Statistics for Industry and Technology

U. Narayan Bhat

An Introduction to Queueing Theory

Modeling and Analysis in Applications

Second Edition

Birkhäuser

Statistics for Industry and Technology

Series Editor

N. Balakrishnan
McMaster University
Hamilton, ON
Canada

Editorial Advisory Board

Zhezhen Jin
Columbia University
New York, NY, USA

Gary C. McDonald
NAO Research & Development Center
Warren, MI, USA

Kazuyuki Suzuki
University of Electro Communications
Chofu-shi, Tokyo
Japan

For further volumes:
http://www.springer.com/series/4982

U. Narayan Bhat

An Introduction to Queueing Theory

Modeling and Analysis in Applications

Second Edition

With guest contributions from:
Professor Srinivas R. Chakravarthy, Kettering University
Professor Krishna M. Kavi, University of North Texas
Professor Andrew Junfang Yu, The University of Tennessee

Birkhäuser

U. Narayan Bhat
Department of Statistical Science
Southern Methodist University
Dallas
Texas
USA

ISSN 2364-6241 ISSN 2364-625X (electronic)
Statistics for Industry and Technology
ISBN 978-0-8176-8420-4 ISBN 978-0-8176-8421-1 (eBook)
DOI 10.1007/978-0-8176-8421-1

Library of Congress Control Number: 2015938808

Mathematics Subject Classification (2010): 60J27, 60K25, 60K30, 68M20, 90B22, 90B36, 91B70

Springer Boston, MA Heidelberg New York Dordrecht London
© Springer Science+Business Media New York 2015
This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.
The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.
The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made.

Printed on acid-free paper

Springer Science+Business Media LLC New York is part of Springer Science+Business Media (www.springer.com)

Dedicated to
N. U. Prabhu
Professor Emeritus
Cornell University
USA

Preface to the First Edition

There are several books on queueing theory available for students as well as researchers. At the low end of mathematical sophistication, some provide useable formulas in a recipe fashion. At the high end there are research monographs on specific topics and books with emphasis on theoretical analysis. In between there are a few textbooks with one common feature. All of them require an adequate background knowledge on probability and Markov processes that can be acquired normally with a semester-length graduate course. Consequently, most of those who deal with the modeling and analysis of queueing systems either do not take a course on the subject because they have to spend an extra semester, or take a course on queueing systems without the necessary background and learn only how to use the results. This book is addressed to remedy this situation by providing a one semester foundational introduction to the theory necessary for modeling and analysis of systems while developing the essential Markov process concepts and techniques with queueing processes as examples.

Some of the key features of the book also distinguish it from others. Its introductory chapter includes a historical perspective on the growth of queueing theory in the last 100 years. With emphasis on modeling and analysis it deals with topics such as identification of models, collection of data, and tests for stationarity and independence of observations. It provides a rigorous treatment of basic models commonly used in applications with appropriate references for advanced topics. It gives a comprehensive discussion of statistical inference techniques useable in the modeling of queueing systems and an introduction to decision problems in their management. The book also includes a chapter, written by computer scientists, on the use of computational tools and simulation in solving queueing theory problems.

The book can be used as a text for first year graduate students in the applied science areas such as computer science, operations research, and industrial and/or systems engineering, and allied fields such as manufacturing and communication engineering. It can also serve as a text for upper level undergraduate students in mathematics, statistics, and engineering who have a reasonable background in calculus and basic probability theory. It is the product of the author's experience in teaching queueing theory for 40 years at various levels with or without the necessary background in stochastic processes.

The mathematical background assumed for the coverage of topics is a two or three semester course in calculus, some exposure to transforms and matrices, and an introductory course in probability and statistics, all at the undergraduate level. No familiarity with measure theoretic terminology is assumed. An appendix on mathematical results provides some of the essential theorems for reference.

The book does not advocate any specific software in the numerical analysis of queueing problems. The one chapter on the modeling and analysis using computational tools employs matrix laboratory (MATLAB) for the purpose and we believe students can benefit more by using mathematical software such as MATLAB and Mathematica rather than system specific software because of their limited scope.

