) in which customers get serviced in the order C_2 , C_{10} , and finally C_1 .

Figure 5.11a shows that the contour of any sub-busy period is identical with the contour of the main busy period over the same time interval and is merely shifted down by a constant amount; this shift, in fact, is equal to the summed service time of all those customers who arrived during C_1 's service time and who have not yet been allowed to generate their own sub-busy periods. The details of customer history are shown in Figure 5.11c and the total number in the system at any time under this discipline is shown in Figure 5.11b. Thus, as far as the queueing system is concerned, it is strictly a last-come-first-served system from start to finish. However, our analysis is simplified *if* we focus upon the sub-busy periods and observe that *each behaves statistically in a fashion identical to the major busy period* generated by C_1 . This is clear since all the sub-busy periods as well as the major busy period

* This is a "push-down" slack. This is only one of many permutations that "work"; it happens that LCFS is convenient for pedagogical purposes.

are each initiated by a single customer whose service times are all drawn from the same distribution independently; each sub-busy period continues until the system catches up to the work load, in the sense that the unfinished work function $U(t)$ drops to zero. Thus we recognize that the random variables $\{X_k\}$ are each independent and identically distributed and have the same distribution as Y , the duration of the major busy period.

In Figure S.11e the reader may follow the customer history in detail; the solid black region in this figure identifies the customer being served during that time interval. At each customer departure the server "floats up" to the top of the customer contour to engage the most recent arrival at that time; occasionally the server "floats down" to the customer directly below him such as at the departure of C_6 . The server may truly be thought of as floating up to the highest customer there to be held by him until his departure, and so on. Occasionally, however, we see that our server "falls down" through a gap in order to pick up the most recent arrival to the system, for example, at the departure of C_5 . It is at such instants that new sub-busy periods begin and only when the server falls down to hit the horizontal axis does the major busy period terminate.

Our point of view is now clear: the duration of a busy period Y is the sum of $1 + \tilde{v}$ random variables, the first of which is the service time for C , and the remainder of which are each random variables describing the duration of the sub-busy periods, each of which is distributed as a busy period itself. $\tilde{v}$ is a random variable equal to the number of customer arrivals during C 's service interval. Thus we have the important relation

$$Y = x_1 + X_{\tilde{v}+1} + X_2 + \dots + X_{\tilde{v}} + X_1 \quad (5.134)$$

We define the busy-period distribution as $G(y)$:

$$G(y) \triangleq \text{pr}\{Y \leq y\} \quad (5.135)$$

We also know that x_1 is distributed according to $B(x)$ and that X_k is distributed as $G(y)$ from our earlier comments. We next derive the Laplace transform for the pdf associated with Y , which we define, as usual, by

$$G^*(s) \triangleq \int_0^\infty e^{-sy} dG(y) \quad (5.136)$$

Once again we remind the reader that these transforms may also be expressed as expectation operators, namely:

$$G^*(s) \triangleq E[e^{-sY}]$$

Let us now take advantage of the powerful technique of conditioning used so often in probability theory; this technique permits one to write down the probability associated with a complex event by conditioning that event on

enough given conditions, so that the conditional probability may be written down by inspection. The unconditional probability is then obtained by multiplying by the probability of each condition and summing over all mutually exclusive and exhaustive conditions. In our case we choose to condition Y on two events: the duration of C_1 's service and the number of customer arrivals during his service. With this point of view we then calculate the following conditional transform:

$$\begin{aligned} E[e^{-sY} | X = x, \tilde{v} = k] &= E[e^{-s(x+X_{k+1}+\dots+X_2)}] \\ &= E[e^{-sx} e^{-sX_{k+1}} \dots e^{-sX_2}] \end{aligned}$$

Since the sub-busy periods have durations that are independent of each other, we may write this last as

$$E[e^{-sY} | X = x, \tilde{v} = k] = E[e^{-sx}] E[e^{-sX_{k+1}}] \dots E[e^{-sX_2}]$$

Since x is a given constant we have $E[e^{-sx}] = e^{-sx}$, and further, since the sub-busy periods are identically distributed with corresponding transforms $G^*(s)$, we have

$$E[e^{-sY} | X = x, \tilde{v} = k] = e^{-sx} [G^*(s)]^k$$

Since $\tilde{v}$ represents the number of arrivals during an interval of length x , then $\tilde{v}$ must have a Poisson distribution whose mean is λx . We may therefore remove the condition on $\tilde{v}$ as follows:

$$\begin{aligned} E[e^{-sY} | X = x] &= \sum_{k=0}^{\infty} E[e^{-sY} | X = x, \tilde{v} = k] P[\tilde{v} = k] \\ &= \sum_{k=0}^{\infty} e^{-sx} [G^*(s)]^k \frac{(\lambda x)^k}{k!} e^{-\lambda x} \\ &= e^{-x[s + \lambda - \lambda G^*(s)]} \end{aligned}$$

Similarly, we may remove the condition on X by integrating with respect to $B(x)$, finally to obtain $G^*(s)$ thusly

$$G^*(s) = \int_0^{\infty} e^{-x[s + \lambda - \lambda G^*(s)]} dB(x)$$

This last we recognize as the transform of the pdf for service time evaluated at a value equal to the bracketed term in the exponent, that is,

$$G^*(s) = B^*[s + \lambda - \lambda G^*(s)] \quad (5.137)$$

This major result gives the transform for the *MIGII* busy-period distribution (for any order of service) expressed as a functional equation (which is usually impossible to invert). It was obtained by identifying sub-busy periods within the busy period all of which had the same distribution as the busy period itself.

Later in this chapter we give an explicit expression for the busy period PDF $G(y)$, but unfortunately it is not in closed form [see Eq. (5.169)]. We point out, however, that it is possible to solve Eq. (5.137) numerically for $G^*(s)$ at any given value of s through the following iterative equation:

$$G_{n+1}^*(s) = B^*[s + \lambda - i.G_n^*(s)] \quad (5.138)$$

in which we choose $0 \leq G_0^*(s) \leq 1$; for $\rho = \lambda\bar{x} < 1$ the limit of this iterative scheme will converge to $G^*(s)$ and so one may attempt a numerical inversion of these calculated values if so desired.

In view of these inversion difficulties we obtain what we can from our functional equation, and one calculation we can make is for the *moments* of the busy period. We define

$$g_k \triangleq E[y^k] \quad (5.139)$$

as the k th moment of the busy-period distribution, and we intend to express the first few moments in terms of the moments of the service-time distribution, namely, x'' . As usual we have

$$\begin{aligned} g_k &= (-1)^k G^{*(k)}(0) \\ x^k &= (-1)^k B^*(k)(0) \end{aligned} \quad (5.140)$$

From Eq. (5.137) we then obtain directly

$$\begin{aligned} g_1 &= -G^*(1)(0) = -B^{*(1)}(0) \frac{d}{ds} [s + \lambda - \lambda G^*(s)] \Big|_{s=0} \\ &= -B^*(1)(0) [1 - \lambda G^{*(1)}(0)] \end{aligned}$$

[note, for $s = 0$, that $s + \lambda - \lambda G^*(s) = 0$] and so

$$g_1 = \bar{x}(1 + i.g_1)$$

Solving for g_1 and recalling that $\rho = \lambda\bar{x}$, we then have

$$g_1 = \frac{\bar{x}}{1 - \rho} \quad (5.141)$$

If we compare this last result with Eq. (3.26), we find that the *average length of a busy period for the system M/G/1 is equal to the average time a customer spends in an M/M/1 system and depends only on λ and $\bar{x}$*

Let us now chase down the second moment of the busy period. Proceeding from Eq. (5.140) and (5.137) we obtain

$$\begin{aligned} g_2 &= G^{*(2)}(0) = \frac{d}{ds} [B^*(1)(s + \lambda - \lambda G^*(s))] \Big|_{s=0} \\ &= B^*(2)(0) [1 - \lambda G^{*(1)}(0)]^2 + B^*(1)(0) [-\lambda G^{*(2)}(0)] \end{aligned}$$

and so

$$g_2 = x^2(1 + \lambda g_1)^2 + \bar{x}\lambda g_2$$

Solving for gz and using our result for g'' we have

$$gz = \frac{x^2(1 + \lambda g_1)^2}{1 - \lambda \bar{x}} - \frac{x\lambda(1 + \lambda \bar{x}/(1 - \rho))z}{1 - \rho}$$

and so finally

$$gz = (1 - \rho)3 \quad - (5.142)$$

This **last** result gives the second moment of the busy period and it is interesting to note the cube in the denominator; this effect does not occur when one calculates the second moment of the wait in the system where only a square power appears [see Eq. (5.114)]. We may now easily calculate the variance of the busy period, denoted by σ_g^2 , as follows:

$$\begin{aligned} \sigma_g^2 &= g_2 - g_1^2 \\ &= \frac{x^2}{(1 - \rho)3} - (\bar{x})^2 \end{aligned}$$

and so

$$\sigma_g^2 = \frac{\sigma_b^2 + \rho(\bar{x})^2}{(1 - \rho)3} \quad - (5.143)$$

where σ_b^2 is the variance of the service-time distribution.

Proceeding as above we find that

$$\begin{aligned} g_3 &= \frac{x^3}{(1 - \rho)^4} + \frac{3\lambda(\bar{x})^2}{(1 - \rho)^5} \\ g_4 &= \frac{x^4}{(1 - \rho)^5} + \frac{10\lambda\bar{x}x^3}{(1 - \rho)^6} + \frac{15\lambda^2(\bar{x})^3}{(1 - \rho)^7} \end{aligned}$$

We observe that the factor $(1 - \rho)$ goes up in powers of 2 for the dominant term of each succeeding moment of the busy period and this determines the behavior as $\rho \rightarrow 1$.

We now consider some examples of inverting Eq. (5.137). We begin with the M/M/1 queueing system. We have

$$B^*(s) = \frac{\mu}{s + \mu}$$

which we apply to Eq. (5.137) to obtain

$$G^*(s) = \frac{\mu}{s + \lambda - \lambda G^*(s) + \mu}$$

or

$$\lambda[G^*(s)]^2 - (\mu + \lambda + s)G^*(s) + \mu = 0$$

Solving for $G^*(s)$ and restricting our solution to the required (stable) case for which $|G^*(s)| \leq 1$ for $\text{Re}(s) \geq 0$, gives

$$G^*(s) = \frac{\mu + \lambda + s - [(\mu + \lambda + s)^2 - 4\mu\lambda]^{1/2}}{2\lambda} \quad (5.144)$$

This equation may be inverted (by referring to transform tables) to obtain the pdf for the busy period, namely,

$$g(y) \triangleq \frac{dG(y)}{dy} = \frac{1}{y(p)^{1/2}} e^{-(\lambda+\mu)y} I_1[2y(\lambda\mu)^{1/2}] \quad (5.145)$$

where I_1 is the modified Bessel function of the first kind of order one.

Consider the limit

$$\lim_{0 < s \rightarrow 0} G^*(s) = \lim_{0 < s \rightarrow 0} \int_0^\infty e^{-sy} y dG(y) \quad (5.146)$$

Examining the right side of this equation we observe that this limit is merely the probability that the busy period is finite, which is equivalent to the probability of the busy period ending. Clearly, for $\rho < 1$ the busy period ends with probability one, but Eq. (5.146) provides information in the case $\rho > 1$. We have

$$P[\text{busy period ends}] = G^*(0)$$

Let us examine this computation in the case of the system $M/M/1$. We have directly from Eq. (5.144)

$$G^*(0) = \frac{\mu + \lambda - [(\mu + \lambda)^2 - 4\mu\lambda]^{1/2}}{2\lambda}$$

and so

$$G^*(0) = \frac{1}{\rho}$$

Thus

$$P[\text{busy period ends in } M/M/1] = \begin{cases} 1 & \rho < 1 \\ \frac{1}{\rho} & \rho > 1 \end{cases} \quad (5.147)$$

The busy period pdf given in Eq. (5.145) is much more complex than we would have wished for this simplest of interesting queueing systems! It is indicative of the fact that Eq. (5.137) is usually uninvertible for more general service-time distributions.

As a second example, let's see how well we can do with our $M/H_2/1$ example. Using the expression for $B^*(s)$ in our functional equation for the busy period we get

$$G^*(s) = \frac{8\lambda^2 + 7\lambda[s + \frac{1}{2}] - \lambda G^*(s)}{4[s + \lambda - \frac{1}{2}]G^*(s) + \lambda[s + \lambda - \frac{1}{2}]G^*(s) + 2\lambda}$$

which leads directly to the cubic equation

$$4[G^*(s)]^3 - 4(2s + 5)[G^*(s)]^2 + (4s^2 + 20s + 31)G^*(s) - (15 + 7s) = 0$$

This last is not easily solved and so we stall at this point in our attempt to invert $G^*(s)$. We will return to the functional equation for the busy period when we discuss priority queueing in Chapter 3, Volume II. This will lead us to the concept of a *delay cycle*, which is a slight generalization of the busy-period analysis we have just carried out and greatly simplifies priority queueing calculations.

5.9. THE NUMBER SERVED IN A BUSY PERIOD

In this section we discuss the distribution of the number of customers served in a busy period. The development parallels that of the previous section very closely, both in the spirit of the derivation and in the nature of the result we will obtain.

Let N_{bp} be the number of customers served in a busy period. We are interested in its probability distribution *defined as*

$$P_n = P[N_{bp} = n] \quad (5.148)$$

The best we can do is to obtain a functional equation for its z -transform defined as

$$F(z) \triangleq E[z^{N_{bp}}] \triangleq \sum_{n=1}^{\infty} f_n z^n \quad (5.149)$$

The term for $n = 0$ is omitted from this definition since at least one customer must be served in a busy period. We recall that the random variable $\tilde{v}$ represents the number of arrivals during a service period and its z -transform $V(z)$ obeys the equation derived earlier, namely,

$$V(z) = B^*(\lambda - \lambda z) \quad (5.150)$$

Proceeding as we did for the duration of the busy period, we condition our argument on the fact that $\tilde{v} = k$, that is, we assume that k customers arrive

during the service of C_1 . Moreover, we recognize immediately that each of these arrivals will generate a sub-busy period and the number of customers served in each of these sub-busy periods will have a distribution given by fn . Let the random variable M_i denote the number of customers served in the i th sub-busy period. We may then write down immediately

$$E[z^{N_{bp}} | \tilde{v} = k] = E[z^{1+M_1+M_2+\dots+M_k}]$$

and since the M_i are independent and identically distributed we have

$$E[z^{N_{bp}} | \tilde{v} = k] = z \prod_{i=1}^k E[z^{M_i}]$$

But each of the M_i is distributed exactly the same as N_{bp} and, therefore,

$$E[z^{N_{bp}} | \tilde{v} = k] = z[F(z)]^k$$

Removing the condition on the number of arrivals we have

$$\begin{aligned} F(z) &= \sum_{k=0}^{\infty} E[z^{N_{bp}} | \tilde{v} = k] P[\tilde{v} = k] \\ &= z \sum_{k=0}^{\infty} P[\tilde{v} = k] [F(z)]^k \end{aligned}$$

From Eq. (5.44) we recognize this last summation as $V(z)$ (the z -transform associated with $\tilde{v}$) with transform variable $F(z)$; thus we have

$$F(z) = zV[F(z)] \quad (5.151)$$

But from Eq. (5.150) we may finally write

$$F(z) = zB^*[\lambda - \rho]F(z) \quad - \quad (5.152)$$

This functional equation for the z -transform of the number served in a busy period is not unlike the equation given earlier in Eq. (5.137).

From this fundamental equation we may easily pick off the moments for the number served in a busy period. We define the k th moment of the number served in a busy period as h_k . We recognize then

$$\begin{aligned} h_1 &= F^{(1)}(1) \\ &= 8^*(1)(0)[- \lambda F^{(1)}(1)] + 8^*(0) \end{aligned}$$

Thus

$$h_1 = \lambda \bar{x} h_1 + 1$$

which immediately gives us

$$h_1 = \frac{1}{1 - \rho} \quad - \quad (5.153)$$

We further recognize

$$F(2) = h_2 - h_1$$

Carrying out this computation in the usual way, we obtain the second moment and variance of the number served in the busy period:

$$\mu_2 = \frac{2\rho(1-\rho) + \lambda^2 \bar{x}^2}{(1-\rho)^3} + \frac{1}{1-\rho} \quad (5.154)$$

$$\sigma_h^2 = \frac{\rho(1-\rho) + \lambda^2 \bar{x}^2}{(1-\rho)^3} \quad (5.155)$$

As an example we again use the simple case of the M/M/1 system to solve for $F(z)$ from Eq. (5.152). Carrying this out we find

$$F(z) = z \frac{\mu}{\mu + \lambda - \lambda F(z)}$$

$$\lambda F^2(z) - (\mu + \lambda)F(z) + \mu z = 0$$

Solving,

$$F(z) = \frac{1+\rho}{2\rho} \left[1 - \left(1 - \frac{4\rho z}{(1+\rho)^2} \right)^{1/2} \right] \quad (5.156)$$

Fortunately, it turns out that the equation (5.156) can be inverted to obtain f_n , the probability of having n served in the busy period:

$$f_n = \frac{1}{n} \binom{2n-2}{n-1} \rho^{n-1} (1+\rho)^{1-2n} \quad (5.157)$$

As a second example we consider the system M/D/1. For this system we have $h(x) = u\alpha(x - \bar{x})$ and from entry three in Table 1.4 we have immediately that

$$B^*(s) = e^{-s\bar{x}}$$

Using this in our functional equation we obtain

$$F(z) = ze^{-\rho p F(z)} \quad (5.158)$$

where as usual $p = \lambda \bar{x}$. It is convenient to make the substitution $u = zpe^{-\rho F(z)}$ and $H(u) = \rho F(z)$, which then permits us to rewrite Eq. (5.158) as

$$u = H(u)e^{-H(u)}$$

The solution to this equation may be obtained [RIOR 62] and then our original function may be evaluated to give

$$F(z) = \sum_{n=1}^{\infty} \frac{(n\rho)^{n-1}}{n!} e^{-n\rho z}$$

From this power series we recognize immediately that the distribution for the number served in the *MIDI* busy period is given explicitly by

$$I_n = \frac{(np)^{n-1}}{n!} e^{-np} \quad - (5.159)$$

For the case of a constant service time we know that if the busy period serves n customers then it must be of duration ni , and therefore we may immediately write down the solution for the *MIDFI* busy-period distribution as

$$G(y) = \sum_{n=1}^{\lfloor y/\bar{x} \rfloor} \frac{(np)^{n-1}}{n!} e^{-np} \quad - (5.160)$$

where $\lfloor y/\bar{x} \rfloor$ is the largest integer not exceeding $y/\bar{x}$.

5.10. FROM BUSY PERIODS TO WAITING TIMES

We had mentioned in the opening paragraphs of this chapter that waiting times could be obtained from the busy-period analysis. We are now in a position to fulfill that claim. As the reader may be aware (and as we shall show in Chapter 3, Volume II), whereas the distribution of the busy-period duration is independent of the queueing discipline, the distribution of waiting time is strongly dependent upon order of service. Therefore, in this section we consider only first-come-first-served *MIG!* systems. Since we restrict ourselves to this discipline, the reordering of customers used in Section 5.8 is no longer permitted. Instead, we must now decompose the busy period into a sequence of intervals whose lengths are *dependent* random variables as follows. Consider Figure 5.12 in which we show a single busy period for the first-come-first-served system [in terms of the unfinished work *Vet*]. Here we see that customer C_1 initiates the busy period upon his arrival at time τ_1 . The first interval we consider is his service time x_1 , which we denote by X_0 ; during this interval more customers arrive (in this case C_2 and C_3). All those customers who arrive during X_0 are served during the next interval, whose duration is X_1 , and which equals the sum of the service times of all arrivals during X_0 (in this case C_2 and C_3). At the expiration of X_1 we then create a new interval of duration X_2 in which all customers arriving during X_1 are served, and so on. Thus X_i is the length of time required to service all those customers who arrive during the previous interval whose duration is X_{i-1} . If we let n_i denote the number of customer arrivals during the interval X_{i-1} then n_i customers are served during the interval X_i . We let n_0 equal the number of customers who arrive during X_0 (the first customer's service time).

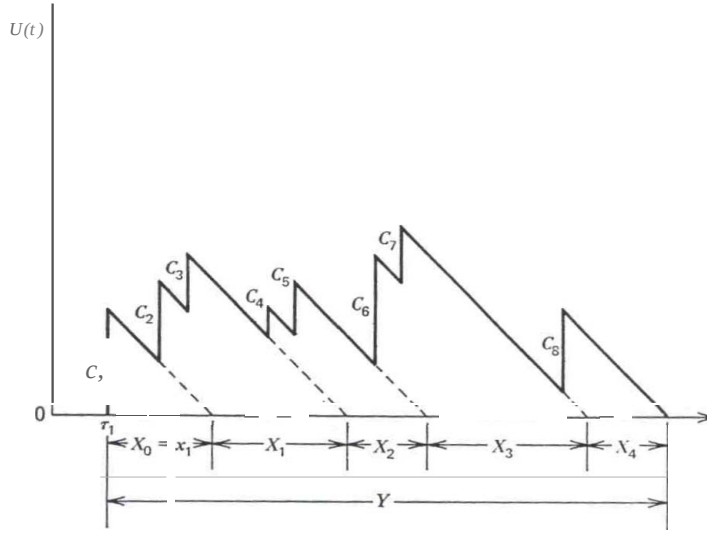


Figure 5.12 The busy period: first-come-first-served.

Thus we see that Y , the duration of the total busy period, is given by

$$Y = \sum_{i=0}^{\infty} X_i$$

where we permit the possibility of an infinite sequence of such intervals. Clearly, we define $X_i = 0$ for those intervals that fall beyond the termination of this busy period; for $p < 1$ we know that with probability 1 there will be a finite i_0 for which X_{i_0} (and all its successors) will be 0. Furthermore, we know that X_{i+1} will be the sum of n_i service intervals [each of which is distributed as $B(x)$].

We now define $X_i(y)$ to be the PDF for X_i that is,

$$X_i(y) \triangleq P[X_i \leq y]$$

and the corresponding Laplace transform of the associated pdf to be

$$\begin{aligned} X_i^*(s) &\triangleq \int_0^{\infty} e^{-sy} dX_i(y) \\ &= E[e^{-sX_i}] \end{aligned}$$

We wish to derive a recurrence relation among the $X_i^*(s)$. This derivation is much like that in Section 5.8, which led up to Eq. (5.137). That is, we first condition our transform sufficiently so that we may write it down by inspection; the conditions are on the interval length X_{i-1} and on the number of

arrivals n_{i-1} during that interval, that is, we may write

$$E[e^{-sX_i}; | X_{i-1} = Y, n_{i-1} = n] = [B^*(s)]^n$$

This last follows from our convolution property leading to the multiplication of transforms in the case when the variables are independent; here we have n independent service times, all with identical distributions. We may uncondition first on n :

$$E[e^{-sX_i}; | X_{i-1} = Y] = \sum_{n=0}^{\infty} \frac{(\lambda y)^n}{n!} e^{-\lambda y} [B^*(s)]^n$$

and next on Y :

$$E[e^{-sX_i}] = \int_{y=0}^{\infty} \sum_{n=0}^{\infty} \frac{(\lambda y)^n}{n!} e^{-\lambda y} [B^*(s)]^n dX_{i-1}(y)$$

Clearly, the left-hand side is $X_i^*(s)$; evaluating the sum on the right-hand side leads us to

$$X_i^*(s) = \int_{y=0}^{\infty} e^{-[B^*(s)]y} dX_{i-1}(y)$$

This integral is recognized as the transform of the pdf for X_{i-1} , namely,

$$X_i^*(s) = X_{i-1}^*[\lambda - \lambda B^*(s)] \quad (5.161)$$

This is the first step.

We now condition our calculations on the event that a new ("tagged") arrival occurs during the busy period and, in particular, while the busy period is in its i th interval (of duration X_i). From our observations in Section 4.1, we know that Poisson arrivals find the system in a given state with a probability equal to the equilibrium probability of the system being in that state. Now we know that if the system is in a busy period, then the fraction of time it spends in the interval of duration X_i is given by $E[X_i]/E[Y]$ (this can be made rigorous by renewal theory arguments). Consider a customer who arrives during an interval of duration X_i . Let his waiting time in system be denoted by $\tilde{w}$; it is clear that this waiting time will equal the sum of the remaining time (residual life) of the i th interval plus the sum of the service times of all jobs who arrived before he did during the i th interval. We wish to calculate $E[e^{-s\tilde{w}} | i]$, which is the transform of the waiting time pdf for an arrival during the i th interval; again, we perform this calculation by conditioning on the three variables X_i, Y_i (defined to be the residual life of this i th interval) and on N_i (defined to be the number of arrivals during the i th interval but prior to our customer's arrival—that is, in the interval $X_i - Y_i$). Thus, using our convolution property as before, we may write

$$E[e^{-s\tilde{w}} | i, X_i = y, Y_i = v, N_i = n] = e^{-sv'} [B^*(s)]^n$$

Now since we assume that n customers have arrived during an interval of duration $y - y'$ we uncondition on N , as follows:

$$\begin{aligned} E[e^{-sy} | X_i = y, Y_i = y'] &= e^{-sy} \sum_{n=0}^{\infty} \frac{[\lambda(y - y')]^n}{n!} e^{-\lambda(y-y')} [B^*(s)]^n \\ &= e^{-sy' - \lambda(y-y') + \lambda(y-y')B^*(s)} \end{aligned} \quad (5.162)$$

We have already observed that Y_i is the residual life of the lifetime X_i . Equation (5.9) gives the joint density for the residual life Y and lifetime X ; in that equation Y and X play the roles of Y_i and X_i in our problem. Therefore, replacing dx/ex in Eq. (5.9) by $dX_i(y)$ and noting that y and y' have replaced x and y in that development, we see that the joint density for X_i and Y_i is given by $dX_i(y) dy'/E[X_i]$ for $0 \leq y' \leq y \leq \infty$. By means of this joint density we may remove the condition on X_i and Y_i in Eq. (5.162) to obtain

$$\begin{aligned} E[e^{-sy} | i] &= \int_{y=0}^{\infty} \int_{y'=0}^y e^{-[s-\lambda+\lambda B^*(s)]y'} e^{-[\lambda-\lambda B^*(s)]y} dX_i(y) dy'/E[X_i] \\ &= \int_{y=0}^{\infty} \frac{[e^{-sy} - e^{-[\lambda-\lambda B^*(s)]y}]}{[s - \lambda + \lambda B^*(s)]E[X_i]} dX(y) \end{aligned}$$

These last integrals we recognize as transforms and so

$$E[e^{-sy} | i] = \frac{X_i^*(s) - X_i^*(\lambda - \lambda B^*(s))}{[s - \lambda + \lambda B^*(s)]E[X_i]} \quad (5.163)$$

But now Eq. (5.161) permits us to rewrite the second of these transforms to obtain

$$E[e^{-sW} | i] = \frac{X_{i+1}^*(s) - X_i^*(s)}{[s - \lambda + \lambda B^*(s)]E[X_i]}$$

Now we may remove the condition on our arrival entering during the i th interval by weighting this last expression by the probability that we have formerly expressed for the occurrence of this event (still conditioned on our arrival entering during a busy period), and so we have

$$\begin{aligned} E[e^{-sW} | \text{enter in busy period}] &= \sum_{i=0}^{\infty} E[e^{-sW} | i] \frac{E[X_i]}{E[Y]} \\ &= \frac{1}{[s - \lambda + \lambda B^*(s)]E[Y]} \sum_{i=0}^{\infty} [X_{i+1}^*(s) - X_i^*(s)] \end{aligned}$$

This last sum nicely collapses to yield $1 - X_0^*(s)$ since $X_i^*(s) = 1$ for those intervals beyond the busy period (recall $X_i = 0$ for $i \geq i_0$); also, since $X_0 = x_1$, a service time, then $X_0^*(s) = B^*(s)$, and so we arrive at

$$E[e^{-s\tilde{w}} | \text{enter in busy period}] = \frac{1 - B^*(s)}{[s - \lambda + \lambda B^*(s)]E[Y]}$$

From previous considerations we know that the probability of an arrival entering during a busy period is merely $p = \lambda\bar{x}$ (and for sure he must wait for service in such a case); further, we may evaluate the average length of the busy period $E[Y]$ either from our previous calculation in Eq. (5.141) or from elementary considerations¹ to give $E[Y] = \bar{x}/(1 - p)$. Thus, unconditioning on an arrival finding the system busy, we finally have

$$\begin{aligned} E[e^{-s\tilde{w}}] &= (1 - p)E[e^{-s\tilde{w}} | \text{enter in idle period}] + pE[e^{-s\tilde{w}} | \text{enter in busy period}] \\ &= (1 - p) + p \frac{[1 - B^*(s)](1 - p)}{[s - \lambda + \lambda B^*(s)]\bar{x}} \\ &= \frac{s(1 - p)}{s - \lambda + \lambda B^*(s)} \end{aligned} \quad (5.163)$$

Voila! This is exactly the P-K transform equation for waiting time, namely, $W^*(s) \triangleq E[e^{-s\tilde{w}}]$ given in Eq. (5.105).

Thus we have shown how to go from a busy-period analysis to the calculation of waiting time in the system. This method is reported upon in [CONW 67] and we will have occasion to return to it in Chapter 3, Volume 11.

S.U. COMBINATORIAL METHODS

We had mentioned in the opening remarks of this chapter that consideration of random walks and combinatorial methods was applicable to the study of the M/G/1 queue. We take this opportunity to indicate some aspects of those methods. In Figure 5.13 we have reproduced *Vet* from Figure 5.10a. In addition, we have indicated the "random walk" $R(t)$, which is the same as *Vet* except that it does not saturate at zero but rather continues to decline at a rate of 1 sec/sec below the horizontal axis; of course, it too takes vertical jumps at the customer-arrival instants. We introduce this diagram in order to define what are known as *ladder indices*. The k th (descending) ladder index

• The following simple argument enables us to calculate $E[Y]$. In a long interval (say, I) the server is busy a fraction p of the time. Each idle period in M/G/1 is of average length $1/\lambda$ sec and therefore we expect to have $(1 - p)/(1/\lambda)$ idle periods. This will also be the number of busy periods, approximately; therefore, since the time spent in busy periods is pI , the average duration of each must be $pI/\lambda I(1 - p) = \bar{x}/(1 - p)$. As $I \rightarrow \infty$, this argument becomes exact.

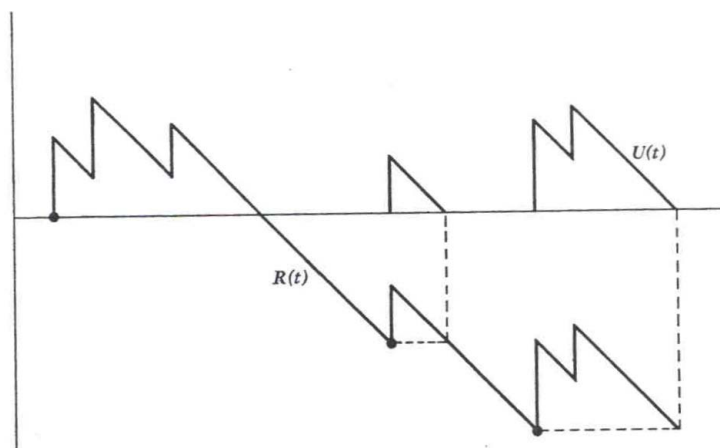


Figure 5.13 The descending ladder indices.

is defined as the instant when the random walk $R(t)$ rises from its k th new minimum (and the value of this minimum is referred to as the ladder height). In Figure 5.13 the first three ladder indices are indicated by heavy dots. Fluctuation theory concerns itself with the distribution of such ladder indices and is amply discussed both in Feller [FELL 66] and in Prabhu [PRAB 65] in which they consider the applications of that theory to queueing processes. Here we merely make the observation that each ladder index identifies the arrival instants for *those customers who begin new busy periods* and it is this observation that makes them interesting for queueing theory. Moreover, whenever $R(t)$ drops below its previous ladder height then a busy period terminates as shown in Figure 5.13. Thus, between the occurrence of a ladder index and the first time $R(t)$ drops below the corresponding ladder height, a busy period ensues and both $R(t)$ and $U(t)$ have exactly the same shape, where the former is shifted down from the latter by an amount exactly equal to the accumulated idle time since the end of the first busy period. One sees that we are quickly led into methods from combinatorial theory when we deal with such indices.

In a similar vein, Takacs has successfully applied combinatorial theory to the study of the busy period. He considers this subject in depth in his book [TAKA 67] on combinatorial methods as applied to queueing theory and develops, as his cornerstone, a generalization of the *classical ballot theorem*. The classical ballot theorem concerns itself with the counting of votes in a two-way contest involving candidate A and candidate B. If we assume that A scores a votes and B scores b votes and that $a \geq mb$, where m is a non-negative integer and if we let P be the probability that throughout the

counting of votes A continually leads B by a factor greater than m and further, if all possible sequences of voting records are equally likely, then the classical ballot theorem states that

$$P = \frac{a - mb}{a + b} \quad (5.164)$$

This theorem originated in 1887 (see [TAKA 67] for its history). Takács generalized this theorem and phrased it in terms of cards drawn from an urn in the following way. Consider an urn with n cards, where the cards are marked with the nonnegative integers $k_1, k_2, \dots, k_n$; and where

$$\sum_{i=1}^n k_i = k \leq n$$

(that is, the i th card in the set is marked with the integer k_i). Assume that all n cards are drawn without replacement from the urn. Let o_r ($r = 1, \dots, n$) be the number on the card drawn at the r th drawing. Let

$$\tilde{N}_r = v_1 + v_2 + \dots + v_r \quad r = 1, 2, \dots, n$$

$\tilde{N}_r$ is thus the sum of the numbers on all cards drawn up through the r th draw. Takacs' generalization of the classical ballot theorem states that

$$P[\tilde{N}_r < r \text{ for all } r = 1, 2, \dots, n] = \frac{n - k}{n} \quad (5.165)$$

The proof of this theorem is not especially difficult but will not be reproduced here. Note the simplicity of the theorem and, in particular, that the probability expressed is independent of the particular set of integers k_i and depends only upon their sum k . We may identify o_r as the number of customer arrivals during the service of the r th customer in a busy period of an M/G/1 queueing system. Thus $\tilde{N}_r + 1$ is the cumulative number of arrivals up to the conclusion of the r th customer's service during a busy period. We are thus involved in a race between $\tilde{N}_r + 1$ and r : As soon as r equals $\tilde{N}_r + 1$ then the busy period must terminate since, at this point, we have served exactly as many as have arrived (including the customer who initiated the busy period) and so the system empties. If we now let N_{bp} be the number of customers served in a busy period it is possible to apply Eq. (5.165) and obtain the following result [TAKA 67]:

$$P[N_{bp} = n] = \frac{1}{n} P[\tilde{N}_n = n - 1] \quad (5.166)$$

It is easy to calculate the probability on the right-hand side of this equation since we have Poisson arrivals: All we need do is condition this number of

arrivals on the duration of the busy period, multiply by the probability that n service intervals will, in fact, sum to this length and then integrate over all possible lengths. Thus

$$P[\tilde{N}_n = n - 1] = \int_0^\infty \frac{(\lambda y)^{n-1}}{(n-1)!} e^{-\lambda y} b_{1..l}(y) dy \quad (5.167)$$

where $b_{1..l}(y)$ is the n -fold convolution of $b(y)$ with itself [see Eq. (5.110)] and represents the pdf for the sum of n independent random variables, where each is drawn from the common density $b(y)$. Thus we arrive at an explicit expression for the probability distribution for the number served in a busy period:

$$P[N_{bp} = n] = \int_0^\infty \frac{(\lambda y)^{n-1}}{(n-1)!} e^{-\lambda y} b_{1..l}(y) dy \quad - (5.168)$$

We may go further and calculate $G(y)$, the distribution of the busy period, by integrating in Eq. (5.168) only up to some point y (rather than ∞) and then summing over all possible numbers served in the busy period, that is,

$$G(y) = \int_0^y \sum_{n=1}^\infty P[N_{bp} = n \mid Y = x] b_{(n)}(x) dx$$

and so,

$$G(y) = \int_0^y \sum_{n=1}^\infty e^{-\lambda x} \frac{(\lambda x)^{n-1}}{(n-1)!} b_{1..l}(x) dx \quad - (5.169)$$

Thus Eq. (5.169) is an explicit expression in terms of known quantities for the distribution of the busy period and in fact may be used in place of the expression given in Eq. (5.137), the Laplace transform of $dG(y)/dy$. This is the expression we had promised earlier, although we have expressed it as an infinite summation; nevertheless, it does provide the ability to approximate the busy-period distribution numerically in any given situation. Similarly, Eq. (5.168) gives an explicit expression for the number served in the busy period.

The reader may have observed that our study of the busy period has really been the study of a transient phenomenon and this is one of the reasons that the development bogged down. In the next section we consider certain aspects of the transient solution for M/G/1 a bit further.

5.12. THE TAKACS INTEGRODIFFERENTIAL EQUATION

In this section we take a closer look at the unfinished work and derive the forward Kolmogorov equation for its time-dependent behavior. A moment's reflection will reveal the fact that the unfinished work $U(t)$ is a continuous-time continuous-state Markov process that is subject to discontinuous

changes. It is a Markov process since the entire past history of its motion is summarized in its current value as far as its future behavior is concerned. That is, its vertical discontinuities occur at instants of customer arrivals and for M/G/1 these arrivals form a Poisson process (therefore, we need not know how long it has been since the last arrival), and the current value for $V(t)$ tells us exactly how much work remains in the system at each instant.

We wish to derive the probability distribution function for $V(t)$, given its initial value at time $t = 0$. Accordingly we define

$$F(w, t; w_0) \triangleq P[U(t) \leq w | U(0) = w_0] \quad (5.170)$$

This notation is a bit cumbersome and so we choose to suppress the initial value of the unfinished work and use the shorthand notation $F(w, t) \triangleq F(w, 1; 1; 0)$ with the understanding that the initial value is 1; 0. We wish to relate the probability $F(w, t + \Delta t)$ to its possible values at time t . We observe that we can reach this state from t if, on the one hand, there had been no arrivals during this increment in time [which occurs with probability $1 - \lambda \Delta t + o(\Delta t)$] and the unfinished work was no larger than $w + \Delta t$ at time t ; or if, on the other hand, there had been an arrival in this interval [with probability $\lambda \Delta t + o(\Delta t)$] such that the unfinished work at time t , plus the new increment of work brought in by this customer, together do not exceed w . These observations lead us to the following equation:

$$\begin{aligned} F(w, 1 + \Delta t) \\ = (1 - \lambda \Delta t)F(w + \Delta t, 1) + \lambda \Delta t \int_{x=0}^w B(w - x) \frac{\partial F(x, 1)}{\partial x} dx + o(\Delta t) \end{aligned} \quad (5.171)$$

Clearly, $(\partial F(x, t) / \partial x) dx \triangleq dF(x, t)$ is the probability that at time t we have $x < V(t) \leq x + dx$. Expanding our distribution function on its first variable we have

$$F(w + \Delta t, 1) = F(w, t) + \frac{\partial F(w, t)}{\partial w} \Delta t + o(\Delta t)$$

Using this expansion for the first term on the right-hand side of Eq. (5.171) we obtain

$$\begin{aligned} F(w, t + \Delta t) = F(w, t) + \frac{\partial F(w, t)}{\partial w} \Delta t - \lambda \Delta t \left[F(w, t) + \frac{\partial F(w, t)}{\partial w} \Delta t \right] \\ + \lambda \Delta t \int_{x=0}^w B(w - x) dx F(x, t) + o(\Delta t) \end{aligned}$$

Subtracting $F(w, t)$, dividing by Δt , and passing to the limit as $\Delta t \rightarrow 0$ we finally obtain the Takacs integrodifferential equation for $V(t)$:

$$\frac{\partial F(w, t)}{\partial t} = \frac{\partial F(w, t)}{\partial w} - \lambda F(w, t) + \lambda \int_{x=0}^w B(w - x) dx F(x, t) \quad (5.172)$$

Takács [TAKA 55] derived this equation for the more general case of a nonhomogeneous Poisson process, namely, where the arrival rate $\lambda(t)$ depends upon t . He showed that this equation is good for *almost* all $w \geq 0$ and $t \geq 0$; it does *not* hold at those w and t for which $of(t, 1)/\partial t$ has an accumulation of probability (namely, an impulse). This occurs, in particular, at $t = 0$ and would give rise to the term $F(0, t)u\phi(w)$ in $of(t, t)/\partial w$, whereas no other term in the equation contains such an impulse.

We may gain more information from the Takacs integrodifferential equation if we transform it on the variable w (and not on t); thus using the transform variable r we define

$$W^{*'}(r, t) \triangleq \int_0^\infty e^{-rw} dF(w, t) \quad (5.173)$$

We use the notation $(*)'$ to denote transformation on the first, but not the second argument. The symbol W is chosen since, as we shall see, $\lim_{t \rightarrow \infty} W^{*'}(r, t) = W^*(r)$ as $t \rightarrow \infty$, which is our former transform for the waiting-time pdf [see, for example, Eq. (5.103)].

Let us examine the transform of each term in Eq. (5.172) separately. First we note that since $F(w, t) = \int_{-\infty}^w dF(i, t)$, then from entry 13 in Table 1.3 of Appendix I (and its footnote) we must have

$$\int_0^\infty F(w, t) e^{-rw} dw = \frac{W^{*'}(r, t) + F(0^-, t)}{r}$$

and, similarly, we have

$$\int_0^\infty B(w) e^{-rw} dw = \frac{B^*(r) + B(0^-)}{r}$$

However, since the unfinished work and the service time are both nonnegative random variables, it must be that $F(0^-, t) = B(0^-) = 0$ always. We recognize that the last term in the Takacs integrodifferential equation is a convolution between $B(w)$ and $\partial F(w, t)/\partial w$, and therefore the transform of this convolution (including the constant multiplier λ) must be (by properties 10 and 13 in that same table) $\lambda W^{*'}(r, t)[B^*(r) - B(0^-)]/r = \lambda W^{*'}(r, t)B^*(r)/r$. Now it is clear that the transform for the term $of(w, t)/\partial w$ will be $W^{*'}(r, t)$; but this transform includes $F(0^+, t)$, the transform of the impulse located at the origin for this partial derivative, and since we know that the Takács integrodifferential equation does not contain that impulse it must be subtracted out. Thus, we have from Eq. (5.172),

$$\left(\frac{\partial}{\partial t}\right) W^{*'}(r, t) = W^{*'}(r, t) - F(0^+, t) - \frac{\lambda W^{*'}(r, t)}{r} + \lambda \frac{W^{*'}(r, t)B^*(r)}{r} \quad (5.174)$$

which may be rewritten as

$$\frac{\partial W^{*'}(r, t)}{\partial t} = [r - \lambda + \lambda B^{*}(r)]W^{*'}(r, t) - rF(O^{+}, 1) \quad (5.175)$$

Takacs gives the solution to this equation {p, 51, Eq. (8) in [TAKA 62b]}.

We may now transform on our second variable 1 by first defining the double transform

$$F^{**}(r, s) \triangleq \int_0^{\infty} e^{-st} W^{*'}(r, t) dt \quad (5.176)$$

We also need the definition

$$r,^{*}(s) \triangleq \int_0^{\infty} e^{-st} F(O^{+}, t) dt \quad (5.177)$$

We may now transform Eq. (5.175) using the transform property given as entry II in Table I.3 (and its footnote) to obtain

$$sF^{**}(r, s) - W^{*'}(r, O-) = [r - \lambda + \lambda B^{*}(r)]F^{**}(r, s) - rF_O^{*}(s)$$

From this we obtain

$$F^{**}(r, s) = \frac{W^{*'}(r, O-) - rF_O^{*}(s)}{s - r + \lambda - \lambda B^{*}(r)} \quad (5.178)$$

The unknown function $F_O^{*}(s)$ may be determined by insisting that the transform $F^{**}(r, s)$ be analytic in the region $\text{Re}(s) > 0$, $\text{Re}(r) > 0$. This implies that the zeroes of the numerator and denominator must coincide in this region; Beneš [BENE 56] has shown that in this region $1 = 1](s)$ is the unique root of the denominator in Eq. (5.178). Thus $W^{*'}(1, 0-) = 1]F_O^{*}(s)$ and so (writing $0-$ as 0), we have

$$F^{**}(r, s) = \frac{W^{*'}(r, 0) - (r/1]W^{*'}(1, 0)}{s - r + \lambda - \lambda B^{*}(r)} \quad (5.179)$$

Now we recall that $V(O) = 1V_O$ with probability one, and so from Eq. (5.173) we have $W^{*'}(r, O) = e^{-rw_O}$. Thus $F^{**}(r, s)$ takes the final form

$$F^{**}(r, s) = \frac{(r/\eta)e^{-\eta w_O} - e^{-rw_O}}{s - r + \lambda - \lambda B^{*}(r)} \quad (5.180)$$

We will return to this equation later in Chapter 2, Volume II, when we discuss the diffusion approximation.

For now it behooves us to investigate the steady-state value of these functions; in particular, it can be shown that $F(w, t)$ has a limit as $t \rightarrow \infty$ so long as $\rho < 1$, and this limit will be independent of the initial condition

$F(0, w)$: we denote this limit by $F(w) = \lim_{t \rightarrow \infty} F(w, t)$, and from Eq. (5.172) we find that it must satisfy the following equation:

$$\frac{dF(w)}{dw} = \lambda F(w) - \lambda \int_0^w B(w-x) dF(x) \quad (5.181)$$

Furthermore, for $\rho < 1$ then $W^*(r) \triangleq \lim_{t \rightarrow \infty} W^*(r, t)$ will exist and be independent of the initial distribution. Taking the transform of Eq. (5.181) we find as we did in deriving Eq. (5.174)

$$W^*(r) - F(0^+) = \frac{i W^*(r)}{r} - \frac{B^*(r) W^*(r)}{r}$$

where $F(0^+) = \lim_{t \rightarrow \infty} F(0^+, t)$ and equals the probability that the unfinished work is zero. This last may be rewritten to give

$$W^*(r) = \frac{r F(0^+)}{r - B^*(r)}$$

However, we require $W^*(0) = 1$, which requires that the unknown constant $F(0^+)$ have a value $F(0^+) = 1 - \rho$. Finally we have

$$W^*(r) = \frac{r(1 - \rho)}{r - \lambda + \lambda B^*(r)} \quad (5.182)$$

which is exactly the Pollaczek-Khinchin transform equation for waiting time as we promised!

This completes our discussion of the system M/G/1 (for the time being). Next we consider the "companion" system, G/M/m.

REFERENCES

- BENE 56 Benes, V. E., "On Queues with Poisson Arrivals," *Annals of Mathematical Statistics*, 28, 670-677 (1956).
- cONW67 Conway, R. W., W. L. Maxwell, and L. W. Miller, *Theory of Scheduling*, Addison-Wesley (Reading, Mass.) 1967.
- COX 55 Cox, D. R., "The Analysis of Non-Markovian Stochastic Processes by the Inclusion of Supplementary Variables," *Proc. Camb. Phil. Soc. (Math. and Phys. Sci.)*, 51, 433-441 (1955).
- COX 62 Cox, D. R., *Renewal Theory*, Methuen (London) 1962.
- FELL 66 Feller, W., *Probability Theory and its Applications* Vol. II, Wiley (New York), 1966.
- GAVE 59 Gaver, D. P., Jr., "Imbedded Markov Chain Analysis of a Waiting-Line Process in Continuous Time," *Annals of Mathematical Statistics* 30, 698-720 (1959).

- HEND 72 Henderson, W., "Alternative Approaches to the Analysis of the M/G/1 and G/M/1 Queues," *Operations Research*, 15, 92-101 (1972).
- KEIL 65 Keilson, J., "The Role of Green's Functions in Congestion Theory," *Proc. Symposium on Congestion Theory*, Univ. of North Carolina Press, 43-71 (1965).
- KEND 51 Kendall, D. G., "Some Problems in the Theory of Queues," *Journal of the Royal Statistical Society, Ser. B*, 13, 151-185 (1951).
- KEND 53 Kendall, D. G., "Stochastic Processes Occurring in the Theory of Queues and their Analysis by the Method of the Imbedded Markov Chain," *Annals of Mathematical Statistics*, 24, 338-354 (1953).
- KHIN 32 Khinchin, A. Y., "Mathematical Theory of Stationary Queues," *Mat. Sbornik*, 39, 73-84 (1932).
- LIND 52 Lindley, D. Y., "The Theory of Queues with a Single Server," *Proc. Cambridge Philosophical Society*, 48, 277-289 (1952).
- PALM 43 Palm, C., "Intensitätsschwankungen im Fernspreverkehr," *Ericsson Technics*, 6, 1-189 (1943).
- POLL 30 Pollaczek, F., "Über eine Aufgabe der Wahrscheinlichkeitstheorie," *I-II Math. Zeitschrift.*, 32, 64-100, 729-750 (1930).
- PRAB 65 Prabhu, N. U., *Queues and Inventories*, Wiley (New York) 1965.
- RIO 62 Riordan, J., *Stochastic Service Systems*, Wiley (New York) 1962.
- SMIT 58 Smith, W. L., "Renewal Theory and its Ramifications," *Journal of the Royal Statistical Society, Ser. B*, 20, 243-302 (1958).
- TAKA 55 Takacs, L., "Investigation of Waiting Time Problems by Reduction to Markov Processes," *Acta Math Acad. Sci. Hung.*, 6, 101-129 (1955).
- TAKA 62a Takacs, L., *Introduction to the Theory of Queues*, Oxford University Press (New York) 1962.
- TAKA 62b Takács, L., "A Single-Server Queue with Poisson Input," *Operations Research*, 10, 388-397 (1962).
- TAKA 67 Takács, L., *Combinatorial Methods in the Theory of Stochastic Processes*, Wiley (New York) 1967.

EXERCISES

- 5.1. Prove Eq. (5.14) from Eq. (5.11).
- 5.2. Here we derive the residual lifetime density $j(x)$ discussed in Section 5.2. We use the notation of Figure 5.1.
- (a) Observing that the event $\{Y \leq y\}$ can occur if and only if $t < \tau_k \leq t + y < \tau_{k+1}$ for some k , show that

$$\begin{aligned}
 t, (y) &\triangleq \text{pr } y \leq Y | t] \\
 &= \sum_{k=1}^{\infty} \int_0^{t+y} [1 - F(t + y - x)] dP[\tau_k \leq x]
 \end{aligned}$$

- (b) Observing that $\tau_k \leq x$ if and only if $\alpha(x)$, the number of "arrivals" in $(0, x)$, is at least k , that is, $P[\tau_k \leq x] = P[\alpha(x) \geq k]$, show that

$$\sum_{k=1}^{\infty} P[\tau_k \leq x] = \sum_{k=1}^{\infty} k P[\alpha(x) = k]$$

- (c) For large x , the mean-value expression in (b) is x/mI . Let $F(y) = \lim F_t(y)$ as $t \rightarrow \infty$ with corresponding pdf fey . Show that we now have

$$fey = \frac{1 - F(y)}{m_i}$$

- 5.3. Let us rederive the **P-K** mean-value formula (5.72).
- (a) Recognizing that a new arrival is delayed by one service time for each queued customer plus the residual service time of the customer in service, write an expression for W in terms of $\bar{N}_q$, ρ , $\bar{x}$, $\bar{c}$ and $P[\tilde{w} > 0]$.
- (b) Use Little's result in (a) to obtain Eq. (5.72).
- 5.4. Replace $1 - \rho$ in Eq. (5.85) by an unknown constant and show that $Q(I) = V(I) = I$ easily gives us the correct value of $1 - \rho$ for this constant.
- 5.5. (a) From Eq. (5.86) form $Q(I)(I)$ and show that it gives the expression for $\bar{q}$ in Eq. (5.63). Note that L'Hospital's rule will be required twice to remove the indeterminacies in the expression for $Q(I)(I)$.
- (b) From Eq. (5.105), find the first two moments of the waiting time and compare with Eqs. (5.113) and (5.114).
- 5.6. We wish to prove that the limiting probability r_k for the number of customers found by an arrival is equal to the limiting probability d_k for the number of customers left behind by a departure, in any queueing system in which the state changes by unit step values only (positive or negative). Beginning at $t = 0$, let x_n be those instants when $N(t)$ (the number in system) increases by one and Y_n be those instants when $N(t)$ decreases by unity, $n = 1, 2, \dots$. Let $N(x_n^-)$ be denoted by α_n and $N(Y_n^+)$ by β_n . Let $N(0) = i$.
- (a) Show that if $\beta_n H \leq k$, then $\alpha_{n+k+1} \leq k$.
- (b) Show that if $\alpha_{n+k+1} \leq k$, then $\beta_{n+i} \leq k$.
- (c) Show that (a) and (b) must therefore give, for any k ,

$$\lim P[\beta_n \leq k] = \lim P[\alpha_n \leq k]$$

which establishes that $r_k = d_k$.

- 5.7. In this exercise, we explore the method of supplementary variables as applied to the M/G/1 queue. As usual, let $P_k(t) = P[N(t) = k]$. Moreover, let $P_k(t, x_0) dx_0 = P[N(t) = k, x_0 < X_0(t) \leq x_0 + dx_0]$ where $X_0(t)$ is the service already received by the customer in service at time t .

(a) Show that

$$\frac{dP_0(t)}{dt} = -\lambda P_0(t) + \int_0^\infty \lambda l(t, x_0) r(x_0) dx_0$$

where

$$r(x_0) = \frac{h(x_0)}{1 - B(x_0)}$$

- (b) Let $p_k = \lim_{t \rightarrow \infty} P_k(t)$ and $p_k(x_0) = \lim_{t \rightarrow \infty} P_k(t, x_0)$ as $t \rightarrow \infty$. From (a) we have the equilibrium result

$$\lambda p_0 = \int_0^\infty \lambda l(x_0) r(x_0) dx_0$$

Show the following equilibrium results [where $P_0(x_0) \triangleq 0$]:

- (i) $\frac{dP_k(x_0)}{dx_0} = -[\lambda + r(x_0)]p_k(x_0) + \lambda p_{k-1}(x_0) \quad k \geq 1$
- (ii) $p_k(0) = \int_0^\infty \lambda p_{k-1}(x_0) r(x_0) dx_0 \quad k > 1$
- (iii) $\lambda l(0) = \int_0^\infty \lambda p_2(x_0) r(x_0) dx_0 + \lambda p_0$
- (e) The four equations in (b) determine the equilibrium probabilities when combined with an appropriate normalization equation. In terms of p_0 and $p_k(x_0)$ ($k = 1, 2, \dots$) give this normalization equation.
- (d) Let $R(z, x_0) = \sum_{k=1}^\infty p_k(x_0) z^k$. Show that

$$\frac{\partial R(z, x_0)}{\partial x_0} = [\lambda z - \lambda - r(x_0)]R(z, x_0)$$

and

$$zR(z, 0) = \int_0^\infty r(x_0)R(z, x_0) dx_0 + (z - 1)p_0$$

- (e) Show that the solution for $R(z, x_0)$ from (d) must be

$$R(z, x_0) = R(z, 0)e^{-\lambda x_0(1-z) - \int_0^{x_0} r(y) dy}$$

$$R(z, 0) = \frac{\lambda z(z - 1)p_0}{z - B^*(\lambda - \lambda z)}$$

- (f) Defining $R(z) \triangleq \int_0^\infty R(z, x_0) dx_0$, show that

$$R(z) = R(z, 0) \frac{1 - B^*(\lambda - \lambda z)}{\lambda(1 - z)}$$

- (g) From the normalization equation of (c), now show that

$$P_0 = 1 - p \quad (p = \lambda \bar{x})$$

- (h) Consistent with Eq. (5.78) we now define

$$Q(z) = P_0 + R(z)$$

Show that $Q(z)$ expressed this way is identical to the P-K transform equation (5.86). (See [COX 55] for additional details of this method.)

- 5.8. Consider the **M/G/∞** queue in which each customer always finds a free server; thus

$$s(y) = b\theta y \text{ and } T = \bar{x}. \text{ Let } P_k(l) = P[N(l) = k] \quad -$$

and assume $P_0(0) = 1$.

- (a) Show that

$$P_k(l) = \sum_{n=k}^{\infty} e^{-\lambda l} \frac{(\lambda l)^n}{n!} \binom{n}{k} \left[\frac{1}{l} \int_0^l [1 - B(x)] dx \right]^k \left[\frac{1}{l} \int_0^l B(x) dx \right]^{n-k} Ink$$

[HINT: $(1/l) \int_0^l B(x) dx$ is the probability that a customer's service terminates by time l , given that his arrival time was uniformly distributed over the interval $(0, l)$. See Eq. (2.137) also.]

- (b) Show that $P_k \triangleq \lim_{t \rightarrow \infty} P_k(l)$ as $t \rightarrow \infty$ is

$$P_k = \frac{(\lambda \bar{x})^k}{k!} e^{-\lambda \bar{x}} \quad -$$

regardless of the form of $B(x)$!

- 5.9. Consider **M/E./1**.

- (a) Find the polynomial for $G^*(s)$.
 (b) Solve for $S(y) = P[\text{time in system} \leq y]$.

- 5.10. Consider an **M/D/1** system for which $\bar{x} = 2$ sec.

- (a) Show that the residual service time pdf $\hat{b}(x)$ is a rectangular distribution.
 (b) For $p = 0.25$, show that the result of Eq. (5.111) with four terms may be used as a good approximation to the distribution of queueing time.

- 5.11. Consider an M/G/I queue in which bulk arrivals occur at rate λ and with a probability g_r that r customers arrive together at an arrival instant.
- Show that the z-transform of the number of customers arriving in an interval of length t is $e^{-\lambda t[1-G(z)]}$ where $G(z) = \sum g_r z^r$.
 - Show that the z-transform of the random variables U_n , the number of arrivals during the service of a customer, is $B^*[\lambda - i.G(z)]$.
- 5.12. Consider the M/G/I bulk arrival system in the previous problem. Using the method of imbedded Markov chains:
- Find the expected queue size. [HINT: show that $\bar{v} = \rho$ and

$$\bar{v}^2 - \rho = \frac{d^2 V(z)}{dz^2} \Big|_{z=1} = \rho^2 (C_g^2 + 1) + \frac{\lambda}{\mu} \left(C_g^2 + 1 - \frac{1}{\bar{g}} \right) (\bar{g})^2$$

where C_g is the coefficient of variation of the bulk group size and $\bar{g}$ is the mean group size.]

- Show that the generating function for queue size is

$$Q(z) = \frac{(1 - \rho)(1 - z)B^*[\lambda - \lambda G(z)]}{B^*[\lambda - \lambda G(z)] - z}$$

Using Little's result, find the ratio $\bar{W}/\bar{x}$ of the expected wait on queue to the average service time.

- Using the same **method** (imbedded Markov chain) find the expected number of groups in the queue (averaged over departure times). [HINTS: Show that $D(z) = \beta^*(\lambda - \lambda z)$, where $D(z)$ is the generating function for the number of *groups* arriving during the service time for an entire *group* and where $\beta^*(s)$ is the Laplace transform of the service-time density for an entire group. Also note that $\beta^*(s) = G[B^*(s)]$, which allows us to show that $\tau^2 = (\bar{x})^2(g^2 - \bar{g}) + \bar{x}^2\bar{g}$, where τ^2 is the second moment of the *group* service time.]
- Using Little's result, find $\bar{W}_g$, the expected wait on queue for a group (measured from the arrival time of the group until the start of service of the *first* member of the group) and show that

$$\bar{W}_g = \frac{\bar{x}\bar{g}}{2(1 - \rho)} \left[1 + \frac{C_g^2}{\bar{g}} + C_g^2 \right]$$

- If the customers within a group arriving together are served in random order, show that the ratio of the mean waiting time for a single customer to the average service time for a single customer is $\bar{W}_g/\bar{x}$ from (d) increased by $(1/2)g(1 + C_g^2) - 1/2$.

- 5.13. Consider an *M/G/1* system in which service is instantaneous but is only available at "service instants," the intervals between successive service instants being independently distributed with PDF $F(x)$. The maximum number of customers that can be served at any service instant is m . Note that this is a bulk service system.

- (a) Show that if q_n is the number of customers in the system just before the n th service instant, then

$$q_{n+1} = \begin{cases} q_n + v_n - m & q_n \geq m \\ v_n & q_n < m \end{cases}$$

where v_n is the number of arrivals in the interval between the n th and $(n+1)$ th service instants.

- (b) Prove that the probability generating function of u , is $F^*(\lambda - \lambda z)$. Hence show that $Q(z)$ is

$$Q(z) = \frac{\sum_{k=0}^{m-1} p_k (zm - k)}{z^m [F^*(\lambda - \lambda z)]^{-1} - 1}$$

where $p_k = p[\tilde{q} = k]$ ($k = 0, \dots, m-1$).

- (c) The $\{p_k\}$ can be determined from the condition that within the unit disk of the z -plane, the numerator must vanish when the denominator does. Hence show that if $F(x) = 1 - \rho x^2$,

$$Q(z) = \frac{zm - 1}{zm - z}$$

where Zm is the zero of $zm[1 + \lambda(1 - z)/\mu] - 1$ outside the unit disk.

- 5.14. Consider an *M/G/1* system with bulk service. Whenever the server becomes free, he accepts *two* customers from the queue into service simultaneously, or, if only one is on queue, he accepts that one; in either case, the service time for the group (of size 1 or 2) is taken from $B(x)$. Let q_n be the number of customers remaining after the n th service instant. Let v_n be the number of arrivals during the n th service. Define $B^*(s)$, $Q(z)$, and $V(z)$ as transforms associated with the random variables $\tilde{x}$, $\tilde{q}$, and $\tilde{v}$ as usual. Let $\rho = \lambda \bar{x}/2$.

- (a) Using the method of imbedded Markov chains, find

$$E(\tilde{q}) = \lim_{n \rightarrow \infty} E(q_n)$$

in terms of ρ , G_b^2 , and $P(\tilde{q} = 0) \triangleq P_0$.

- (b) Find $Q(z)$ in terms of $B^*(\cdot)$, P_0 , and ρ , $\triangleq P(\tilde{q} = 1)$.

- (c) Express ρ , in terms of P_0 .

- 5.15. Consider an *MIGI* queueing system with the following variation. The server refuses to serve any customers unless at least two customers are ready for service, at which time both are "taken into" service. These two customers are served individually and independently, one after the other. The instant at which the second of these two is finished is called a "critical" time and we shall use these critical times as the points in an imbedded Markov chain. Immediately following a critical time, if there are two more ready for service, they are both "taken into" service as above. If one or none are ready, then the server waits until a pair is ready, and so on. Let

q_n = number of customers left behind in the system
immediately following the n th critical time

v_n = number of customers arriving during the combined
service time of the n th pair of customers

- (a) Derive a relationship between q_{n+1} , q_n , and v_n .
(b) Find

$$V(z) = \sum_{k=0}^{\infty} P[v_n = k] z^k$$

- (c) Derive an expression for $Q(z) = \lim_{n \rightarrow \infty} Q_n(z)$ in terms of $p_0 = P[\bar{q} = 0]$, where

$$Q_n(z) = \sum_{k=0}^{\infty} P[q_n = k] z^k$$

- (d) How would you solve for p_0 ?
(e) Describe (do not calculate) two methods for finding $\bar{q}$.
- 5.16. Consider an *M/G/I* queueing system in which service is given as follows. Upon entry into service, a coin is tossed, which has probability p of giving Heads. If the result is Heads, then the service time for that customer is zero seconds. If Tails, his service time is drawn from the following exponential distribution:

$$x \geq 0$$

- (a) Find the average service time $\bar{x}$.
(b) Find the variance of service time σ_b^2 .
(c) Find the expected waiting time W .
(d) Find $W^*(s)$.
(e) From (d), find the expected waiting time W .
(f) From (d), find $W(t) = P[\text{waiting time} \leq t]$.
- 5.17. Consider an *M/G/I* queue. Let E be the event that T_{sec} have elapsed since the arrival of the last customer. We begin at a random time and

measure the time w until event E next occurs. This measurement may involve the observation of many customer arrivals before E occurs.

- (a) Let $\hat{A}(t)$ be the interarrival-time distribution for those intervals during which E does not occur. Find $\hat{A}(t)$.
- (b) Find $\hat{A}^*(s) = \int_0^\infty e^{-st} dA(t)$.
- (c) Find $W^*(s|n) = \int_0^\infty e^{-sw} dW(w|n)$, where $W(w|n) = P[\text{time to event } E \leq w \mid n \text{ arrivals occur before } E]$.
- (d) Find $W^*(s) = \int_0^\infty e^{-sw} dW(w)$, where $W(w) = P[\text{time to event } E \leq w]$.
- (e) Find the mean time to event E .

- 5.18. Consider an $M/G/1$ system in which time is divided into intervals of length q sec each. Assume that arrivals are **Bernoulli**, that is,

$$\begin{aligned} P[1 \text{ arrival in any interval}] &= \lambda q \\ P[0 \text{ arrivals in any interval}] &= 1 - \lambda q \\ P[> 1 \text{ arrival in any interval}] &= 0 \end{aligned}$$

Assume that a customer's service time $\tilde{x}$ is some multiple of q sec such that

$$P[\text{service time} = nq \text{ sec}] = K \quad n = 0, 1, 2, \dots,$$

- (a) Find $E[\text{number of arrivals in an interval}]$.
- (b) Find the average arrival rate.
- (c) Express $E[\tilde{x}] \triangleq \bar{x}$ and $E[\tilde{x}(\tilde{x} - q)] \triangleq \bar{x}^2 - \bar{x}q$ in terms of the moments of the g_n distribution (i.e., let $g^k \triangleq \sum_{n=0}^\infty n^k g_n$).
- (d) Find $Y_m = P[m \text{ customers arrive in } nq \text{ sec}]$.
- (e) Let $v_m = P[m \text{ customers arrive during the service of a customer}]$ and let

$$V(z) = \sum_{m=0}^{\infty} v_m z^m \quad \text{and} \quad G(z) = \sum_{m=0}^{\infty} g_m z^m$$

Express $V(z)$ in terms of $G(z)$ and the system parameters λ and q .

- (f) Find the mean number of arrivals during a customer service time from (e).
- 5.19. Suppose that in an $M/G/1$ queueing system the cost of making a customer wait t sec is $c(t)$ dollars, where $c(t) = \alpha e^{\beta t}$. Find the average cost of queueing for a customer. Also determine the conditions necessary to keep the average cost finite,
- 5.20. We wish to find the *interdeparture* time probability **density** function $d(t)$ for an *MIG/1* queueing system,
- (a) Find the Laplace transform $D^*(s)$ of this density conditioned first on a nonempty queue left behind, and second on an empty queue left behind by a departing customer. Combine these results

to get the Laplace transform of the interdeparture time density and from this find the density itself.

- (b) Give an explicit form for the probability distribution $D(l)$, or density $d(l) = dD(t)/dl$, of the interdeparture time when we have a constant service time, that is

$$B(x) = \begin{cases} 0 & x < T \\ 1 & x \geq T \end{cases}$$

- 5.21. Consider the following modified order of service for *MfGfl*. Instead of LCFS as in Figure 5.11, assume that after the interval x'' the sub-busy period generated by C_2 occurs, which is followed by the sub-busy period generated by C_0 , and so on, until the busy period terminates. Using the sequence of arrivals and service times shown in the upper contour of Figure 5.11a, redraw parts *a*, *b*, and *c* to correspond to the above order of service.
- 5.22. Consider an *MfGfl* system in which a departing customer immediately joins the queue again with probability p or departs forever with probability $q = 1 - p$. Service is FCFS, and the service time for a returning customer is independent of his previous service times. Let $B^*(s)$ be the transform for the service time pdf and let $BT^*(s)$ be the transform for a customer's total service time pdf.
- (a) Find $BT^*(s)$ in terms of $B^*(s)$, p , and q .
- (b) Let $\bar{x}_T^n$ be the n th moment of the total service time. Find $\bar{x}_T$ and $\bar{x}_T^2$ in terms of $\bar{x}$, $\bar{x}^2$, p , and q .
- (c) Show that the following recurrence formula holds:

$$\bar{x}_T^n = \bar{x}^n + \frac{p}{q} \sum_{k=1}^n \binom{n}{k} \bar{x}^k \bar{x}_T^{n-k}$$

- (d) Let

$$QT(z) = \sum_{k=0}^{\infty} p_{kT} z^k$$

where $p_{kT} = P[\text{number in system} = k]$. For $\lambda\bar{x} < q$ prove that

$$QT(z) = \left(1 - \frac{\lambda\bar{x}}{q}\right) \frac{q(1-z)B^*[\lambda(1-z)]}{(q + pz)B^*[\lambda(1-z)] - z}$$

- (e) Find $\bar{N}$, the average number of customers in the system.
- 5.23. Consider a first-come-first-served *MfGfl* queue with the following changes. The server serves the queue as long as someone is in the system. Whenever the system empties the server goes away on vacation for a certain length of time, which may be a random variable. At the end of his vacation the server returns and begins to serve customers again; if he returns to an empty system then he goes away on vacation

again. Let $F(z) = \sum_{j=1}^{\infty} f_j z^j$ be the z -transform for the number of customers awaiting service when the server returns from vacation to find at least one customer waiting (that is, f_j is the probability that at the initiation of a busy period the server finds j customers awaiting service).

- (a) Derive an expression which gives $qn+l$ in terms of qn , $vn+l'$ and j (the number of customer arrivals during the server's vacation).
 - (b) Derive an expression for $Q(z)$ where $Q(z) = \lim_{n \rightarrow \infty} E[z^{qn}]$ as $n \rightarrow \infty$ in terms of P_0 (equal to the probability that a departing customer leaves 0 customers behind). (HINT: condition on j .)
 - (c) Show that $p_0 = (1 - \rho)/F(1)$ where $F^{(1)}(1) = \partial F(z)/\partial z|_{z=1}$ and $\rho = \lambda \bar{x}$.
 - (d) Assume now that the service vacation will end whenever a new customer enters the empty system. For this case find $F(z)$ and show that when we substitute it back into our answer for (b) then we arrive at the classical *MIGII* solution.
- 5.24. We recognize that an arriving customer who finds k others in the system is delayed by the remaining service time for the customer in service plus the sum of $(k - 1)$ complete service times.
- (a) Using the notation and approach of Exercise 5.7, show that we may express the transform of the waiting time pdf as

$$\begin{aligned}
 1V^*(s) = & P_0 + \int_0^{\infty} I \, pk(X_0) [B^*(s)]^{k-1} \\
 & \times \int_0^{\infty} e^{-s'y} r(y + x_0) e^{-\int_0^{y+x_0} r(u) du} dy \\
 & \times \int_0^{x_0} r(u) du \, dX_0
 \end{aligned}$$

- (b) Show that the expression in (a) reduces to $W^*(s)$ as given in Eq. (5.106).
- 5.25. Let us relate s^k , the k^{th} moment of the time in system to N'' , the k^{th} moment of the number in system.
- (a) Show that Eq. (5.98) leads directly to Little's result, namely

$$\bar{N} = \lambda \bar{s} \triangleq J T$$

- (b) From Eq. (5.98) establish the second-moment relationship

$$N - 2 - \bar{N} = \lambda^2 s^2$$

- (c) Prove that the general relationship is

$$N(N - 1)(N - 2) \dots (N - k + 1) = \lambda^k s^k$$

The Queue G/M/rn

We have so far studied systems of the type MfM/I and its variants (elementary queueing theory) and MfG/I (intermediate queueing theory). The next natural system to study is GfM/I , in which we have an arbitrary interarrival time distribution $A(t)$ and an exponentially distributed service time. It turns out that the m -server system $GfMfm$ is almost as easy to study as is the single-server system GfM/I , and so we proceed directly to the m -server case. This study falls within intermediate queueing theory along with MfG/I , and it too may be solved using the method of the imbedded Markov chain, as elegantly presented by Kendall [KEND 51].

6.1. TRANS_mON PROBABILITIES FOR THE IMBEDDED MARKOV CHAIN (G/M/m)

The system under consideration contains m servers, who render service in order of arrival. Customers arrive singly with interarrival times identically and independently distributed according to $A(t)$ and with a mean time between arrivals equal to $1/\lambda$. Service times are distributed exponentially with mean $1/\mu$, the same distribution applying to each server independently. We consider steady-state results only (see discussion below).

As was the case in $M/G/I$, where the state variable became a continuous variable, so too in the system GfM/m we have a continuous-state variable in which we are required to keep track of *the elapsed time since the last arrival*, as well as the number in system. This is true since the probability of an arrival in any particular time interval depends upon the elapsed time (the "age") since the last arrival. It is possible to proceed with the analysis by considering the two-dimensional state description consisting of the age since the last arrival and the number in system; such a procedure is again referred to as the method of supplementary variables. A second approach, very much like that which we used for $MfGfI$, is the method of the imbedded Markov chain, which we pursue below. We have already seen a third approach, namely, the method of stages from Chapter 4.

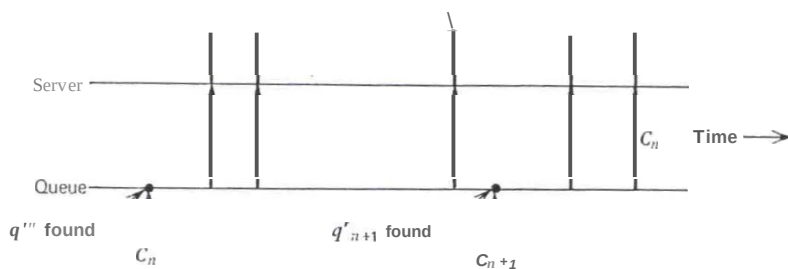


Figure 6.1 The imbedded Markov points.

If we are to use the imbedded Markov chain approach then it must be that the points we select as the regeneration points implicitly inform us of the elapsed time since the last arrival in an analogous way as for the expended service time in the case M/G/I. The natural set of points to choose for this purpose is the set of *arrival* instants. It is certainly clear that at these epochs the elapsed time since the last arrival is zero. Let us therefore define

q_n = number of customers found in the system immediately prior to the arrival of C_n ;

We use q_n' for this random variable to distinguish it from q_n , the number of customers left behind by the departure of C_n in the M/G/I system. In Figure 6.1 we show a sequence of arrival times and identify them as critical points imbedded in the time axis. It is clear that the sequence $\{q_n\}$ forms a discrete-state Markov chain. Defining

v'_{n+1} = the number of customers served between the arrival of C_n and C_{n+1} ,

we see immediately that the following fundamental relation must hold:

$$q'_{n+1} = q_n + 1 - v'_{n+1} \quad - (6.1)$$

We must now calculate the transition probabilities associated with this Markov chain, and so we define

$$p_{ij} = P[q'_{n+1} = j \mid q_n = i] \quad (6.2)$$

It is clear that p_{ij} is merely the probability that $i + 1 - j$ customers are served during an interarrival time. It is further clear that

$$p_{ij} = 0 \quad \text{for} \quad j > i + 1 \quad - (6.3)$$

since there are at most $i + 1$ present between the arrival of C_n and C_{n+1} . The Markov state-transition-probability diagram has transitions such as shown in Figure 6.2; in this figure we show only the transitions out of state E_i .

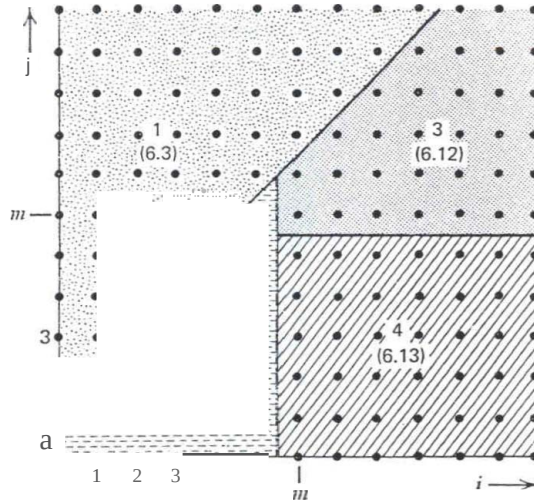


Figure 6.3 Range of validity for p_{ij} equations (equation numbers are also given in parentheses).

see that $i + I - j$ customers will complete their service. Since service times are exponentially distributed, the probability that any given customer will depart within t sec after the arrival of C_n is given by $1 - e^{-\mu t}$; similarly the probability that a given customer will not depart by this time is $e^{-\mu t}$. Therefore, in this region we have

$$P[i + I - j \text{ departures within } t \text{ sec after } C_n \text{ arrives} | q_n = i, j] \\ = \binom{i + I - j}{i + I - j} [1 - e^{-\mu t}]^{i+1-j} [e^{-\mu t}]^j \quad (6.8)$$

where the binomial coefficient

$$\binom{i + I - j}{i + I - j} = \binom{i + 1}{j}$$

merely counts the number of ways in which we can choose the $i + I - j$ customers to depart out of the $i + I$ that are available in the system. With t_{n+1} as the interarrival time between C_n and C_{n+1} , Eq. (6.8) gives $P[q_{n+1} = j | q_n = i, t_{n+1} = t]$. Removing the condition on t_{n+1} we then have the one-step transition probability in this range, namely,

$$P_{0j} = \int_0^\infty \binom{i + I}{j} (1 - e^{-\mu t})^{i+1-j} e^{-\mu t j} dA(t) \quad j \leq i + I \leq m \quad (6.9)$$

Next consider the range $m \leq j \leq i + I$, $i \geq m$ (region 3),* which corresponds to the simple case in which all m servers are busy throughout the

* The point $i = m - 1, j = m$ can properly lie either in region 2 or region 3.

interarrival interval. Under this assumption (that all m servers remain busy), since each service time is exponentially distributed (memoryless), then the number of customers served during this interval will be Poisson distributed (in fact it is a pure Poisson death process) with parameter $m\mu$; that is defining " t , all m busy" as the event that $t_n^+ = t$ and all m servers remain busy during t_n^+ , we have

$$P[k \text{ customers served} \mid t, \text{ all } m \text{ busy}] = \frac{(m\mu t)^k}{k!} e^{-m\mu t}$$

As pointed out earlier, if we are to go from state i to state j , then exactly $i + 1 - j$ customers must have been served during the interarrival time; taking account of this and removing the condition on t_n^+ we have

$$p_{ij} = \int_{t=0}^{\infty} P[i + 1 - j \text{ served} \mid t, \text{ all } m \text{ busy}] dA(t)$$

or

$$p_{ij} = \int_{t=0}^{\infty} \frac{(m\mu t)^{i+1-j}}{(i+1-j)!} e^{-m\mu t} dA(t) \quad m \leq j \leq i+1 \quad (6.10)$$

Note that in Eq. (6.10) the indices i and j appear only as the difference $i + 1 - j$, and so it behooves us to define a new quantity with a single index

$$\beta_{i+1-j} \triangleq p_{ij} \quad i+1-j \leq j \leq i+1, m \leq i \quad (6.11)$$

where β_n = the probability of serving n customers during an interarrival time given that all m servers remain busy during this interval; thus, with $n = i + 1 - j$, we have

$$\beta_n = p_{i,i+1-n} = \int_{t=0}^{\infty} \frac{(m\mu t)^n}{n!} e^{-m\mu t} dA(t) \quad 0 \leq n \leq i+1 - m, m \leq i \quad (6.12)$$

The last case we must consider (region 4) is $j < m < i + 1$, which describes the situation where C_j arrives to find m customers in service and $i - m$ waiting in queue (which he joins); upon the arrival of C_{i+1} there are exactly j customers, all of whom are in service. If we assume that it requires y sec until the queue empties then one may calculate P_{ij} in a straightforward manner to yield (see Exercise 6.1)

$$P_{ij} = \int_0^{\infty} \binom{m}{j} e^{-j\mu t} \times \left[\int_0^t \frac{(m\mu y)^{i-m}}{(i-m)!} (e^{-\mu y} - e^{-\mu t})^{m-j} m\mu dy \right] dA(t) \quad j < m < i+1 \quad (6.13)$$

Thus Eqs. (6.3), (6.9), (6.12), and (6.13) give the complete description of the one-step transition probabilities for the G/M/m system.

Having established the form for our one-step transition probabilities we may place them in the transition matrix

$$\mathbf{p} = \begin{bmatrix}
 p_{00} & p_{01} & 0 & 0 \\
 p_{10} & p_{11} & p_{12} & 0 \\
 p_{20} & p_{21} & p_{22} & p_{23} \\
 \\ \\
 p_{m-2,0} & p_{m-2,1} & & p_{m-2,m-1} & 0 & 0 & 0 \\
 p_{m-1,0} & p_{m-1,1} & & p_{m-1,m-1} & p_{00} & 0 & 0 \\
 p_{m,0} & p_{m,1} & & p_{m,m-1} & p_{10} & p_{11} & 0 \\
 \\ \\
 p_{m+n,0} & p_{m+n,1} & \dots & p_{m+n,m-1} & p_{m+n,0} & p_{m+n,1} & \dots & p_{00} & 0 & \dots \\
 & & & & & & & & \vdots & \\
 & & & & & & & & \vdots & \\
 & & & & & & & & \vdots &
 \end{bmatrix}$$

In this matrix all terms above the upper diagonal are zero, and the terms p_{ij} are given through Eq. (6.12). The "boundary" terms denoted in this matrix by their generic symbol p_{ij} are given either by Eqs. (6.9) or (6.13) according to the range of subscripts i and j . Of most importance to us are the transition probabilities p_{11} .

6.2. CONDITIONAL DISTRIBUTION OF QUEUE SIZE

Now we are in a position to find the equilibrium probabilities r_k , which must satisfy the system of linear equations given in Eq. (6.6). At this point we perhaps could guess at the form for r_k that satisfies these equations, but rather than that we choose to motivate the results that we obtain by the following intuitive arguments. In order to do this we define

$$\begin{aligned}
 N_k(t) = & \text{number of arrival instants in the interval } (0, t) \text{ in which} \\
 & \text{the arriving customer finds the system in state } E_k, \text{ given} \\
 & 0 \text{ customers at } t = 0
 \end{aligned} \tag{6.14}$$

Note from Figure 6.2 that the system can move up by at most one state, but may move down by many states in any single transition. We consider this motion between states and define (for $m - 1 \leq k$)

$$u_k = E[\text{number of times state } E_{k+1} \text{ is reached between two successive visits to state } E_k] \quad (6.15)$$

We have that the probability of reaching state E_{k+1} no times between returns to state E_k is equal to $1 - P_0$ (that is, given we are in state E_k the only way we can reach state E_{k+1} before our next visit to state E_k is for no customers to be served, which has probability P_0 , and so the probability of not getting to E_{k+1} first is $1 - P_0$, the probability of serving at least one). Furthermore, let

$$\begin{aligned} y &= P[\text{leave state } E_{k+1} \text{ and return to it some time later without passing through state } E_j; \text{ where } j \leq k] \\ &= P[\text{leave state } E_{k+1} \text{ and return to it later without passing through state } E_k] \end{aligned}$$

This last is true since a visit to state E_j , for $j \leq k$ must result in a visit to state E_k before next returning to state E_{k+1} (we move up only *one* state at a time). We note that y is independent of k so long as $k \geq m - 1$ (i.e., all m servers are busy). We have the simple calculation on

$$\begin{aligned} P[n \text{ occurrences of state } E_{k+1} \text{ between two successive visits to state } E_k] \\ = y^{n-1}(1 - y)P_0 \end{aligned}$$

This last equation is calculated as the probability (P_0) of reaching state E_{k+1} at all, times the probability (y^{n-1}) of returning to E_{k+1} a total of $n - 1$ times without first touching state E_k , times the probability ($1 - y$) of then visiting state E_k without first returning to state E_{k+1} . From this we may calculate

$$u_k = \sum_{n=1}^{\infty} n y^{n-1} (1 - y) P_0$$

as the average number of visits to E_{k+1} between successive visits to state E_k . Thus

$$u_k = \frac{P_0}{1 - y} \quad \text{for } k \geq m - 1$$

Note that u_k is *independent* of k and so we may drop the subscript, in which case we have

$$u \triangleq u_k = \frac{P_0}{1 - y} \quad \text{for } k \geq m - 1 \quad (6.16)$$

From the definition in Eq. (6.15), u must be the limit of the ratio of the number of times we find ourselves in state E_{k+1} to the number of times we find

ourselves in state E_k ; thus we may write

$$\sigma = \lim_{t \rightarrow \infty} \frac{N_{kH}(t)}{N_k(t)} = \frac{f l_0}{1 - Y} \quad k \geq m-1 \quad (6.17)$$

However, the limit is merely the ratio of the steady-state probability of finding the system in state E_{kH} to the probability of finding it in state E_k . Consequently, we have established

$$r_{k+1} = \sigma r_k \quad k \geq m-1 \quad (6.18)$$

The solution to this last set of equations is clearly

$$r_k = K \sigma^k \quad k \geq m-1 \quad (6.19)$$

for some constant K . This is a basic result, which says that the distribution of number of customers found at the arrival instants is *geometric* for the case $k \geq m-1$. It remains for us to find a and K , as well as r_k for $k < m-1$.

Our intuitive reasoning (which may easily be made rigorous by results from renewal theory) has led us to the basic equation (6.19). We could have "pulled this out of a hat" by guessing that the solution to Eq. (6.6) for the probability vector $\mathbf{r} \triangleq [r_0, r_1, r_2, \dots]$ might perhaps be of the form

$$\mathbf{r} = [r_0, r_1, r_2, \dots, r_{m-2}, K \sigma^{m-1}, K \sigma^m, K \sigma^{m+1}, \dots] \quad (6.20)$$

This flash of brilliance would, of course, have been correct (as our calculations have just shown); once we suspect this result we may easily verify it by considering the k th equation ($k \geq m$) in the set (6.6), which reads

$$\begin{aligned} r_k &= K \sigma^k = \sum_{i=0}^{\infty} r_i p_{ik} \\ &= \sum_{i=k-1}^{\infty} r_i p_{ik} \\ &= \sum_{i=k-1}^{\infty} K a^i f l_{i+1-k} \end{aligned}$$

Canceling the constant K as well as common factors of a we have

$$\sigma = \sum_{i=k-1}^{\infty} \sigma^{i+l-k} \rho_{i+1-k}$$

Changing the index of summation we finally have

$$\sigma = \sum_{n=0}^{\infty} \sigma^n \beta_n$$

Of course we know β_n from Eq. (6.12), which permits the following calculation:

$$\begin{aligned}\sigma &= \sum_{n=0}^{\infty} a n \int_{t=0}^{\infty} \frac{(m\mu t)^n}{n!} e^{-m\mu t} dA(t) \\ &= \int_0^{\infty} e^{-(m\mu - m\mu\sigma)t} dA(t)\end{aligned}$$

This equation must be satisfied if our assumed ("calculated") guess is to be correct. However, we recognize this last integral as the Laplace transform for the pdf of interarrival times evaluated at a special point; thus we have

$$\sigma = A^*(m\mu - m\mu\sigma) \quad - (6.21)$$

This functional equation for σ must be satisfied if our assumed solution is to be acceptable. It can be shown [TAKA 62] that so long as $\rho < 1$ then there is a unique real solution for σ in the range $0 < \sigma < 1$, and it is this solution which we seek; note that $\sigma = 1$ must always be a solution of the functional equation since $A^*(0) = 1$.

We now have the defining equation for σ and it remains for us to find the unknown constant K as well as r_k for $k = 0, 1, 2, \dots, m-2$. Before we settle these questions, however, let us establish some additional important results for the G/M/m system using Eq. (6.19), our basic result so far. This basic result establishes that the distribution for number in system is geometrically distributed in the range $k \geq m-1$. Working from there let us now calculate the probability that an arriving customer must wait for service. Clearly

$$\begin{aligned}P[\text{arrival queues}] &= \sum_{k=m}^{\infty} r_k \\ &= \sum_{k=m}^{\infty} K\sigma^k \\ &= \frac{1 - \sigma}{1 - \sigma^m}\end{aligned} \quad (6.22)$$

(This operation is permissible since $0 < \sigma < 1$ as discussed above.) The conditional probability of finding a queue length of size n , given that a customer must queue, is

$$P[\text{queue size} = n \mid \text{arrival queues}] = \frac{r_{m+n}}{P[\text{arrival queues}]}$$

and so

$$\begin{aligned}P[\text{queue size} = n \mid \text{arrival queues}] &= \frac{Ka^{n+m}}{Ka^m/(1-a)} \\ &= (1-a)a^n \quad n \geq 0 \quad - (6.23)\end{aligned}$$

Thus we conclude that the conditional queue length distribution (given that a queue exists) is geometric for any G/M/m system.

6.3. CONDITIONAL DISTRIBUTION OF WAITING TIME

Let us now seek the distribution of queueing time, given that a customer must queue. From Eq. (6.23), a customer who queues will find $m + n$ in the system with probability $(1 - \sigma)\sigma^n$. Under such conditions our arriving customer must wait until $n + 1$ customers depart from the system before he is allowed into *service*, and this interval will constitute *his* waiting time. Thus we are asking for the distribution of an interval whose length is made up of the sum of $n + 1$ independently and exponentially distributed random variables (each with parameter $m\mu$). The resulting convolution is most easily expressed as a transform, which gives rise to the usual product of transforms. Thus defining $W^*(s)$ to be the Laplace transform of the queueing time as in Eq. (5.103) (i.e., as $E[e^{-sw}]$), and defining

$$W^*(s | n) = E[e^{-sw} | \text{arrival queues and queue size} = n] \quad (6.24)$$

we have

$$W^*(s | n) = \left(\frac{m\mu}{s + m\mu} \right)^{n+1} \quad (6.25)$$

But clearly

$$W^*(s | \text{arrival queues}) = \sum_{n=0}^{\infty} W^*(s | n) P[\text{queue size} = n | \text{arrival queues}]$$

and so from Eqs. (6.25) and (6.23) we have

$$\begin{aligned} W^*(s | \text{arrival queues}) &= \sum_{n=0}^{\infty} (1 - \sigma)\sigma^n \left(\frac{m\mu}{s + m\mu} \right)^{n+1} \\ &= (1 - \sigma) \frac{m\mu}{s + m\mu - m\mu\sigma} \end{aligned}$$

Luckily, we recognize the inverse of this Laplace transform by inspection, thereby yielding the following conditional pdf for queueing time,

$$w(y | \text{arrival queues}) = (1 - \sigma)m\mu e^{-m\mu(1-\sigma)y} \quad y \geq 0 \quad - (6.26)$$

Quite a surprise! The conditional pdf for queueing time is *exponentially distributed* for the system G/M/m!

Thus far we have two principal results: first, that the conditional queue size is geometrically distributed *with* parameter σ as given in Eq. (6.23); and second, that the conditional pdf for queueing time is exponentially distributed with parameter $m\mu(1 - \sigma)$ as given in Eq. (6.26). The parameter σ is found as

the unique root in the range $0 < \sigma < 1$ of the functional equation (6.21). We are still searching for the distribution r_k and have carried that solution to the point of Eq. (6.20); we have as yet to evaluate the constant K as well as the first $m - 1$ terms in that distribution. Before we proceed with these last steps let us study an important special case.

6.4. THE QUEUE G/M/1

This is perhaps the most important system and forms the "dual" to the system M/G/1. Since $m = 1$ then Eq. (6.19) gives us the solution for r_k for all values of k , that is,

$$r_k = K\sigma^k \quad k = 0, 1, 2, \dots$$

K is now easily evaluated since these probabilities must sum to unity. From this we obtain immediately

$$r_k = (1 - \sigma)\sigma^k \quad k = 0, 1, 2, \dots \quad (6.27)$$

where, of course, σ is the unique root of

$$\sigma = A^*(\mu - \mu\sigma) \quad (6.28)$$

in the range $0 < \sigma < 1$. Thus the system G/M/1 gives rise to a geometric distribution for number of customers found in the system by an arrival; this applies as an unconditional statement regardless of the form for the interarrival distribution. We have already seen an example of this in Eq. (4.42) for the system E_r/M/1. We comment that the state probabilities, $P_k = P[k \text{ in system}]$, differ from Eq. (6.27) in that $P_0 = 1 - \rho$ whereas $r_0 = (1 - \sigma)$ and $p_k = p(1 - \sigma)\sigma^{k-1} = pr_{k-1}$ for $k = 1, 2, \dots$ {see Eq. (3.24), p. 209 of [COHE 69]}; in the M/G/1 queue we found $P_k = r_k$.

A customer will be forced to wait for service with probability $1 - r_0 = \sigma$, and so we may use Eq. (6.26) to obtain the unconditional distribution of waiting time as follows (where we define A to be the event "arrival queues" and A' , the complementary event):

$$\begin{aligned} W(y) &= P[\text{queueing time} \leq y] \\ &= 1 - P[\text{queueing time} > y | A]P[A] \\ &\quad - P[\text{queueing time} > y | A']P[A'] \end{aligned} \quad (6.29)$$

Clearly, the last term in this equation is zero; the remaining conditional probability in this last expression may be obtained by integrating Eq. (6.26) from y to infinity for $\mu = 1$; this computation gives $e^{-\mu(1-\sigma)y}$ and since σ is the

probability of queueing we have immediately from Eq. (6.29) that

$$W(y) = 1 - \sigma e^{-\mu(1-\sigma)y} \quad y \geq 0 \quad - (6.30)$$

We have the remarkable conclusion that the unconditional waiting-time distribution is exponential (with a jump of size $1 - \sigma$ at the origin) for the system G/M/1. If we compare this result to (5.123) and Figure 5.9, which gives the waiting-time distribution for M/M/1, we see that the results agree with ρ replacing a . That is, the queueing-time distribution for G/M/1 is of the *same form* as for M/M/1!

By straightforward calculation, we also have that the mean wait in G/M/1 is

$$\sigma \quad - (6.31)$$

Example

Let us now illustrate this method for the example M/M/1. Since $A(t) = 1 - e^{-\lambda t}$ ($t \geq 0$) we have immediately

$$A^*(s) = \frac{\lambda}{s + \lambda} \quad (6.32)$$

Using Eq. (6.28) we find that σ must satisfy

$$a = \frac{\lambda}{\mu - \sigma a + \lambda}$$

or

$$\mu\sigma^2 - (\mu + \lambda)\sigma + \lambda = 0$$

which yields

$$(a - 1)(\mu\sigma - \lambda) = 0$$

Of these two solutions for a , the case $\sigma = 1$ is unacceptable due to stability conditions ($0 < \sigma < 1$) and therefore the only acceptable solution is

$$a = \frac{\lambda}{\mu} = \rho \quad \text{M/M/1} \quad (6.33)$$

which yields from Eq. (6.27)

$$r_k = (1 - \rho)\rho^k \quad (6.34)$$

This, of course, is our usual solution for M/M/1. Further, using $\sigma = \rho$ as the value for σ in our waiting time distribution [Eq. (6.30)] we come up immediately with the known solution given in Eq. (5.123).

Example

As a second (slightly more Interesting) example let us consider a G/M/I system, with an interarrival time distribution such that

$$A^*(s) = \frac{2\mu^2}{(s + \mu)(s + 2\mu)} \quad (6.35)$$

Note that this corresponds to an $E_2/M/I$ system in which the two arrival stages have different death rates; we choose these rates to be linear multiples of the service rate μ . As always our first step is to evaluate σ from Eq. (6.28) and so we have

$$\sigma = \frac{2\mu^2}{(\mu - \mu\sigma + \mu)(\mu - \mu\sigma + 2\mu)}$$

This leads directly to the cubic equation

$$\sigma^3 - 5\sigma^2 + 6\sigma - 2 = 0$$

We know for sure that $\sigma = 1$ is always a root of Eq. (6.28), and this permits the straightforward factoring

$$(\sigma - 1)(\sigma - 2 - \sqrt{2})(\sigma - 2 + \sqrt{2}) = 0$$

Of these three roots it is clear that only $\sigma = 2 - \sqrt{2}$ is acceptable (since $0 < \sigma < 1$ is required). Therefore Eq. (6.27) immediately gives the distribution for number in system (seen by arrivals)

$$r_k = (\sqrt{2} - 1)(2 - \sqrt{2})^k \quad k = 0, 1, 2, \dots \quad (6.36)$$

Similarly we find

$$W(y) = 1 - (2 - \sqrt{2})e^{-\mu(\sqrt{2}-1)y} \quad y \geq 0 \quad (6.37)$$

for the waiting-time distribution.

Let us now return to the more general system G/M/m.

6.5. THE QUEUE G/M/m

At the end of Section 6.3 we pointed out that the only remaining unknowns for the general G/M/m solution were: K , an unknown constant, and the $m - 1$ "boundary" probabilities $r_0, r_1, \dots, r_{m-2}$. That is, our solution

appears in the form of Eq. (6.20); we may factor out the term $K\sigma^{m-1}$ to obtain

$$\mathbf{r} = K\sigma^{m-1}[R_0, R_1, \dots, R_{m-2}, 1, \sigma, \sigma^2, \sigma^3, \dots] \quad (6.38)$$

where

$$R_k = \frac{r_k \sigma^{1-m}}{K} \quad k = 0, 1, \dots, m-2 \quad (6.39)$$

Furthermore, for convenience we define

$$J = Ka^{m-1} \quad (6.40)$$

We have as yet not used the first $m-1$ equations represented by the matrix equation (6.6). We now require them for the evaluation of our unknown terms (of which there are $m-1$). In terms of our one-step transition probabilities p_{ij} we then have

$$e_k = \sum_{i=k-1}^{\infty} R_i p_{ik} \quad k = 0, 1, \dots, m-2$$

where we may extend the definition for R_k in Eq. (6.39) beyond $k = m-2$ by use of Eq. (6.19), that is, $R_i = \sigma^{i-m+1}$ for $i \geq m-1$. The tail of the sum above may be evaluated to give

$$e_k = \sum_{i=k-1}^{m-2} R_i p_{ik} + \sum_{i=m-1}^{\infty} \sigma^{i+1-m} p_{ik}$$

Solving for R_{k-1} , the lowest-order term present, we have

$$R_{k-1} = \frac{R_k - \sum_{i=k}^{m-2} R_i p_{ik} - \sum_{i=m-1}^{\infty} \sigma^{i+1-m} p_{ik}}{p_{k-1,k}} \quad (6.41)$$

for $k = 1, 2, \dots, m-1$. The set of equations (6.41) is a triangular set in the unknowns R_k ; in particular we may start with the fact that $R_{m-1} = 1$ [see Eq. (6.38)] and then solve recursively over the range $k = m-1, m-2, \dots, 1, 0$ in order. Finally we may use the conservation of probability to evaluate the constant J (this being equivalent to evaluating K) as

$$J \sum_{k=0}^{m-2} n_k + J \sum_{k=m-1}^{\infty} \sigma^{k-m+1} = 1$$

or

$$J = \frac{1}{1-a} + \sum_{k=0}^{m-2} R_k \quad (6.42)$$

This then provides a complete prescription for evaluating the distribution of the number of customers in the system. We point out that Takács [TAKA 62] gives an explicit (albeit complex) expression for these boundary probabilities.

Let us now determine the distribution of waiting time in this system [we already have seen the conditional distribution in Eq. (6.26)]. First we have the probability that an arriving customer need not queue, given by

$$W(0) = \sum_{k=0}^{m-1} r_k = J \sum_{k=0}^{m-1} s_k, \quad (6.43)$$

On the other hand, if a customer arrives to find $k \geq m$ others in the system he must wait until exactly $k - m + 1$ customers depart before he may enter service. Since there are m servers working continuously during his wait then the interdeparture times must be exponentially distributed with parameter $m\mu$, and so his waiting time must be of the form of a $(k - m + 1)$ -stage Erlangian distribution as given in Eq. (2.147). Thus for this case ($k \geq m$) we may write

$$P[\tilde{w} \leq Y \mid \text{customer finds } k \text{ in system}] = \int_0^y \frac{m\mu(m\mu x)^{k-m}}{(k-m)!} e^{-m\mu x} dx$$

If we now remove the condition on k we may write the unconditional distribution as

$$\begin{aligned} W(y) &= W(0) + J \sum_{k=m}^{\infty} \int_0^y \frac{(m\mu)(m\mu x)^{k-m} \sigma^{k-m+1}}{(k-m)!} e^{-m\mu x} dx \\ &= W(0) + J \int_0^y m\mu e^{-m\mu x(1-\sigma)} dx \end{aligned} \quad (6.44)$$

We may now use the expression for J in Eq. (6.42) and for $W(0)$ in Eq. (6.43) and carry out the integration in Eq. (6.44) to obtain

$$y \geq 0 \quad (J) \quad - \quad (6.45)$$

This is the final solution for our waiting-time distribution and shows that *in the general case GIM/m we still have the exponential distribution with an accumulation point at the origin for waiting time!*

We may calculate the average waiting time either from Eq. (6.45) or as follows. As we saw, a customer who arrives to find $k \geq m$ others in the system must wait until $k - m + 1$ services are complete, each of which takes on the average $1/m\mu$ sec. We now sum over all those cases where our

customer must wait to obtain

$$E[\tilde{w}] \triangleq W = \sum_{k=m}^{\infty} \frac{1}{m\mu} (k - m + 1)r_k$$

But in this range we know that $r_k = K\sigma^k$ and so

$$W = \frac{K}{m\mu} \sum_{k=m}^{\infty} (k - m + 1)\sigma^k$$

and this is easily calculated to yield

$$W = \frac{K\sigma^m}{m\mu(1-\sigma)^2} - \frac{J\sigma}{m\mu(1-\sigma)^2}$$

6.6. THE QUEUE G/M/2

Let us see how far we can get with the system G/M/2. From Eq. (6.19) we have immediately

$$r_k = K\sigma^k \quad k = 1, 2, \dots$$

Conserving probability we find

$$\sum_{k=0}^{\infty} r_k = 1 = r_0 + \sum_{k=1}^{\infty} K\sigma^k$$

This yields the following relationship between K and r_0 :

$$K = \frac{(1 - r_0)(1 - \sigma)}{\sigma} \quad (6.46)$$

Our task now is to find another relation between K and r_0 . This we may do from Eq. (6.41), which states

$$R_1 = \frac{\sum_{i=1}^{\infty} \sigma^{i-1} P_n}{P_{Ol}} \quad (6.47)$$

But $R_1 = 1$. The denominator is given by Eq. (6.9), namely,

$$P_{Ol} = \int_0^{\infty} \binom{1}{1} [1 - e^{-\mu t}]^0 e^{-\mu t} dA(t)$$

This we recognize as

$$P_{Ol} = A^*(\mu) \quad (6.48)$$

Regarding the one-step transition probabilities in the numerator sum of Eq. (6.47) we find they break into two regions: the term P_u must be calculated from Eq. (6.9) and the terms P_i for $i = 2, 3, 4, \dots$ must be calculated from

Eq. (6.13). Proceeding we have

$$P_u = \int_0^\infty \binom{2}{1} [1 - e^{-\mu t}] e^{-\mu t} dA(t)$$

Again we recognize this as the transform

$$P_a = 2A^*(\mu) - 2A^*(2\mu) \quad (6.49)$$

Also for $i = 2, 3, 4, \dots$, we have

$$P_{i1} = \int_0^\infty \binom{2}{1} e^{-\mu t} \left[\int_0^t \frac{(2\mu y)^{i-2}}{(i-2)!} (e^{-\mu y} - e^{-\mu t}) 2\mu dy \right] dA(t) \quad (6.50)$$

Substituting these last equations into Eq. (6.47) we then have

$$R_0 = \frac{1}{A^*(\mu)} \left[1 - 2A^*(\mu) + 2A^*(2\mu) - \sum_{i=2}^\infty \sigma^{i-1} P_{i1} \right] \quad (6.51)$$

The summation in this equation may be carried out within the integral signs of Eq. (6.50) to give

$$\sum_{i=2}^\infty \sigma^{i-1} P_{i1} = 2A^*(2\mu) + \frac{2A^*(2\mu - 2\mu\sigma) - 4aA^*(\mu)}{2\sigma' - 1} \quad (6.52)$$

But from Eq. (6.21) we recognize that $a = A^*(2\mu - 2\mu\sigma)$ and so we have

$$\sum_{i=2}^\infty \sigma^{i-1} P_{i1} = 2A^*(2\mu) + \frac{2\sigma'}{2\sigma' - 1} [1 - 2A^*(\mu)]$$

Substituting back into Eq. (6.51) we find

$$R_0 = \frac{2A^*(\mu) - 1}{(2\sigma' - 1)A^*(\mu)}$$

However from Eq. (6.39) we know that

$$R_0 = \frac{r_0}{Ka}$$

and so we may express r_0 as

$$r_0 = \frac{Ka[1 - 2A^*(\mu)]}{(1 - 2\sigma)A^*(\mu)} \quad (6.53)$$

Thus Eqs. (6.46) and (6.53) give us two equations in our two unknowns K and r_0 , which when solved simultaneously lead to

$$r_0 = \frac{(1 - \sigma')[1 - 2A^*(\mu)]}{1 - \sigma - A^*(\mu)}$$

$$K = A^*(\mu) \frac{(1 - \sigma')(1 - 2\sigma')}{\sigma[1 - a - A^*(\mu)]}$$

Comparing Eq. (6.56) with our results from Chapter 3 [Eqs. (3.37) and (3.39)] we find that they agree for $m = 2$.

This completes our study of the $G[M|m]$ queue. Some further results of interest may be found in [DESM 73]. In the next chapter, we view transforms as probabilities and gain considerable reduction in the analytic effort required to solve equilibrium and transient queueing problems.

REFERENCES

- COHE 69 Cohen, J. W., *The Single Server Queue*, Wiley (New York) 1969.
 DESM 73 De Smit, J. H. A., "On the Many Server Queue with Exponential Service Times," *Advances in Applied Probability*, 5, 170-182 (1973).
 KEND 51 Kendall, D. G., "Some Problems in the Theory of Queues," *Journal of the Royal Statistical Society, Ser. E.*, 13, 151-185 (1951).
 TAKA 62 Takács, L., *Introduction to the Theory of Queues*, Oxford University Press (New York) 1962.

EXERCISES

- 6.1. Prove Eq. (6.13). [HINT: condition on an interarrival time of duration t and then further condition on the time ($\leq t$) it will take to empty the queue.]
- 6.2. Consider $E_2[M|I]$ (with infinite queueing room).
 - (a) Solve for r_k in terms of σ .
 - (b) Evaluate σ explicitly.
- 6.3. Consider $M[M|m]$.
 - (a) How do p_k and r_k compare?
 - (b) Compare Eqs. (6.22) and (3.40).
- 6.4. Prove Eq. (6.31).
- 6.5. Show that Eq. (6.52) follows from Eq. (6.50).
- 6.6. Consider an H_2/Mfl system in which $\lambda_1 = 2$, $\lambda_2 = 1$, $\mu = 2$, and $\alpha_1 = 5/8$.
 - (a) Find σ .
 - (b) Find r_k .
 - (c) Find $w(y)$.
 - (d) Find W .
- 6.7. Consider a $D[Mfl]$ system with $\mu = 2$ and with the same p as in the previous exercise.
 - (a) Find σ (correct to two decimal places).

- (b) Find r_k .
- (c) Find $w(y)$.
- (d) Find W .

- 6.8. Consider a G/M/I queueing system with room for at most two customers (one in service plus one waiting). Find r_k ($k = 0, 1, 2$) in terms of μ and $A^{*(5)}$.
- 6.9. Consider a G/M/I system in which the cost of making a customer wait y sec is

$$c(y) = ae^{by}$$

- (3) Find the average cost of queueing for a customer.
- (b) Under what conditions will the average cost be finite?

The Method of Collective Marks

When one studies stochastic processes such as in queueing theory, one finds that the work divides into two parts. The first part typically requires a careful *probabilistic argument* in order to arrive at expressions involving the random variables of interest.* The second part is then one of analysis in which the *formal manipulation* of symbols takes place either in the original domain or in some transformed domain. Whereas the probabilistic arguments typically must be made with great care, they nevertheless leave one with a comfortable feeling that the "physics" of the situation are constantly within one's understanding and grasp. On the other hand, whereas the analytic manipulations that one carries out in the second part tend to be rather straightforward (albeit difficult) formal operations, one is unfortunately left with the uneasy feeling that these manipulations relate back to the original problem in no clearly understandable fashion. This "nonphysical" aspect to problem solving typically is taken on when one moves into the domain of transforms, (either Laplace or z-transforms).