For this author writing this book has been a retirement project. He is indebted to Southern Methodist University and the Institute for the Study of Earth and Man for providing necessary resources and facilities even after his retirement. He acknowledges his gratitude to Professors Krishna Kavi and Robert Akl of the University of North Texas for contributing a chapter on numerical analysis of queueing systems in which the author's expertise does not go very far. Special acknowledgment of indebtedness is also made of the reviewers' comments that have helped in improving the organization and contents of the book. The author also wishes to thank Professor N. Balakrishnan for recommending this book for inclusion in the Statistics for Industry and Technology Series of Birkhauser. Thanks are due to Professor Junfang Yu of the Department of Engineering Management, Information and Systems of Southern Methodist University for using the prepublication copy of the book in his class and pointing out some of the typographical errors in it. Thanks are also due to Ms. Sheila Crain of the Department of Statistical Science, for setting the manuscript in LaTex with care and perseverance.

The author's wife Girija, son Girish, and daughter Gouri, have supported and encouraged him throughout his academic career. They deserve all the credit for his success.

U. Narayan Bhat
Dallas, Texas
July 2007

Preface to the Second Edition

After the publication of the first edition of this book, like most authors in the academic world, the author felt that some improvements could have been made in it. So with this revised edition, the author has taken the opportunity to make changes, which hopefully, will increase the usefulness of the book. One major change is the inclusion of additional topics with the help of contributing guest authors to broaden the scope of methodology of analysis and the applicability of queueing models.

This edition includes all topics covered in the first edition with one major rearrangement. The short chapter on renewal models has been absorbed in two other chapters, the theoretical portion in Chapter 3 covering basic concepts in stochastic processes, and the modeling portion in Chapter 6 along with extended Markov models.

Another change made in the narrative of topics covered in the first edition is to warn the reader when an analysis or derivation requires a mathematical background beyond what is stated in the preface to the first edition. Under those circumstances it is suggested that, the reader may skip such analyses or derivations without sacrificing the understanding of the subject.

Professor Srinivas R. Chakravarthy, a contributing guest author, has contributed a chapter on the matrix-analytic method as an alternative method of analysis of queueing systems. The matrix-analytic method was introduced by Professor Marcel Neuts in the 1970s and he expanded its scope along with his associates including the author of this chapter, in the 1980s. Since then its reach in the analysis of queueing systems has grown far and wide. At this time it would not be an exaggeration to say that the majority of new research being done in the application of queueing theory uses this method.

In order to broaden the appeal of the book to applied scientists, a chapter on queueing theory applications in the analysis of manufacturing systems and another on applications in the computer and communication systems have been included. The first is a new chapter authored by Professor Andrew Junfang Yu. The second chapter, authored by Professor Krishna M. Kavi, is an expanded version of a chapter which also included simulation in the first edition. In

this edition simulation of queueing systems gets a separate chapter by itself. Among all the application areas of queueing theory at this time, computer and communication systems, and manufacturing systems stand out because of their breadth and usefulness. For this reason these two chapters have been included as areas of application of queueing theory in this edition.

Most of the statements made in the preface to the first edition of the book stand true for this edition as well. A few additional acknowledgments are also in order. The author is grateful to the three contributing authors for adding their expertise in three different areas. The author is indebted to Southern Methodist University and the Institute for the Study of Earth and Man for their continuing support for his retirement projects. Some of the changes in the earlier material have come about in response to comments made by the reviewers of the first edition. The author wishes to thank them. He also wishes to thank the editors of Birkhauser/Springer, and the Statistics for Industry and Technology series editor Professor N. Balakrishnan for initiating and supporting this revision. Thanks are due to Ms. Sheila Crain for assisting in the preparation of the manuscript with skill and patience.

The author is indebted to his wife Girija, for supporting him in this project. Professor N. U. Prabhu of Cornell University introduced the author to queueing theory in the early 1960s while they were at the University of Western Australia. This book, therefore, is dedicated to Professor Prabhu in recognition of the role he has played in the scholastic career of the author.

Instructors may request a guide to the solutions of exercises via the Springer website at http://www.springer.com/gp/book/9780817684204

U. Narayan Bhat
Dallas, Texas
May 2015

Contents

Chapter 1

Introduction

1.1 Basic System Elements

Queues (or waiting lines) help facilities or businesses provide service in an orderly fashion. Forming a queue being a social phenomenon, it is beneficial to the society if it can be managed so that both the unit that waits and the one that serves get the most benefit. For instance, there was a time when in airline terminals passengers formed separate queues in front of check-in counters. But now we see invariably only one line feeding into several counters. This is the result of the realization that a single line policy serves better for the passengers as well as the airline management. Such a conclusion has come from analyzing the mode by which a queue is formed and the service is provided. The analysis is based on building a mathematical model representing the process of arrival of passengers who join the queue, the rules by which they are allowed into service, and the time it takes to serve the passengers. Queueing theory embodies the full gamut of such models covering all perceivable systems which incorporate characteristics of a queue.