In this chapter we demonstrate that one may deal with transforms and still maintain a handle on the probabilistic arguments taking place as these transforms are manipulated. There are two separate operations involved: the "marking" of customers; and the observation of "catastrophe" processes. Together these methods are referred to as the method of *collective marks*. Both operations need not necessarily be used simultaneously, and we study them separately below. This material is drawn principally from [RUNN 65]; these ideas were introduced by van Dantzig [VAN 48] in order to expose the probabilistic interpretation for transforms.

7.1. THE MARKING OF CUSTOMERS

Assume that, at the entrance to a queueing system, there is a gremlin who marks (i.e., tags) arriving customers with the following probabilities:

$$P[\text{customer is marked}] = 1 - z \quad (7.1)$$

$$P[\text{customer is not marked}] = z \quad (7.2)$$

* As, for example, the arguments leading up to Eqs. (5.31) and (6.1).

where $0 \leq z \leq 1$. We assume that the gremlin marks customers with these probabilities independent of *all* other aspects of the queueing process. As we shall see below, this marking process allows us to create generating functions in a very natural way.

It is most instructive if we illustrate the use of this marking process by examples:

Example 1: Poisson Arrivals

We first consider a Poisson arrival process with a mean arrival rate of λ customers per second. Assume that customers are marked as above. Let us consider the probability

$$q(z, t) \triangleq P[\text{no marked customers arrive in } (0, t)] \quad (7.3)$$

It is clear that k customers will arrive in the interval $(0, t)$ with the probability $(\lambda t)^k e^{-\lambda t} / k!$. Moreover, with probability z^k , none of these k customers will be marked; this last is true since marking takes place independently among customers. Now summing over all values of k we have immediately that

$$\begin{aligned} q(z, t) &= \sum_{k=0}^{\infty} \frac{(\lambda t)^k e^{-\lambda t}}{k!} z^k \\ &= e^{\lambda t(z-1)} \end{aligned} \quad (7.4)$$

Going back to Eq. (2.134) we see that Eq. (7.4) is merely the generating function for a Poisson arrival process. We thus conclude that the generating function for this arrival process may also be interpreted as the probabilistic quantity expressed in Eq. (7.3). This will not be the first time we may give a probabilistic interpretation for a generating function!

Example 2: $M/M/\infty$

We consider the birth-death queueing system with an infinite number of servers. We also assume at time $t = 0$ that there are i customers present. The parameters of our system as usual are λ and μ [i.e., $A(t) = 1 - e^{-\lambda t}$ and $B(x) = 1 - e^{-\mu x}$].

We are interested in the quantity

$$P_k(t) = P[k \text{ customers in the system at time } t] \quad (7.5)$$

and we define its generating function as we did in Eq. (2.153) to be

$$P(z, t) = \sum_{k=0}^{\infty} P_k(t) z^k \quad (7.6)$$

Once again we mark customers according to Eqs. (7.1) and (7.2). In analogy with Example I, we recognize that Eq. (7.6) may be interpreted as the probability that the system contains no marked customers at time t (where the term z^k again represents the probability that none of the k customers present is marked). Here then is our crucial observation: We may calculate $P(z, t)$ *directly* by finding the probability that there are no marked customers in the system at time t , rather than calculating $P_k(t)$ and then finding its z -transform!

We proceed as follows: We need merely find the probability that none of the customers still present in the system at time t is marked and this we do by accounting for all customers present at time 0 as well as all customers who arrive in the interval $(0, t)$. For any customer present at time 0 we may calculate the probability that he is still present at time t *and* is marked as $(1 - z)[1 - B(t)]$ where the first factor gives the probability that our customer was marked in the first place and the second factor gives the probability that his service time is greater than t . Clearly, then, this quantity subtracted from unity is the probability that a customer originally present is not a marked customer present at time t ; and so we have

$$\begin{aligned} P[\text{customer present initially is not a marked customer present at time } t] \\ = 1 - (1 - z)e^{-\mu t} \end{aligned}$$

Now for the new customers who enter in the interval $(0, t)$, we have as before $P[k \text{ arrivals in } (0; t)] = (\lambda t)^k e^{-\lambda t} / k!$. Given that k have arrived in this interval then their arrival instants are uniformly distributed over this interval [see Eq. (2.136)]. Let us consider one such arriving customer and assume that he arrives at a time $\tau < t$. Such a customer will not be a marked customer present at time t with probability

$$\begin{aligned} P[\text{new arrival is not a marked customer present at time } t \text{ given he} \\ \text{arrived at } \tau \leq t] = 1 - (1 - z)[1 - B(t - \tau)] \quad (7.7) \end{aligned}$$

However, we have that

$$P[\text{arrival time} \leq \tau] = \frac{\tau}{t} \quad \text{for } 0 \leq \tau \leq t$$

and so

$$\begin{aligned} P[\text{new arrival still in system at } t] &= \int_{\tau=0}^t e^{-\mu(t-\tau)} \frac{d\tau}{t} \\ &= \frac{1 - e^{-\mu t}}{\mu t} \quad (7.8) \end{aligned}$$

Unconditioning the arrival time from Eq. (7.7) as shown in Eq. (7.8) we have

$$P[\text{new arrival is not a marked customer present at } t] = 1 - (1 - z) \frac{1 - e^{-\mu t}}{\mu t}$$

Thus we may calculate the probability that there are no marked customers at time t as follows:

$$P(z, t) = \sum_{k=0}^{\infty} P[k \text{ arrive in } (0, t)] \\ \times \{P[\text{new arrival is not a marked customer present at } t]\}^k \\ \times \{P[\text{initial customer is not a marked customer present at } t]\}^i$$

Using our established relationships we arrive at

$$P(z, t) = \sum_{k=0}^{\infty} \frac{(\lambda t)^k}{k!} e^{-\lambda t} [1 - (1 - z)]^{1 - \frac{e^{-\mu t}}{\mu t}} [1 - (1 - z)e^{-\mu t}]^i$$

which then gives the known result

$$P(z, t) = [1 - (1 - z)e^{-\mu t}]^i e^{-(\lambda/\mu)(1-z)[1-e^{-\mu t}]} \quad (7.9)$$

It should be clear to the student that the usual method for obtaining this result would have been extremely complex.

Example 3: MJGI

In this example we consider the FCFS MJGI system. Recall that the random variables $w, t, t_1, t_2, \dots, X, X_1, \dots$ are all independent of each other. As usual, we define $B^*(s)$ and $W_n^*(s)$ as the Laplace transform's for the service-time pdf $b(x)$ and the waiting-time pdf $w_n(y)$ for C_n , respectively. We define the event

$$\{no \text{ M in } w.\} \triangleq \{no \text{ customers who arrive during the waiting time of } C_n \text{ are marked}\} \quad (7.10)$$

We wish to find the probability of this event, that is, $P[no \text{ M in } w.]$. Conditioning on the number of arriving customers and on the waiting time w_n , and then removing these conditions, we have

$$P[no \text{ M in } w.] = \sum_{k=0}^{\infty} \int_0^{\infty} \frac{(\lambda y)^k}{k!} e^{-\lambda y} z^k dW_n(y) \\ = \int_0^{\infty} e^{-\lambda y(1-z)} dW_n(y)$$

We recognize the integral as $W_n^*(\lambda - \lambda z)$ and so

$$P[no \text{ M in } w.] = W_n^*(\lambda - \lambda z) \quad (7.11)$$

Thus once again we have a very simple probabilistic interpretation for the (Laplace) transform of an important distribution. By identical arguments we may arrive at

$$P[no \text{ M in } x.] = B^*(\lambda - \lambda z) \quad (7.12)$$

This last gives us another interpretation for an old expression we have seen in Chapter 5.

Now comes a startling insight! It is clear that the arrival of customers during the waiting time of C_n and the arrival of customers during the service time of C ; must be independent events since these are nonoverlapping intervals and our arrival process is memoryless. Thus the events of no marked customers arriving in each of these two disjoint intervals of time must be independent, and so the probability that no marked customers arrive in the union of these two disjoint intervals must be the product of the probabilities that none such arrive in each of the intervals separately. Thus we may write

$$P[\text{no } M \text{ in } w_n + x_n] = P[\text{no } M \text{ in } w_n]P[\text{no } M \text{ in } x_n] \quad (7.13)$$

$$= W_n^*(\lambda - \lambda z)B^*(\lambda - \lambda z) \quad (7.14)$$

This last result is pleasing in two ways. First, because it says that the probability of two independent joint events is equal to the product of the probabilities of the individual events [Eq. (7.13)]. Second, because it says that the transform of the pdf of the sum of two independent random variables is equal to the product of the transforms of the pdf of the individual random variables [Eq. (7.14)]. Thus two familiar results (regarding disjoint events and regarding sums of independent random variables) have led to a meaningful new insight, namely, that multiplication of transforms implies not only the sum of two independent random variables, but also implies the product of the probabilities of two independent events! Not often are we privileged to see such fundamental principles related.

Let us now continue with the argument. At the moment we have Eq. (7.14) as one means for expressing the probability that no marked customers arrive during the interval $w_n + x_n$. We now proceed to calculate this probability by a second argument. Of course we have

$$P[\text{no } M \text{ in } w_n + x_n] = P[\text{no } M \text{ in } w_n + x_n \text{ and } C_{nH} \text{ marked}] \\ + P[\text{no } M \text{ in } w_n + x_n \text{ and } C_{nH} \text{ not marked}] \quad (7.15)$$

Furthermore, we have

$$P[\text{no } M \text{ in } w_n + x_n \text{ and } C_{n+1} \text{ marked}] = 0 \quad \text{if } w_{n+1} > 0$$

since if C_{nH} must wait, then he must arrive in the interval $w_n + x_n$ and it is impossible for him to be marked and still to have the event $\{\text{no } M \text{ in } w_n + x_n\}$. Thus the first term on the right-hand side of Eq. (7.15) must be $P[w_{nH} = 0](1 - z)$ where this second factor is merely the probability that C_{n+1} is marked. Now consider the second term on the right-hand side of Eq. (7.15); as shown in Figure 7.1 it is clear that no customers arrive between C_n and C_{nH} , and therefore the customers of interest (namely, those arriving after

Now, for $p < 1$, we have an ergodic process with $W_O(s) = \lim_{n \rightarrow \infty} W_n^*(s)$ as $n \rightarrow \infty$. Equation (7.17) then reduces to

$$1V^*(i - \lambda z)B^*(\lambda - iz) = (1 - z)W^*(\lambda)B^*(\lambda) + zW^*(\lambda - iz)$$

If we make the change variable $s = \lambda - \lambda z$ and solve for $W^*(s)$, we obtain

$$W_O(s) = \frac{sW^*(i)B^*(i)}{s - \lambda + \lambda B^*(s)}$$

Since $W^*(0) = 1$, we evaluate $W^*(i)B^*(i) = (1 - p)$ and arrive at

$$W_O(s) = \frac{s(1 - p)}{s - \lambda + iB^*(s)} \quad (7.18)$$

which, of course, is the P-K transform equation for waiting time.

We have demonstrated three examples where the marking of customers has allowed us to argue purely with probabilistic reasoning to derive expressions relating transforms. What we have here traded has been straightforward but tedious analysis for deep but physical probabilistic reasoning. We now consider the catastrophe process.

7.2. THE CATASTROPHE PROCESS

Let us pursue the method of collective marks a bit further by observing "catastrophe" processes. Measuring from time 0 let us consider that some event occurs at time t ($t \geq 0$), where the pdf associated with the time of occurrence of this event is given by $f(t)$. Furthermore, let there be an independent "catastrophe" process taking place simultaneously which generates catastrophes at a rate γ according to a Poisson process.

We wish to calculate the probability that the event at time t takes place before the first catastrophe (measuring from time 0). Conditioning on t and integrating over all t , we get

$$\begin{aligned} \text{Prevent occurs before catastrophe} &= \int_0^\infty e^{-Y} f(t) dt \\ &= F^*(\gamma) \end{aligned} \quad (7.19)$$

where, as usual, $f(t) \leftrightarrow F^*(s)$ are Laplace transform pairs. Thus we have a probabilistic interpretation for the Laplace transform (evaluated at the point γ) of the pdf for the time of occurrence of the event, namely, it is the probability that an event with this pdf occurs before a Poisson catastrophe at rate γ occurs.

† A catastrophe is merely an impressive name given to these generated times to distinguish them from the "event" of interest at time t .

As a second illustration using catastrophe processes, consider a sequence of events (that is, a point process) on the interval $(0, \infty)$. Measuring from time 0 we would like to calculate the pdf of the time until the n th event, which we denote by $f(n, I)$, and with distribution $F(n, I)$, where the time between events is given as before with density $f(I)$. That is,

$$F(n, I) = P[\text{nth event has occurred by time } I]$$

We are interested in deriving an expression for the renewal function $H(I)$, which we recall from Section 5.2 is equal to the expected number of events (renewals) in an interval of length I . We proceed by defining

$$P_n(I) = P[\text{exactly } n \text{ events occur in } (0, I)] \quad (7.20)$$

The renewal function may therefore be calculated as

$$\begin{aligned} H(I) &= E[\text{number of events in } (0, I)] \\ &= \sum_{n=0}^{\infty} n P_n(I) \end{aligned}$$

But from its definition we see that $P_n(I) = F(n, I) - F(n+1, I)$ and so we have

$$\begin{aligned} H(I) &= \sum_{n=0}^{\infty} n [F(n, I) - F(n+1, I)] \\ &= \sum_{n=1}^{\infty} F(n, I) \end{aligned} \quad (7.21)$$

If we now permit a Poisson catastrophe process (at rate γ) to develop we may ask for the expectation of the following random variable:

$$N; \triangleq \text{number of events occurring before the first catastrophe} \quad (7.22)$$

With probability $\gamma e^{-\gamma t} dt$ the first catastrophe will occur in the interval $(t, t + dt)$ and then $H(I)$ will give the expected number of events occurring before this first catastrophe, that is,

$$H(t) = E[N_e | \text{first catastrophe occurs in } (t, t + dt)]$$

Summing over all possibilities we may then write

$$E[N_e] = \int_0^{\infty} H(t) \gamma e^{-\gamma t} dt \quad (7.23)$$

In Section 5.2 we had defined $H^*(s)$ to be the Laplace transform of the renewal density $h(I)$ defined as $h(I) \triangleq dH(I)/dI$, that is,

$$H^*(s) \triangleq \int_0^{\infty} h(I) e^{-sI} dI \quad (7.24)$$

[†] We use the subscript (n) to remind the reader of the definition in Eq. (5.110) denoting the n -fold convolution. We see that $f^{(n)}(t)$ is indeed the n -fold convolution of the lifetime density $J(t)$.

If we integrate this last equation by parts, we see that the right-hand side of Eq. (7.24) is merely $\int_0^\infty sH(t)e^{-t} dt$ and so from Eq. (7.23) we have (making the substitution $s = y$)

$$E[N_e] = H^*(y) \quad (7.25)$$

Let us now calculate $E[N_e]$ by an alternate means. From Eq. (7.19) we see that the catastrophe will occur before the first event with probability $1 - F^*(y)$ and in this case $N_e = 0$. On the other hand, with probability $F^*(y)$ we will get at least one event occurring before the catastrophe. Let N'_e be the random variable $N_e - 1$ conditioned on at least one event; then we have $N_e = 1 + N'_e$. Because of the memoryless property of the Poisson process as well as the fact that the event occurrences generate an imbedded Markov process we see that N'_e must have the same distribution as N_e itself. Forming expectations on N_e we may therefore write

$$E[N_e] = 0[1 - F^*(y)] + \{1 + E[N_e]\}F^*(y)$$

This gives immediately

$$E[N_e] = \frac{F^*(y)}{1 - F^*(y)} \quad (7.26)$$

We now have two expressions for $E[N_e]$ and so by equating them (and making the change of variable $s = y$) we have the final result

$$H^*(s) = \frac{F^*(s)}{1 - F^*(s)} \quad (7.27)$$

This last we recognize as the transform expression for the integral equation of renewal theory [see Eq. (5.21)]; its integral formulation is given in Eq. (5.22).

It is fair to say that the method of collective marks is a rather elegant way to get some useful and important results in the theory of stochastic processes. On the other hand, this method has as yet yielded no results that were not previously known through the application of other methods. Thus at present its principal use lies in providing an alternative way for viewing the fundamental relationships, thereby enhancing one's insight into the probabilistic structure of these processes.

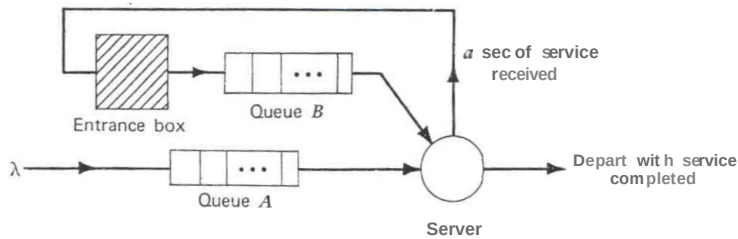
Thus ends our treatment of intermediate queueing theory. In the next part, we venture into the kingdom of the *GIGI* queue.

REFERENCES

- RUNN 65 Runnenburg, J. Th., "On the Use of the Method of Collective Marks in Queueing Theory," *Proc. Symposium on Congestion Theory*, eds, W. L. Smith and W. E. Wilkinson, University of North Carolina Press (1965).
- VAN 48 van Dantzig, D., "Sur la methode des fonctions génératrices," *Colloques internationaux du CNRS*, **13**, 29-45 (1948).

EXERCISES

- 7.1. Consider the M/G/1 system shown in the figure below with average arrival rate λ and service-time distribution $= B(x)$. Customers are served first-come-first-served from queue A until they either leave or receive a sec of service, at which time they join an entrance box as shown in the figure. Customers continue to collect in the entrance box forming



a group until queue A empties and the server becomes free. At this point, the entrance box "dumps" all it has collected as a *bulk arrival* to queue B. Queue B will receive service until a new arrival (to be referred to as a "starter") joins queue A at which time the server switches from queue B to serve queue A and the customer who is preempted returns to the head of queue B. The entrance box then begins to fill and the process repeats. Let

$$g_n = P[\text{entrance box delivers bulk of size } n \text{ to queue } B]$$

$$G(z) = \sum_{n=0}^{\infty} g_n z^n$$

- Give a probabilistic interpretation for $G(z)$ using the method of collective marks.
- Given that the "starter" reaches the entrance box, and using the method of collective marks find [in terms of λ , a , $B(\cdot)$, and $G(z)$]

$$P_k = P[k \text{ customers arrive to queue A during the "starter's" service time and no marked customers arrive to the entrance box from the } k \text{ sub-busy periods created in queue A by each of these customers}]$$

- Given that the "starter" does *not* reach the entrance box, find P_k as defined above.
- From (b) and (c), give an expression (involving an integral) for $G(z)$ in terms of λ , a , $B(\cdot)$, and itself.
- From (d) find the average bulk size $\bar{n} = \sum_{n=0}^{\infty} n g_n$.

- 7.2.** Consider the $M/G/\infty$ system. We wish to find $P(z, I)$ as defined in Eq. (7.6). Assume the system contains $i = 0$ customers at $t = 0$. Let $p(I)$ be the probability that a customer who arrived in the interval $(0, I)$ is still present at I . Proceed as in Example 2 of Section 7.1.
- Express $p(I)$ in terms of $B(x)$.
 - Find $P(z, t)$ in terms of λ, I, z , and $p(I)$.
 - From (b) find $P_k(I)$ defined in Eq. (7.5).
 - From (c), find $\lim P_k(t) = P_k$ as $t \rightarrow \infty$.
- 7.3.** Consider an $M/G/1$ queue, which is idle at time 0. Let $p = P[\text{no catastrophe occurs during the time the server is busy with those customers who arrived during } (0, I)]$ and let $q = P[\text{no catastrophe occurs during } (0, t + U(I))] \text{ where } U(I) \text{ is the unfinished work at time } t$. Catastrophes occur at a rate γ .
- Find p .
 - Find q .
 - Interpret $p - q$ as a probability and find an independent expression for it. We may then use (a) and (b) to relate the distribution of unfinished work to $\mathcal{B}^*(s)$.
- 7.4.** Consider the $G/M/m$ system. The root σ , which is defined in Eq. (6.21) plays a central role in the solution. Examine Eq. (6.21) from the viewpoint of collective marks and give a probabilistic interpretation for σ .

PART IV

ADVANCED MATERIAL

We find ourselves in difficult terrain as we enter the foothills of $G/G/I$. Not even the average waiting time is known for this queue! In Chapter 8, we nevertheless develop a "spectral" method for handling these systems which often leads to useful results. The difficult part of this method reduces to locating the roots of a function, as we have so often seen before. The spectral method suffers from the disadvantage of not providing one with the general behavior pattern of the system; each new queue must be studied by itself. However, we do discuss Kingman's algebra for queues, which so nicely exposes the common framework for all of the various methods so far used to attack the $G/G/I$ queue. Finally, we introduce the concept of a dual queue, and express some of our principal results in terms of idle times and dual queues.

The Queue $G/G/1$

We have so far made effective use of the Markovian property in the queueing systems $M/M/1$, $M/G/1$, and $G/M/m$. We must now leave behind many (but not all) of the simplifications that derive from the Markovian property and find new methods for studying the more difficult system $GIG/1$.

In this chapter we solve the $G/G/1$ system equations by *spectral* methods, making use of transform and complex-variable techniques. There are, however, numerous other approaches: In Section 5.11 we introduced the ladder indices and pointed out the way in which they were related to important events in queueing systems; these ideas can be extended and applied to the general system $GIG/1$. *Fluctuations* of sums of random variables (i.e., the ladder indices) have been studied by Andersen [ANDE 53a, ANDE 53b, ANDE 54] and also by Spitzer [SPIT 56, SPIT 60], who simplified and expanded Andersen's work. This led, among other things, to *Spitzer's identity*, of great importance in that approach to queueing theory. Much earlier (in the 1930's) Pollaczek considered a formalism for solving these systems and his approach (summarized in 1957 [POLL 57]) is now referred to as *Pollaczek's method*. More recently, Kingman [KING 66] has developed an *algebra for queues*, which places all these methods in a common framework and exposes the underlying similarity among them; he also identifies where the problem gets difficult and why, but unfortunately he shows that this method does not extend to the multiple server system. Keilson [KEIL 65] applies the method of *Green's function*. Benes [BENE 63] studied $G/G/1$ through the *unfinished work* and its "relatives."

Let us now establish the basic equations for this system.

8.1. LINDLEY'S INTEGRAL EQUATION

The system under consideration is one in which the interarrival times between customers are independent and are given by an arbitrary distribution $A(t)$. The service times are also independently drawn from an arbitrary distribution given by $B(x)$. We assume there is one server available and that service is offered in a first-come-first-served order. The basic relationship

among the pertinent random variables is derived in this section and leads to Lindley's integral equation, whose solution is given in the following section.

We consider a sequence of arriving customers indexed by the subscript n and remind the reader of our earlier notation :

- C_n = the n th customer arriving to the system
- $t_n = \tau_n - \tau_{n-1}$ = interarrival time between C_{n-1} and C_n
- x_n = service time for C_n
- W_n = waiting time (in queue) for C_n

We assume that the random variables $\{t_n\}$ and $\{x_n\}$ are independent and are given, respectively, by the distribution functions $A(t)$ and $R(x)$ independent of the subscript n . As always, we look for a Markov process to simplify our analysis. Recall for M/G/I, that the unfinished work $U(t)$ is a Markov process for all t . For G/G/I, it should be clear that although $U(t)$ is no longer Markovian, imbedded within $U(t)$ is a crucial Markov process defined at the *customer-arrival times*. At these regeneration points, all of the past history that is pertinent to future behavior is completely summarized in the current value of $U(t)$. That is, for FCFS systems, the value of the unfinished work just prior to the arrival of C_n is exactly equal to his waiting time (W_n) and *this Markov process is the object of our study*. In Figures 8.1 and 8.2 we use the time-diagram notation for queues (as defined in Figure 2.2) to illustrate the history of W_n in two cases: Figure 8.1 displays the case where C_{n+1} arrives to the system before C_n departs from the service facility; and Figure 8.2 shows the case in which C_{n+1} arrives to an empty system. For the conditions of Figure 8.1 it is clear that

$$t_{n+1} + w_{n+1} = w_n + x_n$$

That is,

$$w_{n+1} = w_n + x_n - t_{n+1} \quad \text{if} \quad w_n + x_n - t_{n+1} \geq 0 \quad (8.1)$$

The condition expressed in Eq. (8.1) assures that C_{n+1} arrives to find a busy

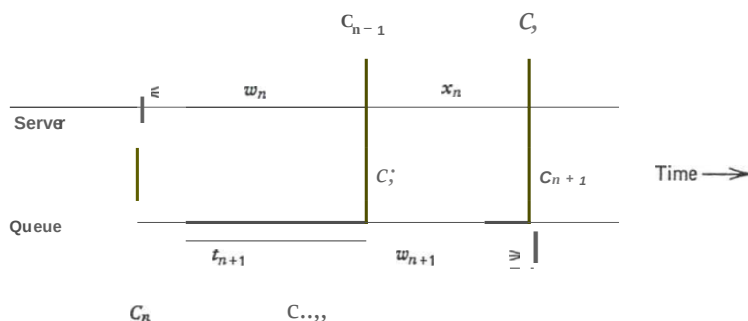


Figure 8.1 The case where C_{n+1} arrives to find a busy system.

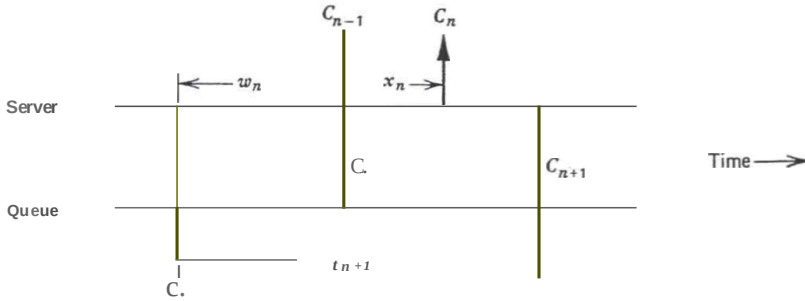


Figure 8.2 The case where C_{n+1} arrives to find an idle system.

system. From Figure 8.2 we see immediately that

$$w_{n+1} = 0 \quad \text{if} \quad w_n + x_n - t_{n+1} \leq 0 \quad (8.2)$$

where the condition in Eq. (8.2) assures that C_{n+1} arrives to find an idle system. For convenience we now define a new (key) random variable u_n as

$$u_n \triangleq x_n - t_{n+1} \quad (8.3)$$

This random variable is merely the difference between the service time for C ; and the interarrival time between C_{n+1} and C_n (for a stable system we will require that the expectation of u_n be negative). We may thus combine Eqs. (8.1)–(8.3) to obtain the following fundamental and yet elementary relationship, first established by Lindley [LIND 52];

$$w_{n+1} = \begin{cases} w_n + u_n & \text{if } w_n + u_n \geq 0 \\ 0 & \text{if } w_n + u_n \leq 0 \end{cases} \quad (8.4)$$

The term $w_n + u_n$ is merely the sum of the unfinished work (w_n) found by C ; plus the service time (x_n), which he now adds to the unfinished work, less the time duration (t_{n+1}) until the arrival of the next customer C_{n+1} ; if this quantity is nonnegative then it represents the amount of unfinished work found by C_{n+1} and therefore represents his waiting time w_{n+1} . However, if this quantity goes negative it indicates that an interval of time has elapsed since the arrival of C_n , which exceeds the amount of unfinished work present in the system just after the arrival of C_n thereby indicating that the system has gone idle by the time C_{n+1} arrives.

We may write Eq. (8.4) as

$$w_{n+1} = \max [0, w_n + u_n] \quad (8.5)$$

We introduce the notation $(x)^+ \triangleq \max [0, x]$; we then have

$$w_{n+1} = (w_n + u_n)^+ \quad (8.6)$$

Since the random variables $\{I_n\}$ and $\{x_n\}$ are independent among themselves and each other, then one observes that the sequence of random variables $\{I_0, w_1, IV_Z, \dots\}$ forms a Markov process with stationary transition probabilities. This can be seen immediately from Eq. (8.4) since the new value IV_{n+1} depends upon the previous sequence of random variables w_i ($i = 0, 1, \dots, n$) only through the most recent value IV_n plus a random variable I_{n+1} which is independent of the random variables w_i for all $i \leq n$.

Let us solve Eq. (8.5) recursively beginning with w_0 as an initial condition. We have (defining Co to be our initial arrival)

$$\begin{aligned} IV_1 &= (IV_0 + I_1)^+ \\ IV_Z &= (w_1 + I_1)^+ = \max [0, w_1 + I_1] \\ &= \max [0, I_1 + \max (0, w_0 + I_0)] \\ &= \max [0, I_1, I_1 + I_0 + IV_0] \\ w_3 &= (w_Z + I_Z)^+ = \max [0, w_Z + I_Z] \\ &= \max [0, I_Z + \max (0, u_1, I_1 + I_0 + IV_0)] \\ &= \max [0, I_Z, I_Z + I_1, I_Z + I_1 + I_0 + w_0] \end{aligned}$$

$$\begin{aligned} w_n &= (IV_{n-1} + I_{n-1})^+ = \max [0, w_{n-1} + u_{n-1}] \\ &= \max [0, u_{n-1}, u_{n-1} + u_{n-2}, \dots, u_{n-1} + \dots + u_1, \\ &\quad I_{n-1} + \dots + I_1 + I_0 + IV_0] \end{aligned} \quad (8.7)$$

However, since the sequence of random variables $\{u_i\}$ is a sequence of independent and identically distributed random variables, then they are "interchangeable" and we may consider a new random variable w_n' with the same distribution as w_n , where

$$\begin{aligned} w_n' &\triangleq \max [0, I_0, I_0 + u_1, I_0 + I_1 + I_Z, \dots, I_0 + I_1 + \dots + I_{n-2} \\ &\quad I_0 + I_1 + \dots + I_{n-2} + I_{n-1} + w_0] \end{aligned} \quad (8.8)$$

Equation (8.8) is obtained from Eq. (8.7) by relabeling the random variables u_i . It is now convenient to define the quantities U_i as

$$U_i = \sum_{j=0}^{n-1} u_j \quad (8.9)$$

$$U_0 = 0$$

We thus have from Eq. (8.8)

$$w_n' = \max [U_0, U_1, U_2, \dots, U_{n-1}, U_n + w_0] \quad (8.10)$$

From this last form we see for $w_0 = 0$ that w_n' can only increase with n . Therefore the limiting random variable $\lim w_n'$ as $n \rightarrow \infty$ must converge to the (possibly infinite) random variable $\tilde{w}$

$$\tilde{w} = \sup_{n \geq 0} U'; \quad - (8.1!)$$

Our imbedded Markov chain is ergodic if, with probability one, w_n is finite, and if so, then the distribution of w_n and of w_n' both converge to the distribution of $\tilde{w}$; in this case, the distribution of $\tilde{w}$ is the waiting-time distribution. Lindley [LIND 52] has shown that for $0 < E[un] < \infty$ then the system is stable if and only if $E[un] < 0$. Therefore, we will henceforth assume

$$E[un] < 0 \quad (8.12)$$

Equation (8.12) is our usual condition for stability as may be seen from the following:

$$\begin{aligned} E[un] &= E[x_n - t_{n+1}] \\ &= E[x_n] - E[t_{n+1}] \\ &= \bar{x} - \bar{t} \\ &= i(p - 1) \end{aligned} \quad (8.13)$$

where as usual we assume that the expected service time is $\bar{x}$ and the expected interarrival time is $\bar{t}$ (and we have $p = \bar{x}/\bar{t}$). From Eqs. (8.12) and (8.13) we see we have required that $p < 1$, as is our usual condition for stability. Let us denote (as usual) the stationary distribution for w_n (and also therefore for w_n') by

$$\lim_{n \rightarrow \infty} P[w_n \leq y] = \lim_{n \rightarrow \infty} P[w_n' \leq y] = W(y) \quad (8.14)$$

which must exist for $p < 1$ [LIND 52]. Thus $W(y)$ will be our assumed stationary distribution for time spent in queue; we will not dwell upon the proof of its existence but rather upon the method for its calculation. As we know for such Markov processes, this limiting distribution is independent of the initial state w_0 .

Before proceeding to the formal derivation of results let us investigate the way in which Eq. (8.7) in fact produces the waiting time. This we do by example; consider Figure 8.3, which represents the unfinished work $U(t)$. For the sequence of arrivals and departures given in this figure, we present the table below showing the interarrival times t_{n+1} , service times x_n' , the random variables u_n' and the waiting time w_n as measured from the diagram; in the last row of this table we give the waiting times w_n as calculated from Eq. (8.7) as follows.

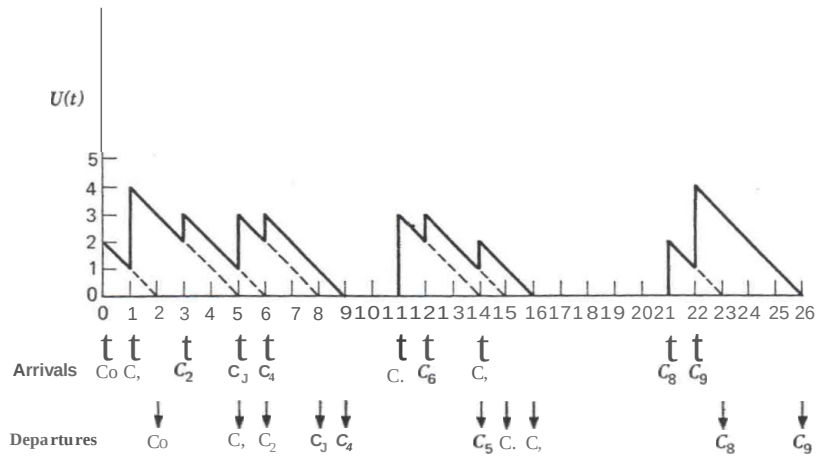
Figure 8.3 Unfinished work $V(t)$ showing sequence of arrivals and departures.

Table of values from Figure 8.3.

n	0	2	3	4	5	6	7	8	9	
t_{n+1}		2	2		5		2	7		
X_n	2	3		2		3		2	3	
I_n		-1		-4	2	-1	-6			
W_n	0	2		2	0	2		0		measured from Fig. 8.3
W_n	0	2		2	0	2		0		calculated from Eq. 8.7

$$IV_0 = 0$$

$$IV_1 = \max(0, w_0 + u_0) = \max(0, 1) = 1$$

$$IV_2 = \max(0, u_1, u_1 + u_0 + I_{10}) = \max(0, 1, 2) = 2$$

$$IV_a = \max(0, u_2, u_2 + u_1, u_2 + u_1 + u_0 + w_0) = \max(0, -1, 0, 1) = 1$$

$$w_4 = \max(0, u_a, I_a + u_2, I_a + u_2 + I_{11}, u_a + I_2 + I_{11} + u_0 + IV_0) \\ = \max(0, 1, 0, 1, 2) = 2$$

$$IV_5 = \max(0, u_4, u_4 + u_a, I_4 + u_2, u_4 + u_a + I_2 + I_{11}, \\ u_4 + u_a + u_2 + I_{11} + I_{10} + IV_0) \\ = \max(0, -4, -3, -4, -3, -2) = 0$$

$$IV_7 = \max(0, 2, -2, -1, -2, -1, 0) = 2$$

$$IV_8 = \max(0, -1, 1, -3, -2, -3, -1) = 1$$

$$IV_9 = \max(0, -6, -7, -5, -9, -8, -9, -8, -7) = 0$$

$$IV_{10} = \max(0, 1, -5, -6, -4, -8, -7, -8, -7, -6) = 1$$

These calculations are quite revealing. For example, whenever we find an m for which $w_m = 0$, then the m rightmost calculations in Eq. (8.7) need be made no more in calculating w_n for all $n > m$; this is due to the fact that a busy period has ended and the service times and interarrival times from that busy period cannot affect the calculations in future busy periods. Thus we see the isolating effect of idle periods which ensue between busy periods. Furthermore, when $w_m = 0$, then the rightmost term $(U_m + 1)^0$ gives the (negative of the) total accumulated idle time of the system during the interval $(0, \tau_m)$.

Let us now proceed with the theory for calculating $W(y)$. We define $C_n(u)$ as the PDF for the random variable u_n , that is,

$$C_n(u) \triangleq P[u_n = x_n - t_{n+1} \leq u] \quad (8.15)$$

and we note that u_n is not restricted to a half line. We now derive the expression for $C_n(u)$ in terms of $A(r)$ and $B(x)$:

$$\begin{aligned} C_n(u) &= P[x_n - t_{n+1} \leq u] \\ &= \int_{t=0}^{\infty} P[x_n \leq u + t \mid t_{n+1} = t] dA(t) \end{aligned}$$

However, the service time for C_n is independent of t_{n+1} and therefore

$$C_n(u) = \int_{t=0}^{\infty} B(u + t) dA(t) \quad (8.16)$$

Thus, as we expected, $C_n(u)$ is independent of n and we therefore write

$$C_n(u) \triangleq c(u) = \int_{t=0}^{\infty} B(u + t) dA(t) \quad (8.17)$$

Also, let $\tilde{u}$ denote the random variable

$$\tilde{u} = \tilde{x} - \tilde{t}$$

Note that the integral given in Eq. (8.17) is very much like a convolution form for $a(t)$ and $B(x)$; it is not quite a straight convolution since the distribution $C(u)$ represents the *difference* between x_n and t_{n+1} rather than the *sum*. Using our convolution notation $(*)$, and defining $cn(u) \triangleq dC_n(u)/du$ we have

$$Cn(u) = c(u) = a(-u) * b(u) \quad (8.18)$$

It is (again) convenient to define the waiting-time distribution for customer C ; as

$$W_n(y) = P[w_n \leq y] \quad (8.19)$$

For $y \geq 0$ we have from Eq. (8.4)

$$\begin{aligned} W_{n+1}(Y) &= P[w_n + \text{lin} \leq y] \\ &= \int_{0^-}^{\infty} P[u_n \leq Y - w \mid w_n = w] dW_n(w) \end{aligned}$$

And now once again, since u_n is independent of w_n we have

$$W_{n+1}(Y) = \int_{0^-}^{\infty} C(y - w) dW_n(w) \quad \text{for } Y \geq 0 \quad (8.20)$$

However, as postulated in Eq. (8.14) this distribution has a limit $W(y)$ and therefore we have the following integral equation, which defines the limiting distribution of waiting time for customers in the system G/G/1:

$$W(y) = \int_{0^-}^{\infty} C(y - w) dW(w) \quad \text{for } y \geq 0$$

Further, it is clear that

$$W(y) = 0 \quad \text{for } y < 0$$

Combining these last two we have *Lindley's integral equation* [LIND 52], which is seen to be an integral equation of the Wiener-Hopf type [SPIT 57].

$$W(y) = \begin{cases} \int_0^\infty C(y - w) dW(w) & y \geq 0 \\ 0 & y < 0 \end{cases} \quad (8.21)$$

Equation (8.21) may be rewritten in at least two other useful forms, which we now proceed to derive. Integrating by parts, we have (for $y \geq 0$)

$$\begin{aligned} W(y) &= C(y - w)W(w) \Big|_{w=0^-}^{\infty} - \int_{0^-}^{\infty} W(w) dC(y - w) \\ &= \lim_{w \rightarrow \infty} C(y - w)W(w) - C(y)W(0^-) - \int_0^\infty W(w) dC(y - w) \end{aligned}$$

We see that $\lim_{w \rightarrow \infty} C(y - w)W(w) = 0$ as $w \rightarrow \infty$ since the limit of $C(u)$ as $u \rightarrow -\infty$ is the probability that an interarrival time approaches infinity, which clearly must go to zero if the interarrival time is to have finite moments. Similarly, we have $W(0^-) = 0$ and so our form for Lindley's integral equation may be rewritten as

$$W(y) = \begin{cases} - \int_0^\infty W(w) dC(y - w) & y \geq 0 \\ 0 & y < 0 \end{cases} \quad (8.22)$$

Let us now show a third form for this equation. By the simple variable change $u = y - w$ for the argument of our distributions we finally arrive at

$$W(y) = \begin{cases} \int_{-\infty}^y W(y - u) dC(u) & y \geq 0 \\ 0 & y < 0 \end{cases} \quad - (8.23)$$

Equations (8.21), (8.22), and (8.23) all describe the basic integral equation which governs the behavior of $G/G//$. These integral equations, as mentioned above, are Wiener-Hopf-type integral equations and are not unfamiliar in the theory of stochastic processes.

One observes from these forms that Lindley's integral equation is almost, but not quite, a convolution integral. The important distinction between a convolution integral and that given in Lindley's equation is that the latter integral form holds only when the variable is nonnegative; the distribution function is identically zero for values of negative argument. Unfortunately, since the integral holds only for the half-line we must borrow techniques from the theory of complex variables and from contour integration in order to solve our system. We find a similar difficulty in the design of optimal linear filters in the mathematical theory of communication; there too, a Wiener-Hopf integral equation describes the optimal solution, except that for linear filters, the unknown appears as one factor in the integrand rather than as in our case in queueing theory, where the unknown appears on both sides of the integral equation. Nevertheless, the solution techniques are amazingly similar and the reader acquainted with the theory of optimal realizable linear filters will find the following arguments familiar.

In the next section, we give a fairly general solution to Lindley's integral equation by the use of spectral (transform) methods. In Exercise 8.6 we examine a solution approach by means of an example that does not require transforms; the example chosen is the system DE/I considered by Lindley. In that (direct) approach it is required to assume the solution form. We now consider the spectral solution to Lindley's equation in which such assumed solution forms will not be necessary.

8.2. SPECTRAL SOLUTION TO LINDLEY'S INTEGRAL EQUATION

In this section we describe a method for solving Lindley's integral equation by means of spectrum factorization [SMIT 53]. Our point of departure is the form for this equation given by (8.23). As mentioned earlier it would be rather straightforward to solve this equation if the right-hand side were a true convolution (it is, in fact, a convolution for the nonnegative half-line on the

variable y but not so otherwise). In order to get around this difficulty we use the following ingenious device whereby we define a "complementary" waiting time, which completes the convolution, and which takes on the value of the integral for negative y only, that is,

$$W_-(y) \triangleq \begin{cases} 0 & y \geq 0 \\ \left(\int_{-\infty}^y W(y-u) dC(u) \right) & y < 0 \end{cases} \quad (8.24)$$

Note that the left-hand side of Eq. (8.23) might consistently be written as $W_+(y)$ in the same way in which we defined the left-hand side of Eq. (8.24). We now observe that if we add Eqs. (8.23) and (8.24) then the right-hand side takes on the integral expression for all values of the argument, that is,

$$W(y) + W_-(y) = \int_{-\infty}^y W(y-u)c(u) du \quad \text{for all real } y \quad (8.25)$$

where we have denoted the pdf for $\tilde{u}$ by $c(u) [\triangleq dC(u)/du]$.

To proceed, we assume that the pdf of the interarrival time is $O(e^{-Dt})$ as $t \rightarrow \infty$ (where D is any real number greater than zero), that is,

$$\lim_{t \rightarrow \infty} \frac{aCt}{e^{-Dt}} < \infty \quad (8.26)$$

The condition (8.26) really insists that the pdf associated with the interarrival time drops off at least as fast as an exponential for very large interarrival times. From this condition it may be seen from Eq. (8.17) that the behavior of $C(u)$ as $u \rightarrow -\infty$ is governed by the behavior of the interarrival time; this is true since as u takes on large negative values the argument for the service-time distribution can be made positive only for large values of t , which also appears as the argument for the interarrival time density. Thus we can show

$$\lim_{u \rightarrow -\infty} \frac{C(u)}{e^{Du}} < \infty$$

That is, $C(u)$ is $O(e^{Du})$ as $u \rightarrow -\infty$. If we now use this fact in Eq. (8.24) it is easy to establish that $W_-(y)$ is also $O(e^{Dy})$ as $y \rightarrow -\infty$.

• The notation $O(g(x))$ as $x \rightarrow x_0$ refers to any function that (as $x \rightarrow x_0$) decays to zero at least as rapidly as $g(x)$ [where $g(x) > 0$], that is,

$$\lim_{x \rightarrow x_0} \frac{f(x)}{g(x)} = K < \infty$$

Let us now define some (bilateral) transforms for various of our functions. For the Laplace transform of $W_-(y)$ we define

$$\Phi_-(s) \triangleq \int_{-\infty}^{\infty} W_-(y) e^{-sy} dy \quad (8.27)$$

Due to the condition we have established regarding the asymptotic property of $W_-(y)$, it is clear that $\Phi_-(s)$ is analytic in the region $\text{Re}(s) < D$. Similarly, for the distribution of our waiting time $W(y)$ we define

$$\Phi_+(s) \triangleq \int_{-\infty}^{\infty} W(y) e^{-sy} dy \quad (8.28)$$

Note that $\Phi_+(s)$ is the Laplace transform of the PDF for waiting time, whereas in previous chapters we have defined $W(s)$ as the Laplace transform of the pdf for waiting time; thus by entry II of Table 1.3, we have

$$s\Phi_+(s) = W(s) \quad (8.29)$$

Since there are regions for Eqs. (8.23) and (8.24) in which the functions drop to zero, we may therefore rewrite these transforms as

$$\Phi_-(s) = \int_{-\infty}^{0^-} W_-(y) e^{-sy} dy \quad (8.30)$$

$$\Phi_+(s) = \int_{0^-}^{\infty} W(y) e^{-sy} dy \quad (8.31)$$

Since $W(y)$ is a true distribution function (and therefore it remains bounded as $y \rightarrow \infty$) then $\Phi_+(s)$ is analytic for $\text{Re}(s) > 0$. As usual, we define the transform for the pdf of the interarrival time and for the pdf of the service time as $A^*(s)$ and $S^*(s)$, respectively. Note for the condition (8.26) that $A^*(-s)$ is analytic in the region $\text{Re}(s) < D$ just as was $\Phi_-(s)$.

From Appendix I we recall that the Laplace transform for the convolution of two functions is the product of the transforms of each. Equation (8.18) is almost the convolution of the service-time density with the interarrival-time density; the only difficulty is the negative argument for the interarrival-time density. Nevertheless, the above-mentioned fact regarding products of transforms goes through merely with the negative argument (this is Exercise 8.1). Thus for the Laplace transform of $c(u)$ we find

$$C^*(s) = A^*(-s)S^*(s) \quad (8.32)$$

Let us now return to Eq. (8.25), which expresses the fundamental relationship among the variables of our problem and the waiting-time distribution $W(y)$. Clearly, the time spent in queue must be a nonnegative random variable, and so we recognize the right-hand side of Eq. (8.25) as a convolution between the waiting time PDF and the pdf for the random variable i . The Laplace

transform of this convolution must therefore give the product of the Laplace transform $\Phi_+(s)$ (for the waiting-time distribution) and $C^*(s)$ (for the density on $\tilde{u}$). The transform of the left-hand side we recognize from Eqs. (8.30) and (8.31) as being $\Phi_+(s) + \Phi_-(s)$, thus

$$\Phi_+(s) + \Phi_-(s) = \Phi_+(s)C^*(s)$$

From Eq. (8.32) we therefore obtain

$$\Phi_+(s) + \Phi_-(s) = \Phi_+(s)A^*(-s)B^*(s)$$

which gives us

$$\Phi_-(s) = \Phi_+(s)[A^*(-s)B^*(s) - 1] \quad (8.33)$$

We have already established that both $\Phi_-(s)$ and $A^*(-s)$ are analytic in the region $\text{Re}(s) < D$. Furthermore, since $\Phi_+(s)$ and $B^*(s)$ are transforms of bounded functions of nonnegative variables then both functions must be analytic in the region $\text{Re}(s) > 0$.

We now come to the *spectrum factorization*. The purpose of this factorization is to find a suitable representation for the term

$$A^*(-s)B^*(s) - 1 \quad (8.34)$$

in the form of two factors. Let us pause for a moment and recall the method of stages whereby Erlang conceived the ingenious idea of approximating a distribution by means of a collection of series and parallel exponential stages. The Laplace transform for the pdf's obtainable in this fashion was generally given in Eq. (4.62) or Eq. (4.64); we immediately recognize these to be rational functions of s (that is, a ratio of a polynomial in s divided by a polynomial in s). We may similarly conceive of approximating the Laplace transforms $A^*(-s)$ and $B^*(s)$ each in such forms; if we so approximate, then the term given by Eq. (8.34) will also be a rational function of s . We thus choose to consider those queueing systems for which $A^*(s)$ and $B^*(s)$ may be suitably approximated with (or which are given initially as) such rational functions of s , in which case we then propose to form the following spectrum factorization

$$A^*(-s)B^*(s) - 1 = \frac{\Psi_+(s)}{\Psi_-(s)} \quad (8.35)$$

Clearly $\Psi_+(s)/\Psi_-(s)$ will be some rational function of s , and we are now desirous of finding a particular factored form for this expression. We specifically wish to find a factorization such that:

- For $\text{Re}(s) > 0$, $\Psi_+(s)$ is an analytic function of s with no zeroes in this half-plane.
- For $\text{Re}(s) < D$, $\Psi_-(s)$ is an analytic function of s with no zeroes in this half-plane.

(8.36)

Furthermore, we wish to find these functions with the *additional* properties:

$$\bullet \quad \text{For } \operatorname{Re}(\zeta) > 0, \quad \lim_{|\zeta| \rightarrow \infty} \frac{\Psi_+(\zeta)}{\zeta} = 1.$$

$$\text{For } \operatorname{Re}(\zeta) < D, \quad \lim_{|\zeta| \rightarrow \infty} \frac{\Psi_-(\zeta)}{\zeta} = -1.$$
(8.37)

The conditions in (8.37) are convenient and must have opposite polarity in the limit since we observe that as ζ runs off to infinity along the imaginary axis, both $A^*(-\zeta)$ and $B^*(\zeta)$ must decay to 0 [if they are to have finite moments and if $A(t)$ and $B(x)$ do not contain a sequence of discontinuities, which we will not permit] leaving the left-hand side of Eq. (8.35) equal to -1 , which we have suitably matched by the ratio of limits given by Conditions (8.37). We shall find that this spectrum factorization, which requires us to find $\Psi_+(\zeta)$ and $\Psi_-(\zeta)$ with the appropriate properties, contains the difficult part of this method of solution. Nevertheless, assuming that we have found such a factorization it is then clear that we may write Eq. (8.33) as

$$\Phi_-(\zeta) = \Phi_+(\zeta) \frac{\Psi_+(\zeta)}{\Psi_-(\zeta)}$$

or

$$\Phi_-(\zeta)\Psi_-(\zeta) = \Phi_+(\zeta)\Psi_+(\zeta) \quad (8.38)$$

where the common region of analyticity for both sides of Eq. (8.38) is within the strip

$$0 < \operatorname{Re}(\zeta) < D \quad (8.39)$$

That this last is true may be seen as follows. We have already assumed that $\Psi_+(\zeta)$ is analytic for $\operatorname{Re}(\zeta) > 0$ and it is further true that $\Phi_+(\zeta)$ is analytic in this same region since it is the Laplace transform of a function that is identically zero for negative arguments; the product of these two must therefore be analytic for $\operatorname{Re}(\zeta) > 0$. Similarly, $\Psi_-(\zeta)$ has been given to be analytic for $\operatorname{Re}(\zeta) < D$ and we have that $\Phi_-(\zeta)$ is analytic here as explained earlier following Eq. (8.27); thus the product of these two will be analytic in $\operatorname{Re}(\zeta) < D$. Thus the common region is as stated in Eq. (8.39). Now, Eq. (8.38) establishes that these two functions are equal in the common strip and so they must represent functions which, when continued in the region $\operatorname{Re}(\zeta) < 0$, are analytic and when continued in the region $\operatorname{Re}(\zeta) > D$, are also analytic; therefore their analytic continuation contains no singularities in the entire finite ζ -plane. Since we have established the behavior of the function $\Phi_+(\zeta)\Psi_+(\zeta) = \Phi_-(\zeta)\Psi_-(\zeta)$ to be analytic and bounded in the finite ζ -plane, and since we assume Condition (8.37), we may then apply Liouville's theorem*

* Liouville's theorem states, "If $f(z)$ is analytic and bounded for all finite values of z , then $f(z)$ is a constant."

[TITC 52], which immediately establishes that this function must be a constant (say, K). We thus have

$$\Phi_-(s)\Psi_-(s) = \Phi_+(s)\Psi_+(s) = K \quad (8.40)$$

This immediately yields

$$\Phi_+(s) = \frac{K}{\Psi_+(s)} \quad (8.41)$$

The reader should recall that what we are seeking in this development is an expression for the distribution of queueing time whose Laplace transform is exactly the function $\Phi_+(s)$, which is now given through Eq. (8.41). It remains for us to demonstrate a method for evaluating the constant K .

Since $s\Phi_+(s) = W^*(s)$, we have

$$s\Phi_+(s) = W^*(s) \triangleq \int_0^\infty e^{-sy} dW(y)$$

Let us now consider the limit of this equation as $s \rightarrow 0$; working with the right-hand side we have

$$\lim_{s \rightarrow 0} \int_0^\infty e^{-sy} dW(y) = \int_0^\infty dW(y) = 1$$

We have thus established

$$\lim_{s \rightarrow 0} s\Phi_+(s) = 1 \quad (8.42)$$

This is nothing more than the final value theorem (entry 18, Table 1.3) and comes about since $W(\infty) = 1$. From Eq. (8.41) and this last result we then have

$$\lim_{s \rightarrow 0} s\Phi_+(s) = \lim_{s \rightarrow 0} \frac{sK}{\Psi_+(s)} = 1$$

and so we may write

$$K = \lim_{s \rightarrow 0} \frac{\Psi_+(s)}{s} \quad (8.43)$$

Equation (8.43) provides a means of calculating the constant K in our solution for $\Phi_+(s)$ as given in Eq. (8.41). If we make a Taylor expansion of the function $\Psi_+(s)$ around $s = 0$ [viz., $\Psi_+(s) = \Psi_+(0) + s\Psi_+^{(1)}(0) + (s^2/2!)\Psi_+^{(2)}(0) + \dots$] and note from Eqs. (8.35) and (8.36) that $\Psi_+(0) = 0$, we then recognize that this limit may also be written as

$$K = \lim_{s \rightarrow 0} \frac{d\Psi_+(s)}{ds} \quad (8.44)$$

and this provides us with an alternate way for calculating the constant K . We may further explore this constant K by examining the behavior of $\Phi_+(s)\Psi_+(s)$ anywhere in the region $\operatorname{Re}(s) > 0$ [i.e., see Eq. (8.40)]; we choose to examine this behavior in the limit as $s \rightarrow \infty$ where we know from Eq. (8.37) that $\Psi_+(s)$ behaves as s does; that is,

$$\begin{aligned} K &= \lim_{s \rightarrow \infty} \Phi_+(s)\Psi_+(s) \\ &= \lim_{s \rightarrow \infty} \Phi_+(s)s \\ &= \lim_{s \rightarrow \infty} s \int_0^\infty e^{-sy} W(y) dy \end{aligned}$$

Making the change of variable $sy = x$ we have

$$K = \lim_{s \rightarrow \infty} \int_0^\infty e^{-x} W\left(\frac{x}{s}\right) dx$$

As $s \rightarrow \infty$ we may pull the constant term $W(0^+)$ outside the integral and then obtain the value of the remaining integral, which is unity. We thus obtain

$$K = W(0^+) \quad (8.45)$$

This establishes that the constant K is merely the probability that an arriving customer need not queue].

In conclusion then, assuming that we can find the appropriate spectrum factorization in Eq. (8.35) we may immediately solve for the Laplace transform of the waiting-time distribution through Eq. (8.41), where the constant K is given in either of the three forms Eq. (8.43), (8.44), or (8.45). Of course it then remains to invert the transform but the problems involved in that calculation have been faced before in numerous of our other solution forms.

It is possible to carry out the solution of this problem by concentrating on $\Psi_-(s)$ rather than $\Psi_+(s)$, and in some cases this simplifies the calculations. In such cases we may proceed from Eq. (8.35) to obtain

$$\Psi_+(s) = \Psi_-(s)[A^*(-s)\theta^*(s) - 1] \quad (8.46)$$

From Eq. (8.41) we then have

$$\Phi_+(s) = \frac{K}{[A^*(-s)\theta^*(s) - 1]\Psi_-(s)} \quad (8.47)$$

† Note that $W(0^+)$ is not necessarily equal to $1 - \rho$, which is the fraction of time the server is idle. (These two are equal for the system $M\{G\{1\}$.)

In order to evaluate the constant K in this case we differentiate Eq. (8.46) at $s = 0$, that is,

$$\Psi_+^{(1)}(0) = [\lambda A^*(0)B^*(0) - 1]\Psi_-(0) + \lambda/\lambda(0)[\lambda A^*(0)B^{*(1)}(0) - A^{*(1)}(0)B^*(0)] \quad (8.48)$$

From Eq. (8.44) we recognize the left-hand side of Eq. (8.48) as the constant K and we may now evaluate the right-hand side to obtain

$$K = 0 + \Psi_-(0)[-x + \lambda]$$

giving

$$K = \Psi_-(0)(1 - \rho)i \quad (8.49)$$

Thus, if we wish to use $\Psi_-(s)$ in our solution form, we obtain the transform of the waiting-time distribution from Eq. (8.47), where the unknown constant K is evaluated in terms of $\Psi_-(s)$ through Eq. (8.49).

Summarizing then, once we have carried out the spectrum factorization as indicated in Eq. (8.35), we may proceed in one of two directions in solving for $\Phi_+(s)$, the transform of the waiting-time distribution. The first method gives us

$$\Phi_+(s) = \frac{1}{\Psi_+(s)} \lim_{s \rightarrow 0} \frac{\Psi_+(s)}{s} = \frac{W(0^+)}{\Psi_+(s)} \quad (8.50)$$

and the second provides us with

$$\Phi_+(s) = \frac{\Psi_-(0)(1 - \rho)i}{[\lambda A^*(s)B^*(s) - 1]\Psi_-(s)} \quad (8.51)$$

We now proceed to demonstrate the use of these results in some examples.

Example 1: *MIMI*

Our old friend *MIMI* is extremely straightforward and should serve to clarify the meaning of spectrum factorization. Since both the interarrival time and the service time are exponentially distributed random variables, we immediately have $A^*(s) = \lambda/(s + \lambda)$ and $B^*(s) = \mu/(s + \mu)$, where $\bar{x} = 1/\mu$ and $i = 1/\lambda$. In order to solve for $\Phi_+(s)$ (the transform of the waiting time distribution), we must first form the expression given in Eq. (8.34), that is,

$$\lambda A^*(s)B^*(s) - 1 = \left(\frac{\lambda}{s + \lambda}\right)\left(\frac{\mu}{s + \mu}\right) - 1 = \frac{s^2 + s(\mu - \lambda)}{(s + \lambda)(s + \mu)}$$

Thus, from Eq. (8.35), we obtain

$$\begin{aligned}\frac{\Psi_+(s)}{\Psi_-(s)} &= A^*(-s)B^*(s) - 1 \\ &= \frac{s(s + \mu - \lambda)}{(s + \mu)(\lambda - s)}\end{aligned}\quad (8.52)$$

In Figure 8.4 we show the location of the zeroes (denoted by a circle) and poles (denoted by a cross) in the complex s -plane for the function given in Eq. (8.52). Note that in this particular example the roots of the numerator (zeroes of the expression) and the roots of the denominator (poles of the expression) are especially simple to find; in general, one of the most difficult parts of this method of spectrum factorization is to solve for the roots. In order to factorize we require that conditions (8.36) and (8.37) maintain. Inspecting the pole-zero plot in Figure 8.4 and remembering that $\Psi_+(s)$ must be analytic and zero-free for $\text{Re}(s) > 0$, we may collect together the two zeroes (at $s = 0$ and $s = -\mu + \lambda$) and one pole (at $s = -\mu$) and still satisfy this required condition. Similarly, $\Psi_-(s)$ must be analytic and free from zeroes for the $\text{Re}(s) < D$ for some $D > 0$; we can obtain such a condition if we allow this function to contain the remaining pole (at $s = \lambda$) and choose $D = \lambda$. This we show in Figure 8.5.

Thus we have

$$\Psi_+(s) = \frac{s(s + \mu - \lambda)}{s + \mu} \quad (8.53)$$

$$\Psi_-(s) = \lambda - s \quad (8.54)$$

Note that Condition (8.37) is satisfied for the limit as $s \rightarrow \infty$.

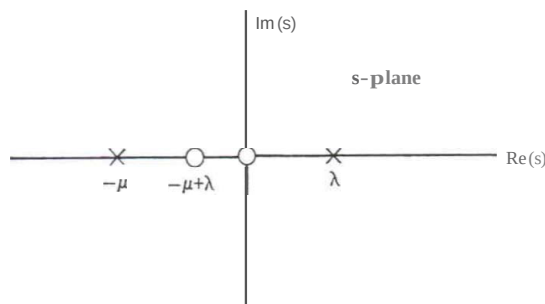
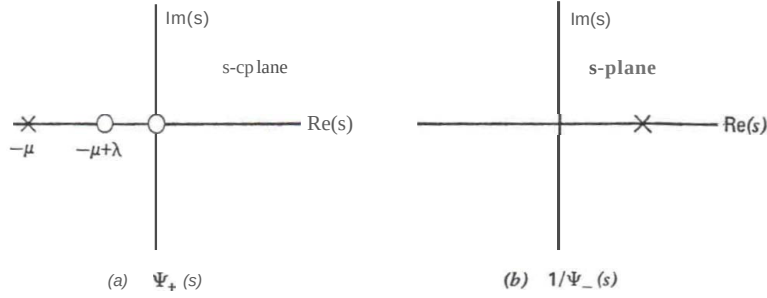


Figure 8.4 Zeroes (o) and poles (x) of $\Psi_+(s)/\Psi_-(s)$ for $MIMI$.

Figure 8.5 Factorization into $\Psi_+(s)$ and $1/\Psi_-(s)$ for *MIMI!*.

We are now faced with finding K . From Eq. (8.43) we have

$$\begin{aligned}
 K &= \lim_{s \rightarrow 0} \frac{\Psi_+(s)}{s} \\
 &= \lim_{s \rightarrow 0} \frac{s + \mu - \lambda}{s + \mu} \\
 &= 1 - \rho
 \end{aligned} \tag{8.55}$$

Our expression for the Laplace transform of the waiting time PDF for *MIMI!* is therefore from Eq. (8.41),

$$\Phi_+(s) = \frac{(1 - \rho)(s + \mu)}{s(s + \mu - \lambda)} \tag{8.56}$$

At this point, typically, we attempt to invert the transform to get the waiting-time distribution. However, for this *M/M/I* example, we have already carried out this inversion for $W^*(s) = s\Phi_+(s)$ in going from Eq. (5.120) to Eq. (5.123). The solution we obtain is the familiar form,

$$W(y) = 1 - \rho e^{-\mu(1-\rho)y} \quad y \geq 0 \tag{8.57}$$

Example 2: *GIMI!*

In this case $B^*(s) = \mu/(s + \mu)$ but now $A^*(s)$ is completely arbitrary, giving us

$$A^*(-s)B^*(s) - 1 = \frac{A^*(-s)\mu}{s + \mu} - 1$$

† This example forces us to locate roots using Rouché's theorem in a way often necessary for specific *G/G/I!* problems when the spectrum factorization method is used. Of course, we have already studied this system in Section 6.4 and will compare the results for both methods.

and so we have

$$\frac{\Psi_+(s)}{\Psi_-(s)} = \frac{\mu A^*(-s) - s - \mu}{s + \mu} \quad (8.58)$$

In order to factorize we must find the roots of the numerator in this equation. We need not concern ourselves with the poles due to $A^*(-s)$ since they must lie in the region $\operatorname{Re}(s) > 0$ [i.e., $A(t) = 0$ for $t < 0$] and we are attempting to find $\Psi_+(s)$, which cannot include any such poles. Thus we only study the zeroes of the function

$$s + \mu - \mu A^*(-s) = 0 \quad (8.59)$$

Clearly, one root of this equation occurs at $s = 0$. In order to find the remaining roots, we make use of Rouché's theorem (given in Appendix I but which we repeat here):

Rouché's Theorem If $f(s)$ and $g(s)$ are analytic functions of s inside and on a closed contour C , and also if $|g(s)| < |f(s)|$ on C , then $f(s)$ and $f(s) + g(s)$ have the same number of zeroes inside C .

In solving for the roots of Eq. (8.59) we make the identification

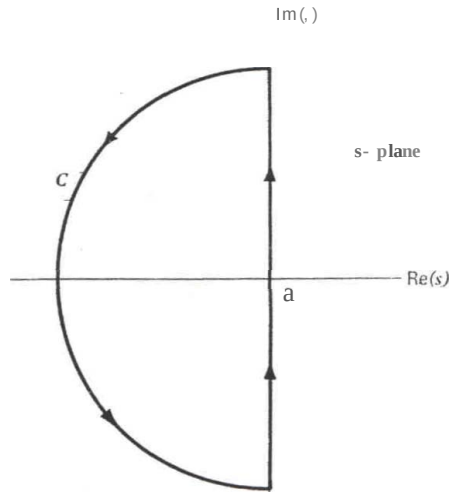
$$\begin{aligned} f(s) &= s + \mu \\ g(s) &= -\mu A^*(-s) \end{aligned}$$

We have by definition

$$A^*(-s) = \int_0^\infty C e^{st} dA(t)$$

We now choose C to be the contour that runs up the imaginary axis and then forms an infinite-radius semicircle moving counterclockwise and surrounding the left half of the s -plane, as shown in Figure 8.6. We consider this contour since we are concerned about all the poles and zeroes in $\operatorname{Re}(s) < 0$ so that we may properly include them in $\Psi_+(s)$ [recall that $\Psi_-(s)$ may contain none such]; Rouché's theorem will give us information concerning the number of zeroes in $\operatorname{Re}(s) < 0$, which we must consider. As usual, we assume that the real and imaginary parts of the complex variable s are given by σ and ω , respectively, that is, for $j = \sqrt{-1}$

$$s = \sigma + j\omega$$

Figure 8.6 The contour C for $G/M/1$.

Now for the $\text{Re}(s) = a \leq 0$ we have $e^{\sigma t} \leq 1$ ($t \geq 0$) and so

$$\begin{aligned}
 |g(s)| &= \left| \mu \int_{0^-}^{\infty} e^{st} dA(t) \right| \\
 &= \left| \mu \int_{0^-}^{\infty} e^{\sigma t} e^{j\omega t} dA(t) \right| \\
 &\leq \left| \mu \int_{0^-}^{\infty} e^{j\omega t} dA(t) \right| \\
 &\leq \left| \mu \int_{0^-}^{\infty} dA(t) \right| \\
 &= \mu
 \end{aligned} \tag{8.60}$$

Similarly we have

$$|l(s)| = |s + \mu| \tag{8.61}$$

Now, examining the contour C as shown in Figure 8.6, we observe that for all points *on* the contour, except at $s = 0$, we have from Eqs. (8.60) and (8.61) that

$$|l(s)| = |s + \mu| > \mu \geq |g(s)| \tag{8.62}$$

This follows since $s + \mu$ (for s on C) is a vector whose length is the distance from the point $-\mu$ to the point on C where s is located. We are almost in a

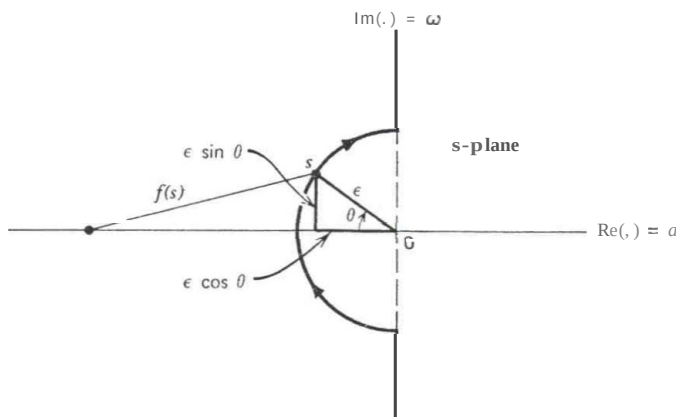


Figure 8.7 The excursion around the origin.

position to apply Rouché's theorem; the only remaining consideration is to show that $|f(s)| > |g(s)|$ in the vicinity $S = 0$. For this purpose we allow the contour C to make a small semicircular excursion to the left of the origin as shown in Figure 8.7. We note at $s = 0$ that $|g(0)| = |f(0)| = \mu$, which does not satisfy the conditions for Rouché's theorem. The small semicircular excursion of radius ϵ ($\epsilon > 0$) that we take to the left of the origin overcomes this difficulty as follows. Considering an arbitrary point s on this semicircle (see the figure), which lies at an angle θ with the a -axis, we may write $s = \sigma + j\omega = -\epsilon \cos \theta + j\epsilon \sin \theta$ and so we have

$$|f(s)|^2 = |s + \mu|^2 = |-\epsilon \cos \theta + j\epsilon \sin \theta + \mu|^2$$

Forming the product of $(s + \mu)$ and its complex conjugate, we get

$$\begin{aligned} |f(s)|^2 &= (\mu - \epsilon \cos \theta)^2 + \epsilon^2 \sin^2 \theta \\ &= \mu^2 - 2\mu\epsilon \cos \theta + \epsilon^2 \end{aligned} \quad (8.63)$$

Note that the smallest value for $|f(s)|$ occurs for $\theta = 0$. Evaluating $g(s)$ on this same semicircular excursion we have

$$|g(s)|^2 = \mu^2 \left| \int_0^\infty e^{(-\epsilon \cos \theta + j\epsilon \sin \theta)t} dA(t) \right|^2$$

From the power-series expansion of the exponential inside the integral we have

$$|g(s)|^2 = \mu^2 \left| \int_0^\infty [1 + (-\epsilon \cos \theta + j\epsilon \sin \theta)t + \dots] dA(t) \right|^2$$

We recognize the integrals in this series as proportional to the moments of the interarrival time, and so

$$|g(s)|^2 = \mu^2 [1 - \epsilon i \cos \theta + j \epsilon \bar{i} \sin \theta + o(\epsilon)]^2$$

Forming $|g(s)|^2$ by multiplying $g(s)$ by its complex conjugate, we have

$$\begin{aligned} |g(s)|^2 &= \mu^2 (1 - 2\epsilon \bar{i} \cos \theta + o(\epsilon)) \\ &= \mu^2 - \frac{2\mu\epsilon}{\rho} \cos \theta + o(\epsilon) \end{aligned} \quad (8.64)$$

where, as usual, $\rho = \bar{x}/\bar{i} = 1/\mu\bar{i}$. Now since θ lies in the range $-\pi/2 \leq \theta \leq \pi/2$, which gives $\cos \theta \geq 0$, we have as $\epsilon \rightarrow 0$ that on the shrinking semi-circle surrounding the origin

$$\mu^2 - 2\mu\epsilon \cos \theta > \mu^2 - \frac{2\mu\epsilon}{\rho} \cos \theta \quad (8.65)$$

This last is true since $\rho < 1$ for our stable system. The left-hand side of Inequality (8.65) is merely the expression given in Eq. (8.63) for $|I(s)|^2$ correct up to the first order in ϵ , and the right-hand side is merely the expression in Eq. (8.64) for $|g(s)|^2$, again correct up to the first order in ϵ . Thus we have shown that in the vicinity $s = 0$, $|I(s)| > |g(s)|$. This fact now having been established for all points on the contour C , we may apply Rouché's theorem and state that $I(s)$ and $I(s) + g(s)$ have the same number of zeroes inside the contour C . Since $I(s)$ has only one zero (at $s = -\mu$) it is clear that the expression given in Eq. (8.59) $[I(s) + g(s)]$ has only one zero for $\text{Re}(s) < 0$; let this zero occur at the point $s = -S_1'$. As discussed above, the point $s = 0$ is also a root of Eq. (8.59).

We may therefore write Eq. (8.58) as

$$\frac{\Psi_+(s)}{\Psi_-(s)} = \left[\frac{\mu A^*(-5) - 5 - \mu}{5(5 + 51)} \right] \frac{[5(5 + 51)]}{5 + \mu} \quad (8.66)$$

where the first bracketed term contains no poles and no zeroes in $\text{Re}(s) \leq 0$ (we have divided out the only two zeroes at $s = 0$ and $s = -51$ in this half-plane). We now wish to extend the region $\text{Re}(s) \leq 0$ into the region $\text{Re}(s) < D$ and we choose $D (> 0)$ such that no new zeroes or poles of Eq. (8.59) are introduced as we extend to this new region. The first bracket qualifies for $[\Psi_-(s)]^{-1}$, and we see immediately that the second bracket qualifies for $\Psi_+(s)$ since none of its zeroes ($s = 0, s = -51$) or poles ($s = -\mu$) are in

$\operatorname{Re}(s) > 0$. We may then factorize Eq. (8.66) in the following form:

$$\Psi_+(s) = \frac{5(5 + s)}{s + \mu} \quad (8.67)$$

$$\Psi_-(s) = \frac{-5(5 + s)}{5 + \mu - \mu A^*(-s)} \quad (8.68)$$

We have now assured that the functions given in these last two equations satisfy Conditions (8.36) and (8.37). We evaluate the unknown constant K as follows:

$$\begin{aligned} K &= \lim_{s \rightarrow 0} \frac{\Psi_+(s)}{s} = \lim_{s \rightarrow 0} \frac{5 + s}{s + \mu} \\ &= \frac{5}{\mu} = W(0+) \end{aligned} \quad (8.69)$$

Thus we have from Eq. (8.41)

$$\Phi_+(s) = \frac{s_1(\mu + s)}{\mu s(s + s_1)}$$

The partial-fraction expansion for this last function gives us

$$\Phi_+(s) = \frac{1}{s} - \frac{1 - s_1/\mu}{s + s_1} \quad (8.70)$$

Inverting by inspection we obtain the final solution for $G\{M\}$:

$$W(y) = 1 - \left(1 - \frac{s_1}{\mu}\right)e^{-s_1 y} \quad y \geq 0 \quad (8.71)$$

The reader is urged to compare this last result with that given in Eq. (6.30), also for the system $G\{M\}$; the comparison is clear and in both cases there is a single constant that must be solved for. In the solution given here that constant is solved as the root of Eq. (8.59) with $\operatorname{Re}(s) < 0$; in the equation given in Chapter 6, one must solve Eq. (6.28), which is equivalent to Eq. (8.59).

Example 3:

The example for $G\{M\}$ can be carried no further in the general case. We find it instructive therefore to consider a more specific $G\{M\}$ example and finish the calculations; the example we choose is the one we used in Chapter 6, for which $A^*(s)$ is given in Eq. (6.35) and corresponds to an $Ez\{M\}$ system, where the two arrival stages have different death rates. For that example we

s-plane

Figure 8.8 Pole-zero pattern for $E_2/M/1$ example.

note that the poles of $A^*(-s)$ occur at the points $s = \mu$, $s = 2\mu$, which as promised lie in the region $\text{Re}(s) > 0$. As our first step in factorizing we form

$$\begin{aligned} \frac{\Psi_+(s)}{\Psi_-(s)} &= \frac{A^*(-s)B^*(s)}{1} \\ &= \frac{\left[-\frac{2\mu^2}{(\mu-s)(2\mu-s)} \right] \left(-\frac{\mu}{s+\mu} \right)}{-\frac{s(s-\mu+\mu\sqrt{2})(s-\mu-\mu\sqrt{2})}{(s+\mu)(\mu-s)(2\mu-s)}} - 1 \end{aligned} \quad (8.72)$$

The spectrum factorization is considerably simplified if we plot these poles and zeroes in the complex plane as shown in Figure 8.8. It is clear that the two poles and one zero in the right half-plane must be associated with $\Psi_-(s)$. Furthermore, since the strip $0 < \text{Re}(s) < \mu$ contains no zeroes and no poles we choose $D = \mu$ and identify the remaining two zeroes and the single pole in the region $\text{Re}(s) < D$ as being associated with $\Psi_+(s)$. Note well that the zero located at $s = (1 - \sqrt{2})\mu$ is in fact the single root of the expression $\mu A^*(-s) - s - \mu$ located in the left half-plane, as discussed above, and therefore $s = -(1 - \sqrt{2})\mu$. Of course, we need go no further to solve our problem since the solution is now given through Eq. (8.71); however, let us continue identifying various forms in our solution to clarify the remaining steps. With this factorization we may rewrite Eq. (8.72) as

$$\frac{\Psi_+(s)}{\Psi_-(s)} = \frac{-(s-\mu-\mu\sqrt{2})}{(\mu-s)(2\mu-s)} \left[\frac{s(s-\mu+\mu\sqrt{2})}{s+\mu} \right]$$

In this form we recognize the first bracket as $\Pi F_-(s)$ and the second bracket as $\Psi_+(s)$. Thus we have

$$\Psi_+(s) = \frac{s(s - \mu + \mu\sqrt{2})}{s + \mu} \quad (8.73)$$

We may evaluate the constant K from Eq. (8.69) to find

$$K = \frac{s}{\mu} = -1 + \sqrt{2} \quad (8.74)$$

and this of course corresponds to $W(0^+)$, which is the probability that a new arrival must wait for service. Finally then we substitute these values into Eq. (8.71) to find

$$W(y) = 1 - (2 - \sqrt{2})e^{-\mu(\sqrt{2}-1)y} \quad y \geq 0 \quad (8.75)$$

which as expected corresponds exactly to Eq. (6.37).

This method of spectrum factorization has been used successfully by Rice [RICE 62], who considers the busy period for the *GfGf* system. Among the interesting results available, there is one corresponding to the limiting distribution of long waiting times in the heavy-traffic case (which we develop in Section 2.1 of Volume II); Rice gives a similar approximation for the duration of a busy period in the heavy traffic case.

8.3. KINGMAN'S ALGEBRA FOR QUEUES

Let us once again state the fundamental relationships underlying the *GfGf* queue. For $u_i = x_n - t_{n+i}$ we have the basic relationship

$$w_{n+1} = (w_n + u_n)^+ \quad (8.76)$$

and we have also seen that

$$w_n = \max [0, u_{n-1}, u_{n-1} + u_{n-2}, \dots, u_{n-1} + \dots + u_1, u_{n-1} + \dots + u_0 + w_0]$$

We observed earlier that $\{W_n\}$ is a Markov process with stationary transition probabilities; its total stochastic structure is given by $P[W_{m+n} \leq y \mid w_m = x]$, which may be calculated as an n -fold integral over the n -dimensional joint distribution of the n random variables $w_{m+1}, \dots, w_{m+n}$ over that region of the space which results in $w_{m+n} \leq y$. This calculation is much too complicated and so we look for alternative means to solve this problem. Pollaczek [POLL 57] used a spectral approach and complex integrals to carry out the solution. Lindley [LIND 52] observed that W_n has the same distribution as W_n^* , defined earlier as

$$w_n^* = \max [U_0, U_1, \dots, U_{n-1}, U_n + w_0]$$

If we have the case $E[u_n] < 0$, which corresponds to $p = \bar{x}/\bar{t} < 1$, then a stable solution exists for the limiting random variable $\tilde{w}$ such that

$$\tilde{w} = \sup_{n \geq 0} U'_n; \quad (8.77)$$

independent of n . The method of spectrum factorization given in the previous section is Smith's [SMIT 53] approach to the solution of Lindley's Wiener-Hopf integral equation. Another approach due to Spitzer using combinatorial methods leads to Spitzer's identity [SPIT 57]. Many proofs for this identity exist and Wendel [WEND 58] carried it out by exposing the underlying algebraic structure of the problem. Keilson [KEIL 65] demonstrated the application of Green's functions to the solution of G/G/1. Beneš [BENE 63] also considered the G/G/1 system by investigating the unfinished work and its variants. These many approaches, each of which is rather complicated, forces one to inquire whether or not there is a larger underlying structure, which places these solution methods in a common framework. In 1966 Kingman [KING 66] addressed this problem and introduced his algebra for queues to expose the common structure; we study this algebra briefly in this section.

From Eq. (8.76) we clearly could solve for the pdf of w_{n+1} iteratively starting with $n = 0$ and with a given pdf for w_0 ; recall that the pdf for u ; [i.e., $c(u)$] is independent of n . Our iterative procedure would proceed as follows. Suppose we had already calculated the pdf for w_n , which we denote by $W_n(y) \triangleq dW_n(y)/dy$, where $W_n(y) = P[W_n \leq y]$. To find $w_{n+1}(y)$ we follow the prescription given in Eq. (8.76) and begin by forming the pdf for the sum $w_n + u_n$, which, due to the independence of these two random variables, is clearly the convolution $W_n(y) \otimes c(y)$. This convolution will result in a density function that has nonnegative values for negative as well as positive values of its argument. However Eq. (8.76) requires that our next step in the calculation of $W_{n+1}(y)$ is to calculate the pdf associated with $(w_n + u_n)^+$; this requires that we take the total probability associated with all negative arguments for this density just found [i.e., for $w_n(y) \otimes c(y)$] and collect it together as an impulse of probability located at the origin for $w_{n+1}(y)$. The value of this impulse will just be the integral of our former density on the negative half line. We say in this case that "we sweep the probability in the negative half line up to the origin." The values found from the convolution on the positive half line are correct for w_{n+1} in that region. The algebra that describes this operation is that which Kingman introduces for studying the system G/G/1. Our iterative procedure continues by next forming the convolution of $w_{n+1}(y)$ with $c(y)$, sweeping the probability in the negative half line up to the origin to form $w_{n+2}(y)$ and then proceeds to form $w_{n+3}(y)$ in a like fashion, and so on.

The elements of this algebra consist of all finite signed measures on the real line (for example, a pdf on the real line). For any two such measures, say h_1 and h_2 , the sum $h_1 + h_2$ and also all scalar multiples of either belong to this algebra. The product operation $h_1 \circledast h_2$ is defined as the convolution of h_1 with h_2 . It can be shown that this algebra is a real commutative algebra. There also exists an identity element denoted by e such that $e \circledast h = h$ for any h in the algebra, and it is clear that e will merely be a unit impulse located at the origin. We are interested in operators that map real functions into other real functions and that are measurable. Specifically we are interested in the operator that takes a value x and maps it into the value $(x)^+$, where as usual we have $(x)^+ \triangleq \max [0, x]$. Let us denote this operator by π , which is not to be confused with the matrix of the transition probabilities used in Chapter 2; thus, if we let A denote some event which is measurable, and let $h(A) = P\{w: X(w) \in A\}$ denote the measure of this event, then π is defined through

$$11'h(A) = P\{w: X^+(w) \in A\}$$

We note the linearity of this operator, that is, $\pi(ah) = a11'(h)$ and $\pi(h_1 + h_2) = \pi(h_1) + \pi(h_2)$. Thus we have a commutative algebra (with identity) along with the linear operator π that maps this algebra into itself.

Since $[(x)^+]^+ = (x)^+$ we see that an important property of this operator π is that

$$\pi^2 = \pi$$

A linear operator satisfying such a condition is referred to as a *projection*. Furthermore a projection whose range and null space are both subalgebras of the underlying algebra is called a *Wendel projection*; it can be shown that π has this property, and it is this that makes the solution for $G/G/1$ possible.

Now let us return to considerations of the queue $G/G/1$. Recall that the random variable u_n has pdf $c(u)$ and that the waiting time for the n th customer w_n has pdf $11'(Y)$. Again since u_n and w_n are independent then $w_n + u_n$ has pdf $c(y) \circledast 11'(Y)$. Furthermore, since $w_{n+1} = (w_n + u_n)^+$ we have therefore

$$w_{n+1}(y) = \pi(c(y) \circledast w_n(y)) \quad n=0, 1, \dots \quad (8.78)$$

and this equation gives the pdf for waiting times by induction. Now if $\rho < 1$ the limiting pdf $11'(Y)$ exists and is independent of $11'0'$. That is, $\tilde{w}$ must have the same pdf as $(\tilde{w} + u)^+$ (a remark due to Lindley [LIND 52]). This gives us the basic equation defining the stationary pdf for waiting time in $G/G/1$:

$$w(y) = \pi(c(y) \circledast 11'(Y)) \quad (8.79)$$

The solution of this equation is of main interest in solving $G/G/1$. The remaining portion of this section gives a succinct summary of some elegant results involving this algebra; only the courageous are encouraged to continue.

The particular formalism used for constructing this algebra and carrying out the solution of Eq. (8.79) is what distinguishes the various methods we have mentioned above. In order to see the relationship among the various approaches we now introduce *Spitzer's identity*. In order to state this identity, which involves the recurrence relation given in Eq. (8.78), we must introduce the following z -transform:

$$X(z, y) = \sum_{n=0}^{\infty} w_n(y) z^n \quad (8.80)$$

Addition and scalar multiplication may be defined in the obvious way for this power series and "multiplication" will be defined as corresponding to convolution as is the usual case for transforms. Spitzer's identity is then given as

$$X(z, y) = e^{-\pi(y)} \pi(w_0(y) e^{\pi(y)-y}) \quad (8.81)$$

where

$$y \triangleq \log [e - zc(y)] \quad (8.82)$$

Thus $w_n(y)$ may be found by expanding $X(z, y)$ as a power series in z and picking out the coefficient of z^n . It is not difficult to show that

$$X(z, y) = w_0(y) + z\pi(QY) \otimes X(z, y) \quad (8.83)$$

We may also form a generating function on the sequence $E[e^{-sw_n}] \triangleq W_n^*(s)$, which permits us to find the transform of the limiting waiting time; that is,

$$\lim_{n \rightarrow \infty} W_n^*(s) = W^*(s) \triangleq E[e^{-s\tilde{w}}]$$

so long as $\rho < 1$. This leads us to the following equation, which is also referred to as Spitzer's identity and is directly applicable to our queueing problem:

$$W^*(s) = \exp \left(- \sum_{n=1}^{\infty} \frac{1}{n} E[I - e^{-sIU_n}]^+ \right) \quad (8.84)$$

We never claimed it would be simple!]

If we deal with $W_n^*(s)$ it is possible to define another real commutative algebra (in which the product is defined as multiplication rather than convolution as one might expect). The algebraic solution to our basic equation (8.79) may be carried out in either of these two algebras; in the transformed

† From this identity we easily find that

$$E[I^j] \triangleq W = \sum_{n=1}^{\infty} \frac{1}{n} E[(U \cdot)^+)^j]$$

case one deals with the power series

$$X^*(Z, S) \triangleq \sum_{n=0}^{\infty} W_n^*(s) z^n \quad (8.85)$$

rather than with the series given in Eq. (8.80).

Pollaczek considers this latter case and for G/G/I obtains the following equation which serves to define the system behavior:

$$X^*(z, s) = W_0^*(s) + \frac{zS}{2\pi j} \int_{je-\infty}^{je+\infty} \frac{C^*(s') X^*(z, s') ds'}{s(s-s')} \quad (8.86)$$

and he then shows after considerable complexity that this solution must be of the form

$$X^*(z, s) = e^{-\hat{\pi}(\hat{\gamma}(s))} \hat{\pi}(W_0^*(s)) e^{\hat{\pi}(\hat{\gamma}(s)) - \hat{\gamma}(s)} \quad (8.87)$$

where

$$\hat{\gamma}(s) \triangleq \log(1 - zC^*(s))$$

and

$$\hat{\pi}(\hat{\gamma}(s)) \triangleq \frac{s}{2\pi j} \int_{je-\infty}^{je+\infty} \frac{\hat{\gamma}(s') ds'}{s'(s-s')}$$

When $C^*(s)$ is simple enough then these expressions can be evaluated by contour integrals.

On the other hand, the method we have described in the previous section using spectrum factorization may be phrased in terms of this algebra as follows. If we replace $s\Phi_+(s)$ by $W^*(s)$ and $s\Phi_-(s)$ by $W_-^*(s)$ then our basic equation reads

$$W^*(s) + W_-^*(s) = C^*(s)W^*(s)$$

Corresponding to Eq. (8.83) the transformed version becomes

$$\hat{\pi}(X^*(z, s) - W_0^*(s) - zC^*(s)X^*(z, s)) = 0$$

and the spectrum factorization takes the form

$$1 - zC^*(s) = e^{\hat{\pi}(\hat{\gamma}(s))} e^{\hat{\gamma}(s) - \hat{\pi}(\hat{\gamma}(s))} \quad (8.88)$$

This spectrum factorization, of course, is the critical step.

This unification as an algebra for queues is elegant but as yet has provided little in the way of extending the theory. In particular, Kingman points out that this approach does not easily extend to the system G/G/m since whereas the range of this algebra is a subalgebra, its null space is not; therefore, we do not have a Wendel projection. Perhaps the most enlightening aspect of this discussion is the significant equation (8.79), which gives the basic condition that must be satisfied by the pdf of waiting time. We take advantage of its recurrence form, Eq. (8.78), in Chapter 2, Volume II.

8.4. THE IDLE TIME AND DUALITY

Here we obtain an expression for $W^*(s)$ in terms of the transform of the idle-time pdf and interpret this result in terms of duality in queues.

Let us return to the basic equation given in (8.5), that is,

$$w_{n+1} = \max [0, w_n + u_n]$$

We now define a new random variable which is the "other half" of the waiting time, namely,

$$Y_n = -\min [0, w_n + u_n] \quad (8.89)$$

This random variable in some sense corresponds to the random variable whose distribution is $W_-(y)$, which we studied earlier. Note from these last two equations that when $Y_n > 0$ then $w_{n+1} = 0$ in which case Y_n is merely the length of the *idle* period, which is terminated with the arrival of C_{n+1} . Moreover, since either w_{n+1} or Y_n must be 0, we have that

$$w_{n+1}y_n = 0 \quad (8.90)$$

We adopt the convention that in order for an idle period to exist, it must have nonzero length, and so if Y_n and w_{n+1} are both 0, then we say that the busy period continues (an annoying triviality).

From the **definitions** we observe the following to be true in all cases:

$$w_{n+1} - y_n = w_n + u_n \quad (8.91)$$

From this last equation we may obtain a number of important results and we proceed here as we did in Chapter 5, where we derived the expected queue size for the system *MjGjl* using the imbedded Markov chain approach. In particular, let us take the expectation of both sides of Eq. (8.91) to give

$$E[w_{n+1}] - E[Y_n] = E[w_n] + E[u_n]$$

We assume $E[u_n] < 0$, which (except for $D\{D\}$ where ≤ 0 will do) is the necessary and sufficient condition for there to be a stationary (and unique) waiting-time distribution independent of n ; this is the same as requiring $\rho = \bar{x}/\bar{t} < 1$. In this case we have*

$$\lim_{n \rightarrow \infty} E[w_{n+1}] = \lim_{n \rightarrow \infty} E[w_n]$$

* One must be cautious in claiming that

$$\lim_{n \rightarrow \infty} E[w_n^k] = \lim_{n \rightarrow \infty} E[w_{n+1}^k]$$

since these are distinct random variables. We permit that step here, but refer the interested reader to Wolff [WOLF 70] for a careful treatment.

and so our earlier equation gives

$$E[\tilde{y}] = -E[\tilde{u}] \quad (8.92)$$

where $Y_n \rightarrow \tilde{y}$ and $u_n \rightarrow \tilde{u}$. (We note that the *idle* periods are independent and identically distributed, but the duration of an idle period does depend upon the duration of the previous busy period.) Now from Eq. (8.13) we have $E[\tilde{u}] = i(p - 1)$ and so

$$E[y] = i(1 - p) \quad (8.93)$$

Let us now square Eq. (8.91) and then take expected values as follows:

$$w_{n+1}^2 - 2w_{n+1}y_n + y_n^2 = w_n^2 + 2w_nu_n + u_n^2$$

Using Eq. (8.90) and recognizing that the moments of the limiting distribution on w_n must be independent of the subscript we have

$$E[(y)^2] = 2E[\tilde{w}\tilde{u}] + E[(\tilde{u})^2]$$

We now revert to the *simpler* notation for moments, $w^k \triangleq E[(\tilde{w})^k]$, etc. Since w_n and u_n are independent random variables we have $E[\tilde{w}\tilde{u}] = \tilde{w}\tilde{u}$; using this and Eq. (8.92) we find

$$E[y^2] = W = \frac{u^2}{2i} - \frac{y^2}{2y} \quad (8.94)$$

Recalling that the mean residual life of a random variable X is given by $X/2$, we observe that W is merely the mean residual life of $-\tilde{u}$ less the mean residual life of y ! We must now evaluate the second moment of $\tilde{u}$. Since $\tilde{u} = \tilde{x} - i$; then $\tilde{u}^2 = (\tilde{x} - i)^2$, which gives

$$\overline{\tilde{u}^2} = \sigma_a^2 + \sigma_b^2 + (i)^2(1 - \rho)^2 \quad (8.95)$$

where σ_a^2 and σ_b^2 are the variance of the interarrival-time and service-time densities, respectively. Using this expression and our previous result for $\tilde{u}$ we may thus convert Eq. (8.94) to

$$W = \frac{\sigma_a^2 + \sigma_b^2 + (i)^2(1 - \rho)^2}{2i(1 - \rho)} - \frac{y^2}{2y} \quad (8.96)$$

We must now calculate the first two moments of $\tilde{y}$ [we already know that $\tilde{y} = i(1 - p)$ but wish to express it differently to eliminate a constant]. This we do by conditioning these moments with respect to the occurrence of an idle period. That is, let us define

$$\begin{aligned} G_0 &= P[y > 0] \\ &= P[\text{arrival finds the system idle}] \end{aligned} \quad (8.97)$$

It is clear that we have a stable system when $a_0 > 0$. Furthermore, since we have defined an idle period to occur only when the system remains idle for a nonzero interval of time, we have that

$$P[\tilde{y} \leq y \mid \tilde{y} > 0] = P[\text{idle period} \leq y] \quad (8.98)$$

and this last is just the idle-period distribution earlier denoted by $F(y)$. We denote by I the random variable representing the idle period. Now we may calculate the following:

$$\begin{aligned} \bar{y} &= E[\tilde{y} \mid \tilde{y} = 0]P[\tilde{y} = 0] + E[\tilde{y} \mid \tilde{y} > 0]P[\tilde{y} > 0] \\ &= 0 + a_0 E[\tilde{y} \mid \tilde{y} > 0] \end{aligned}$$

The expectation in this last equation is merely the expected value of I and so we have

$$\bar{y} = a_0 \bar{I} \quad (8.99)$$

Similarly, we find

$$\overline{y^k} = a_0 \bar{I}^k \quad (8.100)$$

Thus, in particular, $y^2/2\bar{y} = \bar{I}^2/2\bar{I}$ (a_0 cancels!) and so we may rewrite the expression for W in Eq. (8.96) as

$$W = \frac{\sigma_a^2 + \sigma_b^2 + (i)^2(1-p)/2}{2i(1-p)} \frac{1^2}{21} \quad (8.101)$$

Unfortunately this is as far as we can go in establishing W for $G/G/1$. The calculation now involves the determination of the first two moments of the idle period. In general, for $G/G/1$ we cannot easily solve for these moments since the idle period depends upon the particular way in which the previous busy period terminated. However in Chapter 2, Volume II, we place bounds on the second term in this equation, thereby bounding the mean wait W .

As we did for $M/G/1$ in Chapter 5 we now return to our basic equation (8.91) relating the important random variables and attempt to find the transform of the waiting time density $W^*(s) \triangleq E[e^{-s\tilde{w}}]$ for $G/G/1$. As one might expect this will involve the idle-time distribution as well. Forming the transform on both sides of Eq. (8.91) we have

$$E[e^{-s(w_{n+1} - y_n)}] = E[e^{-s(w_n + u_n)}]$$

However since w_n and u_n are independent, we find

$$E[e^{-s(w_{n+1} - y_n)}] = E[e^{-s w_n}] E[e^{-s u_n}] \quad (8.102)$$

In order to evaluate the left-hand side of this transform expression we take advantage of the fact that only one or the other of the random variables w_{n+1} and Y_n may be nonzero. Accordingly, we have

$$E[e^{-s(w_{n+1} - Y_n)}] = E[e^{-s(-Y_n)} | Y_n > 0]P[Y_n > 0] + E[e^{-s w_{n+1}} | S_n^+ = 0]P[Y_n = 0] \quad (8.103)$$

To determine the right-hand side of this last equation we may use the following similar expansion:

$$E[e^{-s w_{n+1}}] = E[e^{-s w_{n+1}} | S_n^+ = 0]P[Y_n = 0] + E[e^{-s w_{n+1}} | S_n^+ > 0]P[Y_n > 0] \quad (8.104)$$

However, since $w_{n+1} Y_n = 0$, we have $E[e^{-s w_{n+1}} | Y_n > 0] = 1$. Making use of the definition for ϕ_0 in Eq. (8.97) and allowing the limit as $\rho \rightarrow \infty$ we obtain the following transform expression from Eq. (8.104):

$$E[e^{-s w} | \tilde{y} = 0]P[\tilde{y} = 0] = w^*(s) - \phi_0$$

We may then write the limiting form of Eq. (8.103) as

$$E[e^{-s(\tilde{w} - \tilde{y})}] = 1^*(-s) + W^*(s) - \phi_0 \quad (8.105)$$

where $1^*(s)$ is the Laplace transform of the idle-time pdf [see Eq. (8.98) for the definition of this distribution]. Thus, from this last and from Eq. (8.102), we obtain immediately

$$W^*(s)C^*(s) = \phi_0 1^*(-s) + W^*(s) - \phi_0$$

where as in the past $C^*(s)$ is the Laplace transform for the density describing the random variable $\tilde{y}$. This last equation finally gives us [MARS 68]

$$W^*(s) = \frac{\phi_0 [1 - 1^*(-s)]}{1 - C^*(s)} \quad (8.106)$$

which represents the generalization of the Pollaczek-Khinchin transform equation given in Chapter 5 and which now applies to the system G/G/1. Clearly this equation holds at least along the imaginary axis of the complex s -plane, since in that case it becomes the characteristic function of the various distributions which are known to exist.

Let us now consider some examples.

Example 1: M/M/1

For this system we know that the idle-period distribution is the same as the interarrival-time distribution, namely,

$$F(y) = P[I \leq y] = 1 - e^{-y}, \quad y \geq 0 \quad (8.107)$$

And so we have the first two moments $1 = 1/\lambda$, $1/2 = 2/\lambda^2$; we also have $a\sigma^2 = 1/\lambda^2$ and $a_b^2 = 1/\mu^2$. Using these values in Eq. (8.101) we find

$$W = \frac{\lambda^2(1/\lambda^2 + 1/\mu^2) + (1-p)^2}{2\lambda(1-p)} \quad 1$$

and so

$$W = \frac{\rho/\mu}{1-p} \quad (8.108)$$

which of course checks with our earlier results for M/M/1.

We know that $I^*(s) = \lambda/(s + \lambda)$ and $C^*(s) = \lambda\mu/(\lambda - s)(s + \mu)$. Moreover, since the probability that a Poisson arrival finds the system empty is the same as the long-run proportion of time the system is empty, we have that $G_0 = 1 - \rho$ and so Eq. (8.106) yields

$$\begin{aligned} W^*(s) &= \frac{(1-p)[1 - \lambda/(\lambda - s)]}{1 - \lambda\mu/(\lambda - s)(s + \mu)} \\ &= \frac{(1-p)s(s + \mu)}{(\lambda - s)(s + \mu) - \lambda\mu} \\ &= \frac{(1-p)(s + \mu)}{s + \mu - \lambda} \end{aligned} \quad (8.109)$$

which is the same as Eq. (5.120).

Example 2: M/G/1

In this case the idle-time distribution is as in M/M/1; however, we must leave the variance for the service-time distribution as an unknown. We obtain

$$\begin{aligned} W &= \frac{\lambda^2[(1/\lambda^2) + \sigma_b^2] + (1-p)^2}{2(1-p)} \quad 1 \\ &= \frac{(1 + C_b^2)}{2\mu(1-p)} \end{aligned} \quad (8.110)$$

which is the P-K formula. Also, $C^*(s) = B^*(s)\lambda/(\lambda - s)$ and again $G_0 = (1 - \rho)$. Equation (8.106) then gives

$$\begin{aligned} W^*(s) &= \frac{(1-p)[1 - \lambda/(\lambda - s)]}{1 - [\lambda/(\lambda - s)]B^*(s)} \\ &= \frac{s(1-p)}{s - \lambda + \lambda B^*(s)} \end{aligned} \quad (8.111)$$

which is the **P-K** transform equation for waiting time!

Example 3: D!D!]

In this case we have that the length of the idle period is a constant and is given by $\bar{i} = i - \bar{x} = \bar{i}(1 - \rho)$; therefore $1 = \bar{i}(1 - \rho)$, and $\bar{i}^2 = (1)^2$. Moreover, $a\alpha^2 = \sigma_b^2 = 0$. Therefore Eq. (8.101) gives

$$W = \frac{0 + (\bar{i})^2(1 - \rho)^2}{2\bar{i}(1 - \rho)} - \frac{1}{2}\bar{i}(1 - \rho)$$

and so

$$W = 0 \quad (8.112)$$

This last is of course correct since the equilibrium waiting time in the (stable) system $D/D/1$ is always zero.

Since $\bar{x}$, I , and J are all constants, we have $B^*(s) = e^{-s\bar{x}}$, $A^*(s) = e^{-sI}$ and $I^*(s) = e^{-sJ}$, $J = e^{-s\bar{i}(1-\rho)}$. Also, with probability one an arrival finds the system empty; thus $a_0 = 1$. Then Eq. (8.106) gives

$$\begin{aligned} W^*(s) &= \frac{1[1 - e^{-sI} - \rho]}{1 - e^{-s\bar{i}}e^{-s\bar{x}}} \\ &= 1 \end{aligned} \quad (8.113)$$

and so $w(y) = \delta(y)$, an impulse at the origin which of course checks with the result that no waiting occurs.

Considerations of the idle-time distribution naturally lead us to the study of *duality* in queues. This material is related to the ladder indices we had defined in Section 5.11. The random walk we are interested in is the sequence of values taken on by U' ; [as given in Eq. (8.9)]. Let us denote by U_{n_k} the value taken on by U' at the k th *ascending* ladder index (instants when the function first drops below its latest *maximum*). Since $\bar{u} < 0$ it is clear that $\lim_{n \rightarrow \infty} U_n = -\infty$ as $n \rightarrow \infty$. Therefore, there will exist a (finite) integer K such that K is the largest ascending ladder index for U_n . Now from Eq. (8.11) repeated below

$$U_n' = \sup_{n \geq 0} U_n$$

It is clear that

$$U_n'' = U_n''$$

Now let us define the random variable t , (which as we shall see is related to an idle time) as

$$t_k = U_{n_k} - U_{n_{k-1}} \quad (8.114)$$

for $k \leq K$. That is t_k is merely the amount by which the new ascending ladder height exceeds the previous ascending ladder height. Since all of the random variables u_i are independent then the random variables i_k conditioned on K are independent and identically distributed. If we now let $1 - \sigma = P[U_n \leq U_n \text{ for all } n > n_k]$ then we may easily calculate the distribution for K as

$$P[K = k] = (1 - \sigma)\sigma^k \quad (8.115)$$

In exercise 8.16, we show that $(1 - \sigma) = P[\tilde{w} = 0]$. Also it is clear that

$$\begin{aligned} I_1 + I_2 + \dots + I_K &= U_{n_1} - U_{n_0} + U_{n_2} - U_{n_1} + \dots + U_{n_g} - U_{n_{K-1}} \\ &= U_{n_K} \end{aligned}$$

where $n_0 \triangleq 0$ and $U_0 \triangleq 0$. Thus we see that $\tilde{w}$ has the same distribution as $I_1 + \dots + I_K$ and so we may write

$$\begin{aligned} E[e^{-s\tilde{w}}] &= E[E[e^{-s(I_1 + \dots + I_K)} | K]] \\ &= E[(I^*(s))^K] \end{aligned} \quad (8.116)$$

where $I^*(s)$ is the Laplace transform for the pdf of each of the t_i (each of which we now denote simply by I). We may now evaluate the expectation in Eq. (8.116) by using the distribution for K in Eq. (8.115) finally to yield

$$W^*(s) = 1 - \frac{\sigma}{1 - I^*(s)} \quad (8.117)$$

Here then, is yet another expression for $W^*(s)$ in the $G/G/1$ system.

We now wish to interpret the random variable $\tilde{w}$ by considering a "dual" queue (whose variables we will distinguish by the use of the symbol $\hat{\cdot}$). The dual queue for the $G/G/1$ system considered above is the queue in which the service times x_n in the original system become the interarrival times i_{n+1} in the dual queue and also the interarrival times I_{n+1} from the original queue become the service times $\hat{x}_n$ in the dual queue.] It is clear then that the random variable $\hat{u}_n$ for the dual queue will merely be $\hat{u}_n = \hat{x}_n - i_{n+1} = I_{n+1} - x_n = -U_n$ and defining $\hat{O}_n = \hat{u}_0 + \dots + \hat{u}_{n-1}$ for the dual queue we have

$$\hat{U}_n = -U_n \quad (8.118)$$

Clearly, if the original queue is stable, the dual must be unstable, and conversely (except that both may be unstable if $\rho = 1$).

as the relationship among the dual and the original queues. It is then clear from our discussion in Section 5.11 that the ascending and descending ladder indices are interchanged for the original and the dual queue (the same is true of the ladder heights). Therefore the first ascending ladder index n in the original queue will correspond to the first descending ladder index in the dual queue; however, we recall that descending ladder indices correspond to the arrival of a customer who terminates an idle period. We denote this customer by e_n . Clearly the length of the idle period that he terminates in the dual queue is the difference between the accumulated interarrival times and the accumulated service times for all customers up to his arrival (these services must have taken place in the first busy period), that is, for the dual queue,

$$\begin{aligned}
 \text{Length of first idle period} \\
 \{\text{following first busy period}\} &= \sum_{n=0}^{n_1-1} t_{n+1} - \sum_{n=0}^{n_1-1} x_n \\
 &= \sum_{n=0}^{n_1-1} x_n - \sum_{n=0}^{n_1-1} I_{n+J} \\
 &= u_0 + u_1 + \cdots + u_{n_1-1} \\
 &= U_{n_1}
 \end{aligned} \tag{8.119}$$

where we have used Eq. (8.114) at the last step. Thus we see that *the random variable j is merely the idle period in the dual queue* and so our Eq. (8.117) relates the transform of the waiting time in the original queue to the transform of the idle time pdf in the dual queue [contrast this with Eq. (8.106), which relates this waiting-time transform to the transform of the idle time in its own queue].

This duality observation permits some rather powerful conclusions to be drawn in simple fashion (and these are discussed at length in [FELL 66], especially Sections VI.9 and XII.5). Let us discuss two of these.

Example 4: GIMII

If we have a stable GIMII queue (with $i = 1/\lambda$ and $\bar{x} = 1/\mu$) then the dual is an unstable queue of the type MIGII (with $\bar{i} = 1/\mu$ and $\bar{x} = 1/\lambda$ and so 1 (the distribution of idle time in the dual queue) will be of exponential form; therefore $1^*(s) = \mu/(s + \mu)$, which gives from Eq. (8.117) the following

result for the original $G/M/1$ queue :

$$W^*(s) = \frac{(1 - \frac{\sigma}{\mu})(s + \mu)}{s + \mu - \sigma\mu}$$

Inverting this and forming the PDF for waiting time we have

$$W(y) = 1 - \sigma e^{-\mu(1-\sigma)y} \quad y \geq 0 \quad (8.120)$$

which corresponds exactly to Eq. (6.30)

Example 5: $M/G/1$

As a second example let the original queue be of the form $M/G/1$ and therefore the dual is of the form $G/M/1$. Since $\sigma = P[\tilde{w} > 0]$ it must be that $\sigma = P$ for $M/G/1$. Now in the dual system, since a busy period ends at a random point in time (and since the service time in this dual queue is memoryless), an idle period will have a duration equal to the residual life of an interarrival time; therefore from Eq. (5.11) we see that

$$I^*(s) = \frac{1 - B^*(s)}{s\bar{x}} \quad (8.121)$$

and when these calculations are applied to Eq. (8.117) we have

$$W^*(s) = \frac{1 - P}{1 - p\{[1 - \frac{P}{B^*(s)}]/sx\}} \quad (8.122)$$

which is the P-K transform equation for waiting time rewritten as in Eq. (5.106).