We identify the unit demanding service, whether it is human or otherwise, as *customer*. The unit providing service is known as the *server*. This terminology of customers and servers is used in a generic sense regardless of the nature of the physical context. Some examples are given below:

(a) In communication systems, voice or data traffic queue up for lines for transmission. A simple example is the telephone exchange.

(b) In a manufacturing system with several work stations, units completing work in one station wait for access to the next.

(c) Vehicles requiring service wait for their turn in a garage.

(d) Patients arrive at a doctor's clinic for treatment.

© Springer Science+Business Media New York 2015

U. N. Bhat, *An Introduction to Queueing Theory*, Statistics for Industry and Technology, DOI 10.1007/978-0-8176-8421-1_1

Numerous examples of this type are of everyday occurrence. While analyzing them we can identify some basic elements of the systems.

Input Process If the occurrence of arrivals and the offer of service are strictly according to schedule, a queue can be avoided. But in practice this does not happen. In most cases, the arrivals are the product of external factors. Therefore, the best one can do is to describe the input process in terms of random variables that represent either the number arriving during a time interval or the time interval between successive arrivals. If customers arrive in groups, their size can be a random variable as well.

Service Mechanism The uncertainties involved in the service mechanism are the number of servers, the number of customers getting served at any time, and the duration and mode of service. Networks of queues consist of more than one server arranged in series and/or parallel. Random variables are used to represent service times, and the number of servers, when appropriate. If service is provided for customers in groups, their size can also be a random variable.

System Capacity The number of customers that can wait at a time in a queueing system is a significant factor for consideration. If the waiting room is large, one can assume that for all practical purposes, it is infinite. But our everyday experience with the telephone systems tells us that the size of the buffer that accommodates our call while waiting to get a free line is important as well.

Queue Discipline All other factors regarding the rules of conduct of the queue can be pooled under this heading. One of these is the rule followed by the server in accepting customers for service. In this context, the rules such as "first-come, first-served" (FCFS), "last-come, first-served" (LCFS), and "random selection for service" (RS) are self-explanatory. Others such as "round robin" and "shortest processing time" may need some elaboration, which is provided in later chapters. In many situations, customers in some classes get priority for service over others. There are many other queue disciplines which have been introduced for the efficient operation of computers and communication systems. Also, there are other factors of customer behavior such as balking, reneging, and jockeying, that require consideration as well.

The identification of these elements provides a taxonomy for symbolically representing queueing systems with a variety of system elements. The basic representation widely used in queueing theory is due to D. G. Kendall (1953) and made up of symbols representing three elements: input, service, and number of servers. For instance, using M for Poisson or exponential, D for deterministic (constant), E_k for the Erlang distribution with scale parameter k, and G for general (also GI, for general independent) we write:

$M/G/1$: Poisson arrivals, general service, single server
$E_k/M/1$: Erlangian arrival, exponential service, single server
$M/D/s$: Poisson arrival, constant service, s servers.
These symbolic representations are modified when other factors are involved.

1.2 Problems in a Queueing System

The ultimate objective of the analysis of queueing systems is to understand the behavior of their underlying processes so that informed and intelligent decisions can be made in their management. Three types of problems can be identified in this process.

Behavioral Problems The study of behavioral problems of queueing systems is intended to understand how they behave under various conditions. The bulk of the results in queueing theory is based on research on behavioral problems. Mathematical models for the probability relationships among the various elements of the underlying process are used in the analysis. To make the ideas concrete let us define a few terms that are defined formally later. A collection or a sequence of random variables that are indexed by a parameter such as time is known as a *stochastic process*; e.g., an hourly record of the number of accidents occurring in a city. In the context of a queueing system, the number of customers with time as the parameter is a stochastic process. Let $Q(t)$ be the number of customers in the system at time t. This number is the difference between the number of arrivals and departures during $(0, t)$. Let $A(t)$ and $D(t)$, respectively, be these numbers. A simple relationship would then be $Q(t) = A(t) - D(t)$. In order to manage the system efficiently, one has to understand how the process $Q(t)$ behaves over time. Since the process $Q(t)$ is dependent on $A(t)$ and $D(t)$, both of which are also stochastic processes, their properties and dependence characteristics between the two should also be understood. All these are idealized models to varied degrees of realism. As done in many other branches of science, they are studied analytically with the hope that the information obtained from such study will be useful in the decision-making process.