This concludes our study of $G/G/1$. Sad to say, we have been unable to give analytic expressions for the waiting-time distribution explicitly in terms of known quantities. In fact, we have not even succeeded for the mean wait W . Nevertheless, we have given a method for handling the rational case by spectrum factorization, which is quite effective. In Chapter 2, Volume II, we return to $G/G/1$ and succeed in extracting many of its important properties through the use of bounds, inequalities, and approximations.

REFERENCES

- ANDE 53a Andersen, S. E., "On Sums of Symmetrically Dependent Random Variables," *Skand. Aktuar.*, 36, 123-138 (1953).
- ANDE 53b Andersen, S. E., "On the Fluctuations of Sums of Random Variables I," *Math. Scand.*, 1, 263-285 (1953).
- ANDE 54 Andersen, S. E., "On the Fluctuations of Sums of Random Variables II," *Math. Scand.*, 2, 195-223 (1954).
- BENE 63 Benes, V. E., *General Stochastic Processes in the Theory of Queues*, Addison-Wesley (Reading, Mass.), 1963.
- FELL 66 Feller, W., *Probability Theory and its Applications*, Vol. II, Wiley (New York), 1966.
- KEIL 65 Keilson, J., "The Role of Green's Functions in Congestion Theory," *Proc. Symp. on Congestion Theory* (edited by W. L. Smith and W. E. Wilkinson) Univ. of North Carolina Press (Chapel Hill), 43-71 (1965).
- KING 66 Kingman, J. F. C., "On the Algebra of Queues," *Journal of Applied Probability*, 3, 285-326 (1966).
- LIND 52 Lindley, D. V., "The Theory of Queues with a Single Server," *Proc. Cambridge Philosophical Society*, 48, 277-289 (1952).
- MARS 68 Marshall, K. T., "Some Relationships between the Distributions of Waiting Time, Idle Time, and Interoutput Time in the $GI/G/1$ Queue," *SIAM Journal Applied Math.*, 16, 324-327 (1968).
- POLL 57 Pollaczek, F., *Problemes Stochastiques Posés par le Phenomene de Formation d'une Queue d'Attente à un Guichet et par de Phénomènes Apparentes*, Gauthiers Villars (Paris), 1957.
- RICE 62 Rice, S. O., "Single Server Systems," *Bell System Technical Journal*, 41, Part I: "Relations Between Some Averages," 269-278 and Part II: "Busy Periods," 279-310 (1962).
- SMIT 53 Smith, W. L., "On the Distribution of Queueing Times," *Proc. Cambridge Philosophical Society*, 49, 449-461 (1953).
- SPIT 56 Spitzer, F. "A Combinatorial Lemma and its Application to Probability Theory," *Transactions of the American Mathematical Society*, 82, 323-339 (1956).
- SPIT 57 Spitzer, E., "The Wiener-Hopf Equation whose Kernel is a Probability Density," *Duke Mathematics Journal*, 24, 327-344 (1957).
- SPIT 60 Spitzer, F. "A Tauberian Theorem and its Probability Interpretation," *Transactions of the American Mathematical Society*, 94, 150-160 (1960).
- SYSK 62 Syski, R., *Introduction to Congestion Theory in Telephone Systems*, Oliver and Boyd (London), 1962.
- TITC 52 Titchmarsh, E. C., *Theory of Functions*, Oxford Univ. Press (London), 1952.
- WEND 58 Wendel, F. G., "Spitzer's Formula; a Short Proof," *Proc. American Mathematical Society*, 9, 905-908 (1958).

- WOLF 70 Wolff, R. W., "Bounds and Inequalities in Queueing," unpublished notes, Department of Industrial Engineering and Operations Research, University of California (Berkeley), 1970.

EXERCISES

- 8.1. From Eq. (8.18) show that $C^*(s) = A^*(-s)B^*(s)$.
- 8.2. Find ceu for *MIMI*!
- 8.3. Consider the system *MIDI* with a fixed service time of $\bar{x}$ sec.
 (3) Find $ceu = P[u_n \leq II]$ and sketch its shape.
 (b) Find $E[u_n]$.
- 8.4. For the sequence of random variables given below, generate the figure corresponding to Figure 8.3 and complete the table.

II	0	2	3	4	5	6	7	8	9
t_{n+1}	2	1	1	5	7	2	2		6
x_n	3	4	2	3	3	4	2		3
II_n									
w_n measured									
w_n calculated									

- 8.5. Consider the case where $\rho = 1 - \epsilon$ for $0 < \epsilon \ll 1$. Let us expand $W(y - II)$ in Eq. (8.23) as

$$W(y - u) = W(y) - uW'(y) + \frac{u^2}{2}W''(y) + R(u, y)$$

where $w^{(n)}(y)$ is the n th derivative of $W(y)$ and $R(II, y)$ is such that $\int_{-\infty}^y R(u, y) dC(u)$ is negligible due to the slow variation of $W(y)$ when $\rho = 1 - \epsilon$. Let II_k denote the k th moment of II .

- (3) Under these conditions convert Lindley's integral equation to a second-order linear differential equation involving $\bar{u}^2$ and II .
- (b) With the boundary condition $W(0) = 0$, solve the equation found in (a) and express the mean wait W in terms of the first two moments of i and $\bar{x}$.
- 8.6. Consider the *DIEI* queueing system, with a constant interarrival time (of $\bar{t}$ sec) and a service-time pdf given as in Eq. (4.16).
 (3) Find ceu .

- (b) Show that Lindley's integral equation yields $W(y - i) = 0$ for $y < i$ and

$$W(y - i) = \int_0^y W(y - w) dB(w) \quad \text{for } y \geq i$$

- (c) Assume the following solution for $W(y)$:

$$W(y) = 1 + \sum_{i=1}^r G_i e^{a_i y} \quad y \geq 0$$

where a , and α_i may both be complex, but where $\text{Re}(\alpha_i) < 0$ for $i = 1, 2, \dots, r$. Using this assumed solution, show that the following equations must hold:

$$e^{-\alpha_i i} = \left(\frac{r\mu}{r\mu + \alpha_i} \right)^r \quad i = 1, 2, \dots, r$$

$$\sum_{i=0}^r (r\mu + \alpha_i)^{j+1} = 0 \quad j = 0, 1, \dots, r-1$$

where $a_0 = 1$ and $\alpha_0 = 0$. Note that $\{\alpha_i\}$ may be found from the first set of (transcendental) equations, and then the second set gives $\{a_i\}$. It can be shown that the α_i are distinct. See [SYSK 62].

- 8.7. Consider the following queueing systems in which *no queue* is permitted. Customers who arrive to find the system busy must leave without service.
- $M/M/1$: Solve for $P_k = P[k \text{ in system}]$.
 - $M/H/1$: As in Figure 4.10 with $\alpha_1 = \alpha$, $\alpha_2 = 1 - \alpha$, $\mu_1 = 2\mu\alpha$ and $\mu_2 = 2\mu(1 - \alpha)$.
 - Find the mean service time $\bar{x}$.
 - Solve for P_0 (an empty system), $P_{\ll}$ (a customer in the $2\mu\alpha$ box) and $P_{1-\alpha}$ (a customer in the $2\mu(1 - \alpha)$ box).
 - $H/M/1$: Where $A(t)$ is hyperexponential as in (b), but with parameters $\mu_1 = 2\lambda\alpha$ and $\mu_2 = 2\lambda(1 - \alpha)$ instead. Draw the state-transition diagram (with labels on branches) for the following four states: E_{ij} is state with "arriving" customer in arrival stage i and j customers in service $i = 1, 2$ and $j = 0, 1$.

- (d) $M/Er/1$: Solve for $P_i = P[j \text{ stages of service left to go}]$.
- (e) $M/D/1$: With all service times equal to $\bar{x}$
- Find the probability of an empty system.
 - Find the fraction of lost customers.
- (O) $E_2/M/1$: Define the four states as E_{ij} where i is the number of "arrival" stages left to go and j is the number of customers in service. Draw the labeled state-transition diagram.
- 8.8. Consider a single-server queueing system in which the interarrival time is chosen with probability α from an exponential distribution of mean $1/\lambda$ and with probability $1 - \alpha$ from an exponential distribution with mean $1/\mu$. Service is exponential with mean $1/\mu$.
- Find $A^*(s)$ and $B^*(s)$.
 - Find the expression for $\Psi_+(s)/\Psi_-(s)$ and show the pole-zero plot in the s -plane.
 - Find $\Psi_+(s)$ and $\Psi_-(s)$.
 - Find $\Phi_+(s)$ and $W(y)$.
- 8.9. Consider a $G/G/1$ system in which

$$A^*(s) = \frac{2}{(s+1)(s+2)}$$

$$B^*(s) = \frac{1}{s+1}$$

- Find the expression for $\Psi_+(s)/\Psi_-(s)$ and show the pole-zero plot in the s -plane.
 - Use spectrum factorization to find $\Psi_+(s)$ and $\Psi_-(s)$.
 - Find $\Phi_+(s)$.
 - Find $W(y)$.
 - Find the average waiting time W .
 - (O) We solved for $W(y)$ by the method of spectrum factorization. Can you describe another way to find $W(y)$?
- 8.10. Consider the system $M/G/1$. Using the spectral solution method for Lindley's integral equation, find
- $\Psi_+(s)$. {HINT: Interpret $[1 - B^*(s)]/sX$.}
 - $\Psi_-(s)$.
 - $s\Phi_+(s)$.

8.11. Consider the queue $E_0/E_r/1$.

(a) Show that

$$\frac{\Psi_+(s)}{\Psi_-(s)} = \frac{F(s)}{1 - F(s)}$$

where $F(s) = 1 - (1 - s/\lambda q)^a(1 + s/\mu r)^r$.

(b) For $\rho < 1$, show that $F(s)$ has one zero at the origin, zeroes $s_1, s_2, \dots, s_r$ in $\operatorname{Re}(s_i) < 0$, and zeroes $s_{r+1}, s_{r+2}, \dots, s_{r+q-1}$ in $\operatorname{Re}(s_i) > 0$.

(c) Express $\Psi_+(s)$ and $\Psi_-(s)$ in terms of s_i .

(d) Express $W^*(s)$ in terms of s_i ($i = 1, 2, \dots, r + q - 1$).

8.12. Show that Eq. (8.71) is equivalent to Eq. (6.30).

8.13. Consider a $M/M/1$ queue with $\rho < 1$. Assume $w_0 = 4f(1 - \rho)$.

(a) Calculate $W_n(y)$ using the procedure defined by Eq. (8.78) for $n = 0, 1, 2, \dots$.

(b) Show that the known solution for

$$w(y) = \lim_{n \rightarrow \infty} w_n(y)$$

satisfies Eq. (8.79).

8.14. Consider an $M/M/1$ queue with $\rho < 1$. Assume $w_0 = 0$.

(a) Calculate $W_1(y)$ using the procedure defined by Eq. (8.78).

(b) Repeat for $w_2(y)$.

(c) Show that our known solution for

$$W(y) = \lim_{n \rightarrow \infty} w_n(y)$$

satisfies Eq. (8.79).

(d) Compare $w_2(y)$ with $W(y)$.

8.15. By first cubing Eq. (8.91) and then forming expectations, express σ_w^2 (the variance of the waiting time) in terms of the first three moments of i , $\tilde{x}$, and l .

8.16. Show that $P[\tilde{w} = 0] = 1 - a$ from Eq. (8.117) by finding the constant term in a power-series expansion of $W^*(s)$.

8.17. Consider a $G/G/1$ system.

(a) Express $1^*(1)$ in terms of the transform of the pdf of idle time in the given system.

(b) Using (a) find $1^*(1)$ when the original system is the ordinary $M/M/1$.

- (C) Using (a), show that the transform of the idle-time pdf in a G/M/1 queue is given by

$$I^*(s) = \frac{1 - A^*(s)}{s\bar{f}}$$

thereby reaffirming Eq. (8.121).

- (d) Since either the original or the dual queue must be unstable (except for $D/D/1$), discuss the existence of the transform of the idle-time pdf for the unstable queue.

Epilogue

We have invested eight chapters (and two appendices!) in studying the theory of queueing systems. Occasionally we have been overjoyed at the beauty and generality of the results, but more often we have been overcome (with frustration) at the lack of real progress in the theory. (No, we never promised you a rose garden.) However, we did seduce you into believing that this study would provide worthwhile methods for practical application to many of today's pressing congestion problems. We confirm that belief in Volume II.

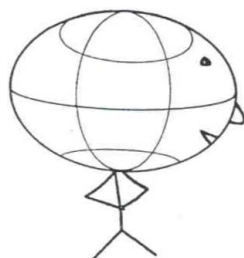
In the next volume, after a brief review of this one, we begin by taking a more relaxed view of *G/G/1*. In Chapter 2, we enter a new world leaving behind the rigor (and pain) of exact solutions to exact problems. Here we are willing to accept the raw facts of life, which state that our models are not perfect pictures of the systems we wish to analyze so we should be willing to accept approximations and bounds in our problem solution. Upper and lower bounds are found for the average delay in *G/G/1* and we find that these are related to a very useful heavy traffic approximation for such queues. This approximation, in fact, predicts that the long waiting times are exponentially distributed. A new class of models is then introduced whereby the discrete arrival and departure processes of queueing systems are replaced first by a fluid approximation (in which these stochastic processes are replaced by their mean values as a function of time), and then secondly by a diffusion approximation (in which we permit a variation about these means). We happily find that these approximations give quite reasonable results for rather general queueing systems. In fact they even permit us to study the transient behavior not only of stable queues but also of saturated queues, and this is the material in the final section of Chapter 2 whereby we give Newell's treatment of the rush-hour approximation—an effective method indeed.

Chapter 3 points the way to our applications in time-shared computer systems by presenting some of the principal results for priority queueing systems. We study general methods and apply them to a number of important queueing disciplines. The conservation law for priority systems is established, preventing the useless search for nonrealizable disciplines.

In the remainder, we choose applications principally from the computer field, since these applications are perhaps the most recent and successful for the theory of queues. In fact, the queueing analysis of allocation of resources and job flow through computer systems is perhaps the only tool available

to computer scientists in understanding the behavior of the complex interaction of users, programs, processes, and resources. In Chapter 4 we emphasize multi-access computer systems in isolation, handling demands of a large collection of competing users. We look for throughput and response time as well as utilization of resources. The major portion of this chapter is devoted to a particular class of algorithms known as *processor-sharing* algorithms, since they are singularly suited to queueing analysis and capture the essence of more difficult and more **complex** algorithms seen in real scheduling problems. Chapter 5 addresses itself to computers in networks, a field that is perhaps the fastest growing in the young computer industry itself (most of the references there are drawn from the last three years—a tell-tale indicator indeed). The chapter is devoted to developing methods of analysis and design for computer-communication networks and identifies many unsolved important problems. A specific existing network, the ARPANET, is used throughout as an example to guide the reader through the motivation and evaluation of the various techniques developed.

Now it remains for you, the reader, to sharpen and apply your new set of tools. The world awaits and you must serve!



Transform Theory Refresher:

z -Transform and Laplace Transform

In this appendix we develop some of the properties and expressions for the z -transform and the Laplace transform as they apply to our studies in queueing theory. We begin with the z -transform since it is easier to visualize its operation. The forms and properties of both transforms are very similar, and we compare them later under the discussion of Laplace transforms.

1.1. WHY TRANSFORMS ?

So often as we progress through the study of interesting physical systems we find that transforms appear in one form or another. These transforms occur in many varieties (e.g., z -transform, Laplace transform, Fourier transform, Mellin transform, Hankel transform, Abel transform) and with a variety of names (e.g., transform, characteristic function, generating function). Why is it that they appear so often? The answer has two parts; first, because they arise *naturally* in the formulation and the solution of systems problems; and second, because when we observe or introduce them into our solution method, they greatly *simplify* the calculations. Moreover, oftentimes they are the only tools we have available for proceeding with the solution at all.

Since transforms do appear naturally, we should inquire as to what gives rise to their appearance. The answer lies in the consideration of *linear time-invariant systems*. A system, in the sense that we use it here, is merely a transformation, or mapping, or input-output relationship between two functions. Let us represent a general system as a "black" box with an input f and an output g , as shown in figure 1.1. Thus the system operates on the function f to produce the function g . In what follows we will assume that these functions depend upon an independent time parameter t ; this arbitrary choice results in no loss of generality but is convenient so that we may discuss certain notions more explicitly. Thus we assume that $f = J(t)$. In order to represent the input-output relationship between the functions $J(t)$ and $get\}$

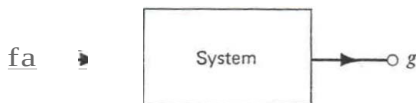


Figure 1.1 A general system.

we use the notation

$$f(t) \rightarrow g(t) \quad (1.1)$$

to denote the fact that $g(t)$ is the output of our system when $f(t)$ is applied as input. A system is said to be *linear* if, when

$$f_1(t) \rightarrow g_1(t)$$

and

$$f_2(t) \rightarrow g_2(t)$$

then also the following is true:

$$af_1(t) + bf_2(t) \rightarrow ag_1(t) + bg_2(t) \quad (1.2)$$

where a and b are independent of the time variable t . Further, a system is said to be *time-invariant* if, when Eq. (1.1) holds, then the following is also true:

$$f(t + T) \rightarrow g(t + T) \quad (1.3)$$

for any T . If the above two properties both hold, then our system is said to be a *linear time-invariant* system, and it is these with which we concern ourselves for the moment.

Whenever one studies such systems, one finds that *complex exponential functions of time* appear throughout the solution. Further, as we shall see, the transforms of interest merely represent ways of decomposing functions of time into sums (or integrals) of complex exponentials. That is, complex exponentials form the *building blocks* of our transforms, and so, we must inquire further to discover why these complex exponentials pervade our thinking with such systems. Let us now pose the fundamental question, namely, which functions of time $f(t)$ may pass through our linear time-invariant systems with no change in form; that is, for which $f(t)$ will $g(t) = Hf(t)$, where H is some scalar multiplier (with respect to t)? If we can discover such functions $f(t)$ we will then have found the "eigenfunctions," or "characteristic functions," or "invariants" of our system. Denoting these eigenfunctions by $f_e(t)$ it will be shown that they must be of the following form (to within an arbitrary scalar multiplier):

$$f_e(t) = e^{st} \quad (1A)$$

where s is, in general, a complex variable. That is, the complex exponentials given in (1.4) form the set of eigenfunctions for *all* linear time-invariant systems. This result is so fundamental that it is worthwhile devoting a few lines to its derivation. Thus let us assume when we apply $f_e(t)$ that the output is of the form $g_e(t)$, that is,

$$f_e(t) = e^{st} \rightarrow g_e(t)$$

But, by the linearity property we have

$$e^{s\tau} f_e(t) = e^{s(t+\tau)} \rightarrow e^{s\tau} g_e(t)$$

where τ and therefore $e^{s\tau}$ are both constants. Moreover, from the time-invariance property we must have

$$f_e(t + \tau) = e^{s(t+\tau)} \rightarrow g_e(t + \tau)$$

From these last two, it must be that

$$e^{s\tau} g_e(t) = g_e(t + \tau)$$

The unique solution to this equation is given by

$$g_e(t) = H e^{st}$$

which confirms our earlier hypothesis that the complex exponentials pass through our linear time-invariant systems unchanged except for the scalar multiplier H . H is independent of t but may certainly be a function of s and so we choose to write it as $H = H(s)$. Therefore, we have the final conclusion that

$$e^{st} \rightarrow H(s)e^{st} \quad (1.5)$$

and this fundamental result exposes the eigenfunctions of our systems.

In this way the complex exponentials are seen to be the basic functions in the study of linear time-invariant systems. Moreover, if it is true that the input to such a system is a complex exponential, then it is a trivial computation to evaluate the output of that system from Eq. (1.5) if we are given the function $H(s)$. Thus it is natural to ask that for any *input* $f(t)$ we would hope to be able to decompose $f(t)$ into a sum (or integral) of complex exponentials, each of which contributes to the overall output $g(t)$ through a computation of the form given in Eq. (1.5). Then the overall output may be found by summing (integrating) these individual components of the output. (The fact that the sum of the individual outputs is the same as the output of the sum of the individual inputs—that is, the complex exponential decomposition—is due to the linearity of our system.) The process of decomposing our input into sums of exponentials, computing the response to each from Eq. (1.5), and then reconstituting the output from sums of exponentials is

referred to as the transform method of analysis. This approach, as we can see, arises very naturally from our foregoing statements. In this sense, transforms arise in a perfectly natural way. Moreover, we know that such systems are described by constant-coefficient linear differential equations, and so the common use of transforms in the solution of such equations is not surprising.

We still have not given a precise definition of the transform itself; be patient, for we are attempting to answer the question "why transforms?" If we were to pursue the line of reasoning that follows from Eq. (1.5), we would quickly encounter Laplace transforms. However, it is convenient at this point to consider only functions of discrete time rather than functions of continuous time, as we have so far been discussing. This change in direction brings us to a consideration of z -transforms and we will return to Laplace transforms later in this appendix. The reason for this switch is that it is easier to visualize operations on a discrete-time axis as compared to a continuous-time axis (and it also delays the introduction of the unit impulse function temporarily).

Thus we consider functions f that are defined only at discrete instants in time, which, let us say, are multiples of some basic time unit T . That is, $f(t) = f(t = nT)$, where $n = \dots, -2, -1, 0, 1, 2, \dots$. In order to incorporate this discrete-time axis into our notation we will denote the function $f(t = nT)$ by f_n . We assume further that our systems are also discrete in time. Thus we are led to consider linear time-invariant systems with inputs f and outputs g (also functions of discrete time) for which we obtain the following three equations corresponding to Eqs. (1.1)-(1.3):

$$f_n \rightarrow g_n \quad (1.6)$$

$$af_n^{(1)} + bf_n^{(2)} \rightarrow ag_n^{(1)} + bg_n^{(2)} \quad (1.7)$$

$$f_{n+m} \rightarrow g_{n+m} \quad (1.8)$$

where m is some integer constant. Here Eq. (1.7) is the expression of linearity whereas Eq. (1.8) is the expression of time-invariance for our discrete systems. We may ask the same fundamental question for these discrete systems and, of course, the answer will be essentially the same, namely, that the eigenfunctions are given by

$$f_n^{(e)} = e^{st} = e^{snT}$$

Once again the complex exponentials are the eigenfunctions. At this point it is convenient to introduce the definition

$$z \triangleq e^{-sT} \quad (1.9)$$

and so the eigenfunctions $f_n^{(e)}$ take the form

$$f_n^{(e)} = z^{-n}$$

Since s is a complex variable, so, too, is z . Following through steps essentially identical to those which led from Eq. (1.4) to Eq. (1.5) we find that

$$z^{-n} \rightarrow H(z)z^{-n} \quad (1.10)$$

where $H(z)$ is a function independent of n . This merely expresses the fact that the set of functions $\{z^{-n}\}$ for any value of z form the set of eigenfunctions for discrete linear time-invariant systems. Moreover, the function (constant) H either in Eq. (1.5) or (1.10) tells us precisely how much of a given complex exponential we get out of our linear system when we insert a unit amount of that exponential at the input. That is, H really describes the effect of the system on these exponentials; for this reason it is usually referred to as the *system (or transfer) function*.

Let us pursue this line of reasoning somewhat further. As we all know, a common way to discover what is inside a system is to kick it hard and quickly. For our systems this corresponds to providing an input only at time $t = 0$ and then observing the subsequent output. Thus let us define the *Kronecker delta function* (also known as the *unit function*) as

$$u_n = \begin{cases} 1 & n = 0 \\ 0 & n \neq 0 \end{cases} \quad (1.11)$$

When we apply u_n to our system it is common to refer to the output as the *unit response*, and this is usually denoted by h_n ; That is,

$$u_n \rightarrow h_n$$

From Eq. (1.8) we may therefore also write

$$u_{n+m} \rightarrow h_{n+m}$$

From the linearity property in Eq. (1.7) we have therefore

$$z^m u_{n+m} \rightarrow z^m h_{n+m}$$

Certainly we may multiply both expressions by unity, and so

$$z^{-n} z^n z^m u_{n+m} \rightarrow z^{-n} z^n z^m h_{n+m} \quad (1.12)$$

Furthermore, if we consider a set of inputs $\{f_n^{(i)}\}$, and if we define the output for each of these by

$$f_n^{(i)} \rightarrow g_n^{(i)}$$

then by the linearity of our system we must have

$$\sum_i f_n^{(i)} \rightarrow \sum_i g_n^{(i)} \quad (1.13)$$

If we now apply this last observation to Eq. (1.12) we have

$$z^{-n} \sum_m u_{n+m} z^{n+m} \rightarrow z^{-n} \sum_m h_{n+m} z^{n+m}$$

where the sum ranges over all integer values of m . From the definition in Eq. (1.11) it is clear that the sum on the left-hand side of this equation has only one nonzero term, namely, for $m = -n$, and this term is equal to unity; moreover, let us make a change of variable for the sum on the right-hand side of this expression, giving

$$z^{-n} \rightarrow z^{-n} \sum_k h_k z^k$$

This last equation is now in the same form as Eq. (1.10); it is obvious then that we have the relationship

$$H(z) = \sum_k h_k z^k \quad (1.14)$$

This last equation relates the system function $H(z)$ to the unit response h_k . Recall that our linear time-invariant system was completely* specified by knowledge of $H(z)$, since we could then determine the output for any of our eigenfunctions; similarly, knowledge of the unit response also completely* determines the operation of our linear time-invariant system. Thus it is no surprise that some explicit relationship must exist between the two, and, of course, this is given in Eq. (1.14).

Finally, we are in a position to answer the question—why transforms? The key lies in the expression (1.14), which is, itself, a transform (in this case a z -transform), which converts† the time function h_k into a function of a complex variable $H(z)$. This transform arose naturally in our study of linear time-invariant systems and was not introduced into the analysis in an artificial way. We shall see later that a similar relationship exists for continuous-time systems, as well, and this gives rise to the Laplace transform. Recalling that continuous-time systems may be described by constant-coefficient linear differential equations and that the use of transforms greatly simplifies the solution of these equations, we are not surprised that discrete-time systems lead to sets of constant-coefficient linear *difference* equations whose solution is simplified by the use of z -transforms. Lastly, we comment that the inputs^f

* Completely specified in the sense that the only additional required information is the initial state of the system (e.g., the initial conditions of all the energy storage elements). Usually, the system is assumed to be in the zero-energy state, in which case we truly have a complete specification.

† Transforms not only change the form in which the information describing a given function is presented, but they also present this information in a simplified form which is convenient for mathematical manipulation.

and the outputs g are easily decomposed into weighted sums of complex exponentials by means of transforms, and of course, once this is done, then results such as (1.5) or (1.10) immediately give us the component-by-component output of our system for each of these inputs; the total output is then formed by summing the output components as in Eq. (1.13).

The fact that these transforms arise naturally in our system studies is really only a partial answer to our basic question regarding their use in analysis. The other and more pragmatic reason is that they greatly simplify the analysis itself; most often, in fact, the analysis can only proceed with the use of transforms leading us to a partial solution from which properties of the system behavior may be derived.

The remainder of this appendix is devoted to giving examples and properties of these two principal transforms which are so useful in queueing theory.

1.2. THE z-TRANSFORM [JURY 64, CADZ.73]

Let us consider a function of discrete time f_n which takes on nonzero values only for the nonnegative integers, that is, for $n = 0, 1, 2, \dots$ (i.e., for convenience we assume that $f_n = 0$ for $n < 0$). We now wish to compress this semi-infinite sequence into a single function in a way such that we can expand the compressed form back into the original sequence when we so desire. In order to do this, we must place a "tag" on each of the terms in the sequence. We choose to tag the term f_n , by multiplying it by z^n ; since n is then unique for each term in the sequence, each tag is also unique. z will be chosen as some complex variable whose permitted range of values will be discussed shortly. Once we tag each term, we may then sum over all tagged terms to form our compressed function, which represents the original sequence. Thus we define the z -transform (also known as the generating function or geometric transform) for f_n as follows:

$$F(z) \triangleq \sum_{n=0}^{\infty} f_n z^n \quad (1.15)$$

$F(z)$ is clearly only a function of our complex variable z since we have summed over the index n ; the notation we adopt for the z -transform is to use a capital letter that corresponds to the lower-case letter describing the sequence, as in Eq. (1.15). We recognize that Eq. (1.14) is, of course, in exactly this form. The z -transform for a sequence will exist so long as the terms in that sequence grow no faster than geometrically, that is, so long as there is some $a > 0$ such that

$$\lim_{n \rightarrow \infty} \frac{|f_n|}{a^n} = 0$$

Furthermore, given a sequence f_n its z -transform $F(z)$ is unique.

If the sum over all terms in the sequence f_n is finite, then certainly the unit disk $|z| \leq 1$ represents a range of analyticity for $F(z)$.^{*} In such a case we have

$$F(l) = \sum_{n=0}^{\infty} f_n \quad (\text{I.16})$$

We now consider some important *examples* of z -transforms. It is convenient to denote the relationship between a sequence and its transform by means of a double-barred, double-headed arrow; thus Eq. (US) may be written as

$$f_n \Leftrightarrow F(z) \quad (\text{I.17})$$

For our first example, let us consider the unit function as defined in Eq. (1.11). For this function and from the definition given in Eq. (1.15) we see that exactly one term in the infinite summation is nonzero, and so we immediately have the transform pair

$$(\text{U8})$$

For a related example, let us consider the unit function shifted to the right by k units, that is,

$$u_{n-k} = \begin{cases} 1 & n=k \\ 0 & n \neq k \end{cases}$$

From Eq. (US) again, exactly one term will be nonzero, giving

$$u_{n-k} \Leftrightarrow z^k$$

As a third example, let us consider the *unit step function* defined by

$$\text{for } n = 0, 1, 2, \dots$$

(recall that all functions are zero for $n < 0$). In this case we have a geometric series, that is,

$$\delta_n \Leftrightarrow \sum_{n=0}^{\infty} 1e^n = \frac{1}{1-z} \quad (\text{I.19})$$

We note in this case that $|z| < 1$ in order for the z -transform to exist. An extremely important sequence often encountered is the geometric series

$$f_n = A\alpha^n \quad n=0, 1, 2, \dots$$

^{*} A function of a complex variable is said to be analytic at a point in the complex plane if that function has a unique derivative at that point. The Cauchy-Riemann necessary and sufficient condition for analyticity of such functions may be found in any text on functions of a complex variable [AHLF 66].

[†] The double bar denotes the transform relationship whereas the double heads on the arrow indicate that the journey may be made in either direction, $f \Rightarrow F$ and $F \Rightarrow f$.

Its z-transform may be calculated as

$$\begin{aligned} F(z) &= \sum_{n=0}^{\infty} A\alpha^n z^n \\ &= A \sum_{n=0}^{\infty} (\alpha z)^n \\ &= \frac{A}{1 - \alpha z} \end{aligned}$$

And so

$$A\alpha^n \longleftrightarrow \frac{A}{1 - \alpha z} \quad (1.20)$$

where, of course, the region of analyticity for this function is $|z| < 1/|\alpha|$; note that α may be greater or less than unity.

Linear transformations such as the z-transform enjoy a number of important *properties*. Many of these are listed in Table 1.1. However, it is instructive for us to derive the *convolution* property which is most important in queueing systems. Let us consider two functions of discrete time f_n and g_n , which may take on nonzero values only for the nonnegative integers. Their respective z-transforms are, of course, $F(z)$ and $G(z)$. Let $\otimes$ denote the convolution operator, which is defined for f_n and g_n as follows:

$$f_n \otimes g_n \triangleq \sum_{k=0}^n f_{n-k} g_k$$

We are interested in deriving the z-transform of the convolution for f_n and g_n , and this we do as follows:

$$\begin{aligned} f_n \otimes g_n &\longleftrightarrow \sum_{n=0}^{\infty} (f_n \otimes g_n) z^n \\ &= \sum_{n=0}^{\infty} \sum_{k=0}^n f_{n-k} g_k z^{n-k+k} \end{aligned}$$

However, since

$$\sum_{n=0}^{\infty} \sum_{k=0}^n = \sum_{k=0}^{\infty} \sum_{n=k}^{\infty}$$

we have

$$\begin{aligned} f_n \otimes g_n &\longleftrightarrow \sum_{k=0}^{\infty} g_k z^k \sum_{n=k}^{\infty} f_{n-k} z^{n-k} \\ &= \left(\sum_{k=0}^{\infty} g_k z^k \right) \left(\sum_{m=0}^{\infty} f_m z^m \right) \\ &= G(z)F(z) \end{aligned}$$

Table 1.1
Some Properties of the z-Transform

SEQUENCE	z-TRANSFORM
1. $f_n \quad n = 0, 1, 2, \dots$	$F(z) = \sum_{n=0}^{\infty} f_n z^n$
2. $af_n + bg_n$	$aF(z) + bG(z)$
3. $a^n f_n$	$F(az)$
4. $f_{nk} \quad n = 0, k, 2k, \dots$	$F(z^k)$
5. f_{n+1}	$\frac{1}{z} [F(z) - f_0]$
6. $f_{n+k} \quad k > 0$	$\frac{F(z)}{z^k} - \sum_{i=1}^k z^{i-k-1} f_{i-1}$
7. f_{n-1}	$zF(z)$
8. $f_{n-k} \quad k > 0$	$z^k F(z)$
9. nf_n	$z \frac{d}{dz} F(z)$
10. $n(n-1)(n-2), \dots, (n-m+1)f_n$	$z^m \frac{d^m}{dz^m} F(z)$
11. $f_n \otimes g_n$	$F(z)G(z)$
12. $f_n - f_{n-1}$	$(1-z)F(z)$
13. $\sum_{k=0}^n f_k \quad n = 0, 1, 2, \dots$	$\frac{F(z)}{1-z}$
14. $\frac{\partial}{\partial a} [\llcorner \quad (a \text{ is a parameter of } n)$	$\frac{\partial}{\partial a} F(z)$
15. Series sum property	$F(1) = \sum_{n=0}^{\infty} f_n$
16. Alternating sum property	$F(-1) = \sum_{n=0}^{\infty} (-1)^n f_n$
17. Initial value theorem	$F(0) = f_0$
18. Intermediate value theorem	$\frac{1}{n!} \left. \frac{d^n F(z)}{dz^n} \right _{z=0} = f_n$
19. Final value theorem	$\lim_{z \rightarrow 1} (1-z)F(z) = f_{\infty}$

2
Transform Pairs

SEQUENCE	Z-TRANSFORM
$n = 0, 1, 2, \dots$	$F(z) = \sum_{n=0}^{\infty} I_n z^n$
$\begin{cases} 1 & n = 0 \\ 0 & n \neq 0 \end{cases}$	z^k
$1/n! \quad n = 0, 1, 2, \dots$	$1 - z$
	z^k
	$1 - z$
	A
	$1 - \alpha z$
	$\frac{\alpha z}{(1 - \alpha z)^2}$
	$\frac{z}{(1 - z)^2}$
	$\frac{\alpha z(1 + \alpha z)}{(1 - \alpha z)^3}$
	$\frac{z(1 + z)}{(1 - z)^3}$
$1/n! \alpha^n$	$\frac{1}{(1 - \alpha z)^2}$
$1/n!$	$\frac{1}{(1 - z)^2}$
$1/(n+m)!(n+m-1) \dots (n+1)\alpha^n$	$\frac{1}{(1 - \alpha z)^{m+1}}$
	e^z

len, we have that the a-transform of the convolution of two equal to the product of the z-transform of each of the sequences

.1 we list a number of important properties of the z-transform, ,g that in Table 1.2 we provide a list of important common

"he
rm
"he
ion
the
itly
vay
sed

her
) is
nial
uor
wer
hen

' to
ress
ins-
.ing
LOW
for
h is
is a
5 in
rms
I to
out
.eed

.ach

UIL

Some comments regarding these tables are in order. First, in the property table we note that Property 2 is a statement of linearity, and Properties 3 and 4 are statements regarding scale change in the transform and time domain, respectively. Properties 5-8 regard translation in time and are most useful. In particular, note from Property 7 that the unit delay (delay by one unit of time) results in multiplication of the transform by the factor z whereas Property 5 states that a unit advance involves division by the factor z , Properties 9 and 10 show multiplication of the sequence by terms of the form $n(n-1)\dots(n-m)$. Combinations of these may be used in order to find, for example, the transform of $n^2 f_n$; this may be done by recognizing that $n^2 = n(n-1) + n$, and so the transform of $n^2 f_n$ is merely $z^2 d^2 F(z)/dz^2 + z dF(z)/dz$. This shows the simple differentiation technique of obtaining more complex transforms. Perhaps the most important, however, is Property 11 showing that the convolution of two time sequences has a transform that is the product of the transform of each time sequence separately. Properties 12 and 13 refer to the difference and summation of various terms in the sequence. Property 14 shows if a is an independent parameter of f_n differentiating the sequence with respect to this parameter is equivalent to differentiating the transform. Property 15 is also important and shows that the transform expression may be evaluated at $z = 1$ directly to give the sum of all terms in the sequence. Property 16 merely shows how to calculate the alternating sum. From the definition of the z -transform, the initial value theorem given in Property 17 is obvious and shows how to calculate the initial term of the sequence directly from the transform. Property 18, on the other hand, shows how to calculate any term in the original sequence directly from its z -transform by successive differentiation; this then corresponds to one method for calculating the sequence given its transform. It can be seen from Property 18 that the sequence f_n forms the coefficients in the Taylor-series expansion of $F(z)$ about the point 0 . Since this power-series expansion is unique, then it is clear that the inversion process is also unique. Property 19 gives a direct method for calculating the final value of a sequence from its z -transform.

Table 1.2 lists some useful transform pairs. This table can be extended considerably by making use of the properties listed in Table 1.1; in some cases this has already been done. For example, Pair 5 is derived from Pair 4 by use of the delay theorem given as entry 8 in Table 1.1. One of the more useful relationships is given in Pair 6 considered earlier.

Thus we see the effect of compressing a time sequence f_n into a single function of the complex variable z . Recall that the use of the variable z was to tag the terms in the sequence f_n so that they could be recovered from the compressed function; that is, f_n was tagged with the factor z^n . We have

see!
pre
 $F(z)$
firs
 $F(z)$
sec
ex
wr
as

(
t
e

seen how to form the z-transform of the sequence [through Eq. (U5)]. The problem confronting us now is to find the sequence f_n given the z-transform $F(z)$. There are basically three methods for carrying out this inversion. The *first* is the *power-series method*, which attempts to take the given function $F(z)$ and express it as a power series in z ; **once** this is done the terms in the sequence f_n may be picked off by inspection since the tagging is now explicitly exposed. The power series may be obtained in one of two ways: the first way we have already seen through our intermediate value theorem expressed as Item 18 in Table I.1, that is,

$$f_n = \frac{1}{n!} \left. \frac{d^n F(z)}{dz^n} \right|_{z=0}$$

(this method is useful if one is only interested in a few terms but is rather tedious if many terms are required); the second way is useful if $F(z)$ is expressible as a rational function of z (that is, as the ratio of a polynomial in z over a polynomial in z) and in this case one may divide the denominator into the numerator to pick off the sequence of leading terms in the power series directly. The power-series expansion method is usually difficult when many terms are required.

The *second* and most useful method for inverting z-transforms [that is, to calculate j' from $F(z)$] is the *inspection method*. That is, one attempts to express $F(z)$ in a fashion such that it consists of terms that are recognizable as transform pairs, for example, from Table I.2. The standard approach for placing $F(z)$ in this form is to carry out a *partial-fraction expansion*,* which we now discuss. The partial-fraction expansion is merely an algebraic technique for expressing rational functions of z as sums of simple terms, each of which is easily inverted. **In** particular, we will attempt to express a rational $F(z)$ as a sum of terms, each of which looks either like a simple pole (see entry 6 in Table I.2) or as a multiple pole (see entry 13). Since the sum of the transforms equals the transform of the sum we may apply Property 2 from Table I.1 to invert each of these now recognizable forms separately, thereby carrying out the required inversion. To carry out the partial-fraction expansion we proceed as follows. We assume that $F(z)$ is in rational form, that is

$$F(z) = \frac{N(z)}{D(z)}$$

where both the numerator $N(z)$ and the denominator $D(z)$ are each

* This procedure is related to the Laurent expansion of $F(z)$ around each pole [GUIL 49].

polynomials in z .^{*} Furthermore we will assume that $D(z)$ is already in factored form, that is,

$$D(z) = \prod_{i=1}^k (1 - \alpha_i z)^{m_i} \quad (1.21)$$

The product notation used in this last equation is defined as

$$\prod_{i=1}^k a_i \triangleq a_1 a_2 \cdots a_k$$

Equation (1.21) implies that the i th root at $z = 1/\alpha_i$ occurs with multiplicity m_i . [We note here that in most problems of interest, the difficult part of the solution is to take an arbitrary polynomial such as $D(z)$ and to find its roots α_i so that it may be put in the factored form given in Eq. (1.21). At this point we assume that that difficult task has been accomplished.] If $F(z)$ is in this form then it is possible to express it as follows [GUIL 49]:

$$\begin{aligned} F(z) = & \frac{A_{11}}{(1 - \alpha_1 z)^{m_1}} + \frac{A_{12}}{(1 - \alpha_1 z)^{m_1-1}} + \cdots + \frac{A_{1m_1}}{(1 - \alpha_1 z)} \\ & + \frac{A_{21}}{(1 - \alpha_2 z)^{m_2}} + \frac{A_{22}}{(1 - \alpha_2 z)^{m_2-1}} + \cdots + \frac{A_{2m_2}}{(1 - \alpha_2 z)} + \cdots \\ & + \frac{A_{k1}}{(1 - \alpha_k z)^{m_k}} + \frac{A_{k2}}{(1 - \alpha_k z)^{m_k-1}} + \cdots + \frac{A_{km_k}}{(1 - \alpha_k z)} \end{aligned} \quad (1.22)$$

This last form is exactly what we were looking for, since each term in this sum may be found in our table of transform pairs; in particular it is Pair 13 (and in the simplest case it is Pair 6). Thus if we succeed in carrying out the partial-fraction expansion, then by inspection we have our time sequence. It remains now to describe the method for calculating the coefficients A_{ij} . The general expression for such a term is given by

$$A_{ij} = \frac{1}{(j-1)!} \left(-\frac{1}{\alpha_i} \right)^{j-1} \frac{d^{j-1}}{dz^{j-1}} \left[(1 - \alpha_i z)^{m_i} \frac{N(z)}{D(z)} \right] \bigg|_{z=1/\alpha_i} \quad (1.23)$$

This rather formidable procedure is, in fact, rather straightforward as long as the function $F(z)$ is not terribly complex.

* We note here that a partial-fraction expansion may be carried out only if the degree of the numerator polynomial is strictly less than the degree of the denominator polynomial: if this is **not** the case, then it is necessary to divide the denominator into the numerator until the remainder is of lower degree than the denominator. This remainder divided by the original denominator may then be expanded in partial fractions by the method shown; the terms generated from the division also may be inverted by inspection making use of transform pair 3 in Table 1.2. An alternative way of satisfying the degree condition is to attempt to factor out enough powers of z from the numerator if possible.

worthwhile at this point to carry out an example in order to demonstrate the method. Let us assume $F(z)$ is given by

$$F(z) = \frac{4z^2(1 - 8z)}{(1 - 4z)(1 - 2z)^2} \quad (1.24)$$

; example the numerator and denominator both have the same degree) it is necessary to bring the expression into proper form (numerator less than denominator degree). In this case our task is simple since we can factor out two powers of z (we are required to factor out only one power of z in order to bring the numerator degree below that of the denominator). It is obviously in this case we may as well factor out both and simplify the calculations). Thus we have

$$F(z) = z \left[\frac{4(1 - 8z)}{(1 - 4z)(1 - 2z)^2} \right]$$

; define the term in square brackets as $G(z)$. We note in this example the denominator has three poles: one at $z = 1/4$; and two (that is a double pole) at $z = 1/2$. Thus in terms of the variables defined in Eq. (1.21) we have $k = 2$, $\alpha_1 = 4$, $\alpha_{11} = 1$, $\alpha_2 = 2$, $\alpha_{21} = 2$. From Eq. (1.22) we are now seeking the following expansion:

$$\begin{aligned} G(z) &\triangleq \frac{4(1 - 8z)}{(1 - 4z)(1 - 2z)^2} \\ &= \frac{A_{11}}{1 - 4z} + \frac{A_{21}}{(1 - 2z)} + \frac{A_{22}}{(1 - 2z)^2} \end{aligned}$$

such as A_{11} (that is, coefficients of simple poles) are easily obtained from Eq. (1.23) by multiplying the original function by the factor corresponding to the pole and then evaluating the result at the pole itself (that is, when z is on a value that drives the factor to 0). Thus in our example we have

$$A_{11} = (1 - 4z)G(z) \Big|_{z=1/4} = \frac{4[1 - (8/4)]}{[1 - (2/4)]^2} = -16$$

and A_{21} may be evaluated in a similar way from Eq. (1.23) as follows:

$$A_{21} = (1 - 2z)^2 G(z) \Big|_{z=1/2} = \frac{4[1 - (8/2)]}{[1 - (4/2)]} = 12$$

Finally, in order to evaluate A_{22} we must apply the differentiation formula given Eq. (1.23) once, that is,

$$\begin{aligned}
 A_{22} &= -\frac{1}{2} \frac{d}{dz} [(1 - 2z)2G(z)] \Big|_{z=1/2} \\
 &= -\frac{1}{2} \frac{d}{dz} \frac{4(1 - 8z)}{(1 - 4z)} \Big|_{z=1/2} \\
 &= -\frac{1}{2} \frac{(1 - 4z)(-32) - 4(1 - 8z)(-4)}{(1 - 4z)^2} \Big|_{z=1/2} \\
 &= 8
 \end{aligned}$$

Thus we conclude that

$$G(z) = \frac{-16}{1 - 4z} + \frac{12}{(1 - 2z)^2} + \frac{8}{1 - 2z}$$

This is easily shown to be equal to the original factored form of $G(z)$ by placing these terms over a common denominator. Our next step is to invert $G(z)$ by inspection. This we do by observing that the first and third terms are of the form given by transform pair 6 in Table 1.2 and that the second term is given by transform pair 13. This, coupled with the linearity property 2 in Table 1.1 gives immediately that

$$G(z) \Leftrightarrow g^n = \begin{cases} 0 & n < 0 \\ \{-16(4)^n + 12(n+1)(2)^n + 8(2)^n\} & n = 0, 1, 2, \dots \end{cases} \quad (1.25)$$

Of course, we must now account for the factor z^2 to give the expression for f_n . As mentioned above we do this by taking advantage of Property 8 in Table 1.1 and so we have (for $n = 2, 3, \dots$)

$$f_n = -16(4)^{n-2} + 12(n-1)(2)^{n-2} + 8(2)^{n-2}$$

and so

$$f_n = \begin{cases} 0 & n < 2 \\ \{(3/16)(2)^n - 4^n\} & n = 2, 3, 4, \dots \end{cases}$$

This completes our example.

The *third* method for carrying out the inversion process is to use the *inversion formula*. This involves evaluating the following integral:

$$f_n = \frac{-1}{2\pi j} \oint_C F(z)z^{-1-n} dz \quad (1.26)$$

where $j = \sqrt{-1}$ and the integral is evaluated in the complex z -plane around a closed circular contour C , which is large enough* to surround all poles of $F(z)$. This method of evaluation works properly when factors of the form z^k are removed from the expression; the reduced expression is then evaluated and the final solution is obtained by taking advantage of Property 9 in Table I.1 as we shall see below. This contour integration is most easily performed by making use of the Cauchy residue theorem [GUIL 49]. This theorem may be stated as follows:

Cauchy Residue Theorem The integral of $g(z)$ over a closed contour C containing *within* it only isolated singular points of $g(z)$ is equal to $2\pi j$ times the sum of the residues at these points, whenever $g(z)$ is analytic *outside* and *within* the closed contour C .

An isolated singular point of an analytic function is a singular point whose neighborhood contains no other singular points; a simple pole (i.e., a pole of order one—see below) is the classical example. If $z = a$ is an isolated singular point of $g(z)$ and if $y(z) = (z - a)^m g(z)$ is analytic at $z = a$ and $y(a) \neq 0$, then $g(z)$ is said to have a *pole of order m* at $z = a$ with the *residue r_a* given by

$$r_a = \frac{1}{(m-1)!} \frac{d^{m-1}}{dz^{m-1}} (z-a)^m g(z) \Big|_{z=a} \quad (1.27)$$

We note that the residue given in this last equation is almost the same as A_i given in Eq. (1.23), the main difference being the form in which we write the pole. Thus we have now defined all that we need to apply the Cauchy residue theorem in order to evaluate the integral in Eq. (1.26) and thereby to recover the time function $f(t)$ from our z -transform. By way of illustration we carry out the calculation of our previous example given in Eq. (1.24). Making use of Eq. (1.26) we have

$$g(t) = \frac{1}{2\pi j} \oint_C \frac{4z^{-1} \cdot (1-8z)}{(1-4z)(1-2z)^2} dz$$

* Since Jordan's lemma (see p. 353) requires that $F(z) \rightarrow 0$ as $z \rightarrow \infty$ if we are to let the contour grow, then we require that any function $F(z)$ that we consider have this property; thus for rational functions of z if the numerator degree is not less than the denominator degree, then we must divide the numerator by the denominator until the remainder is of lower degree than the denominator, as we have seen earlier. The terms generated by this division are easily transformed by inspection, as discussed earlier, and it is only the remaining function which we now consider in this inversion method for the z -transform.

† A function $F(z)$ of a complex variable z is said to be analytic in a region of the complex plane if it is single-valued and differentiable at every point in that region.

where C is a circle large enough to enclose the poles of $F(z)$ at $z = 1/4$ and $z = 1/2$. Using the residue theorem and Eq. (1.27) we find that the residue at $z = 1/4$ is given by

$$\begin{aligned} r_{1/4} &= \left(z - \frac{1}{4} \right) \frac{4z - \frac{1}{2}n(1 - 8z)}{(1 - 4z)(1 - 2z)^2} \Big|_{z=1/4} \\ &= \frac{-(1/4) - \frac{1}{2}n(1 - (8/4))}{(1 - (2/4))^2} = 16(4)^n \end{aligned}$$

whereas the residue at $z = 1/2$ is calculated as

$$\begin{aligned} r_{1/2} &= \frac{d}{dz} \left[\left(z - \frac{1}{2} \right)^2 \frac{4z - \frac{1}{2}n(1 - 8z)}{(1 - 4z)(1 - 2z)^2} \right] \Big|_{z=1/2} \\ &= \frac{d}{dz} \left[\frac{z - \frac{1}{2}n(1 - 8z)}{1 - 4z} \right] \Big|_{z=1/2} \\ &= \frac{(1 - 4z)[(-1 - n)z - \frac{n(1 - 8z)}{2} + \frac{z^{1-n}(-8)}{(1 - 4z)^2}] - (z - \frac{1}{2}n(1 - 8z))(-4)}{(1 - 4z)^2} \Big|_{z=1/2} \\ &= (-1) \left[(-1 - n) \left(\frac{1}{2} \right)^{-2-n} (-3) + \left(\frac{1}{2} \right)^{-1-n} (-8) \right] - \left(\frac{1}{2} \right)^{-1-n} (-3)(-4) \\ &= -12(n + 1)2^n + 16(2)^n - 24(2)^n \end{aligned}$$

Now we must take $2\pi j$ times the sum of the residues and then multiply by the factor preceding the integral in Eq. (1.26) (thus we must take -1 times the sum of the residues) to yield

$$g_n = -16(4)^n + 12(n + 1)2^n + 8(2)^n \quad n = 0, 1, 2, \dots$$

But this last is exactly equal to the form for g_n in Eq. (1.25) found by the method of partial-fraction expansions. From here the solution proceeds as in that method, thus confirming the consistency of these two approaches.

Thus we have reviewed some of the techniques for applying and inverting the z -transform in the handling of discrete-time functions. The application of these methods in the solution of difference equations is carefully described in Sect. 1.4 below.

1.3. THE LAPLACE TRANSFORM [WIDD 46]

The Laplace transform, defined below, enjoys many of the same properties as the z -transform. As a result, the following discussion very closely parallels that given in the previous section.

We now consider functions of continuous time $f(t)$, which take on non-zero values only for nonnegative values of the continuous parameter t . [Again for

convenience we are assuming that $f(t) = 0$ for $t < 0$. For the more general case, most of these techniques apply as discussed in the paragraph containing Eq. (1.38) below.] As with discrete-time functions, we wish to take our continuous-time function and transform it from a function of t to a function of a new complex variable (say, s). At the same time we would like to be able to "untransform" back into the t domain, and in order to do this it is clear we must somehow "tag" $f(t)$ at each value of t . For reasons related to those described in Section 1.1 the tag we choose to use is e^{-st} . The complex variable s may be written in terms of its real and complex parts as $s = \sigma + j\omega$ where, again, $j = \sqrt{-1}$. Having multiplied by this tag, we then integrate over all nonzero values in order to obtain our transform function defined as follows:

$$F^*(s) \triangleq \int_{-\infty}^{\infty} f(t) e^{-st} dt \quad (1.28)$$

Again, we have adopted the notation for general Laplace transforms in which we use a capital letter for the transform of a function of time, which is described in terms of a lower case letter. This is usually referred to as the "two-sided," or "bilateral" Laplace transform since it operates on both the negative and positive time axes. We have assumed that $f(t) = 0$ for $t < 0$, and in this case the lower limit of integration may be replaced by 0^- , which is defined as the limit of $0 - \epsilon$ as $\epsilon (> 0)$ goes to zero; further, we often denote this lower limit merely by 0 with the understanding that it is meant as 0^- (usually this will cause no confusion). There also exists what is known as the "one-sided" Laplace transform in which the lower limit is replaced by 0^+ , which is defined as the limit of $0 + \epsilon$ as $\epsilon (> 0)$ goes to zero; this one-sided transform has application in the solution of transient problems in linear systems. It is important that the reader distinguish between these two transforms with zero as their lower limit since in the former case (the bilateral transform) any accumulation at the origin (as, for example, the unit impulse defined below) will be included in the transform, whereas in the latter case (the one-sided transform) it will be omitted.