In addition to the number of customers in the system, which we call the *queue length*, the time a new arrival has to wait till its service begins (*waiting time*) and the length of time the server is continuously busy (*busy period*) or continuously idle (*idle period*) are major characteristics of interest. It should be noted that the queue length and the waiting time are stochastic processes and the busy period is a random variable. Distribution characteristics of the stochastic processes and random variables are needed to understand their behavior. Since time is a factor, the analysis has to make a distinction between the *time-dependent*, also known as *transient*, and the *limiting*, also known as the *long-term*, behavior. Under certain conditions a stochastic process may settle down to what is commonly

called a *steady state* or a state of *equilibrium*, in which its distribution properties are independent of time.

Statistical Problems Under statistical problems we include the analysis of empirical data in order to identify the correct mathematical model, and validation methods to determine whether the proposed model is appropriate. Chronologically, the statistical study precedes the behavioral study as could be seen from the early papers by A. K. Erlang (as reported in Brockmeyer et al. (1960)) and others. For an insight into the selection of the correct mathematical model, which could be used to derive its properties, a statistical study is fundamental.

In the course of modeling we make several assumptions regarding the basic elements of the model. Naturally, there should be a mechanism by which these assumptions could be verified. Starting with testing the goodness of fit for the arrival and service distributions, one would need to estimate the parameters of the model and/or test hypotheses concerning the parameters or behavior of the system. Other important questions where statistical procedures play a part are in the determination of the inherent dependencies among elements, and dependence of the system on time.

Decision Problems Under this heading we include all problems that are inherent in the operation of queueing systems. Some such problems are statistical in nature. Others are related to the design, control, and the measurement of effectiveness of the systems.

1.3 A Historical Perspective

The history of queueing theory goes back more than 100 years. Johannsen's "Waiting Times and Number of Calls" (an article published in 1907 and reprinted in *Post Office Electrical Engineers Journal*, London, October, 1910) seems to be the first paper on the subject. But the method used in this paper was not mathematically exact and therefore, from the point of view of exact treatment, the paper that has historic importance is A. K. Erlang's, "The Theory of Probabilities and Telephone Conversations" (*Nyt tidsskrift for Matematik, B, 20* (1909), p. 33). In this paper he lays the foundation for the place of Poisson (and hence, exponential) distribution in queueing theory. His papers written in the next 20 years contain some of the most important concepts and techniques; the notion of statistical equilibrium and the method of writing down state balance equations are two such examples. Special mention should be made of his paper "On the Rational Determination of the Number of Circuits" (see Brockmeyer et al. (1960)), in which an optimization problem in queueing theory was tackled for the first time.

It should be noted that in Erlang's work, as well as the work done by others in the twenties and thirties, the motivation has been the practical problem of congestion. See for instance, Molina (1927) and Fry (1928). During the next two

decades, several theoreticians became interested in these problems and developed general models which could be used in more complex situations. Some of the authors with important contributions are Crommelin, Feller, Jensen, Khintchine, Kolmogorov, Palm, and Pollaczek. A detailed account of the investigations made by these authors may be found in books by Syski (1960) and Saaty (1961). Kolmogorov's and Feller's study of purely discontinuous processes laid the foundation for the theory of Markov processes as it developed in later years.

Noting the inadequacy of the equilibrium theory in many queue situations, Pollaczek (1934) began investigations of the behavior of the system during a finite time interval. Since then and throughout his career, he did considerable work in the analytical behavioral study of queueing systems; see Pollaczek (1965). The trend toward the analytical study of the basic stochastic processes of the system continued, and queueing theory proved to be a fertile field for researchers who wanted to do fundamental research on stochastic processes involving mathematical models.