For our assumed case in which $f(t) = 0$ for $t < 0$ we may write our transform as

$$F^*(s) = \int_0^{\infty} f(t) e^{-st} dt \quad (1.29)$$

where, we repeat, the lower limit is to be interpreted as 0^- . This Laplace transform will exist so long as $f(t)$ grows no faster than an exponential, that is, so long as there is some real number σ_a such that

$$\lim_{\tau \rightarrow \infty} \int_0^{\tau} |f(t)| e^{\sigma_a t} dt < \infty$$

The smallest possible value for " σ " is referred to as the abscissa of absolute convergence. Again we state that the Laplace transform $F^*(s)$ for a given function $f(t)$ is unique.

If the integral of $f(t)$ is finite, then certainly the right-half plane $\text{Re}(s) \geq 0$ represents a region of analyticity for $F^*(s)$; the notation $\text{Re}(\)$ reads as "the real part of the complex function within the parentheses." In such a case we have, corresponding to Eq. (1.16),

$$F^*(0) = \int_0^{\infty} f(t) dt \quad (1.30)$$

From our earlier definition in Eq. (1.9) we see that properties for the z-transform when $z = 1$ will correspond to properties for the Laplace transform when $s = 0$ as, for example, in Eqs. (1.16) and (1.30).

Let us now consider some important examples of Laplace transforms. We use notation here identical to that used in Eq. (1.17) for z-transforms, namely, we use a double-barred, double-headed arrow to denote the relationship between a function and its transform; thus, Eq. (1.29) may be written as

$$f(t) \Leftrightarrow F^*(s) \quad (1.31)$$

The use of the double arrow is a statement of the uniqueness of the transform as earlier.

As in the case of z-transforms, the most useful method for finding the inverse [that is, calculating $f(t)$ from $F^*(s)$] is the *inspection method*, namely, looking up the inverse in a table. Let us, therefore, concentrate on the calculation of some Laplace transform pairs. By far the most important Laplace transform pair to consider is for the one-sided exponential function, namely,

$$f(t) = \begin{cases} Ae^{-at}, & t \geq 0 \\ 0, & t < 0 \end{cases}$$

Let us carry out the computation of this transform, as follows:

$$\begin{aligned} f(t) \Leftrightarrow F^*(s) &= \int_0^{\infty} Ae^{-at} e^{-st} dt \\ &= A \int_0^{\infty} e^{-(s+a)t} dt \\ &= A \left[\frac{e^{-(s+a)t}}{-(s+a)} \right]_0^{\infty} \\ &= \frac{A}{s+a} \end{aligned}$$

And so we have the fundamental relationship

$$Ae^{-at} \delta(t) \Leftrightarrow \frac{A}{s+a} \quad (1.32)$$

where we have defined the unit step function in continuous time as

$$\delta(t) = \begin{cases} 1 & t \geq 0 \\ 0 & t < 0 \end{cases} \quad (1.33)$$

In fact, we observe that the unit step function is a special case of our one-sided exponential function when $A = 1$, $a = 0$, and so we have immediately the additional pair

$$\delta(t) \leftrightarrow \frac{1}{s} \quad (1.34)$$

We note that the transform in Eq. (1.32) has an abscissa of convergence $\sigma_a = -G$.

Thus we have calculated analogous z-transform and Laplace-transform pairs: the geometric series given in Eq. (1.20) with the exponential function given in Eq. (1.32) and also the unit step function in Eqs. (1.19) and (1.34), respectively. It remains to find the continuous analog of the unit function defined in Eq. (1.11) and whose z-transform is given in Eq. (1.18). This brings us face to face with the *unit impulse function*. The unit impulse function plays an important part in transform theory, linear system theory, as well as in probability and queueing theory. It therefore behoves us to learn to work with this function. Let us adopt the following notation

$$u\delta(t) \triangleq \text{unit impulse function occurring at } t = 0$$

$u\delta(t)$ corresponds to highly concentrated unit-area pulses that are of such short duration that they cannot be distinguished by available measurement instruments from other perhaps briefer pulses. Therefore, as one might expect, the exact shape of the pulse is unimportant, rather only its time of occurrence and its area matter. This function has been studied and utilized by scientists for many years [GUIL 49], among them Dirac, and so the unit impulse function is often referred to as the *Dirac delta function*. For a long time pure mathematicians have refrained from using $u\delta(t)$ since it is a highly improper function, but years ago Schwartz's theory of distributions [SCHW 59] put the concept of a unit impulse function on firm mathematical ground. Part of the difficulty lies with the fact that the unit impulse function is not a function at all, but merely provides a notational way for handling discontinuities and their derivatives. In this regard we will introduce the unit impulse as the limit of a sequence without appealing to the more sophisticated generalized functions that place much of what we do in a more rigorous framework.

As we mentioned earlier, the exact shape of the pulse is unimportant. Let us therefore choose the following representative pulse shape for our discussion of impulses:

$$f_{\alpha}(t) = \begin{cases} \alpha & |t| \leq \frac{1}{2\alpha} \\ 0 & |t| > \frac{1}{2\alpha} \end{cases}$$

This rectangular wave form has a height and **width** dependent upon the parameter α as shown in Figure 1.2. Note that this function has a constant area of unity (hence the name *unit* impulse function). As α increases, we note that the pulse gets taller and narrower. The limit of this sequence as $\alpha \rightarrow \infty$ (or the limit of any one of an infinite number of other sequences with similar properties, i.e., increasing height, decreasing width, unit area) is what we mean by the unit impulse "function." Thus we are led to the following description of the unit impulse function.

$$u_0(t) = \begin{cases} \infty & t = 0 \\ 0 & t \neq 0 \end{cases}$$

$$\int_{-\infty}^{\infty} u_0(t) dt = 1$$

This function is represented graphically by a vertical arrow located at the instant of the impulse and with a number adjacent to the head of the arrow

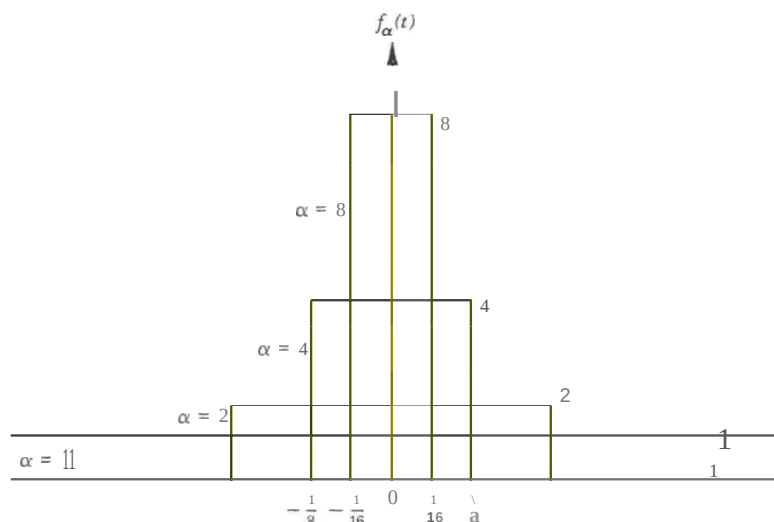


Figure 1.2 A sequence of functions whose limit is the unit impulse function $u_0(t)$.

0

Figure 1.3 Graphical representation of $Au_0(t - a)$.

indicating the area of the impulse; that is, A times a unit impulse function located at the point $t = a$ is denoted as $Au_0(t - a)$ and is depicted as in Figure 1.3.

Let us now consider the integral of the unit impulse function. It is clear that if we integrate from $-\infty$ to a point t where $t < 0$ then the total integral must be 0 whereas if $t > 0$ then we will have successfully integrated past the unit impulse and thereby will have accumulated a total area of unity. Thus we conclude

$$\int_{-\infty}^t u_0(x) dx = \begin{cases} 1 & t \geq 0 \\ 0 & t < 0 \end{cases}$$

But we note immediately that the right-hand side is the same as the definition of the unit step function given in Eq. (1.33). Therefore, we conclude that the unit step function is the integral of the unit impulse function, and so the "derivative" of the unit step function must therefore be a unit impulse function. However, we recognize that the derivative of this discontinuous function (the step function) is not properly defined; once again we appeal to the theory of distributions to place this operation on a firm mathematical foundation. We will therefore assume this is a proper operation and proceed to use the unit impulse function as if it were an ordinary function.

One of the very important properties of the unit impulse function is its *sifting* property; that is, for an arbitrary differentiable function $g(t)$ we have

$$\int_{-\infty}^{\infty} u_0(t - x)g(x) dx = g(t)$$

This last equation merely says that the integral of the product of our function $g(x)$ with an impulse located at $x = t$ "sifts" the function $g(x)$ to produce its value at t , $g(t)$. We note that it is possible also to define the derivative of the unit impulse which we denote by $U_1(t) = dU_0(t)/dt$; this is known as the *unit doublet* and has the property that it is everywhere 0 except in the vicinity of the origin where it runs off to ∞ just to the left of the origin and off to $-\infty$

just to the right of the origin, and, in addition, has a total area equal to zero. Such functions correspond to electrostatic dipoles, for example, used in physics. In fact, an impulse function may be likened to the force placed on a piece of paper when it is laid over the edge of a knife and pressed down whereas a unit doublet is similar to the force the paper experiences when cut with scissors. Higher-order derivatives are possible and in general we may have $u_n(t) = dU_{n-1}(t)/dt$. In fact, as we have seen, we may also go back down the sequence by integrating these functions as, for example, by generating the unit step function as the integral of the unit impulse function; the obvious notation for the unit step function, therefore, would be $U_I(t)$ and so we may write $u_0(t) = dU_I(t)/dt$. [Note, from Eq. (1.33), that we have also reserved the notation $\delta(t)$ to represent the unit step function.] Thus we have defined an infinite sequence of specialized functions beginning with the unit impulse and proceeding to higher-order derivatives such as the doublet, and so on, as well as integrating the unit impulse and thereby generating the unit step function, the ramp, namely,

$$u_{-2}(t) \triangleq \int_{-\infty}^t u_{-1}(x) dx = \begin{cases} t & t \geq 0 \\ 0 & t < 0 \end{cases}$$

the parabola, namely,

$$u_{-3}(t) \triangleq \int_{-\infty}^t u_{-2}(x) dx = \begin{cases} \frac{t^2}{2} & t \geq 0 \\ 0 & t < 0 \end{cases}$$

and in general

$$u_{-n}(t) \triangleq \int_{-\infty}^t u_{-(n-1)}(x) dx = \begin{cases} \frac{t^{n-1}}{(n-1)!} & t \geq 0 \\ 0 & t < 0 \end{cases} \quad (1.35)$$

This entire family is called the *family of singularity functions*, and the most important members are the unit step function and the unit impulse function.

Let us now return to our main discussion and consider the Laplace transform of $u_0(t)$. We proceed directly from Eq. (1.28) to obtain

$$u_0(t) \Leftrightarrow \int_0^{\infty} u_0(t)e^{-st} dt = \frac{1}{s}$$

(Note that the lower limit is interpreted as 0-.) Thus we see that the unit impulse has a Laplace transform equal to the constant unity.

Let us now consider some of the important properties of the transformation. As with the z-transform, the convolution property is the most important and we proceed to derive it here in the continuous time case. Thus, consider two functions of continuous time $f(t)$ and $g(t)$, which take on nonzero values

only for $t \geq 0$; we denote their Laplace transforms by $F^*(s)$ and $G^*(s)$, respectively. Defining $\otimes$ once again as the convolution operator, that is,

$$J(t) \otimes g(t) \triangleq \int_{-\infty}^{\infty} f(t-x)g(x) dx \quad (1.36)$$

which in our case reduces to

$$f(t) \otimes g(t) = \int_0^t f(t-x)g(x) dx$$

we may then ask for the Laplace transform of this convolution. We obtain this formally by plugging into Eq. (1.28) as follows:

$$\begin{aligned} J(t) \otimes g(t) &\longleftrightarrow \int_{t=0}^{\infty} (f(t) \otimes g(t)) e^{-st} dt \\ &= \int_{t=0}^{\infty} \int_{x=0}^t f(t-x)g(x) dx e^{-st} dt \\ &= \int_{x=0}^{\infty} \int_{t=x}^{\infty} f(t-x)e^{-s(t-x)} dt g(x)e^{-sx} dx \\ &= \int_{x=0}^{\infty} g(x)e^{-sx} dx \int_{v=0}^{\infty} f(v)e^{-sv} dv \end{aligned}$$

And so we have

$$f(t) \otimes g(t) \longleftrightarrow F^*(s)G^*(s)$$

Once again we see that the transform of the convolution of two functions equals the product of the transforms of each. In Table 1.3, we list a number of important properties of the Laplace transform, and in Table 1.4 we list some of the important transforms themselves. In these tables we adopt the usual notation as follows:

$$\frac{d^n f(t)}{dt^n} \triangleq f^{(n)}(t) \quad (1.37)$$

$$\underbrace{\int \cdots \int}_{n \text{ times}} f(x) dx \triangleq f^{(-n)}(t)$$

For example, $f^{(-1)}(t) = \int_{-\infty}^t f(x) dx$; when we deal with functions which are zero for $t < 0$, then $f^{(-1)}(0^-) = 0$. We comment here that the one-sided transform that uses 0^+ as a lower limit in its definition is quite commonly used in transient analysis, but we prefer 0^- so as to include impulses at the origin.

The table of properties permits one to compute many transform pairs from a given pair. Property 2 is the statement of linearity and Property 3 describes the effect of a scale change. Property 4 gives the effect of a translation in time,

Table I.3

Some Properties of the Laplace Transform

FUNCTION	TRANSFORM
1. $f(t) \quad t \geq 0$	$F^*(s) = \int_{0^-}^{\infty} f(t)e^{-st} dt$
2. $af(t) + bg(t)$	$aF^*(s) + bG^*(s)$
3. $f\left(\frac{t}{a}\right) \quad (a > 0)$	$aF^*(as)$
4. $f(t - a)$	$e^{-as}F^*(s)$
5. $e^{-at}f(t)$	$F^*(s + a)$
6. $tf(t)$	$-\frac{dF^*(s)}{ds}$
7. $t^n f(t)$	$(-1)^n \frac{d^n F^*(s)}{ds^n}$
8. $\int_0^t f(\tau) d\tau$	$\int_{s_1=s}^{\infty} F^*(s_1) ds_1$
9. $\int_0^l f(l-\tau) g(\tau) d\tau$	$\int_{s_1=s}^{\infty} ds_1 \int_{s_2=s_1}^{\infty} ds_2 \cdots \int_{s_n=s_{n-1}}^{\infty} ds_n F^*(s_n) G^*(s)$
10. $f(l) \otimes g(l)$	$F^*(s)G^*(s)$
11. $\frac{df(t)}{dt}$	$sF^*(s)$
12. $\frac{d^n f(t)}{dt^n}$	$s^n F^*(s)$
B. 13. $\int_{-\infty}^t f(t) dt$	$F^*(s)$
14. $\underbrace{\int_{-\infty}^t \cdots \int_{-\infty}^t f(t) (dt)^n}_{n \text{ times}}$	$\frac{F^*(s)}{s^n}$
15. $\frac{\partial}{\partial a} f(t) \quad [a \text{ is a parameter}]$	$\frac{\partial}{\partial a} F(s)$
16. Integral property	$F^*(0) = \int_{0^-}^{\infty} f(t) dt$
17. Initial value theorem	$\lim_{s \rightarrow \infty} sF^*(s) = \lim_{t \rightarrow 0} f(t)$
18. Final value theorem	$\lim_{s \rightarrow 0} sF^*(s) = \lim_{t \rightarrow \infty} f(t)$ if $sF^*(s)$ is analytic for $\text{Re}(s) \geq 0$

^t To be complete, we wish to show the form of the transform for entries 11-14 in the case when $f(t)$ may have nonzero values for $t < 0$ also:

$$\begin{aligned} \frac{d^n f(t)}{dt^n} &\Leftrightarrow s^n F^*(s) - s^{n-1}f(0^-) - s^{n-2}f'(0^-) - \cdots - f^{(n-1)}(0^-) \\ \underbrace{\int_{-\infty}^t \cdots \int_{-\infty}^t f(t) (dt)^n}_{n \text{ times}} &\Leftrightarrow \frac{F^*(s)}{s^n} + \frac{f^{(-1)}(0^-)}{s^{n-1}} + \frac{f^{(-2)}(0^-)}{s^{n-2}} + \cdots + \frac{f^{(-n)}(0^-)}{s} \end{aligned}$$

Table 1.4
Some Laplace Transform Pairs

FUNCTION	TRANSFORM
1. $f(t) \quad t \geq 0$	$F^*(s) = \int_{0-}^{\infty} f(t)e^{-st} dt$
2. $\delta(t)$ (unit impulse)	1
3. $\delta(t - a)$	
4. $f(t) \triangleq \frac{d}{dt} f(t)$	
5. $u_{-1}(t) \triangleq \langle\langle t \rangle\rangle$ (unit step)	
6. $u_{-1}(t - a)$	s
7. $f^{(n)}(t) \triangleq \frac{d^n f(t)}{dt^n}$	$s^n F^*(s)$
8. $Ae^{-at} \delta(t)$	$\frac{A}{s + a}$
9. $t^n e^{-at} \delta(t)$	$\frac{n!}{(s + a)^{n+1}}$
10. $\frac{t^n}{n!} e^{-at} \delta(t)$	$\frac{1}{(s + a)^{n+1}}$

whereas Property 5, its dual, gives the effect of a parameter shift in the transform domain. Properties 6 and 7 show the effect of multiplication by t (to some power), which corresponds to differentiation in the transform domain; similarly, Properties 8 and 9 show the effect of division by t (to some power), which corresponds to integration. Property 10, a most important property (derived earlier), shows the effect of convolution in the time domain going over to simple multiplication in the transform domain. Properties 11 and 12 give the effect of time differentiation; it should be noted that this corresponds to multiplication by s (to a power equal to the number of differentiations in time) times the original transform. In a similar way Properties 13 and 14 show the effect of time integration going over to division by s in the transform domain. Property 15 shows that differentiation with respect to a parameter $off(t)$ corresponds to differentiation in the transform domain as well. Property 16, the integral property, shows the simple way in which the transform may be evaluated at the origin to give the total integral

of $J(t)$. Properties 17 and 18, the initial and final value theorems, show how to compute the values for $J(t)$ at $t = 0$ and $t = \infty$ directly from the transform.

In Table 1.4 we have a rather short list of important Laplace transform pairs. Much more extensive tables exist and may be found elsewhere [DOET 61]. Of course, as we said earlier, the table shown can be extended considerably by making use of the properties listed in Table 1.3. We note, for example, that the transform pair 3 in Table 1.4 is obtained from transform pair 2 by application of Property 4 in Table 1.3. We point out again that this table is limited in length since we have included only those functions that find relevance to the material contained in this text.

So far in this discussion of Laplace transforms we have been considering only functions $J(t)$ for which $J(t) = 0$ for $t < 0$. This will be satisfactory for most of the work we consider in this text. However, there is an occasional need for transforming a function of time which may be nonzero anywhere on the real-time axis. For this purpose we must once again consider the lower limit of integration to be $-\infty$, that is,

$$F^*(s) = \int_{-\infty}^{\infty} f(t)e^{-st} dt \quad (1.38)$$

One can easily show that this (bilateral) Laplace transform may be calculated in terms of one-sided time functions and their transforms as follows. First we define

$$f_-(t) = \begin{cases} J(t) & t < 0 \\ 0 & t \geq 0 \end{cases}$$

$$J_+(t) = \begin{cases} 0 & t < 0 \\ J(t) & t \geq 0 \end{cases}$$

and so it immediately follows that

$$J(t) = J_-(t) + J_+(t)$$

We now observe that $J_-(-t)$ is a function that is nonzero only for positive values of t , and $J_+(t)$ is nonzero only for nonnegative values of t . Thus we have

$$J_+(t) \Leftrightarrow F_+^*(s)$$

$$J_-(-t) \Leftrightarrow F_-^*(s)$$

where these transforms are defined as in Eq. (1.29). However, we need the transform of $J_-(t)$ which is easily shown to be

$$J_-(t) \Leftrightarrow F_-^*(-s)$$

Thus, by the linearity of transforms, we may finally write the bilateral transform in terms of one-sided transforms:

$$F^*(s) = F_-^*(-s) + F_+^*(s)$$

As always, these Laplace transforms have abscissas of absolute convergence. Let us therefore define σ_+ as the convergence abscissa for $F^*(s)$; this implies that the region of convergence for $F^*(s)$ is $\text{Re}(s) > \sigma_+$. Similarly, $F_-^*(s)$ will have some abscissa of absolute convergence, which we will denote by σ_- , which implies that $F_-^*(s)$ converges for $\text{Re}(s) > \sigma_-$. It then follows directly that $F_-^*(-s)$ will have the same convergence abscissa (σ_-) but will converge for $\text{Re}(s) < \sigma_-$. Thus we have a situation where $F^*(s)$ converges for $\sigma_+ < \text{Re}(s) < \sigma_-$ and therefore we will have a "convergence strip" if and only if $\sigma_+ < \sigma_-$; if such is not the case, then it is not useful to define $F^*(s)$. Of course, a similar argument can be made in the case of z-transforms for functions that take on nonzero values for negative time indices.

So far we have seen the effect of tagging our time function $f(t)$ with the complex exponential e^{-st} and then compressing (integrating) over all such tagged functions to form a new function, namely, the transform $F^*(s)$. The purpose of the tagging was so that we could later "untransform" or, if you will, "unwind" the transform in order to obtain $f(t)$ once again. In principle we know this is possible since a transform and its time function are uniquely related. So far, we have specified how to go in the one direction from $f(t)$ to $F^*(s)$. Let us now discuss the problem of inverting the Laplace transform $F^*(s)$ to recover $f(t)$. There are basically two methods for conducting this inversion: The *inspection method* and the *formal inversion integral method*. These two methods are very similar.

First let us discuss the *inspection method*, which is perhaps the most useful scheme for inverting transforms. Here, as with z-transforms, the approach is to rewrite $F^*(s)$ as a sum of terms, each of which can be recognized from the table of Laplace transform pairs. Then, making use of the linearity property, we may invert the transform term by term, and then sum the result to recover $f(t)$. Once again, the basic method for writing $F^*(s)$ as a sum of recognizable terms is that of the partial-fraction expansion. Our description of that method will be somewhat shortened here since we have discussed it at some length in the z-transform section. First, we will assume that $F^*(s)$ is a rational function of s , namely,

$$F^*(s) = \frac{N(s)}{D(s)}$$

where both the numerator $N(s)$ and denominator $D(s)$ are each polynomials in s . Again, we assume that the degree of $N(s)$ is less than the degree of $D(s)$; if this is not the case, $N(s)$ must be divided by $D(s)$ until the remainder is of degree less than the degree of $D(s)$, and then the partial-fraction expansion is carried out for this remainder, whereas the terms of the quotient resulting from the division will be simple powers of s , which may be inverted by appealing to Transform 4 in Table 1.4. In addition, we will assume that the

"hard" part of the problem has been done, namely, that $D(s)$ has been put in factored form

$$D(s) = \prod_{i=1}^k (s + a_i)^{m_i} \quad (1.39)$$

Once $F^*(s)$ is in this form we may then express it as the following sum:

$$\begin{aligned} F^*(s) = & \frac{B_{11}}{(s + a_1)^{m_1}} + \frac{B_{12}}{(s + a_1)^{m_1-1}} + \cdots + \frac{B_{1m_1}}{(s + a_1)} \\ & + \frac{B_{21}}{(s + a_2)^{m_2}} + \frac{B_{22}}{(s + a_2)^{m_2-1}} + \cdots + \frac{B_{2m_2}}{(s + a_2)} + \cdots \\ & + \frac{B_{kl}}{(s + a_k)^{m_k}} + \frac{B_{k2}}{(s + a_k)^{m_k-1}} + \cdots + \frac{B_{km}}{(s + a_k)} \end{aligned} \quad (1.40)$$

Once we have expressed $F^*(s)$ as above we are then in a position to invert each term in this sum by inspection from Table 1.4. In particular, Pairs 8 (for simple poles) and 10 (for multiple poles) give us the answer directly. As before, the method for calculating the coefficients B_{ij} is given in general by

$$B_{ij} = \frac{1}{(j-1)!} \left[\frac{d^{j-1}}{ds^{j-1}} \left((s + a_i)^{m_i} \frac{N(s)}{D(s)} \right) \right]_{s=-a_i} \quad (1.41)$$

Thus we have a complete prescription for finding $f(t)$ from $F^*(s)$ by inspection in those cases where $F^*(s)$ is rational and where $D(s)$ has been factored as in Eq. (1.39). This method works very well in those cases where $F^*(s)$ is not overly complex.

To elucidate some of these principles let us carry out a simple example. Assume that $F^*(s)$ is given by

$$F^*(s) = \frac{8(s^2 + 3s + 1)}{(s + 3)(s + 1)^3} \quad (1.42)$$

We have already written the denominator in factored form, and so we may proceed directly to expand $F^*(s)$ as in Eq. (1.40). Note that we have $k = 2$, $a_1 = 3$, $m_1 = 1$, $a_2 = 1$, $m_2 = 3$. Since the denominator degree (4) is greater than the numerator degree (2), we may immediately expand $F^*(s)$ as a partial fraction as given by Eq. (1.40), namely,

$$F^*(s) = \frac{B_{11}}{s + 3} + \frac{B_{21}^2}{(s + 1)^3} + \frac{B_{22}^2}{(s + 1)^2} + \frac{B_{23}}{(s + 1)}$$

Evaluation of the coefficients B_{ij} proceeds as follows. B_{11} is especially simple since no differentiations are required, and we obtain

$$B_{11} = (s+3)^{-1} F(s) \Big|_{s=-3} = 8 \frac{(9-9+1)}{(-2)^3} = -1$$

B_{21} is also easy to evaluate:

$$B_{21} = (s+1)^3 F(s) \Big|_{s=-1} = 8 \frac{(1-3+1)}{2^3} = -4$$

For B_{22} we must differentiate once, namely,

$$\begin{aligned} B_{22} &= \frac{d}{ds} \left[\frac{8(s^2 + 3s + 1)}{(s+3)^2} \right] \Big|_{s=-1} \\ &= 8 \frac{(s+3)(2s+3) - (s^2 + 3s + 1)(2)}{(s+3)^3} \Big|_{s=-1} \\ &= 8 \frac{s^2 + 6s + 8}{(s+3)^2} \Big|_{s=-1} = 8 \frac{1 - 6 + 8}{(2)^2} \\ &= 6 \end{aligned}$$

Lastly, the calculation of B_{23} involves two differentiations; however, we have already carried out the first differentiation, and so we take advantage of the form we have derived in B_{22} just prior to evaluation at $s = -1$; furthermore, we note that since $j = 3$, we have for the first time an effect due to the term $(j-1)!$ from Eq. (1.41). Thus

$$\begin{aligned} B_{23} &= \frac{1}{2!} \frac{d^2}{ds^2} \left[\frac{8(s^2 + 3s + 1)}{(s+3)^2} \right] \Big|_{s=-1} \\ &= \frac{1}{2} (8) \frac{d}{ds} \left[\frac{s^2 + 6s + 8}{(s+3)^2} \right] \Big|_{s=-1} \\ &= 4 \frac{(s+3)^2(2s+6) - (s^2 + 6s + 8)(2)(s+3)}{(s+3)^4} \Big|_{s=-1} \\ &= 4 \frac{(2)(2)(4) - (1 - 6 + 8)(2)(2)}{(2)^4} \\ &= 1 \end{aligned}$$

This completes the evaluation of the constants B_{ii} to give the partial-fraction expansion

$$F^*(s) = \frac{-1}{s+3} + \frac{-4}{(s+1)^3} + \frac{6}{(s+1)^2} + \frac{1}{(s+1)} \quad (1.43)$$

This last form lends itself to inversion 'by inspection' as we had promised. In particular, we observe that the first and last terms invert directly according to transform pair 8 from Table 1A, whereas the second and third terms invert directly from Pair 10 of that table; thus we have for $t \geq 0$ the following:

$$f(t) = -e^{-3t} - 2t^2e^{-t} + 6te^{-t} + e^{-t} \quad (1.44)$$

and of course, $f(t) = 0$ for $t < 0$.

In the course of carrying out an inversion by partial-fraction expansions there are two natural points at which one can conduct a test to see if any errors have been made: first, once we have the partial-fraction expansion [as in our example, the result given in Eq. (1.43)], then one can combine this sum of terms into a single term over a common denominator and check that this single term corresponds to the original given $F^*(s)$; the other check is to take the final form for $J(t)$ and carry out the forward transformation and confirm that it gives the original $F^*(s)$ [of course, one then gets $F^*(s)$ expanded directly as a partial fraction].

The *second* method for finding $j'(r)$ from $F^*(s)$ is to use the *inversion integral*

$$J(t) = \frac{1}{2\pi j} \int_{\sigma_c - j\infty}^{\sigma_c + j\infty} F^*(s) e^{st} ds \quad (1.45)$$

for $t \geq 0$ and $\sigma_c > \sigma_a$. The integration in the complex s -plane is taken to be a straight-line integration parallel to the imaginary axis and lying to the right of σ_a , the abscissa of absolute convergence for $F^*(s)$. The usual means for carrying out this integration is to make use of the Cauchy residue theorem as applied to the integral in the complex domain around a closed contour. The closed contour we choose for this purpose is a semicircle of infinite radius as shown in Figure 1A. In this figure we see the path of integration required for Eq. (1.45) is $s_1 - s_2$ and the semicircle of infinite radius closing this contour is given as $s_1 - s_2 - s_3$. If the integral along the path $s_1 - s_2 - s_3$ is 0, then the integral along the entire closed contour will in fact give us $J(t)$ from Eq. (1.45). To establish that this contribution is 0, we need

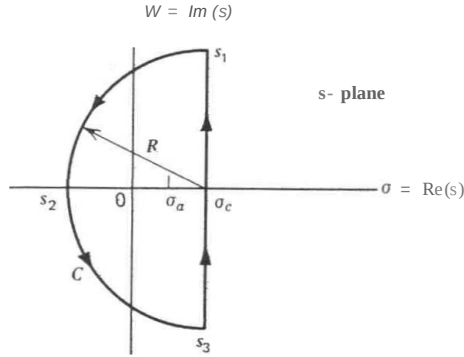


Figure 1.4 Closed contour for inversion integral.

Jordan's Lemma If $|F^*(s)| \rightarrow 0$ as $R \rightarrow \infty$ on $s = s_1 - s_3$, then

$$\int_{s_1 - s_2 - s_3} F^*(s) e^{st} ds = 0 \quad \text{for } t > 0$$

Thus, in order to carry out the complex inversion integral shown in Eq. (1.45), we must first express $F^*(s)$ in a form for which Jordan's lemma applies. Having done this we may then evaluate the integral around the closed contour C by calculating residues and using Cauchy's residue theorem. This is most easily carried out if $F^*(s)$ is in rational form with a factored denominator as in Eq. (1.39). In order for Jordan's lemma to apply, we will require, as we did before, that the degree of the numerator be strictly less than the degree of the denominator, and if this is not so, we must divide the rational function until the remainder has this property. That is all there is to the method. Let us carry this out on our previous example, namely that given in Eq. (1.42). We note this is already in a form for which Jordan's lemma applies, and so we may proceed directly with Cauchy's residue theorem. Our poles are located at $s = -3$ and $s = -1$. We begin by calculating the residue at $s = -3$, thus

$$\begin{aligned} r_{-3} &= (s + 3)F^*(s)e^{st} \Big|_{s=-3} \\ &= \frac{8(s' + 3s + 1)e'}{(s + 1)^3} \Big|_{s=-3} \\ &= \frac{8(9 - 9 + 1)e^{-3t}}{(-2)^3} \\ &= -\frac{1}{2}e^{-3t} \end{aligned}$$

Similarly, we must calculate the residue at $s = -1$, which requires the differentiations indicated in our residue formula Eq. (1.27):

$$\begin{aligned}
 r_{-1} &= \frac{1}{2!} \frac{d^2}{ds^2} (s+1) 3F^*(s) e'' \Big|_{s=-1} \\
 &= \frac{1}{2} \frac{d^2}{ds^2} \frac{8(s^2 + 3s + 1)e''}{(5+3)} \Big|_{s=-1} \\
 &= \frac{1}{2} \frac{d}{ds} \left[\frac{(s+3)[8(2s+3)e't + 8(5^2 + 3s + 1)te'']}{(s+3)^2} - \frac{8(s^2 + 3s + 1)e''}{(s+3)^2} \right] \Big|_{s=-1} \\
 &= \frac{1}{2(s+3)^4} \{ (s+3)^2 [8(2s+3)e'' + 8(5^2 + 3s + 1)e''] \\
 &\quad + (s+3)8[2e'' + (25+3)le''] \\
 &\quad + (5+3)8[(2s+3)le^{st} + (s^2 + 3s + 1)l^2e''] \\
 &\quad - 8(2s+3)e'' - 8(s^2 + 3s + 1)le''] \\
 &\quad - [(s+3)8[(2s+3)e't + (5^2 + 3s + 1)jre''] \\
 &\quad - 8(s^2 + 3s + 1)e''] 2(s+3) \} \Big|_{s=-1} \\
 &= e^{-t} + 6le^{-t} - 2l^2e^{-t}
 \end{aligned}$$

Combining these residues we have

$$f(t) = -e^{-3t} + e^{-t} + 6le^{-t} - 2l^2e^{-t}, \quad t \geq 0$$

Thus we see that our solution here is the same as in Eq. (1.44), as it must be: we have once again that $f(t) = 0$ for $t < 0$.

In our earlier discussion of the (bilateral) Laplace transform we discussed functions of time $f_{-}(t)$ and $f_{+}(t)$ defined for the negative and positive real-time axis, respectively. We also observed that the transform for each of these functions was analytic in a left half-plane and a right half-plane, respectively, as measured from their appropriate abscissas of absolute convergence. Moreover, in our last inversion method [the application of Eq. (1.45)] we observed that closing the contour by a semicircle of infinite radius in a counterclockwise direction gave a result for $t > 0$. We comment now that had we closed the contour in a clockwise fashion to the right, we would have obtained the result that would have been applicable for $t < 0$, assuming that the contribution of this contour could be shown to be 0 by Jordan's lemma. In order to invert a bilateral transform, we proceed by obtaining first $f(t)$ for positive values of t and then for negative values of t . For the first we take a path of integration within the convergence strip defined by $\sigma_- < \sigma_c < \sigma_+$ and then closing the contour with a counterclockwise semicircle; for $t < 0$, we take the same vertical contour but close it with a semicircle to the right.

As may be anticipated from our contour integration methods, it is sometimes necessary to determine exactly how many singularities of a function exist within a closed region. A very powerful and convenient theorem which aids us in this determination is given as follows :

Rouche's Theorem [GUIL 49] If $f(s)$ and $g(s)$ are analytic functions of s inside and on a closed contour C , and also $|f(s)| < |g(s)|$ on C , then $f(s)$ and $f(s) + g(s)$ have the same number of zeroes inside C .

1.4. USE OF TRANSFORMS IN THE SOLUTION OF DIFFERENCE AND DIFFERENTIAL EQUATIONS

As we have already mentioned, transforms are extremely useful in the solution of both differential and difference equations with constant coefficients. In this section we illustrate that technique; we begin with difference equations using z -transforms and then move on to differential equations using Laplace transforms, preparing us for the more complicated differential-difference equations encountered in the text for which we need both methods simultaneously.

Let us consider the following general N th-order linear difference equation with constant coefficients

$$a_N g_{n-N} + a_{N-1} g_{n-N+1} + \dots + a_0 g_n = e_n \quad (1.46)$$

where the a_i are known constant coefficients, g_n are unknown functions to be found, and e_n is a given function of n . In addition, we assume we are given N boundary equations (e.g., initial conditions). As always with such equations, the solution which we are seeking consists of both a homogeneous and a particular solution, namely,

$$g_n = g_n^{(h)} + g_n^{(p)}$$

just as with differential equations. We know that the homogeneous solution must satisfy the homogeneous equation

$$a_N g_{n-N} + \dots + a_0 g_n = 0 \quad (1.47)$$

The general form of solution to Eq. (1.47) is

$$g_n^{(h)} = A \alpha^n$$

where A and α are yet to be determined. If we substitute the proposed solution into Eq. (1.47), we find

$$a_N A \alpha^{n-N} + a_{N-1} A \alpha^{n-N+1} + \dots + a_0 A \alpha^n = 0 \quad (1.48)$$

This N th-order polynomial clearly has N solutions, which we will denote by $\alpha_1, \alpha_2, \dots, \alpha_N$, assuming for the moment that all the α_i are distinct. Associated with each such solution is an arbitrary constant A_i which will be determined from the initial conditions for the difference equation (of which there must be N). By cancelling the common term $A\alpha^{n-N}$ from Eq. (1.48) we finally arrive at the *characteristic equation* which determines the values α_i

$$a_N + a_{N-1}\alpha + a_{N-2}\alpha^2 + \dots + a_0\alpha^N = 0 \quad (1.49)$$

Thus the search for the homogeneous solution is now reduced to finding the N roots of our characteristic equation (1.49). If all N of the α_i are distinct, then the homogeneous solution is

$$g_n^{(h)} = A_1\alpha_1^n + A_2\alpha_2^n + \dots + A_N\alpha_N^n$$

In the case of nondistinct roots, we have a slightly different situation. In particular, let α_1 be a multiple root of order k ; in this case the k equal roots will contribute to the homogeneous solution in the following form:

$$(A_1n^{k-1} + A_2n^{k-2} + \dots + A_k) \alpha_1^n$$

and similarly for any other multiple roots. As far as the particular solution $g_n^{(p)}$ is concerned, we know that it must be found by an appropriate guess from the form of en^r

Let us illustrate some of these principles by means of an example. Consider the second-order difference equation

$$6g_n - 5g_{n-1} + g_{n-2} = 6\left(\frac{1}{5}\right)^n \quad n = 2, 3, 4, \dots \quad (1.50)$$

This equation gives the relationship among the unknown functions g_n for $n = 2, 3, 4, \dots$. Of course, we must give two initial conditions (since the order is 2) and we choose these to be $g_0 = 0$, $g_1 = 6/5$. In order to find the homogeneous solution we must form Eq. (1.49), which in this case becomes

$$6\alpha^2 - 5\alpha + 1 = 0$$

and so the two values of α which solve this equation are

$$\alpha_1 = \frac{1}{2}$$

$$\alpha_2 = \frac{1}{3}$$

and thus we have the homogeneous solution

$$g_n^{(h)} = A_1\left(\frac{1}{2}\right)^n + A_2\left(\frac{1}{3}\right)^n$$

The particular solution must be guessed at, and the correct guess in this case is

$$g_n^{(p)} = B\left(\frac{1}{5}\right)^n$$

If we plug $g_n^{(p)}$ as given back into our basic equation, namely, Eq. (1.50), we find that $B = 1$ and so we are convinced that the particular solution is correct. Thus our *complete* solution is given by

$$g_n = g_n^{(h)} + g_n^{(p)} = A_1\left(\frac{1}{2}\right)^n + A_2\left(\frac{1}{3}\right)^n + \left(\frac{1}{5}\right)^n$$

We use the initial conditions to solve for A_1 and A_2 and find $A_1 = 8$ and $A_2 = -9$. Thus our final solution is

$$g_n = \left(\frac{1}{2}\right)^{n-3} - \left(\frac{1}{3}\right)^{n-2} + \left(\frac{1}{5}\right)^n \quad n = 0, 1, 2, \dots \quad (1.51)$$

This completes the standard way for solving our difference equation.

Let us now describe the method of c-transforms for solving difference equations. Assume once again that we are given Eq. (1.46) and that it is good in the range $n = k, k+1, \dots$. Our approach begins by defining the following c-transform:

$$G(z) = \sum_{n=0}^{\infty} g_n z^n \quad (1.52)$$

From our earlier discussion we know that once we have found $G(z)$ we may then apply our inversion techniques to find the desired solution g_n . Our next step is to multiply the n th equation from Eq. (1.46) by z^n and then form the sum of all such multiplied equations from k to infinity; that is, we form

$$\sum_{n=k}^{\infty} \sum_{i=0}^N a_i g_{n-i} z^n = \sum_{n=k}^{\infty} e_n z^n$$

We then carry out the summations and attempt to recognize $G(z)$ in this single equation. Next we solve for $G(z)$ algebraically and then proceed with our inversion techniques to obtain the solution. This method does not require that we guess at the particular solution, and so in that sense is simpler than the direct method; however, as we shall see, it still has the basic difficulty that we must solve the characteristic equation [Eq. (1.49)] and in general this is the difficult part of the solution. However, even if we cannot solve for the roots α_i it is possible to obtain meaningful properties of the solution g_n from the perhaps unfactored form for $G(z)$.

Let us solve our earlier example using the method of z -transforms. Accordingly we begin with Eq. (1.50), multiply by z^n and then sum; the sum will go from 2 to infinity since this is the applicable range for that equation. Thus

$$\sum_{n=2}^{\infty} 6g_n z^n - \sum_{n=2}^{\infty} 5g_{n-1} z^n + \sum_{n=2}^{\infty} g_{n-2} z^n = \sum_{n=2}^{\infty} 6\left(\frac{1}{5}\right)^n z^n$$

We now factor out enough powers of z from each sum so that these powers match the subscript on g thusly:

$$6 \sum_{n=2}^{\infty} g_n z^n - 5z \sum_{n=2}^{\infty} g_{n-1} z^{n-1} + z^2 \sum_{n=2}^{\infty} g_{n-2} z^{n-2} = \sum_{n=2}^{\infty} 6\left(\frac{1}{5}\right)^n z^n$$

Focusing on the first summation we see that it is almost of the form $G(z)$ except that it is missing the terms for $n = 0$ and $n = 1$ [see Eq. (1.52)]; applying this observation to each of the sums on the left-hand side and carrying out the summation on the right-hand side directly, we find

$$6[G(z) - g_0 - g_1 z] - 5z[G(z) - g_0] + z^2 G(z) = \frac{6(1/5)z^2}{1 - (1/5)z}$$

Observe how the first term in this last equation reflects the fact that our summation was missing the first two terms for $G(z)$. Solving for $G(z)$ algebraically we find

$$G(z) = \frac{6g_0 + 6g_1 z - 5g_0 z + (6/25)z^2/[1 - (1/5)z]}{6 - 5z + z^2}$$

If we now use our given values for g_0 and g_1 , we have

$$G(z) = \frac{1}{5} \frac{z(6 - z)}{[1 - (1/3)z][1 - (1/2)z][1 - (1/5)z]}$$

Proceeding with a partial-fraction expansion of this last form we obtain

$$G(z) = \frac{-9}{1 - (1/3)z} + \frac{8}{1 - (1/2)z} + \frac{1}{1 - (1/5)z}$$

which by our usual inversion methods yields the final solution

$$g_n = -9\left(\frac{1}{3}\right)^n + 8\left(\frac{1}{2}\right)^n + \left(\frac{1}{5}\right)^n \quad n = 0, 1, 2, \dots$$

Note that this is exactly the same as Eq. (LSI) and so our method checks. We comment here that even were we not able to invert the given form for $G(z)$ we could still have found certain of its properties; for example, we could find

that the sum of all terms is given immediately by $G(I)$, that is,

$$G(1) = \sum_{n=0}^{\infty} g_n = \frac{15}{4}$$

Let us now consider the application of the Laplace transform to the solution of constant-coefficient linear differential equations. Consider an N th-order equation of the following form:

$$a_N \frac{d^N f(t)}{dt^N} + a_{N-1} \frac{d^{N-1} f(t)}{dt^{N-1}} + \dots + a_1 \frac{df(t)}{dt} + a_0 f(t) = e(t) \quad (1.53)$$

Here the coefficients a_i are given constants, and $e(t)$ is a given driving function. Along with this equation we must also be given N initial conditions in order to carry out a complete solution; these conditions typically are the values of the first N derivatives at some instant, usually at time zero. It is required to find the function $f(t)$. As usual, we will have a homogeneous solution $f^{(h)}(t)$, which solves the homogeneous equation [when $e(t) = 0$] as well as a particular solution $f(Q(t))$ that corresponds to the nonhomogeneous equation. The form for the homogeneous solution will be

$$f^{(h)}(t) = Ae^{\alpha t}$$

If we substitute this into Eq. (1.53) we obtain

$$a_N A \alpha^N e^{\alpha t} + a_{N-1} A \alpha^{N-1} e^{\alpha t} + \dots + a_1 A \alpha e^{\alpha t} + a_0 A e^{\alpha t} = 0$$

This equation will have N solutions $\alpha_1, \alpha_2, \dots, \alpha_n$, which must solve the *characteristic equation*

$$a_N \alpha^N + a_{N-1} \alpha^{N-1} + \dots + a_1 \alpha + a_0 = 0$$

which is equivalent to Eq. (1.49) with a change in subscripts. If all of the α_i are distinct, then the general form for our homogeneous solution will be

$$f^{(h)}(t) = A_1 e^{\alpha_1 t} + A_2 e^{\alpha_2 t} + \dots + A_N e^{\alpha_N t}$$

The evaluation of the coefficients A_i is carried out making use of the initial conditions. In the case of multiple roots we have the following modification. Let us assume that α_1 is a repeated root of order k ; this multiple root will contribute to the homogeneous solution in the following way:

$$(A_{11} t^{k-1} + A_{12} t^{k-2} + \dots + A_{1, k-1} t + A_{1k}) e^{\alpha_1 t}$$

and in the case of more than one multiple root the modification is obvious. As usual, one must guess in order to find the particular solution $f^{(p)}(t)$. The complete solution then is, of course, the sum of the homogeneous and particular solutions, namely,

$$J(t) = j(h(t) + f^{(p)}(t))$$

Let us apply this method to the solution of the following differential equation for illustrative **purposés**:

$$\frac{d^2 f(t)}{dt^2} - 6 \frac{df(t)}{dt} + 9f(t) = 2t \quad (1.54)$$

with the two initial conditions $f(0^-) = 0$ and $df(0^-)/dt = 0$. Forming the characteristic equation

$$\alpha^2 - 6\alpha + 9 = 0 \quad (1.55)$$

we find the following multiple root:

$$\alpha_1 = \alpha_2 = 3$$

and so the homogeneous solution must be of the form

$$j(h)(t) = (A_1 t + A_2)e^{3t}$$

Making an appropriate guess for the particular solution we try

$$f^{(p)}(t) = B_1 + B_2 t$$

Substituting this back into the basic equation (1.54) we find that $B_1 = 4/27$ and $B_2 = 2/9$. Thus our complete solution takes the form

$$j(t) = (A_1 t + A_2)e^{3t} + \frac{4}{27} + \frac{2}{9}t$$

Since our initial conditions state that *both* $f(t)$ and its first derivative must be zero at $t = 0^-$, we find that $A_1 = 2/9$ and $A_2 = -4/27$, which gives for our final and complete solution

$$J(t) = \frac{2}{9}\left(t - \frac{2}{3}\right)e^{3t} + \frac{2}{9}\left(t + \frac{2}{3}\right) \quad t \geq 0 \quad (1.56)$$

The Laplace transform provides an alternative method for solving constant-coefficient linear differential equations. The method is based upon Properties II and 12 of Table 1.3, which relate the derivative of a **time** function to its Laplace transform. The approach is to make use of these properties to transform both sides of the given differential equation into an equation involving the Laplace transform of the unknown function $f(t)$ itself, which we denote as usual by $F^*(s)$. This algebraic equation is then solved for $F^*(s)$, and is then

inverted by any of our methods in order to immediately yield the complete solution *for* U). No guess is required in order to find the particular solution, since it comes out of the inversion procedure directly.

Let us apply this technique to our previous example. We begin by transforming both sides of Eq. (1.54), which will require that we take advantage of our initial conditions as follows:

$$s^2 F^*(s) - sJ(0^-) - f^{(1)}(0^-) = 6sF^*(s) + 6J(0^-) + 9F^*(s) = \frac{2}{s^2}$$

In carrying out this last operation we have taken advantage of Laplace transform pair 7 from Table IA. Since our initial conditions are both zero, we may eliminate certain terms in this last equation and proceed directly to solve for $F^*(s)$ thusly:

$$F^*(s) = \frac{2/s^2}{s^2 - 6s + 9}$$

We must now factor this last equation, which is the same problem we faced in finding the roots of Eq. (1.55) in the direct method, and as usual forms the basically difficult part of all direct and indirect methods. Carrying this out we have

$$F^*(s) = \frac{2}{s^2(s - 3)^2}$$

We are now in the position to make a **partial-fraction** expansion yielding

$$F^*(s) = \frac{2/9}{s^2} + \frac{4/27}{s} + \frac{2/9}{(s - 3)^2} + \frac{-4/27}{s - 3}$$

Inverting as usual we then obtain, for $t \geq 0$,

$$J(t) = \frac{2}{9} + \frac{4}{27} + \frac{2}{9} te^{3t} - \frac{4}{27} e^{3t}$$

which is identical to our former solution given in Eq. (1.56).

In our study of queueing systems we often encounter not only difference equations and differential equations but also the combination in the form of differential-difference equations. That is, if we refer back to Eq. (1.53) and replace the time functions by time functions that depend upon an index, say n , and if we then display a set of differential equations for various values of n , then we have an infinite set of differential-difference equations. The solution to such equations often requires that we take both the z -transform on the discrete index n and the Laplace transform on the continuous time parameter t . Examples of this type of analysis are to be found in the text itself.

REFERENCES

- AHLF 66 Ahlfors, L. V., *Complex Analysis*, 2nd Edition, McGraw-Hill (New York), 1966.
- CADZ 73 Cadzow, J. A., *Discrete-Time Systems*, Prentice-Hall (Englewood Cliffs, N.J.), 1973.
- DOET 61 Doetsch, G., *Guide to the Applications of Laplace Transforms*, Van Nostrand (Princeton), 1961.
- GU[L 49 Guillemin, E. A., *The Mathematics of Circuit Analysis*, Wiley (New York), 1949.
- JURY 64 Jury, E. 1., *Theory and Application of the z-Transform Method*, Wiley (New York), 1964.
- SCHW 59 Schwartz, L., *Theorie des Distributions*, 2nd printing, Actualités scientifiques et industrielles Nos. [245 and I[22, Hermann et Cie. (Paris), Vol. 1 (1957), Vol. 2 (1959).
- WIDD 46 Widder, D. V., *The Laplace Transform*, Princeton University Press (Princeton), 1946.

Probability Theory Refresher

In this appendix we review selected topics from probability theory, which are relevant to our discussion of queueing systems. Mostly, we merely list the important definitions and results with an occasional example. The reader is expected to be familiar with this material, which corresponds to a good first course in probability theory. Such a course would typically use one of the following texts that contain additional details and derivations: Feller, Volume I [FELL 68]; Papoulis [PAPO 65]; Parzen [PARZ 60]; or Davenport [DAVE 70].

Probability theory concerns itself with describing random events. A typical dictionary definition of a random event is an event lacking aim, purpose, or regularity. Nothing could be further from the truth! In fact, it is the *extreme regularity* that manifests itself in collections of random events, that makes probability theory interesting and useful. The notion of statistical regularity is central to our studies. For example, if one were to toss a fair coin four times, one expects on the average two heads and two tails. Of course, there is one chance in sixteen that no heads will occur. As a consequence, if an unusual sequence came up (that is, no heads), we would not be terribly surprised nor would we suspect the coin was unfair. On the other hand, if we tossed the coin a million times, then once again we expect approximately half heads and half tails, but in this case, if no heads occurred, we would be more than surprised, we would be indignant and with overwhelming assurance could state that this coin was clearly unfair. In fact, the odds are better than 10^{98} to 1 that at least 490,000 heads will occur! This is what we mean by statistical regularity, namely, that we can make some very precise statements about large collections of random events.

ILL RULES OF THE GAME

We now describe the rules of the game for creating a mathematical model for probabilistic situations, which is to correspond to real-world experiments. Typically one examines three features of such experiments:

1. A set of possible experimental *outcomes*.

2. A grouping of these outcomes into classes called *results*.
3. The *relative frequency* of these classes in many independent trials of the experiment.

The relative frequency fe of a class c is merely the number of times the experimental outcome falls into that class, *divided* by the number of *times* the experiment is performed; as the number of experimental trials increases, we expect fe to reach a limit due to our *notion* of *statistical* regularity.

The *mathematical model* we create also has three quantities of interest that are in one-to-one relation with the three quantities listed above in the experimental world. They are, respectively:

1. A *sample space* which is a collection of objects which we denote by S . S corresponds to the set of mutually exclusive exhaustive outcomes of the model of an experiment. Each object (i.e., *possible* outcome) w in the set S is referred to as a *sample point*.
2. A family of *events* $\mathcal{E}$ denoted $\{A, B, C, \dots\}$ in which each event is a set of sample points $\{w\}$. An event corresponds to a class or result in the real world.
3. A *probability measure* P which is an assignment (mapping) of the events defined on S into the set of real numbers. P corresponds to the relative frequency in the experimental *situation*. The notation $P[A]$ is used to denote the real number associated with the event A . This assignment must satisfy the following properties (axioms):

$$(a) \quad \text{For any event } A, 0 \leq P[A] \leq 1. \quad (11.1)$$

$$(b) \quad P[S] = 1. \quad (11.2)$$

$$(c) \quad \text{If } A \text{ and } B \text{ are "mutually exclusive" events [see (11.4) below], then } P[A \cup B] = P[A] + P[B]. \quad (11.3)$$

It is appropriate at this point to define some set theoretic notation [for example, the use of the symbol $\cup$ in property (c)]. Typically, we describe an event A as follows: $A = \{w : w \text{ satisfies the membership property for the event } A\}$; this is read as " A is the set of sample points w such that w satisfies the membership property for the event A ." We further define

$$A^c = \{w : w \text{ not in } A\} = \text{complement of } A$$

$$A \cup B = \{w : w \text{ in } A \text{ or } B \text{ or both}\} = \text{union of } A \text{ and } B$$

$$A \cap B = AB = \{w : w \text{ in } A \text{ and } B\} = \text{intersection of } A \text{ and } B$$

$$\varnothing = S^c = \text{null event (contains no sample points since } S \text{ contains all the points)}$$

If $AB = \varphi$, then A and B are said to be *mutually exclusive* (or disjoint). A set of events whose union forms the sample space S is said to be an *exhaustive* set of events. We are therefore led to the definition of a set of *mutually exclusive exhaustive* events $\{A_1, A_2, \dots, A_n\}$, which have the properties

$$\begin{aligned} A_i A_j &= \varphi & \text{for all } i \neq j \\ A_1 \vee A_2 \vee \dots \vee A_n &= S \end{aligned} \quad (1104)$$

We note further that $A \vee A^c = S$, $AA^c = \varphi$, $AS = A$, $A\varphi = \varphi$, $A \vee S = S$, $A \vee \varphi = A$, $S'' = \varphi$, and $\varphi^c = S$. Also, we comment that the union and intersection operators are commutative, associative, and distributive.

The triplet $(S, \mathcal{C}, P)$ along with Axioms (11.1)-(11.3) form a *probability system*. These three axioms are all that one needs in order to develop an axiomatic theory of probability whenever the number of events that can be defined on the sample space S is finite. [When the number of such events is infinite it is necessary to include an additional axiom which extends Axiom (11.3) to include the infinite union of disjoint events. This leads us to the notion of a Borel field and of infinite additivity of probability measures. We do not discuss the details further in this refresher.] Lastly, we comment that Axiom (11.2) is nothing more than a normalization statement and the choice of unity for this normalization is quite arbitrary (but also very natural).

Two other definitions are now in order. The first is that of *conditional probability*. The conditional probability of the event A given that the event B occurred (denoted as $P[A | B]$) is defined as

$$P[A | B] \triangleq \frac{P[AB]}{P[B]}$$

whenever $P[B] \neq 0$. The introduction of the conditional event B forces us to restrict attention from the original sample space S to a new sample space defined by the event B ; since this new constrained sample space must now have a total probability of unity, we magnify the probabilities associated with conditional events by dividing by the term $P[B]$ as given above.

The second additional notion we need is that of *statistical independence* of events. Two events A, B are said to be statistically independent if and only if

$$P[AB] = P[A]P[B] \quad (11.5)$$

For three events A, B, C we require that each pair of events satisfies Eq. (11.5) and in addition

$$P[ABC] = P[A]P[B]P[C]$$

This definition extends of course to n events requiring the n -fold factoring of the probability expression as well as all the $(n - 1)$ -fold factorings all the way

down to all the pairwise factorings. It is easy to see for two independent events A, B that $P[A | B] = P[A]$, which merely says that knowledge of the occurrence of the event B in no way affects the probability of the occurrence of the independent event A .

The *theorem of total probability* is especially simple and important. It relates the probability of an event B and a set of mutually exclusive exhaustive events $\{A_i\}$ as defined in Eq. (11.4). The theorem is

$$P[B] = \sum_{i=1}^n P[A_i B]$$

which merely says that if the event B is to occur it must occur in conjunction with exactly one of the mutually exclusive exhaustive events A_i . However from the definition of conditional probability we may always write

$$P[A_i B] = P[A_i | B] P[B] = P[B | A_i] P[A_i]$$

Thus we have the second important form of the theorem of total probability, namely,

$$P[B] = \sum_{i=1}^n P[B | A_i] P[A_i]$$

This last equation is perhaps one of the most useful for us in studying queueing theory. It suggests the following approach for finding the probability of some complex event B , namely, first to condition the event B on some event A_i in such a way that the calculation of the occurrence of event B given this condition is less complex, and then of course to multiply by the probability of the conditional event A_i , to yield the joint probability $P[A_i B]$; this having been done for a set of mutually exclusive exhaustive events $\{A_i\}$ we may then sum these probabilities to find the probability of the event B . Of course, this approach can be extended and we may wish to condition the event B on more than one event then uncondition each of these events suitably (by multiplying by the probability of the appropriate condition) and then sum all possible forms of all conditions. We will use this approach many times in the text.

We now come to the well-known *Bayes' theorem*. Once again we consider a set of events $\{A_i\}$, which are mutually exclusive and exhaustive. The theorem says

$$P[A_i | B] = \frac{P[B | A_i] P[A_i]}{\sum_{j=1}^n P[B | A_j] P[A_j]}$$

This theorem permits us to calculate the probability of one event conditioned on a second by calculating the probability of the second conditioned on the first and other similar terms.

simple example is in order here to illustrate some of these ideas. Consider you have just entered a gambling casino in Las Vegas. You approach a man who is known to have an identical twin brother; the twins cannot be distinguished. It is further known that one of the twins is an honest dealer as the second twin is a cheating dealer in the sense that when you play with the honest dealer you lose with probability one-half, whereas when you play with the cheating dealer you lose with probability p (if p is greater than one-half, he is cheating against you whereas if p is less than one-half he is cheating for you). Furthermore, it is equally likely that upon entering the casino you will find one or the other of these two dealers. Consider that you play one game with the particular twin whom you encounter and further you lose. Of course you are disappointed and you would now like to estimate the probability that the dealer you faced was in fact the cheat, for if you can establish that this probability is close to unity, you have a case for closing the casino. Let D_H be the event that you play with the honest dealer; let D_C be the event that you play with the cheating dealer; further let L be the event that you lose. What we are then asking for is $P[D_C | L]$. It is not immediately obvious how to make this calculation; however, if we apply Bayes' theorem the calculation itself is trivial, for

$$P[D_C | L] = \frac{P[L | D_C] P[D_C]}{P[L | D_C] P[D_C] + P[L | D_H] P[D_H]}$$

In the application of Bayes' theorem the collection of mutually exclusive events is the set $\{D_H, D_C\}$, for one of these two events must occur and both cannot occur simultaneously. Our problem is now trivial since the term on the right-hand side is easily calculated and leads us to

$$P[D_C | L] = \frac{p(\frac{1}{2})}{p(\frac{1}{2}) + (\frac{1}{2})(\frac{1}{2})} = \frac{2p}{2p + 1}$$

As the answer we were seeking and we find that the probability of having faced a cheating dealer, given that we lost in one play, ranges from 0 ($p = 0$) to 1 ($p = 1$). Thus, even if we know that the cheating dealer is completely honest ($p = 1$), we can only say that with probability 2/3 we faced this dealer given that we lost one play.

As a final word on elementary topics, let us remind the reader that the number of permutations of N objects taken K at a time is

$$\frac{N!}{(N - K)!} = N(N - 1) \dots (N - K + 1)$$

whereas the number of *combinations* of N things taken K at a time is denoted by $\binom{N}{K}$ and is given by

$$\binom{N}{K} = \frac{N!}{K!(N - K)!}$$

11.2. RANDOM VARIABLES

So far we have described a probability system which consists of the triplet $(S, \mathcal{E}, P)$, that is, a sample space, a set of events, and a probability assignment to the events of that sample space. We are now in a position to define the important concept of a *random variable*. A random variable is a variable whose value depends upon the outcome of a random experiment. Since the outcomes of our random experiments are represented as points $w \in S$ then to each such outcome w , we associate a real number $X(w)$, which is in fact the value the random variable takes on when the experimental outcome is w . Thus our (real) random variable $X(w)$ is nothing more than a function defined on the sample space, or if you will, a mapping from the points of the sample space into the (real) line.

As an example, let us consider the random experiment which consists of one play of a game of blackjack in Las Vegas. The sample space consists of all possible pairs of scores that can be obtained by the dealer and the player. Let us assume that we have grouped all such sample points into three (mutually exclusive) events of interest: lose (L), draw (D), or win (W). In order to complete the probability system we must assign probabilities to each of these events as follows*: $P[L] = 3/8$, $P[D] = 1/4$, $P[W] = 3/8$. Thus our probability system may be represented as in the Venn diagram of Figure 11.1. The numbers in parentheses are of course the probabilities. Now for the random variable $X(w)$. Let us assume that if we win the game we win \$5, if we draw we win \$0, and if we lose we win -\$5 (that is, we lose \$5). Let our winnings on this single play of blackjack be the random variable $X(w)$. We may therefore define this variable as follows:

$$X(w) = \begin{cases} +5 & w \in W \\ 0 & w \in D \\ -5 & w \in L \end{cases}$$

Similarly, we may represent this random variable as the mapping shown in Figure 11.2.

* This is the most difficult step in practice, that is, determining appropriate numbers to use in our model of the real world.

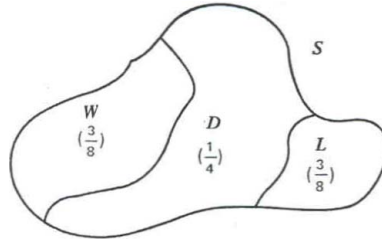


Figure 11.1 The probability system for the blackjack example.

The *domain* of the random variable $X(w)$ is the set of events $\mathcal{E}$ and the values it takes on form its *range*. We note in passing that the probability assignment P may itself be thought of as a random variable since it satisfies the definition; this particular assignment P , however, has further restrictions on it, namely, those given in Axioms (II.1)-(II.3).

We are mainly interested in describing the *probability* that the random variable $X(w)$ takes on certain values. To this end we define the following shorthand notation for events:

$$[X = x] \triangleq \{w : X(w) = x\} \quad (11.6)$$

We may discuss the probability of this event which we define as

$$P[X = x] = \text{probability that } X(w) \text{ is equal to } x$$

which is merely the sum of the probabilities associated with each point w for which $X(w) = x$. For our example we have

$$\begin{aligned} P[X = -5] &= 3/8 \\ P[X = 0] &= 1/4 \\ P[X = 5] &= 3/8 \end{aligned} \quad (11.7)$$

Another convenient form for expressing the probabilities associated with the random variable is the *probability distribution function* (PDF) also known

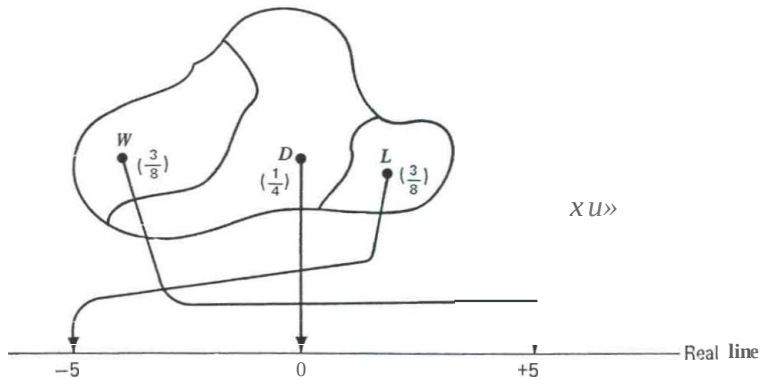


Figure 11.2 The random variable $X(w)$.

as the cumulative distribution function. For this purpose we define notation similar to that given in Eq. (II.6), namely,

$$[X \leq x] = \{w: X(w) \leq x\}$$

We then have that the PDF is defined as

$$F_X(x) \triangleq P[X \leq x]$$

which expresses the probability that the random variable X takes on a value less than or equal to x . The important properties of this function are

$$F_X(x) \geq 0 \quad (\text{II.5})$$

$$F_X(\infty) = 1$$

$$F_X(-\infty) = 0$$

$$F_X(b) - F_X(a) = P[a < X \leq b] \quad \text{for} \quad a < b \quad (\text{II.9})$$

$$F_X(b) \geq F_X(a) \quad \text{for} \quad a \leq b$$

Thus $F_X(x)$ is a nonnegative monotonically nondecreasing function with limits 0 and 1 at $-\infty$ and $+\infty$, respectively. In addition $F_X(x)$ is assumed to be continuous from the right. For our blackjack example we then have the function given in Figure II.3. We note that at points of discontinuity the PDF takes on the upper value (as indicated by the dot) since the function is piecewise continuous from the right. From Property (II.9) we may easily calculate the probability that our random variable lies in a given interval. Thus for our blackjack example, we may write $P[-2 < x \leq 6] = 5/8$, $P[-1 < x \leq 4] = 0$, and so on.

For purposes of calculation it is much more convenient to work with a function closely related to the PDF rather than with the PDF itself. Thus we

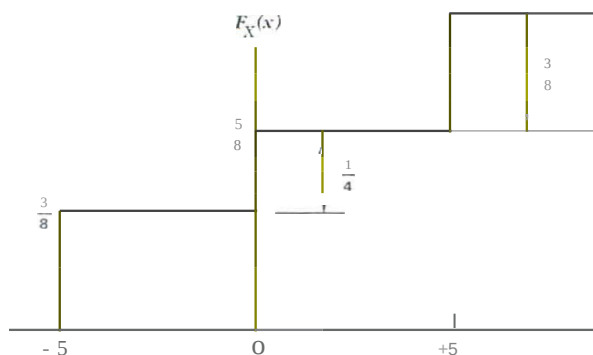


Figure 11.3 The PDF for the blackjack example.

to the definition of the *probability density function* (pdf) defined as follows:

$$f_X(x) \triangleq \frac{dF_X(x)}{dx} \quad (11.10)$$

First, we are immediately faced with the question of whether or not such a function exists and if so over what interval. We temporarily avoid that question and assume that $F_X(x)$ possesses a continuous derivative everywhere (this is false for our blackjack example). As we shall see later, it is possible to define the pdf even when the PDF contains jumps. We may "invert" (11.10) to yield

$$F_X(x) = \int_{-\infty}^x f_X(y) dy \quad (11.11)$$

Using this and Eq. (11.8) we have

$$f_X(x) \geq 0$$

Since $F_X(\infty) = 1$, we have from Eq. (11.11)

$$\int_{-\infty}^{\infty} f_X(x) dx = 1$$

Since the pdf is a function which when integrated over an interval gives the probability that the random variable X lies in that interval, namely, for $a < x < b$, we have

$$P(a < X \leq b) = \int_a^b f_X(x) dx$$

Let $a \rightarrow b$, and the axiom stated in Eq. (11.1) we see that this last equation implies

$$f_X(x) \geq 0$$

For an example, let us consider an exponentially distributed random variable defined as one for which

$$F_X(x) = \begin{cases} 1 - e^{-\lambda x} & 0 \leq x \\ 0 & x < 0 \end{cases} \quad (11.12)$$

where $\lambda > 0$.

The corresponding pdf is given by

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x} & 0 \leq x \\ 0 & x < 0 \end{cases} \quad (11.13)$$

For this example, the probability that the random variable lies between the values $a(>0)$ and $b(>a)$ may be calculated in either of the two following ways:

$$\begin{aligned} P[a < x \leq b] &= F_X(b) - F_X(a) \\ &= e^{-\lambda a} - e^{-\lambda b} \\ P[a < x \leq b] &= \int_a^b f_X(x) dx \\ &= e^{-\lambda a} - e^{-\lambda b} \end{aligned}$$

From our blackjack example we notice that the PDF has a derivative which is everywhere 0 except at the three critical points ($x = -5, 0, +5$). In order to complete the definition for the pdf when the PDF is discontinuous we recognize that we must introduce a function such that when it is integrated over the region of the discontinuity it yields a value equal to the size of the discontinuous jump; that is, in the blackjack example the probability density function must be such that when integrated from $-5 - \epsilon$ to $-5 + \epsilon$ (for small $\epsilon > 0$) it should yield a probability equal to $3/8$. Such a function has already been studied in Appendix I and is, of course, the impulse function (or Dirac delta function). Recall that such a function $u_0(x)$ is given by

$$u_0(x) = \begin{cases} \infty & x = 0 \\ 0 & x \neq 0 \end{cases}$$

$$\int_{-\infty}^{\infty} u_0(x) dx = 1$$

and also that it is merely the derivative of the unit step function as can be seen from

$$\int_{-\infty}^x u_0(y) dy = \begin{cases} 0 & x < 0 \\ 1 & x \geq 0 \end{cases}$$

Using the graphical notation in Figure 1.3, we may properly describe the pdf for our blackjack example as in Figure 11.4. We note immediately that this representation gives exactly the information we had in Eq. (11.7), and therefore the use of impulse functions is overly cumbersome for such problems. In particular if we define a discrete random variable as one that takes on values over a discrete set (finite or countable) then the use of the pdf* is a bit heavy and unnecessary although it does fit into our general definition in the obvious way. On the other hand, in the case of a random variable that takes on values over a continuum it is perfectly natural to use the pdf and in the

* In the discrete case, the function $P[X = x_k]$ is often referred to as the *probability mass function*. The generalization to the pdf leads one to the notion of a *mass density function*.

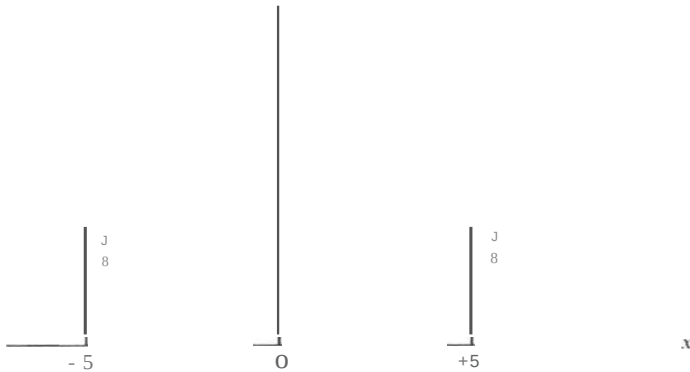


Figure 11.4 The pdf for the blackjack example.

case where there is also a nonzero probability that the random variable takes on a specific value (i.e., that the PDF contains jumps) then the use of the pdf is necessary as well as is the introduction of the impulse function to account for these points of accumulation. We are thus led to distinguish between a *discrete* random variable, a purely *continuous* random variable (one whose PDF is continuous and everywhere differentiable), and the third case of a *mixed* random variable which contains some discrete as well as continuous portions.* So, for example, let us consider a random variable that represents the lifetime of an automobile. We will assume that there is a finite probability, say of value p , that the automobile will be inoperable immediately upon delivery, and therefore will have a lifetime of length zero. On the other hand, if the automobile is operable upon delivery then we will assume that the remainder of its lifetime is exponentially distributed as given in Eqs. (11.12) and (11.13). Thus for this automobile lifetime we have a PDF and a pdf as given in Figure 11.5. Thus we clearly see the need for impulse functions in describing interesting random variables.

We have now discussed the notion of a probability system $(\mathcal{S}, S, P)$ and the notion of a random variable $X(\omega)$ defined upon the sample space $\mathcal{S}$. There is, of course, no reason why we cannot define *many* random variables on the same sample space. Let us consider the case of two random variables X and Y defined for some probability system $(\mathcal{S}, \mathcal{E}, P)$. In this case we have

- It can be shown that any PDF may be decomposed into a sum of three parts, namely, a pure jump function (containing only discontinuous jumps), a purely continuous portion, and a singular portion (which rarely occurs in distribution functions of interest and which will be considered no further in this text).

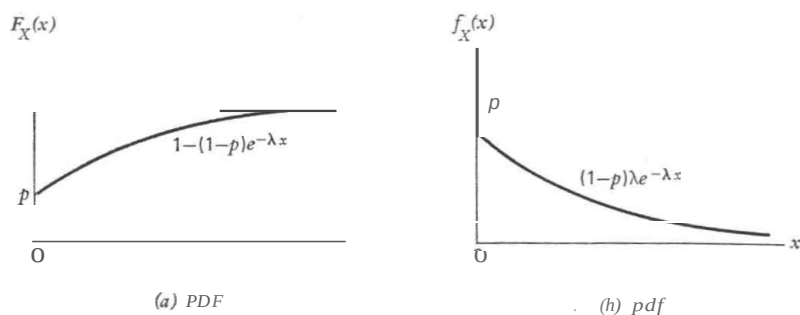


Figure U.5 PDF and pdf for automobile lifetime.

the natural extension of the PDF for two random variables, namely,

$$F_{XY}(x, y) \triangleq P[X \leq x, Y \leq y]$$

which is merely the probability that X takes on a value less than or equal to x at the same time Y takes on a value less than or equal to y ; that is, it is the sum of the probabilities associated with all sample points in the intersection of the two events $\{\omega: X(\omega) \leq x\}$, $\{\omega: Y(\omega) \leq y\}$. $F_{XY}(x, y)$ is referred to as the *joint* PDF. Of course, associated with this function is a joint probability density function defined as

$$f_{XY}(x, y) \triangleq \frac{d^2 F_{XY}(x, y)}{dx dy}$$

Given a joint pdf, one naturally inquires as to the "marginal" density function for one of the variables and this is clearly given by integrating over all possible values of the second variable, thus

$$f_X(x) = \int_{y=-\infty}^{\infty} f_{XY}(x, y) dy \quad (11.14)$$

We are now in a position to define the notion of *independence* between random variables. Two random variables X and Y are said to be independent if and only if

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

that is, if their joint pdf factors into the product of the one-dimensional pdf's. This is very much like the definition for two independent events as given in Eq. (11.5). However, for three or more random variables, the definition is essentially the same as for two, namely, $X_1, X_2, \dots, X_n$ are said to be independent random variables if and only if

$$f_{X_1 X_2 \dots X_n}(x_1, x_2, \dots, x_n) = f_{X_1}(x_1)f_{X_2}(x_2) \dots f_{X_n}(x_n)$$

This last is a much simpler test than that required for multiple *events* to be independent.

With more than one random variable, we can now define conditional distributions and densities as follows. For example, we could ask for the PDF of the random variable X conditioned on some given value of the random variable Y , which would be expressed as $P[X \leq x \mid Y = y]$. Similarly, the conditional pdf on X , given Y , is defined as

$$f_{X|Y}(x|y) \triangleq \frac{d}{dx} P[X \leq x \mid Y = y] = \frac{f_{XY}(x, y)}{f_Y(y)}$$

much as the definition for conditional probability of events.

To review again, we see that a random variable is defined as a mapping from the sample space for some probability system into the real line and from this mapping the PDF may easily be determined. Usually, however, a random variable is not given in terms of its sample space and the mapping, but rather directly in terms of its PDF or pdf.

It is possible to define one random variable Y in terms of a second random variable X , in which case Y would be referred to as a *function* of the random variable X . In its most general form we then have

$$Y = g(X) \quad (11.15)$$

where $g(\cdot)$ is some given function of its argument. Thus, once the value for X is determined, then the value for Y may be computed; **however**, the value for X depends upon the sample point w , and therefore so does the value of Y which we may therefore write as $Y = Y(w) = g(X(w))$. Given the random variable X and its PDF, one should be able to calculate the PDF for the random variable Y , once the function $g(\cdot)$ is known. In principle, the computation takes the following form:

$$F_Y(y) = P[Y \leq y] = P[\{w: g(X(w)) \leq y\}]$$

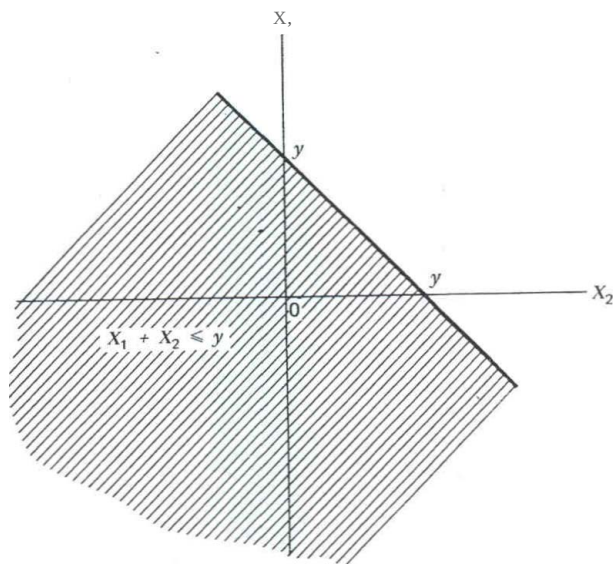
In general, this computation is rather complex.

One random variable may be a function of *many* other random variables rather than just one. A particularly important form which often arises is in fact the *sum* of a collection of independent random variables $\{X_i\}$ namely,

$$Y = \sum_{i=1}^n X_i \quad (11.16)$$

Let us derive the distribution function of the sum of *two* independent random variables ($n = 2$). It is clear that this distribution is given by

$$F_Y(y) = P[Y \leq y] = P[X_1 + X_2 \leq y]$$

Figure I/6 The integration region for $Y = X_1 + X_2 \leq y$.

We have the situation shown in Figure I1.6. Integrating over the indicated region we have

$$F_Y(y) = \int_{-\infty}^{\infty} \int_{-\infty}^{y-x_2} f_{X_1 X_2}(x_1, x_2) dx_1 dx_2$$

Due to the independence of X_1 and X_2 we then obtain the PDF for Y as

$$\begin{aligned} F_Y(y) &= \int_{-\infty}^{\infty} \left[\int_{-\infty}^{y-x_2} f_{X_1}(x_1) dx_1 \right] f_{X_2}(x_2) dx_2 \\ &= \int_{-\infty}^{\infty} F_{X_1}(y - x_2) f_{X_2}(x_2) dx_2 \end{aligned}$$

Finally, forming the pdf from this PDF, we have

$$f_Y(y) = \int_{-\infty}^{\infty} f_{X_1}(y - x_2) f_{X_2}(x_2) dx_2$$

This last equation is merely the *convolution* of the density functions for X_1 and X_2 and, as in Eq. (1.36), we denote this convolution operator (which is both associative and commutative) by an asterisk enclosed within a circle. Thus

$$f_Y(y) = f_{X_1}(y) \circledast f_{X_2}(y)$$

In a similar fashion, one easily shows for the case of arbitrary n that the pdf for Y as defined in Eq. (11.16) is given by the convolution of the pdf's for the X_i 's, that is,

$$f_Y(y) = f_{X_1}(y) \otimes f_{X_2}(y) \otimes \cdots \otimes f_{X_n}(y) \quad (11.17)$$

II.3. EXPECTATION

In this section we discuss certain measures associated with the PDF and the pdf for a random variable. These measures will in general be called *expectations* and they deal with integrals of the pdf. As we saw in the last section, the pdf involves certain difficulties in its definition, and these difficulties were handily resolved by the use of impulse functions. However, in much of the literature on probability theory and in most of the literature on queueing theory the use of impulses is either not accepted, not understood or not known; as a result, special care and notation has been built up to get around the problem of differentiating discontinuous functions. The result is that many of the integrals encountered are Stieltjes integrals rather than the usual Riemann integrals with which we are most familiar. Let us take a moment to define the Stieltjes integral. A *Stieltjes* integral is defined in terms of a non-decreasing function $F(x)$ and a continuous function $\varphi(x)$; in addition, two sets of points $\{f_k\}$ and $\{\xi_k\}$ such that $1_{-1} < \xi_k \leq f_k$ are defined and a limit is considered where $\max |f_k - f_{k-1}| \rightarrow 0$. From these definitions, consider the sum

$$\sum_k \varphi(\xi_k) [F(f_k) - F(f_{k-1})]$$

This sum tends to a limit as the intervals shrink to zero independent of the sets $\{f_k\}$ and $\{\xi_k\}$ and the limit is referred to as the Stieltjes integral of φ with respect to F . This Stieltjes integral is written as

$$\int \varphi(x) dF(x)$$

Of course, we recognize that the PDF may be identified with the function F in this definition and that $dF(x)$ may be identified with the pdf [say, $f(x)$] through

$$dF(x) = f(x) dx$$

by definition. Without the use of impulses the pdf may not exist; however, the Stieltjes integral will always exist and therefore it avoids the issue of impulses. However, in this text we will feel free to incorporate impulse functions and therefore will work with both the Riemann and Stieltjes integrals; when impulses are permitted in the function $f(x)$ we then have the

following identity:

$$\int \varphi(x) dF(x) = \int \varphi(x) f(x) dx$$

We will use both notations throughout the text in order to familiarize the student with the more common Stieltjes integral for queueing theory, as well as with the more easily manipulated Riemann integral with impulse functions. Having said all this we may now introduce the definition of expectation.

The *expectation* of a real random variable $X(w)$ denoted by $E[X]$ and also by $\bar{X}$ is given by the following:

$$E[X] \triangleq \bar{X} \triangleq \int_{-\infty}^{\infty} x dF_X(x) \quad (11.18)$$

This last is given in the form of a Stieltjes integral; in the form of a Riemann integral we have, of course,

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

The expectation of X is also referred to as the *mean* or *average value* of X . We may also write

$$E[X] = \int_0^{\infty} [1 - F_X(x)] dx - \int_{-\infty}^0 F_X(x) dx$$

which, upon integrating by parts, is easily shown to be equal to Eq. (11.18) so long as $E[X] < \infty$. Similarly, for X a nonnegative random variable, this form becomes

$$E[X] = \int_0^{\infty} [1 - F_X(x)] dx \quad X \geq 0$$

In general, the expectation of a random variable is equal to the product of the value the random variable may take on and the probability it takes on this value, summed (integrated) over all possible values.

Now let us consider once again a new random variable Y , which is a function of our first random variable X , namely, as in Eq. (IU5)

$$Y = g(X)$$

We may define the expectation $E_Y[Y]$ for Y in terms of its PDF just as we did for X ; the subscript Y on the expectation is there to distinguish expectation with respect to Y as opposed to any other random variables (in this case X). Thus we have

$$E_Y[Y] = \int_{-\infty}^{\infty} y f_Y(y) dy$$

This last computation requires that we find either $F_Y(Y)$ or $f_Y(y)$, which as mentioned in the previous section, may be a rather complex computation. However, *the fundamental theorem of expectation* gives a much more straightforward calculation for this expectation in terms of distribution of the underlying random variable X , namely,

$$\begin{aligned} E_Y[y] &= E_X[g(X)] \\ &= \int_{-\infty}^{\infty} g(x) f_X(x) dx \end{aligned}$$

We may define the expectation of the *sum* of two random variables given by the following obvious generalization of the one-dimensional case :

$$\begin{aligned} E[X + Y] &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x + y) f_{XY}(x, y) dx dy \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x f_{XY}(x, y) dx dy + \int_{-\infty}^{\infty} y f_{XY}(x, y) dx dy \\ &= \int_{-\infty}^{\infty} x f_X(x) dx + \int_{-\infty}^{\infty} y f_Y(y) dy \\ &= E[X] + E[Y] \end{aligned} \quad (11.19)$$

This may also be written as $(X + Y) = \bar{X} + \bar{Y}$. In going from the second line to the third line we have taken advantage of Eq. (11.14) of the previous section in which the marginal density was defined from the joint density. We have shown the very important result, that *the expectation of the sum of two random variables is always equal to the sum of the expectations of each—this is true whether or not these random variables are independent*. This very nice property comes from the fact that the expectation operator is a linear operator. The more general statement of this property for any number of random variables, independent or not, is that *the expectation of the sum is always equal to the sum of the expectations*, that is,

$$E[X_1 + X_2 + \dots + X_n] = E[X_1] + E[X_2] + \dots + E[X_n]$$

A similar question may be asked about the *product* of two random variables, that is,

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

In the special case where the two random variables X and Y are *independent*, we may write the pdf for this joint random variable as the product of the pdf's for the individual random variables, thus obtaining

$$E[XY] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xy f_X(x) f_Y(y) dx dy = E[X]E[Y] \quad (11.20)$$

This last equation (which may also be written as $\underline{XY} = \bar{X}\bar{Y}$) states that the expectation of the product is equal to the product of the expectations if the random variables are independent. A result similar to that expressed in Eq. (11.20) applies also to *functions* of independent random variables. That is, if we have two independent random variables X and Y and functions of each denoted by $g(X)$ and $h(Y)$, then by arguments exactly the same as those leading to Eq. (11.20) we may show

$$E[g(X)h(Y)] = E[g(X)]E[h(Y)] \quad (11.21)$$

Often we are interested in the expectation of the n th power of a random variable. In fact, this is so common that a special name has been coined so that the expected value of the n th power of a random variable is referred to as its *n th moment*. Thus, by definition (really this follows from the fundamental theorem of expectation), the n th moment of X is given by

$$E[X^n] \triangleq \underline{X^n} \triangleq \int_{-\infty}^{\infty} x^n f_X(x) dx$$

Furthermore, the *n th central moment* of this random variable is given as follows:

$$\overline{(X - \bar{X})^n} \triangleq \int_{-\infty}^{\infty} (x - \bar{X})^n f_X(x) dx$$

The *n th central moment* may be expressed in terms of the first n moments themselves; to show this we first write down the following identity making use of the binomial theorem

$$(X - \bar{X})^n = \sum_{k=0}^n \binom{n}{k} X^k (-\bar{X})^{n-k}$$

Taking expectations on both sides we then have

$$\begin{aligned} \overline{(X - \bar{X})^n} &= \sum_{k=0}^n \binom{n}{k} \overline{X^k (-\bar{X})^{n-k}} \\ &= \sum_{k=0}^n \binom{n}{k} \bar{X}^{n-k} \overline{X^k} \end{aligned} \quad (11.22)$$

In going from the first to the second line in this last equation we have taken advantage of the fact that the expectation of a sum is equal to the sum of the expectations and that the expectation of a constant is merely the constant itself. Now for a few observations. First we note that the 0th moment of a random variable is just unity. Also, the 0th central moment must be one. The first central moment must be 0 since

$$\overline{(X - \bar{X})} = \bar{X} - \bar{X} = 0$$

var

In
(l
fil
tc
o
q
d

The second central moment is extremely important and is referred to as the *variance*; a special notation has been adopted for the variance and is given by

$$\begin{aligned}\sigma_X^2 &\triangleq (X - \bar{X})^2 \\ &\triangleq \overline{X^2} - (\bar{X})^2\end{aligned}$$

In the second line of this last equation we have taken advantage of Eq. (II.22) and have expressed the variance (a central moment) in terms of the first two moments themselves. The square root of the variance σ_X is referred to as the *standard deviation*. The ratio of the standard deviation to the mean of a random variable is a most important quantity in statistics and also in queueing theory; this ratio is referred to as the *coefficient of variation* and is denoted by

$$C_X \triangleq \frac{\sigma_X}{\bar{X}} \quad (11.23)$$

II.4. TRANSFORMS, GENERATING FUNCTIONS, AND CHARACTERISTIC FUNCTIONS

In probability theory one encounters a variety of functions (in particular, expectations) all of which are close relatives of each other. Included in this class is the *characteristic function* of a random variable, its *moment generating function*, the *Laplace transform of its probability density function*, and its *probability generating function*. In this section we wish to define and distinguish these various forms and to indicate a common central property that they share.

The *characteristic function* of a random variable X , denoted by $\phi_X(u)$, is given by

$$\begin{aligned}\phi_X(u) &\triangleq E[e^{juX}] \\ &= \int_{-\infty}^{\infty} e^{jux} f_X(x) dx\end{aligned}$$

where $j = \sqrt{-1}$ and where u is an arbitrary real variable. (Note that except for the sign of the exponent, the characteristic function is the Fourier transform of the pdf for X). Clearly,

$$|\phi_X(u)| \leq \int_{-\infty}^{\infty} |e^{jux}| |f_X(x)| dx$$

and since $|e^{jux}| = 1$, we have

$$|\phi_X(u)| \leq \int_{-\infty}^{\infty} f_X(x) dx$$

which shows that

$$|\phi_X(u)| \leq 1$$

An important property of the characteristic function may be seen by expanding the exponential in the integrand in terms of its power series and then integrating each term separately as follows:

$$\begin{aligned}\phi_X(u) &= \int_{-\infty}^{\infty} f_X(x) \left[1 + ju_x + \frac{(ju_x)^2}{2!} + \dots \right] dx \\ &= 1 + jU\bar{X} + \frac{(ju)^2}{2!} \overline{X^2} + \dots\end{aligned}$$

From this expansion, we see that the characteristic function is expressed in terms of all the moments of X . Now, if we set $u = 0$ we find that $\phi_X(0) = 1$. Similarly, if we first form $d\phi_X(u)/du$ and then set $u = 0$, we obtain $j\bar{X}$. Thus, in general, we have

$$\left. \frac{d^n \phi_X(u)}{du^n} \right|_{u=0} = j^n \bar{X}^n \quad (11.24)$$

This last important result gives a rather simple way for calculating a constant times the n th moment of the random variable X .

Since this property is frequently used, we find it convenient to adopt the following simplified notation (consistent with that in Eq. 11.37) for the n th derivative of an arbitrary function $g(x)$, evaluated at some fixed value $x = x_0$:

$$g^{(n)}(x_0) \triangleq \left. \frac{d^n g(x)}{dx^n} \right|_{x=x_0} \quad (11.25)$$

Thus the result in Eq. (11.24) may be rewritten as

$$\phi_X^{(n)}(0) = j^n \bar{X}^n.$$

The *moment generating function* denoted by $M_X(v)$ is given below along with the appropriate differential relationship that yields the n th moment of X directly.

$$\begin{aligned}M_X(v) &\triangleq E[e^{vx}] \\ &= \int_{-\infty}^{\infty} e^{vx} f_X(x) dx \\ M_X^{(n)}(0) &= \bar{X}^n\end{aligned}$$

where v is a real variable. From this last property it is easy to see where the name "moment generating function" comes from. The derivation of this moment relationship is the same as that for the characteristic function.

Another important and useful function is the *Laplace transform of the pdf* of a random variable X . We find it expedient to use a notation now in which the PDF for a random variable is labeled in a way that identifies the random variable without the use of subscripts. Thus, for example, if we have a

random variable X , which represents, say, the interarrival time between adjacent customers to a system, then we define $A(x)$ to be the PDF for X ;

$$A(x) = P[X \leq x]$$

where the symbol A is keyed to the word "Arrival." Further, the pdf for this example would be denoted $a(x)$. Finally, then, we denote the Laplace transform of $a(x)$ by $A^*(s)$ and it is given by the following:

$$\begin{aligned} A^*(s) &\triangleq E[e^{-sx}] \\ &\triangleq \int_{-\infty}^{\infty} e^{-sx} a(x) dx \end{aligned}$$

where s is a complex variable. Here we are using the "two-sided" transform; however, as mentioned in Section 1.3, since most of the random variables we deal with are nonnegative, we often write

$$A^*(s) = \int_{0^-}^{\infty} e^{-sx} a(x) dx$$

The reader should take special note that the lower limit 0 is defined as 0^- ; that is, the limit comes in from the left so that we specifically mean to include any impulse functions at the origin. In the fashion identical to that for the moment generating function and for the characteristic function, we may find the moments of X through the following formula:

$$A^{*(n)}(0) = (-1)^n n! \mu_n \quad (11.26)$$

For nonnegative random variables

$$|A^*(s)| \leq \int_0^{\infty} |e^{-sx}| |a(x)| dx$$

But the complex variable s consists of a real part $\operatorname{Re}(s) = \sigma$ and an imaginary part $\operatorname{Im}(s) = \omega$ such that $s = \sigma + j\omega$. Then we have

$$\begin{aligned} |e^{-sx}| &= |e^{-\sigma x} e^{-j\omega x}| \\ &\leq |e^{-\sigma x}| |e^{-j\omega x}| \\ &= |e^{-\sigma x}| \end{aligned}$$

Moreover, for $\operatorname{Re}(s) \geq 0$, $|e^{-\sigma x}| \leq 1$ and so we have from these last two equations and from $\int_0^{\infty} a(x) dx = 1$,

$$|A^*(s)| \leq 1 \quad \operatorname{Re}(s) \geq 0$$

It is clear that the three functions $\phi_X(u)$, $M_X(v)$, $A^*(s)$ are all close relatives of each other. In particular, we have the following relationship:

$$\phi_X(sj) = M_X(-s) = A^*(s)$$

Thus we are not surprised that the moment generating properties (by differentiation) are so similar for each; this property is the central property that we will take advantage of in our studies. Thus the n th moment of X is calculable from any of the following expressions:

$$\begin{aligned}\underline{X}^n &= j^{-n} \phi_X^{(n)}(0) \\ \underline{X}^n &= M_X^{(n)}(0) \\ \underline{X}^n &= (-1)^n A^{*(n)}(0)\end{aligned}$$

It is perhaps worthwhile to carry out an example demonstrating these properties. Consider the continuous random variable X , which represents, say, the interarrival time of customers to a system and which is exponentially distributed, that is,

$$f_X(x) = a(x) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0 \\ 0 & x < 0 \end{cases}$$

By direct substitution into the defining integrals we find immediately that

$$\begin{aligned}\phi_X(u) &= \frac{\lambda}{\lambda - ju} \\ M_X(v) &= \frac{\lambda}{\lambda - v} \\ A^*(s) &= \frac{\lambda}{\lambda + s}\end{aligned}$$

It is always true that

$$\phi_X(0) = M_X(0) = A^*(0) = 1$$

and, of course, this checks out for our example as well. Using our expression for the first moment we find through any one of our three functions that

$$\bar{X} = \frac{1}{\lambda}$$

and we may also verify that the second moment may be calculated from any of the three to yield

$$\overline{X^2} = \frac{2}{\lambda^2}$$

and so it goes in calculating all of the moments.

In the case of a *discrete* random variable described, for example, by

$$g_k = P[X = k]$$

we make use of the *probability generating function* denoted by $G(z)$ as follows:

$$\begin{aligned} G(z) &\triangleq E[z^X] \\ &= \sum_k z^k g_k \end{aligned} \quad (11.27)$$

where z is a complex variable. It should be clear from our discussion in Appendix I that $G(z)$ is nothing more than the z -transform of the discrete sequence g_k . As with the continuous transforms, we have for $|z| \leq 1$

$$\begin{aligned} |G(z)| &\leq \sum_k |z|^k |g_k| \\ &\leq \sum_k g_k \end{aligned}$$

and so

$$|G(z)| \leq 1 \quad \text{for } |z| \leq 1 \quad (11.28)$$

Note that the first derivative evaluated at $z = 1$ yields the first moment of X

$$G'(1) = \bar{X} \quad (11.29)$$

and that the second derivative yields

$$G''(1) = \underline{X^2} - \bar{X}$$

in a fashion similar to that for continuous random variables.* Note that in all cases

$$G(1) = 1$$

Let us apply these methods to the blackjack example considered earlier in this appendix. Working either with Eq. (11.7), which gives the probability of various winnings or with the impulsive pdf given in Figure 11.4, we find that the probability generating function for the number of dollars won in a game of blackjack is given by

$$G(z) = \frac{3}{8} z^{-5} + \frac{1}{4} + \frac{3}{8} z^5$$

We note here that, of course, $G(1) = 1$ and further, that the mean winnings may be calculated as

$$\bar{X} = G'(1) = 0$$

Let us now consider the sum of n independent variables X_i , namely, $y = \sum_{i=1}^n X_i$ as defined in Eq. (11.16). If we form the characteristic function

* Thus we have that $\sigma_X^2 = G''(1) + G'(1) - [G'(1)]^2$.

for Y , we have by definition

$$\begin{aligned}\phi_Y(u) &\triangleq E[e^{iuY}] \\ &= E\left[e^{ju \sum_{i=1}^n X_i}\right] \\ &= E[e^{iuX_1} e^{iuX_2} \dots e^{iuX_n}]\end{aligned}$$

Now in Eq. (II.21) we showed that the expectation of the product of functions of independent random variables is equal to the product of the expectations of each function separately; applying this to the above we have

$$\phi_Y(u) = E[e^{iuX_1}] E[e^{iuX_2}] \dots E[e^{iuX_n}]$$

Of course the right-hand side of this equation is just a product of characteristic functions, and so

$$\phi_Y(u) = \phi_{X_1}(u) \phi_{X_2}(u) \dots \phi_{X_n}(u) \quad (11.30)$$

In the case where each of the X_i is *identically distributed*, then, of course, the characteristic functions will all be the same, and so we may as well drop the subscript on X_i and conclude

$$\phi_Y(u) = [\phi_X(u)]^n \quad (11.31)$$

We have thus shown that the characteristic function of a sum of n identically distributed independent random variables is the n th power of the characteristic function of the individual random variable itself. This important result also applies to our other transforms, namely, the moment generating function, the Laplace transform and the z -transform. It is this significant property that accounts, in no small way, for the widespread use of transforms in probability theory and in the theory of stochastic processes.

Let us say a few more words now about sums of independent random variables. We have seen in Eq. (II. 17) that the pdf of a sum of independent variables is equal to the convolution of the pdf for each; also, we have seen in Eq. (11.30) that the transform of the sum is equal to the product of the transforms for each. From Eq. (II. 19) it is clear (regardless of the independence) that the expectation of the sum equals the sum of the expectations, namely,

$$\bar{Y} = \bar{X}_1 + \bar{X}_2 + \dots + \bar{X}_n \quad (11.32)$$

For $n = 2$ we see that the second moment of Y must be

$$\underline{Y}^2 = (X_1 + X_2)^2 = \underline{X_1^2} + 2\underline{X_1 X_2} + \underline{X_2^2}$$

And also in this case

$$(\bar{Y})^2 = (\bar{X}_1 + \bar{X}_2)^2 = (\bar{X}_1)^2 + 2\bar{X}_1\bar{X}_2 + (\bar{X}_2)^2$$

finding the variance of Y and then using these last two equations we have

$$\begin{aligned}\sigma_Y^2 &= Y^2 - (\bar{Y})^2 \\ &= \underline{X_1^2} - (\bar{X}_1)^2 + \underline{X_2^2} - (\bar{X}_2)^2 + 2(\underline{X_1 X_2} - \bar{X}_1 \bar{X}_2) \\ &= \sigma_{X_1}^2 + \sigma_{X_2}^2 + 2(\underline{X_1 X_2} - \bar{X}_1 \bar{X}_2)\end{aligned}$$

if X_1 and X_2 are also independent, then $\underline{X_1 X_2} = \bar{X}_1 \bar{X}_2$, giving the final result

$$\sigma_Y^2 = \sigma_{X_1}^2 + \sigma_{X_2}^2$$

In a similar fashion it is easy to show that the variance of the sum of n independent random variables is equal to the sum of the variances of each, that is,

$$\sigma_Y^2 = \sigma_{X_1}^2 + \sigma_{X_2}^2 + \cdots + \sigma_{X_n}^2$$

Continuing with sums of independent random variables let us now assume that the number of these variables that are to be summed together is itself a random variable, that is, we define

$$Y = \sum_{i=1}^N X_i$$

where $\{X_i\}$ is a set of identically distributed independent random variables, each with mean $\bar{X}$ and variance σ_X^2 , and where N is also a random variable with mean $\bar{N}$ and variance σ_N^2 , respectively; we assume that N is also independent of the X_i . In this case, $F_Y(y)$ is said to be a *compound* distribution. Let us now find $Y^*(s)$, which is the Laplace transform of the pdf of Y . By definition of the transform and due to the independence of all random variables we may write down

$$\begin{aligned}Y^*(s) &= E\left[e^{-s \sum_{i=1}^N X_i}\right] \\ &= \sum_{n=0}^{\infty} E\left[e^{-s \sum_{i=1}^n X_i}\right] P[N = n] \\ &= \sum_{n=0}^{\infty} E[e^{-s X_1}] \cdots E[e^{-s X_n}] P[N = n]\end{aligned}$$

since $\{X_i\}$ is a set of identically distributed random variables, we have

$$Y^*(s) = \sum_{n=0}^{\infty} [X^*(s)]^n P[N = n] \quad (11.33)$$

where we have denoted the Laplace transform of the pdf for each of the X_i by $X^*(s)$. The final expression given in Eq. (11.33) is immediately recognized

as the z -transform for the random variable N , which we choose to denote by $N(z)$ as defined in Eq. (11.27); in Eq. (11.33), z has been replaced by $X^*(s)$. Thus we finally conclude

$$Y^*(s) = N(X^*(s)) \quad (11.34)$$

Thus a *random* sum of identically distributed independent random variables has a transform that is related to the transforms of the sum's random variables and of the number of terms in the sum, as given above. Let us now find an expression similar to that in Eq. (11.32); in that equation for the case of identically distributed X_i we had $\bar{Y} = n\bar{X}$, where n was a given constant. Now, however, the number of terms in the sum is a random quantity and we must find the new mean $\bar{Y}$. We proceed by taking advantage of the moment generating properties of our transforms [Eq. (11.26)]. Thus differentiating Eq. (11.34), setting $s = 0$, and then taking the negative of the result we find

$$\bar{Y} = \bar{N}\bar{X}$$

which is a perfectly reasonable result. Similarly, one can find the variance of this random sum by differentiating twice and then subtracting off the mean squared to obtain

$$\sigma_Y^2 = \bar{N}\sigma_X^2 + (\bar{X})^2\sigma_N^2$$

This last result perhaps is not so intuitive.

11.5. INEQUALITIES AND LIMIT THEOREMS

In this section we present some of the classical inequalities and limit theorems in probability theory.

Let us first consider bounding the probability that a random variable exceeds some value. If we know only the mean value of the random variable, then the following *Markov inequality* can be established for a nonnegative random variable X :

$$P[X \geq x] \leq \frac{\bar{X}}{x}$$

Since only the mean value of the random variable is utilized, this inequality is rather weak. The *Chebyshev inequality* makes use of the mean and variance and is somewhat tighter; it states that for any $x > 0$,

$$P[|X - \bar{X}| \geq x] \leq \frac{\sigma_X^2}{x^2}$$

Other simple inequalities involve moments of two random variables, as follows: First we have the *Cauchy-Schwarz inequality*, which makes a statement about the expectation of a product of random variables in terms of

the second moments of each.

$$(\overline{XY})^2 \leq \overline{X^2} \overline{Y^2} \quad (11.35)$$

A generalization of this last is *Holder's inequality*, which states for $\alpha > 1$, $\beta > 1$, $\alpha^{-1} + \beta^{-1} = 1$, and $X > 0$, $Y > 0$ that

$$\overline{XY} \leq (\overline{X^\alpha})^{1/\alpha} (\overline{Y^\beta})^{1/\beta}$$

whenever the indicated expectations exist. Note that the Cauchy-Schwartz inequality is the (important) special case in which $\alpha = \beta = 2$. The *triangle inequality* relates the expectation of the absolute value of a sum to the sum of the expectations of the absolute values, namely,

$$\overline{|X + Y|} \leq \overline{|X|} + \overline{|Y|}$$

A generalization of the triangle inequality, which is known as the *C.-inequality*, is

$$\overline{|X + Y|^r} \leq C_r [\overline{|X|^r} + \overline{|Y|^r}]$$

where

$$C_r = \begin{cases} 1 & 0 < r \leq 1 \\ 2^{r-1} & 1 < r \end{cases}$$

Next we bound the expectation of a *convex* function g of an arbitrary random variable X (whose first moment $\overline{X}$ is assumed to exist). A convex function $g(x)$ is one that lies on or below all of its chords, that is, for any $x_1 \leq x_2$, and $0 \leq \alpha \leq 1$

$$g(\alpha x_1 + (1 - \alpha)x_2) \leq \alpha g(x_1) + (1 - \alpha)g(x_2)$$

For such convex functions g and random variables X , we have *Jensen's inequality* as follows:

$$g(\overline{X}) \geq g(\overline{X})$$

When we deal with sums of random variables, we find that some very nice limiting properties exist. Let us once again consider the sum of n independent identically distributed random variables X_i but let us now divide that sum by the number of terms n , thusly

$$W_n = \frac{1}{n} \sum_{i=1}^n X_i$$

This arithmetic mean is often referred to as the *sample mean*. We assume that each of the X_i has a mean given by $\overline{X}$ and a variance σ_X^2 . From our earlier discussion regarding means and variances of sums of independent

random variables we have

$$\begin{aligned} W_n &= \bar{X} \\ \sigma_{W_n}^2 &= \frac{\sigma_X^2}{n} \end{aligned}$$

If we now apply the Chebyshev inequality to the random variable W_n and make use of these last two observations, we may express our bound in terms of the mean and variance of the random variable X itself thusly

$$P[|W_n - \bar{X}| \geq x] \leq \frac{\sigma_X^2}{nx^2} \quad (11.36)$$

This very important result says that the arithmetic mean of the sum of n independent and identically distributed random variables will approach its expected value as n increases. This is due to the decreasing value of σ_X^2/nx^2 as n grows (σ_X^2/x^2 remains constant). In fact, this leads us directly to the *weak law of large numbers*, namely, that for any $\epsilon > 0$ we have

$$\lim P[|W_n - \bar{X}| \geq \epsilon] = 0$$

The *strong law of large numbers* states that

$$\lim W_n = \bar{X} \quad \text{with probability one}$$

Once again, let us consider the sum of n independent identically distributed random variables X_i each with mean $\bar{X}$ and variance σ_X^2 . The *central limit theorem* concerns itself with the normalized random variable Z ; defined by

$$Z = \frac{\sum_{i=1}^n X_i - n\bar{X}}{\sigma_X \sqrt{n}} \quad (11.37)$$

and states that the PDF for Z , tends to the standard *normal distribution* as n increases; that is, for any real number x we have

$$\lim_{n \rightarrow \infty} P[Z \leq x] = \Phi(x)$$

where

$$\Phi(x) \triangleq \int_{-\infty}^x \frac{1}{\sqrt{2\pi}} e^{-y^2/2} dy$$

That is, the appropriately normalized sum of a large number of independent random variables tends to a Gaussian, or a normal distribution. There are many other forms of the central limit theorem that deal, for example, with dependent random variables.

A rather sophisticated means for bounding the tail of the sum of a large number of independent random variables is available in the form of the *Chernoff bound*. It involves an inequality similar to the Markov and Chebyshev inequalities, but makes use of the entire distribution of the random variable itself (in particular, the moment generating function). Thus let us consider the sum of n independent identically distributed random variables X_i as given by

$$Y = \sum_{i=1}^n X_i$$

From Eq. (II.31) we know that the moment generating function for Y , $M_Y(v)$, is related to the moment generating function for each of the random variables X_i [namely, $M_X(v)$] through the relationship

$$M_Y(v) = [M_X(v)]^n \quad (11.38)$$

As with our earlier inequalities, we are interested in the probability that our sum exceeds a certain value, and this may be calculated as

$$P(Y \geq y) = \int_y^\infty f_Y(w) dw \quad (11.39)$$

Clearly, for $v \geq 0$ we have that the unit step function [see Eq. (1.33)] is bounded above by the following exponential:

$$u_-(w - y) \leq e^{v(w - y)}$$

Applying this inequality to Eq. (11.39) we have

$$P(Y \geq y) \leq e^{-vy} \int_{-\infty}^\infty e^{vw} f_Y(w) dw \quad \text{for } v \geq 0$$

However, the integral on the right-hand side of this equation is merely the moment generating function for Y , and so we have

$$P[Y \geq y] \leq e^{-vy} M_Y(v) \quad v \geq 0 \quad (11.40)$$

Let us now define the "semi-invariant" generating function

$$\gamma(v) \triangleq \log M(v)$$

(Here we are considering natural logarithms.) Applying this definition to Eq. (11.38) we immediately have

$$\gamma_Y(v) = n\gamma_X(v)$$

and applying these last two to Eq. (11.40) we arrive at

$$P(Y \geq y) \leq e^{-vy + n\gamma_X(v)} \quad v \geq 0$$

Since this last is good for any value of v (≥ 0), we should choose v to create the tightest possible bound; this is simply carried out by differentiating the exponent and setting it equal to zero. We thus find the optimum relationship between v and y as

$$y = n\gamma_X^{(1)}(v) \quad (\text{H.41})$$

Thus the Chernoff bound for the tail of a density function takes the final form*

$$P[Y \geq n\gamma_X^{(1)}(v)] \leq e^{n[\gamma_X(v) - v\gamma_X^{(1)}(v)]} \quad v \geq 0 \quad (\text{H.42})$$

It is perhaps worthwhile to carry out an example demonstrating the use of this last bounding procedure. For this purpose, let us go back to the second paragraph in this appendix, in which we estimated the odds that at least 490,000 heads would occur in a million tosses of a fair coin. Of course, that calculation is the same as calculating the probability that no more than 510,000 heads will occur in the same experiment, assuming the coin is fair. In this example the random variable X may be chosen as follows

$$X = \begin{cases} 1 & \text{heads} \\ 0 & \text{tails} \end{cases}$$

Since Y is the sum of a million trials of this experiment, we have that $n = 10^6$, and we now ask for the complementary probability that Y add up to 510,000 or more, namely, $P[Y \geq 510,000]$. The moment-generating function for X is

$$M_X(v) = \frac{1}{2} + \frac{1}{2}e^v$$

and so

$$\gamma_X(v) = \log \frac{1}{2} (1 + e^v)$$

Similarly

$$\gamma_X^{(1)}(v) = \frac{e^v}{1 + e^v}$$

From our formula (H.41) we then must have

$$nylll(v) = 10^6 \frac{e^v}{1 + e^v} = 510,000 = y$$

Thus we have

$$e^v = \frac{51}{49}$$

and

$$v = \log \frac{51}{49}$$

* The same derivation leads to a bound on the "lower tail" in which all three inequalities from Eq. (H.42) face thusly: $\leq$. For example $v \leq 0$.

Thus we see typically how v might be calculated. Plugging these values back into Eq. (11.42) we conclude.

$$P[Y \geq 510,000] \leq e^{10^6 (10^6 (50/4) - 0.51) \log(51/4)}.$$

This computation shows that the probability of exceeding 510,000 heads in a million tosses of a fair coin is less than 10^{-88} (this is where the number in our opening paragraphs comes from). An alternative way of carrying out this computation would be to make use of the central limit theorem. Let us do so as an example. For this we require the calculation of the mean and variance of X which are easily seen to be $\bar{X} = 1/2$, $\sigma_X^2 = 1/4$. Thus from Eq. (11.37) we have

$$Z = \frac{Y - 10^6(1/2)}{(1/2)10^3}$$

If we require Y to be greater than 510,000, then we are requiring that Z be greater than 20. If we now go to a table of the cumulative normal distribution, we find that

$$P[Z \geq 20] = 1 - \Phi(20) \simeq 25 \times 10^{-4}.$$

Again we see the extreme implausibility of such an event occurring. On the other hand, the Chebyshev inequality, as given in Eq. (11.36), yields the following;

$$p[|W - 1/2| > 0.01] \leq \frac{0.25}{10^6 \cdot 10^{-4}} = 25 \times 10^{-4}$$

This result is twice as large as it should be for our calculation since we have effectively calculated both tails (namely, the probability that more than 510,000 or less than 490,000 heads would occur); thus the appropriate answer for the Chebyshev inequality would be that the probability of exceeding 510,000 heads is less than or equal to 12.5×10^{-4} . Note what a poor result this inequality gives compared to the central limit theorem approximation, which in this case is comparable to the Chernoff bound.

11.6. STOCHASTIC PROCESSES

It is often said that queueing theory is part of the theory of applied stochastic processes. As such, the main portion of this text is really the proper sequel to this section on stochastic processes; here we merely state some of the fundamental definitions and concepts.

We begin by considering a probability system $(S, \mathcal{E}, P)$, which consists of a sample space S , a set of events $\{A, E, \dots\}$, and a probability measure P . In addition, we have already introduced the notion of a random variable

$X(w)$. A *stochastic process* may be defined as follows: For each sample point $w \in S$ we assign a time function $X(t, w)$. This family of functions forms a stochastic process; alternatively, we may say that for each t included in some appropriate parameter set, we choose a random variable $X(t, w)$. This is a collection of random variables depending upon t . Thus a stochastic process (or random function) is a function $X(t)$ whose values are random variables. An example of a random process is the sequence of closing prices for a given security on the New York Stock Exchange; another example is the temperature at a given point on the earth as a function of time.

We are immediately confronted with the problem of completely specifying a random process $X(t)$. For this purpose we define, for each allowed t , a PDF, which we denote by $F_X(x, t)$ and which is given by

$$F_X(x, t) = P[X(t) \leq x]$$

Further we define for each of n allowable t , $\{t_1, t_2, \dots, t_n\}$ a joint PDF, given by

$$F_{X_1 X_2 \dots X_n}(x_1, x_2, \dots, x_n; t_1, t_2, \dots, t_n) \\ \triangleq P[X(t_1) \leq x_1, X(t_2) \leq x_2, \dots, X(t_n) \leq x_n]$$

and we use the vector notation $F_X(x; t)$ to denote this function.

A stochastic process $X(t)$ is said to be *stationary* if all $F_X(x, t)$ are invariant to shifts in time; that is, for any given constant τ the following holds:

$$F_X(x; t + T) = F_X(x; t)$$

where the notation $t + \tau$ implies the vector $(t_1 + \tau, t_2 + \tau, \dots, t_n + \tau)$. Of most interest in the theory of stochastic processes are these stationary random functions.

In order to completely specify a stochastic process, then, one must give $F_X(x; t)$ for all possible subsets of $\{X_i\}$, $\{t_i\}$, and all n . This is a monstrous task in general! Fortunately, for many of the interesting stochastic processes, it is possible to provide this specification in very simple terms.

Some other definitions are in order. The first is the definition of the pdf for a stochastic process, and this is defined by

$$f_X(x; t) \triangleq \frac{\partial F_X(x; t)}{\partial X}$$

Second, we often discuss the mean value of a stochastic process given by

$$\underline{X(t)} = E[X(t)] = \int_{-\infty}^{\infty} x f_X(x; t) dx$$

- Usually we denote $X(t, \omega)$ by $X(t)$ for simplicity.

Next, we introduce the *autocorrelation* of $X(t)$ given by

$$R_{XX}(t_1, t_2) = E[X(t_1)X(t_2)] \\ = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x_1 x_2 f_{X_1 X_2}(x_1, x_2; t_1, t_2) dx_1 dx_2$$

A large theory of stochastic process has been developed, known as *second-order* theory, in which these processes are classified and distinguished only on the basis of their mean $\bar{X}(t)$ and autocorrelation $R_{XX}(t_1, t_2)$. In the case of stationary random processes, we have

$$\bar{X}(t) = \bar{X} \quad (11.43)$$

and

$$R_{XX}(t_1, t_2) = R_{XX}(t_2 - t_1) \quad (11.44)$$

that is, R_{XX} is a function only of the time difference $\tau = t_2 - t_1$. In the stationary case, then, random processes are characterized in the second-order theory *only* by a constant (their mean $\bar{X}$) and a one-dimensional function $R_{XX}(\tau)$. A random process is said to be *wide-sense stationary* if Eqs. (11.43) and (11.44) hold. Note that all stationary processes are wide-sense stationary, but *not* conversely.

REFERENCES

- DAVE 70 Davenport, W. B. Jr., *Probability and Random Processes*, McGraw-Hill (New York), 1970.
- FELL 68 Feller, W., *An Introduction to Probability Theory and Its Applications*, 3rd Edition, Vol. I, Wiley (New York), 1968.
- PAPO 65 Papoulis, A., *Probability, Random Variables, and Stochastic Processes*, McGraw-Hill (New York), 1965.
- PARZ 60 Parzen, E., *Modern Probability Theory and Its Applications*, Wiley (New York), 1960.

Glossary of Notation*

(Only the notation used often in this book is included below.)

NOTATION†	DEFINITION	TYPICAL PAGE REFERENCE
$A.(t) = A(t)$	$P[t. \leq t] = P[i \leq t]$	13
$An^*(s) = A^*(s)$	Laplace transform of aCt	14
a_k	k th moment of aCt	14
$a.(t) = aCt$	$dAn(t)/dt = dA(t)/dt$	14
$Bn(x) = B(x)$	$P[x. \leq x] = P[\tilde{x} \leq x]$	14
$B;^*(s) = B^*(s)$	Laplace transform of $b(x)$	14
b_k	k th moment of $b(x)$	14
$bn(x) = b(x)$	$dBn(x)/dx = dB(x)/dx$	14
C_b^2	Coefficient of variation for service time	187
C_n	n th customer to enter the system	11
$Cn(u) = CCu$	$P[u. \leq u]$	281
$C_n^*(s) = C^*(s)$	Laplace transform of $c_n(l) = c(u)$	285
$c.(l) = c(u)$	$dC.(l)/du = dCCu/du$	281
D	Denotes deterministic distribution	viii
d_k	$P[ij = k]$	176
$E[X] = \bar{X}$	Expectation of the random variable X	378
E_i	System state i	27
E_r	Denotes r -stage Erlangian distribution	124
FCFS	First-come-first-served	8
$F_X(x)$	$P[X \leq x]$	370

* In those few cases where a symbol has more than one meaning, the context (or a specific statement) resolves the ambiguity.

† The use of the notation $Y_n \rightarrow y$ is meant to indicate that $y = \lim Y_n'$ as $n \rightarrow \infty$ whereas $y(t) \rightarrow y$ indicates that $Y = \lim y(r)$ as $t \rightarrow \infty$.

$f_X(x)$	$dF_X(x)/dx$	371
G	Denotes general distribution	viii
$G(y)$	Busy-period distribution	208
$G^*(s)$	Laplace transform of $g(y)$	211
g_k	k th moment of busy-period duration	213
$g(y)$	$dG(y)/dy$	215
H_R	Denotes R-stage hyperexponential distribution	141
$Im(s)$	Imaginary part of the complex variable s	293
$I_n \rightarrow I$	Duration of the (nth) idle period	206
$I^*(s)$	Laplace transform of idle-period density	307
$I^*(s)$	Laplace transform of idle-time density in the dual system	310
K	Size of finite storage	viii
LCFS	Last-come-first-served	8
M	Denotes exponential distribution	viii
M	Size of finite population	viii
M	Number of servers	viii
$N_c(l) \rightarrow N_q$	Number of customers in queue at time t	17
$N(l) \rightarrow N$	Number of customers in system at time t	11
$O(x)$	$\lim_{x \rightarrow 0} O(x)/x = K < \infty$	284
$o(x)$	$\lim_{x \rightarrow 0} o(x)/x = 0$	48
P	Matrix of transition probabilities	31
$PtA]$	Probability of the event A	364
$PtA B]$	Probability of the event A conditioned on the event B	365
PDF	Probability distribution function	369
pdf	Probability density function	371
$P_k(l)$	$P[N(l) = k]$	55
p_{ij}	$P[\text{next state is } t; \text{current state is } E_i]$	27
$P_{ij}(S, I)$	$P[X(t) = j X(es) = i]$	46
p_k	$P[k \text{ customers in system}]$	90
$Q(z)$	c-transform of $P[ij = k]$	192
$q_{ij}(t)$	Transition rates at time t	48
$q_n \quad \tilde{q}$	Number left behind by departure (of C_n)	177
$q_n' \rightarrow \tilde{q}'$	Number found by arrival (of C_n)	242

$Re(s)$	Real part of the complex variable s	340
r_{ij}	$P[\text{next node is } j \mid \text{current node is } i]$	149
r_k	$P[\tilde{q}' = k]$	176
$Sn(Y) \rightarrow S(y)$	$P[sn \leq y] \rightarrow P[\tilde{s} \leq y]$	14
$S;*(s) \rightarrow S^*(s)$	Laplace transform of $sn(Y) \rightarrow sty$	14
s	Laplace transform variable	339
$s_n \rightarrow \tilde{s}$	Time in system (for e_n)	14
$sn(Y) \rightarrow sty$	$dSn(y)/dy \rightarrow dS(y)/dy$	14
$\tilde{s}_n \rightarrow \tilde{s} = T$	Average time in system (for e_n)	14
$sr^k \rightarrow s^k$	k th moment of $sn(Y)$	14
T	Average time in system	14
$ln \rightarrow i$	Interarrival time (between e_{n-1} and e_n)	14
$in = \bar{t} = 1/\lambda$	Average interarrival time	14
$\bar{t}^k$	k th moment of $a()$	14
$U()$	Unfinished work in system at time t	206
$!lo()$	Unit impulse function	341
$u_n \rightarrow \tilde{u}$	$u_n = x_n - t_{n+1} \rightarrow \tilde{u} = \tilde{x} - i$	277
$V(z)$	z -transform of $P[\tilde{v} = k]$	184
$un \rightarrow \tilde{v}$	Number of arrivals during service time (of e_n)	177
W	Average time in queue	14
W_0	Average remaining service time	190
$W_-(y)$	Complementary waiting time	284
$Wn(y) \rightarrow W(y)$	$P[w_n \leq y] \rightarrow P[\tilde{w} \leq y]$	14
$Wn^*(s) \rightarrow W^*(s)$	Laplace transform of $wn(y)$	14
$w_n \rightarrow \tilde{w}$	Waiting time (for e_n) in queue	14
$wn(y) \rightarrow w(y)$	$dWn(y)/dy \rightarrow dW(Y)/dy$	14
$\tilde{w}_n \rightarrow \tilde{w} = W$	Average waiting time (for e_n)	14
$w_n^k \rightarrow w^k$	k th moment of $wn(y)$	14
$X(l)$	State of stochastic process $X(t)$ at time t	19
$x_n \rightarrow \tilde{x}$	Service time (of e_n)	14
x^k	k th moment of $b(x)$	14
$\bar{x}_n = \bar{x} = 1/\mu$	Average service time	14
y	Busy-period duration	206
z	z -transform variable	327

$\alpha(t)$	Number of arrivals in $(0, t)$	15
Y_i	(E)xternal input rate to node i	149
$\delta(t)$	Number of departures in $(0, t)$	16
λ	Average arrival rate	14
$Z,$	Birth (arrival) rate when $N = k$	53
μ	Average service rate	14
μ_k	Death (service) rate when $N = k$	54
$\pi^{(n)} \rightarrow \pi$	Vector of state probabilities $\pi_k^{(n)}$	31
$\pi_k^{(n)} \rightarrow \pi_k$	$P[\text{system state (at } n\text{th step) is } E_k]$	29
$\prod_{i=1}^k a_i$	$a_1 \cdot \dots \cdot a_k$ (Product notation)	334
ρ	Utilization factor	18
σ	Root for G/M/m	249
σ_a^2	Variance of interarrival time	305
σ_b^2	Variance of service time	305
τ_n	Arrival time of e_n	12
$\Phi_+(s)$	Laplace transform of $W(y)$	285
$\Phi_-(s)$	Laplace transform of $W_-(y)$	285
$\triangleq$	Equals by definition	11
$(0, t)$	The interval from 0 to t	15
$\bar{X} = E[X]$	Expectation of the random variable X	378
$(y)^+$	$\max [0, y]$	277
$\binom{n}{k}$	Binomial coefficient $= \frac{n!}{k!(n-k)!}$	368
A/B/m/K/M	»z-Server queue with $A(t)$ and $B(x)$ identified by A and B, respectively, with storage capacity of size K , and with a customer population of size M (if any of the last two descriptors are missing they are assumed to be infinite)	viii
$FCnl(a)$	$dnF(y)/dyn _{y=a}$	382
$f \text{Ckl}(x)$	$f(x) \circledast \dots \circledast f(x)$ k-fold convolution	200
$\circledast$	Convolution operator	376
$f \rightarrow g$	Input f gives output g	322
$A \leftrightarrow B$	Statement A implies statement B and conversely	68
$f \Leftrightarrow F$	f and F form a transform pair	328

Summary of Important Results

Following is a collection of the basic results (those marked by –) from this text in the form of a list of equations. To the right of each equation is the page number where it first appears in a meaningful way; this is to aid the reader in locating the descriptive text and theory relevant to that equation.

GENERAL SYSTEMS

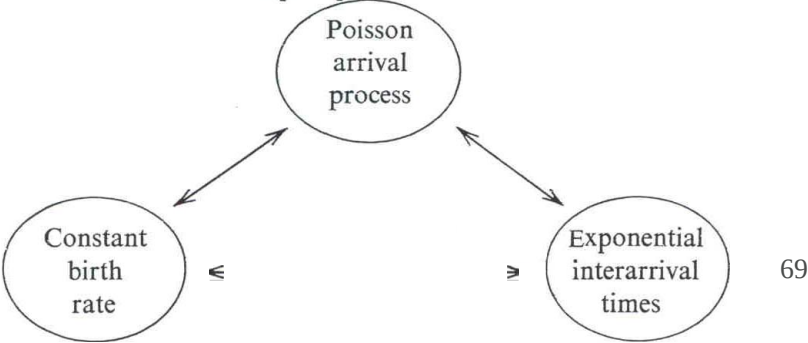
$P = \lambda \bar{x}$	(G/G/1)	18
$\rho \triangleq \lambda \bar{x}/m$	(G/G/m)	18
$T = \bar{x} + W$		18
$\bar{N} = \lambda T$	(Little's result)	17
$\bar{N}_q = \lambda W$		17
$\bar{N}_q = \bar{N} - \rho$		188
$dPk(t)/dt$	= flow rate into E_k -flow rate out of E_k	59
$P_{\gg} = r_k$	(for Poisson arrivals)	176
$r_k = d_k$	$[N(t)$ makes unit changes]	176

MARKOV PROCESSES

For a summary of discrete state Markov chains, see the table on pp. 402-403.

POISSON PROCESSES

$P_k(t) = \frac{(\lambda t)^k}{k!} e^{-\lambda}$	$k \geq 0, t \geq 0$	60
$N(t) = \lambda t$		62
$\sigma_{N(t)}^2 = \lambda t$		62
$E[z^{N(t)}] = e^{\lambda t(z-1)}$		63



BIRTH-DEATH SYSTEMS

$$\frac{dP_k(t)}{dt} = -(\lambda_k + \mu_k)P_k(t) + \lambda_{k-1}P_{k-1}(t) + \mu_{k+1}P_{k+1}(t) \quad k \geq 1 \quad 57$$

$$\frac{dP_0(t)}{dt} = -\lambda_0P_0(t) + \mu_1P_1(t) \quad k=0 \quad 57$$

$$p_k = P_0 \prod_{i=0}^{k-1} \frac{\lambda_i}{\mu_{i+1}} \quad (\text{equilibrium solution}) \quad 92$$

$$P_0 = \frac{1}{1 + \sum_{k=1}^{\infty} \prod_{i=0}^{k-1} \frac{\lambda_i}{\mu_{i+1}}} \quad 92$$

MIMII

$$P_k(t) = e^{-(\lambda+\mu)t} \left[\rho^{(k-i)/2} I_{k-i}(at) + p l k^{-1/2} I_k^{+H}(at) + (1 - \rho) \rho^k \sum_{j=k+i+2}^{\infty} \rho^{-j/2} I_j(at) \right] \quad 77$$

$$P_k = (1 - \rho) \rho^k \quad 96$$

$$\bar{N} = p l (l - p) \quad 96$$

$$\sigma_N^2 = p l (l - p) 2 \quad 97$$

$$W = \frac{\rho/\mu}{1 - p} \quad 191$$

$$T = \frac{1/\mu}{1 - p} \quad 98$$

$$P[\geq k \text{ in system}] = \rho^k \quad 99$$

$$s(y) = \mu(1 - \rho)e^{-\mu(1-\rho)y} \quad y \geq 0 \quad 202$$

$$S(y) = 1 - e^{-\mu(1-\rho)y} \quad y \geq 0 \quad 202$$

$$w(y) = (1 - p)u_0(Y) + \lambda(1 - p)e^{-p l (1-p)y} \quad y \geq 0 \quad 203$$

$$W(y) = 1 - \rho e^{-\mu(1-\rho)y} \quad y \geq 0 \quad 203$$

Summary of Discrete-State Markov Chains

	DISCRETE-TIME		CONTINUOUS-TIME	
	HOMOGENEOUS	NON HOMOGENEOUS	HOMOGENEOUS	NONHOMOGENEOUS
One-step transition probability Matrix of one-step transition probabilities	$p_{ij}^{\Delta} \triangleq P[X_{n+1} = j X_n = i]$ $\mathbf{P} \triangleq [P_{ij}]$	$P_{ij}(n, n+1) \triangleq P[X_{n+1} = j X_n = i]$ $\mathbf{P}(n) \triangleq [p_{ij}(n, n+1)]$	$P_{ij}^0 \triangleq P[X(t + \Delta t) = j X(t) = i]$ $\mathbf{P} \triangleq [P_{ij}]$	$P_{ij}^{0, I + \Delta t} \triangleq P[X(I + \Delta t) = j X(I) = i]$ $\mathbf{P}(t) \triangleq [p_{ij}(t, I + \Delta t)]$
Multiple-step transition probabilities Matrix of multiple-step transition probabilities	$p_{ij}^{(m)} \triangleq P[X_{n+m} = j X_n = i]$ $\mathbf{P}^{(m)} \triangleq [p_{ij}^{(m)}]$	$P_{ij}(III, n) \triangleq P[X_n = j X_m = i]$ $\mathbf{H}(III, n) \triangleq [p_{ij}(m, II)]$	$P_{ij}^{0(I)} \triangleq P[X(s + t) = j X(es) = i]$ $\mathbf{H}(t) \triangleq [P_{ij}(I)]$	$P_{ij}^{0C, I} \triangleq P[X(t) = j X(es) = i]$ $\mathbf{H}(s, I) \triangleq [p_{ij}(S, I)]$
Chapman-Kolmogorov equation	$p_{ij}^{(m)} = \sum_k p_{ik}^{(m-q)} p_{kj}^{(q)}$ $\mathbf{P}^{(m)} = \mathbf{P}^{(m-q)} \mathbf{P}^{(q)}$	$p_{ij}^{0(m, n)} = \sum_k p_{ik}(m, q) p_{kj}(q, n)$ $\mathbf{H}(III, n) = \mathbf{H}(m, q) \mathbf{H}(q, n)$	$P_{ij}(1) = \sum_k P_{ik}(1-s) p_{kj}(s)$ $\mathbf{H}(t) = \mathbf{H}(I-s) \mathbf{H}(s)$	$P_{ij}(es, I) = \sum_k P_{ik}(S, II) p_{kj}(II, I)$ $\mathbf{H}(s, I) = \mathbf{H}(s, II) \mathbf{H}(II, I)$

Table (continued)

$$P[\text{interdeparture time} \leq t_j = 1 - e^{-\lambda t_j} \quad t_j \geq 0 \quad 148$$

$$g(y) \triangleq \frac{(1/y)!}{y^p} e^{-(\lambda+\mu)y} I_1[2y(\lambda\mu)^{1/2}] \quad 215$$

$$in = \frac{1}{n} \binom{2n-2}{n-1} p n (1+p)^{I-2n} \quad 218$$

$$P_k = \begin{cases} \frac{1 - \lambda/\mu}{1 - (\lambda/\mu)^{K+1}} \left(\frac{\lambda}{\mu}\right)^k & 0 \leq k \leq K \\ 0 & \text{otherwise} \end{cases} \quad (M/M/I/K) \quad 104$$

$$P_k = \frac{M!}{(M-k)!} \frac{(\lambda/\mu)^k}{\mu} \quad (M/M/IIIM) \quad 107$$

$$\sum_{i=0}^M \frac{M!}{(M-i)!} \left(\frac{\lambda}{\mu}\right)^i$$

$$P(z) = \frac{\mu(1 - \rho)(1 - z)}{\mu(1 - \rho) - \lambda z[1 - G(z)]} \quad (M/M/I \text{ bulk arrival}) \quad 136$$

$$p_k = \left(1 - \frac{1}{z_0}\right) \left(\frac{1}{z_0}\right)^k \quad k = 0, 1, 2, \dots \quad (MIMI \text{ bulk service}) \quad 139$$

M/M/rn

$$p_k = \begin{cases} p_0 \frac{(m\rho)^k}{k!} & k \leq m \\ p_0 \frac{(\rho)^k m^m}{m!} & k \geq m \end{cases} \quad 102$$

$$P_0 = \left[\sum_{k=0}^{m-1} \frac{(m\rho)^k}{k!} + \frac{(m\rho)^m}{m!} \frac{1}{1 - \rho} \right]^{-1} \quad 103$$

$$P[\text{queueing}] = \frac{\left(\frac{(m\rho)^m}{m!}\right) \left(\frac{1}{1 - \rho}\right)}{\left[\sum_{k=0}^{m-1} \frac{(m\rho)^k}{k!} + \frac{(m\rho)^m}{m!} \frac{1}{1 - \rho} \right]} \quad (\text{Erlang C formula}) \quad 103$$

$$P_k = \frac{(\lambda/\mu)^k}{k!} \bigg/ \sum_{i=0}^{\infty} \frac{(\lambda/\mu)^i}{i!} \quad (M/M/m/m) \quad 105$$

$$P_m = \frac{(\lambda/\mu)^m}{m!} \bigg/ \sum_{i=0}^{\infty} \frac{(\lambda/\mu)^i}{i!} \quad (M/M/m/m) \quad (\text{Erlang's loss formula}) \quad 106$$

MIDI

$$\bar{q} = \frac{\bar{P}}{1 - p} - \frac{\bar{\rho}^2}{2(1 - p)} \quad 188$$

$$W = \frac{\bar{\rho}\bar{x}}{2(1 - p)} \quad 191$$

$$G(y) = \sum_{n=1}^{\lfloor y/\bar{x} \rfloor} \frac{(np)^{n-1}}{n!} e^{-np} \quad 219$$

$$f = \frac{(np)^{n-1}}{n!} e^{-np} \quad 219$$

E_r (r-stage Erlang Distribution)

$$b(x) = \frac{r\mu(r\mu x)^{r-1}e^{-r\mu x}}{(r-1)!} \quad x \geq 0 \quad 124$$

$$\sigma_b = \frac{1}{\mu(r)^{1/2}} \quad 124$$

M/E_r/1

$$P_j = (1 - p) \sum_{i=1}^r A_i(z_i)^{-j} \quad j = 1, 2, \dots, r \quad 129$$

$$P_k = \begin{cases} 1 - p & k = 0 \\ \{p(ZOT - 1)ZOT - rk\} & k > 0 \end{cases} \quad 133$$

HR (R-stage Hyperexponential Distribution)

$$b(x) = \sum_{i=1}^R \alpha_i \mu_i e^{-\mu_i x} \quad x \geq 0 \quad 141$$

$$C_b^2 \geq 1 \quad 143$$

MARKOVIAN NETWORKS

$$\dot{y}_i = y_i + \sum_{j=1}^S \lambda_j r_{ji} \quad 149$$

$$p(k_1 \bullet k_2 \bullet \dots \bullet k_N) = P_1(k_1) p_2(k_2) \dots p_N(k_N) \quad 150$$

(open) where $P_i C(k_i)$ is solution to isolated $M/M/m$,

$$p(k_1, k_2, \dots, k_N) = \frac{1}{G(K)} \prod_{i=1}^N \frac{x_i^{k_i}}{\beta(k_i)} \quad (\text{closed}) \quad 152$$

LIFE AND RESIDUAL LIFE

$$f_X(x) = \frac{x f(x)}{m} \quad (\text{lifetime density of sampled interval}) \quad 171$$

$$f''(y) = \frac{1 - F(y)}{m} \quad (\text{residual life density}) \quad 172$$

$$I^*(s) = \frac{1 - P(s)}{sm} \quad (\text{residual life transform}) \quad 172$$

$$r_n = \frac{m_{n+1}}{(n+1)m} \quad (\text{nth moment of residual life}) \quad 173$$

$$r_1 = \frac{m_2}{2m} \quad (\text{mean residual life}) \quad 173$$

$$r(x) = \frac{f(x)}{1 - F(x)} \quad (\text{failure rate}) \quad 173$$

M/G/1

$$r_k = p_k = d_k \quad 176$$

$$q_{n+1} = q_n - \Delta_{q_n} + v_{n+1} \quad 181$$

$$\bar{v} = \rho \quad 183$$

$$v^2 - \bar{v} = \lambda^2 x^2 = \rho^2(1 + C_b^2) \quad 187$$

$$V(z) = B^*(\lambda - \lambda z) \quad 184$$

$$\frac{1}{q} - \rho = \frac{1}{p} + \frac{1}{2} \frac{(1 + C_b^2)}{(1 - \rho)} \quad (\text{P-K mean value formula}) \quad 187$$

$$\frac{T}{\bar{x}} = 1 + p \frac{(1 + C_b^2)}{2(1 - \rho)} \quad (\text{P-K mean value formula}) \quad 191$$

$$\frac{W}{\bar{x}} = p \frac{(1 + C_b^2)}{2(1 - \rho)} \quad (\text{P-K mean value formula}) \quad 191$$

$$W = \frac{W_0}{1 - \rho} \quad (\text{P-K mean value formula}) \quad 190$$

$$W_0 \triangleq \frac{\lambda x^2}{2} \quad 190$$

$$Q(z) = B^*(\lambda - \lambda z) \frac{(1 - \rho)(1 - z)}{B^*(\lambda - \lambda z) - z} \quad (\text{P-K transform equation}) \quad 194$$

$$W^*(s) = \frac{s(I - \rho)}{s - i + \lambda B^*(s)} \quad (\text{P-K transform equation}) \quad 200$$

$$5^*(s) = \frac{B^*(s)}{s - i + i B^*(s)} \quad 199$$

$$P[J \leq y] = 1 - e^{-\lambda y} \quad y \geq 0 \quad 208$$

$$G^*(s) = B^*(s + \lambda - \lambda G^*(s)) \quad 212$$

$$G(y) = \int_0^y \sum_{n=1}^{\infty} e^{-\lambda x} \frac{(\lambda x)^{n-1}}{n!} b(n)(x) dx \quad 226$$

$$g_1 = -\frac{\bar{x}}{-P} \quad 213$$

$$g_2 = \overline{(1 - \rho)^3} \quad 214$$

$$\sigma_*^2 = \frac{\sigma_b^2 + \rho(\bar{x})^2}{(1 - \rho)^3} \quad 214$$

$$g_3 = \frac{x^3}{(1 - \rho)^5} + \frac{3\lambda(\bar{x}^2)^2}{(1 - \rho)^5} \quad 214$$

$$g_4 = \frac{x'}{(1 - \rho)^5} + \frac{10\lambda x^2 x^3}{(1 - \rho)^6} + \frac{15\lambda^2(\bar{x}^2)^3}{(1 - \rho)^7} \quad 214$$

$$F(z) = zB^*[\lambda - \lambda F(z)] \quad 217$$

$$P[N_{bp} = n] = \int_0^\infty \frac{(\lambda y)^{n-1}}{n!} e^{-\lambda y} b(n)(y) dy \quad 226$$

$$h_1 = \frac{1}{1 - \rho} \quad 217$$

$$h_i = \frac{2\rho(1 - \rho) + \lambda^2 \bar{x}^2}{(1 - \rho)^3} + \frac{1}{1 - \rho} \quad 218$$

$$\sigma_h^2 = \frac{\rho(1 - \rho) + \lambda^2 \bar{x}^2}{(1 - \rho)^3} \quad 218$$

$$\frac{\partial F(w, t)}{\partial t} = \frac{\partial F(w, t)}{\partial w} - \lambda F(w, t) + \lambda \int_{x=0}^w B(w - x) d_x F(x, t) \quad 227$$

(Takacs integrodifferential equation)

$$F^{**}(r, s) = \frac{(r/\eta)e^{-\eta w_0} e^{-r\mathbb{C}O}}{\lambda B^*(s) - i + r - s} \quad 229$$

$$Q(z) = \frac{(1 - \rho)(I - z)B^*[\lambda - iG(z)]}{B^*(.) - \lambda G(z) - z} \quad (\text{bulk arrival}) \quad 235$$

M/G/ ∞

$$p_k = \frac{\rho^k}{k!} e^{-\rho} \quad 234$$

$$T = \bar{x} \quad 234$$

$$s(y) = b(y) \quad 234$$

G/M/1

$$r_k = (1 - \sigma)\sigma^k \quad k = 0, 1, 2, \dots \quad 251$$

$$\sigma = A^*(\mu - \mu\sigma) \quad 251$$

$$W(y) = 1 - \sigma e^{-\mu(1-\sigma)y} \quad y \geq 0 \quad 252$$

$$\sigma \quad 252$$

G[M/m]

$$q'_{n+1} = q'_n + 1 - v'_{n+1} \quad 242$$

$$\sigma = A^*(m\mu - m\mu a) \quad 249$$

$$P[\text{queue size} = n \mid \text{arrival queues}] = (1 - \sigma)\sigma^n \quad n \geq 0 \quad 249$$

$$\mathbf{r} = J[R_0, R_1, \dots, R_{m-2}, 1, \sigma, \sigma^2, \sigma^3, \dots] \quad 254$$

$$\frac{R_k - \sum_{i=k}^{m-2} R_i P_{ik} - \sum_{i=m-1}^{\infty} \sigma^{i+1-m} P_{ik}}{P_{k-1,k}} \quad 254$$

$$P_U = 0 \quad \text{for} \quad j > i + 1 \quad 242$$

$$p_{ij} = \int_0^\infty \left(1 + \frac{1}{j}\right) [1 - e^{-\mu t}]^{i+1-j} e^{-\mu t j} dA(t) \quad j \leq i + 1 \leq m \quad 244$$

$$f_{jn} = P_{i,i+1-n} = \int_{t=0}^\infty \frac{(m\mu t)^n}{n!} e^{-m\mu t} dA(t) \quad 0 \leq n \leq i + 1 - m, m \leq i \quad 245$$

$$p_{ij} = \int_0^\infty \binom{m}{j} e^{-j\mu t} \times \left[\int_0^t \frac{(m\mu y)^{i-m}}{(i-m)!} (e^{-\mu y} - e^{-\mu t})^{m-j} m\mu dy \right] dA(t) \quad j < m < i + 1 \quad 245$$

$$J = \frac{1}{1 - \sigma} + \sum_{k=0}^{m-2} R_k \quad 254$$

$$W = \frac{J\sigma}{m\mu(1 - \sigma)^2} \quad 256$$

$$w(y \text{ arrival queues}) = (1 - \sigma)m\mu e^{-m\mu(1-\sigma)y} \quad y \geq 0 \quad 250$$

$$W(y) = 1 - \frac{\sigma}{1 + (1 - \sigma) \sum_{k=0}^{m-2} R_k} e^{-m\mu(1-\sigma)y} \quad y \geq 0 \quad 255$$

GIGII

$$W_{+1} = (w_{+1})_{+} \quad 277$$

$$C(II) = a(-II) \otimes b(II) \quad 281$$

$$W(y) = \begin{cases} \int_{-\infty}^y W(y - II) dC(II) & y \geq 0 \\ 0 & y < 0 \end{cases} \quad \begin{matrix} \text{(Lindley's} \\ \text{integral} \\ \text{equation)} \end{matrix} \quad 283$$

$$A^* (-s) B^* (s) - 1 = \frac{\Psi_+^*(s)}{\Psi_-^*(s)} \quad 286$$

$$\Phi_+(s) = \frac{1}{\Psi_+(s)} \lim_{s \rightarrow 0} \frac{\Psi_+(s)}{s} = \frac{W(0+)}{\Psi_+(s)} \quad 290$$

$$\Phi_+(s) = \frac{\Psi^*(0)(1-p)f}{[A^*(-s)B^*(s) - 1]\Psi_-(s)} \quad 290$$

$$W = \frac{\sigma_a^2 + \sigma_b^2 + (i)^2(1-p)2}{2f(1-p)} \quad \begin{matrix} [2 \\ 21 \end{matrix} \quad 306$$

$$W = -\frac{112}{2\bar{u}} - \frac{y^2}{2\bar{y}} \quad 305$$

$$\tilde{w} = \sup_{n \geq 0} U; \quad 279$$

$$W(y) = \pi(c(y) \otimes w(y)) \quad 301$$

$$W^*(s) = \frac{a\phi[1 - \phi^*(-s)]}{1 - C^*(s)} \quad 307$$

$$W^*(s) = \frac{1}{1 - \sigma \hat{I}^*(s)} \quad 310$$

Index

- Abel transform, 321
- Abscissa of absolute convergence, 340, 349
- Absorbing state, 28
- Age, 170
- Algebra, real commutative, 301
- Algebra for queues, 229-303
- Alternating sum property, 330
- Analytic continuation, 287
- Analytic function, 328, 337
 - isolated singular point, 337
- Analyticity, 328
 - common region of, 287
- Aperiodic state, 28
- Approximations, 319
- ARPANET, 320
- Arrival rate, 4
 - average, 13
- Arrival time, 12
- Arrivals, 15
- Autocorrelation, 395
- Availability, 9
- Average value, 378
- Axiomatic theory of probability, 365
- Axioms of probability theory, 364

- Backward Chapman-Kolmogorov equation, 42, 47, 49
- Balking, 9
- Ballot theorem, classical, 224-225
 - generalization, 225
- Baricentric coordinates, 34
- Bayes' theorem, 366
- Birth-death process, 22, 25, 42, 53-78
 - assumption, 54
 - equilibrium solution, 90-94
 - existence of, 93-94
 - infinitesimal generator, 54
 - linear, 82
 - probability transition matrix, 43
 - summary of results, 401
 - transitions, 55-56
- Birth rate, 53
- Borel field, 365
- Bottleneck, 152
- Bound, 319
 - Chernoff, 391
- Bribing, 9
- Bulk arrivals, 134-136, 162-163, 235, 270
- Bulk service, 137-139, 163, 236
- Burke's theorem, 149

- Capacity, 4, 5, 15
- Catastrophe process, 267-269
- Cauchy inequality, 143
- Cauchy residue theorem, 337, 352
- Cauchy-Riemann, 328
- Central limit theorem, 390
- Chain, 20
- Channels, 3
- Chapman-Kolmogorov equation, 41, 47, 51
- Characteristic equation, 356, 359
- Characteristic function, 321, 381
- Cheating, 9
- Chernoff bound, 391
- Closed queueing network, 150
- Closed subset, 28
- Coefficient of variation, 381
- Combinatorial methods, 223-226
- Commodity, 3
- Complement, 364
- Complete solution, 357, 360
- Complex exponentials, 322-324
- Complex s-plane, 291
- Complex variable, 325
- Compound distribution, 387
- Computer center example, 6
- Computer-communication networks, 320
- Computer network, 7
- Conditional pdf, 375

- Conditional PDF, 375
- Conditional probability, 365
- Continuous-parameter process, 20
- Continuous-state process, 20
- Contour, 293
- Convergence strip, 354
- Convex function, 389
- Convolution, density functions, 376
 - notation, 376
 - property, 329, 344-345
- Cumulative distribution function, 370
- Cut, 5
- Cyclic queue, 113, 156-158

- 0/0/1, example, 309
- Death process, 245
- Death rate, 54
- Decomposition, 119, 323, 327
- Defections, 9
- Delta function, Dirac, 341
 - Kronecker, 325
- Departures, 16, 174-176
- D/E, I queueing system, 314
- Difference equations, 355-359
 - standard solution, 355-357
 - z-transform solution, 357-359
- Differential-difference equation, 57, 361
- Differential equations, 359-361
 - Laplace transform solution, 360-361
 - linear constant coefficient, 324
 - standard solution, 359-360
- Differential matrix, 38
- Diffusion approximation, 319
- Dirac delta function, 341
- Discouraged arrivals, 99
- Discrete-parameter process, 20
- Discrete-state process, 20
- Disjoint, 365
- Domain, 369
- Duality, 304, 309
- Dual queue, 310-311

- E2/M/I, 259
 - spectrum factorization, 297
- Eigenfunctions, 322-324
- Engset distribution, 109
- Equilibrium equation, 91
- Equilibrium probability, 30
- Ergodic Markov chain, 30, 52
 - process, 94
 - state, 30
- Erlang, 119, 286
 - B formula, 106
 - C formula, 103
 - distribution, 72, 124
 - loss** formula, 106
- E, M/I, 130-133, 405
- E, (r-stage Erlang Distribution), 405
- Event, 364
- Exhaustive set of events, 365
- Expectation, 13, 377-381
 - fundamental theorem of, 379
- Exponential distribution, 65-71
 - coefficient of variation, 71
 - Laplace transform, 70
 - mean, 69
 - memory less property, 66-67
 - variance, 70
- Exponential function, 340

- FCFS, 8
- Fig flow example, 5
- Final **value** theorem, 330, 346
- Finite capacity, 4
- Flow, 58
 - conservation, 91-92
 - rate, 59, 87, 91
 - system, 3
 - time, 12
- Fluid approximation, 319
- Forward Chapman-Kolmogorov equation, 42, 47, 49, 90
- Foster's criteria, 30
- Fourier transform, 321, 381
- Function of a random variable, 375, 380

- Gaussian distribution, 390
- Generating function, 321, 327
 - probabilistic interpretation, 262
- Geometric distribution, 39
- Geometric series, 328
- Geometric transform, 327
- G/G/I, 19, 275-312
 - defining equation, 277
 - mean wait, 306
 - summary of results, 409
 - waiting time transform, 307, 310
- G/G/m, II
- Global-balance equations, 155
- Glossary of Notation, 396-399

- G!M I, 251-253
 - dual queue, 311
 - mean wait, 252
 - spectrum factorization, 292
 - summary of results, 408
 - waiting time distribution, 252
- G/M/2, 256 - 259
 - distribution of number of customers, 258
 - distribution of waiting time, 258
- G/M/m, 241 - 259
 - conditional pdf for queueing time, 250
 - conditional queue length distribution, 249
 - functional equation, 249
 - imbedded Markov chain, 241
 - mean wait, 256
 - summary of results, 408 - 409
 - transition probabilities, 241 - 246
 - waiting-time distribution, 255
- Gremlin, 261
- Group arrivals, *see* Bulk arrivals
- Group service, *see* Bulk service

- Heavy traffic approximation, 319
- Hippie example, 26-27, 30-38
- Homogeneous Markov chain, 27
- Homogeneous solution, 355,
- HR (R-stage Hyperexponential Distribution), 141, 405

- Idle period, 206, 305, 311
 - isolating effect, 281
- Idle time, 304, 309
- Imbedded Markov chain, 23, 167, 169, 241
 - G/G/I, 276-279
 - G/M/m, 241-246
 - M/G/I, 174-177
- Independence, 374
- Independent process, 21
- Independent random variables, product of
 - functions, 386
 - sums, 386
- Inequality, Cauchy-Schwarz, 388
 - Chebyshev, 388
 - Cr. 389
 - Holder, 389
 - Jensen, 389
 - Markov, 388
 - triangle, 389
- Infinitesimal generator, 48
- Initial value theorem, 330, 346

- Input-output relationship, 321
- Input variables, 12
- Inspection technique, 58
- Integral property, 346
- Interarrival time, 8, 12
- Interchangeable random variables, 278
- Interdeparture time distribution, 148
- Intermediate value theorem, 330
- Intersection, 364
- Irreducible Markov chain, 28

- Jackson's theorem, ISO
- Jockeying, 9
- Jordan's lemma, 353

- Kronecker delta function, 325

- Labeling algorithm, 5
- Ladder height, 224
- Ladder index, 223
 - ascending, 309
 - descending, 311
- Laplace transform, 321, 338- 355
 - bilateral, 339, 348
 - one-sided, 339
 - probabilistic interpretation, 264, 267
 - table of properties, 346
 - table of transform pairs, 347
 - two-sided, 339
- Laplace transform inversion, inspection**
 - method, 340, 349
 - inversion integral, 352-354
- Laplace transform of the pdf, 382
- Laurent expansion, 333
- Law of large numbers, strong, 390
 - weak, 390
- LCFS, 8, 210
- Life, summary of results, 406
- Lifetime, 170
- Limiting probability, 90
- Lindley's integral equation, 282-283, 314
- Linearity, 332, 345
- Linear system, 321, 322, 324
- Liouville's theorem, 287
- Little's result, 17
 - generalization, 240
- Local-balance equations, 155- 160
- Loss system, IOS

- Marginal density function, 374

- Marking of customers, 261-267
- Markov chain, 21
 - continuous-time, 22, 44-53
 - definition, 44
 - discrete-time, 22, 26-44
 - definition, 27
 - homogeneous, 27, 51
 - nonhomogeneous, 39-42
 - summary of results, 402-403
 - theorems, 29
 - transient behavior, 32, 35-36
- Markovian networks, summary of results, 405
- Markovian population processes, 155
- Markov process, 21, 25, 402-403
- Markov property, 22
- Max-flow-min-cut theorem, 5
- MIDI*, 188
 - busy period, number served, 219
 - pdf, 219
 - mean wait, 191
 - summary of results, 405
- MEI*/1, 234
- Mean, 378
- Mean recurrence time, 28
- Measures, 301
 - finite signed, 301
- Memoryless property, 45
- MEI*/1, 126-130
 - summary of results, 405
- Merged Poisson stream, 79
- Method of stages, 119-126
- Method of supplementary variables, 169
- Method of z-transform, 74-75
- MG*/1, 167-230
 - average time in system, 190
 - average waiting time, 190
 - busy period, 206-216
 - distribution, 226
 - moments, **213-214**, 217-218
 - number served, 217
 - transform; 212
 - discrete time, 238
 - dual queue, 312
 - example, 308
 - feedback system, 239
 - idle-time distribution, 208
 - interdeparture time, 238
 - mean queue length, 180-191
 - probabilistic interpretation, 264
 - state description, 168
 - summary of results, 406
 - system time, moments, 202
 - transform, 199
 - time-dependent behavior, 264-267
 - transition probabilities, 177-180
 - waiting time, moments, 201
 - pdf, 201
- MG*/∞, 234
 - summary of results, 408
 - time dependent behavior, 271
- MH*2/1 example, 189, 205
- MM*/c, 101
 - time-dependent behavior, 262
- M/M/∞/M***, 107-108
- M/M*/1, 73-78, 94-99, 401, 404
 - busy period, number served, 218
 - pdf, 215
 - discrete time, 112
 - example, 307
 - feedback, 113
 - mean number, 96
 - mean system time, 98
 - mean wait, 191
 - spectrum factorization, 290
 - summary of results, 401, 404
 - system time pdf, 202
 - transient analysis, 77, 84-85
 - variance of number, 97
 - waiting time pdf, 203
- MM*/1/*K*, 103-105
- MM*/1/*M*, 106-107
- MM/m*.102-103, 259
 - summary of results, 404
- MM/m/K/M*, 108-109
- MM/m/m*, 105-106
- MM*/2, 110
- Moment generating function, 382
- Moment generating properties, 384
- Moments, 380
 - central, 380
- Multi-access computer systems, 320
- Mutually exclusive events, 365
- Mutually exclusive exhaustive events, 365-366
- Nearest neighbors, 53, 58
- Network, 4
 - closed, 150
 - computer, 7
 - open, 149

- Network **flow** theory, 5
- Networks of Markovian **queues**, 147-160, 405
- Non-nearest-neighbor, 116
- No queue, 161-162, 315-316
- Normal distribution, 390
- Notation, 10-15, 396-399
- Null event, 364
- Number of customers, **II**

- Open queueing networks, 149

- Paradox of residual life, 169-173
- Parallel stages, 140-143
- Parameter shift, 347
- Partial-fraction expansion, 333-336, 349-352
- Particular solution, 355
- Periodic state, 28
- Permutations, 367
- Pineapple factory example, 4
- Poisson, catastrophe, 267
 - distribution, 60, 63
 - process, 61-65
 - mean, 62
 - probabilistic interpretation, 262
 - summary of results, 400
 - variance, 62
- pure death process, 245
- Pole, 291
 - multiple, 350
 - simple, 350
- Pole-zero pattern, 292, 298
- Pollaczek-Khinchin (P-K) mean value
 - formula, 187, 191, 308
- Pollaczek-Khinchin (P-K) transform equation, 194, 199, 200, 308
- Power-iteration, 160
- Priority queueing, 8, 319
- Probabilistic arguments, 261
- Probability density function (pdf), 13, 371, 374
- Probability distribution function (PDF), 13, 369
- Probability generating function, 385
- Probability measure, 364
- Probability system, 365
- Probability theory, 10, 363-395
- Processor-sharing algorithm, 320
- Product notation, 334
- Projection, 301

- Pure birth process, 60, 72, 81
- Pure death process, 72

- Queueing discipline, 8
- Queueing system, 8-9, **II**
- Queue size, average, 188

- Random event, 363
- Random sum, 388
- Random variables, 368-377
 - continuous, 373
 - discrete, 373
 - expectation of product, 379
 - expectation of sum, 379
 - mixed, 373
- Random walk, 23, 25, 223-224
- Range, 369
- Recurrent, nonnull, 29
 - null, 29, 94
 - process, 24
 - state, 28
- Reducible, 28
- Regularity, 363
- Relative frequency, 364
- Renewal, density, 173
 - function, 173, 268
 - process, 24, 25
 - theorem, 174
 - theory, 169-174
 - integral equation, 174, 269
- Residual life, 169, 170, 222
 - density, 172, 231
 - mean, 173, 305
 - moments, 173
 - summary of results, 406
- Residual service time, 200
- Responsive server, 101
- Riemann integral, 377
- Root, multiple, 356, 359
- Rouché's theorem, 293, 355

- Sample, mean, 389
 - path, 40
 - point, 364
 - space, 364
- Scale change, 332, 345
- Schwartz's theory of distributions, 341
- Second-order theory, 395
- Semi-invariant generating function, 391
- Semi-Markov process, 23, 25, 175

- Series-parallel stages, 139- 147
- Series stages, 119-126, 140
- Series sum property, 330
- Service time, 8, 12
- Set theoretic notation, 364
- Sifting property, 343
- Single channel, 4
- Singularity, 355
- Singularity functions, family of, 344
- Spectrum factorization, 283, 286- 290
 - examples, 290- 299
 - solution, 290
- Spitzer's identity, 302
- Stable flow, 4
- Stages, 119, 126
- Standard deviation, 381
- State space, 20
- State-transition diagram, 30
- State-transition-rate diagram, 58
- State vector, 167
- Stationary distribution, 29
- Stationary process, 21
- Statistical independence, 365
- Steady flow, 4
- Step-function, 151, 181
- Stieltjes integral, 377
- Stochastic flow, 6
- Stochastic processes, 20, 393-395
 - classification, 19- 26
 - stationary**, 394
- Stochastic sequence, 20
- Storage capacity, 8
- Sub-busy period, 210
- Supplementary variables, method of, 233
- Sweep (probability), 300
- System function, 325
- System time, 12
- Takács** integrodifferential equation, 226- 230
- Tandem network, 147- 149
- Taylor-series expansion, 332
- Time-diagram notation, 14
- Time-invariant system, 322, 324
- Time-shared computer systems, 319
- Total probability, theorem of, 366
- Traffic intensity, 18
- Transfer function, 325
- Transform, 321
 - Abel, 321
 - bilateral, 354
 - Fourier, 381
 - Hankel, 321
 - Laplace, 338-355
 - Mellin, 321
 - method of analysis, 324
 - two-sided, 383
 - z-transform, 327- 338
- Transient process, 94
- Transient state, 28
- Transition probability, $G/M/m$, 241-246
 - $M/G/1$, 177-180
 - matrix, 31
 - m-step, 27
 - one-step, 27
- Transition-rate matrix, 48
- Translation, 332, 345
- Unfinished work, **11**, 206-208, 276
 - time-dependent transform, 229
- Union, 364
- Unit advance, 332
- Unit delay, 332
- Unit doublet, 343
- Unit function, 325, 328
- Unit impulse function, 301, 341-344
- Unit response, 325
- Unit step function, 328, 341, 343
- Unsteady flow, 4
- Utilization factor, 18
- Vacation, 239
- Variance, 381
- Vector transform, 35
- Virtual waiting time, 206
 - see also* Unfinished work
- Waiting time, 12
 - complementary, 284
 - transform, 290
- Wendel projection, 301, 303
- Wide-sense stationarity, 21, 395
- Wiener-Hopf integral equation, 282
- Work, 18
 - see also* Unfinished work

Zeros. 291

z-Transform, 321, 327- 338, 385
inversion, inspection method. 333
inversion formula. 336

power-series method, 333
method of, 74- 75
table of properties. 330
table of transform pairs. 331