A concept that plays a significant role in the analysis of stochastic systems is *statistical equilibrium*. This is a state of the stochastic process which signifies that its behavior is independent of time and the initial state. Suppose we define

$$P_{ij}(s,t) = P[Q(t) = j | Q(s) = i] \qquad s < t$$

as the *transition probability* of the process $\{Q(t),\ t \geq 0\}$, which is a statement of the probability distribution of the state of the process at time t, conditional on its state at time s, $s < t$. The statement that the process attains statistical equilibrium implies that

$$\lim_{t \to \infty} P_{ij}(s,t) = p_j$$

which does not depend on time t and the initial state i.

Even though Erlang did not explicity state his results in these terms, he used this basic concept in his results. To this day a large majority of queueing theory results used in practice are those derived under the assumption of statistical equilibrium. Nevertheless, to understand the underlying processes fully, a time-dependent analysis is essential. But the processes involved are not simple and for such an analysis sophisticated mathematical procedures become necessary. Thus, the growth of queueing theory can be traced on two parallel tracks:

(i) Using existing mathematical techniques or developing new ones for the analysis of the underlying processes

(ii) Incorporating various system characteristics to make the model closely represent the real-world phenomenon

Queueing theory as an identifiable body of literature was essentially defined by the foundational research of the 1950s and 1960s. For a complete bibliography of research in this period, see Syski (1960), Saaty (1961, 1966), and Bhat (1969). Here we mention only a few papers and books that, in the opinion of this author, have made a profound impact in the direction of research in queueing theory.

The queue $M/M/1$ (Poisson arrival, exponential service, single server) is one of the earliest systems to be analyzed. Under statistical equilibrium, the state balance equations are simple and the limiting distribution of the queue size is obtained by recursive arguments. But for a time-dependent solution, more advanced mathematical techniques become necessary. The first such solution was given by Bailey (1954) using generating functions for the differential equations governing the underlying process, while Lederman and Reuter (1956) used spectral theory in their solution. Laplace transforms were used later for the same problem, and their use together with generating functions has been one of the standard and popular procedures in the anlaysis of queueing systems ever since.

A probabilistic approach to the analysis was initiated by Kendall (1951, 1953) when he demonstrated that imbedded Markov chains can be identified in the queue length process in systems $M/G/1$ and $GI/M/s$. Lindley (1952) derived integral equations for waiting time distributions defined at imbedded Markov points in the general queue $GI/G/1$. These investigations led to the use of renewal theory in queueing systems analysis in the 1960s. Identification of the imbedded Markov chains also facilitated the use of combinatorial methods by considering the queue length at Markov points as a random walk. See Prabhu and Bhat (1963) and Takàcs (1967).

Mathematical modeling of a random phenomenon is a process of approximation. A probabilistic model brings it a little bit closer to reality; nevertheless it cannot completely represent the real-world phenomenon because of involved uncertainties. Therefore, it is a matter of convenience where one can draw the line between the simplicity of the model and the closeness of the representation. In the 1960s several authors initiated studies on the role of approximations in the analysis of queueing systems. Because of the need for useable results in applications, various types of approximations have appeared in the literature. For an extensive bibliography, see Bhat et al. (1979). To mention a few, one approach to approximation is the analysis under heavy traffic (when the traffic intensity, the ratio of the rates of input to output, approaches 1) and investigations under this topic were initiated by Kingman (for an extensive bibliography, see Kingman (1965)) with the objective of deriving a simpler expression for the final result. The heavy traffic assumption also led to diffusion approximation as well as weak convergence results by researchers such as Iglehart (see Iglehart and Whitt (1970a, b)). Also see Whitt (2000) with an extensive bibliography. Gaver's analysis (1968) of the virtual waiting time of an $M/G/1$ queue is one of the initial efforts using diffusion approximation for a queueing system. Fluid approximation, as suggested by Newell (1968, 1971) considers the arrival and departure processes in the system as a fluid flowing in and out of a reservoir, and their properties are derived using applied mathematical techniques. For a recent survey of some fluid models see Kulkarni (1997).

By the end of 1960s most of the basic queueing systems that could be considered as reasonable models of real-world phenomena had been analyzed and the papers coming out dealt with only minor variations of the systems without

contributing much to methodology. There were even statements made to the effect that queueing theory was at the last stages of its life. But such predictions were made without knowing what advances in computer technology would mean to queueing theory. Advances inspired or assisted by computer technology have come in two dimensions: methodology and applications. Given below are some of the prominent topics explored in such advances. Since in applied probability, methodology, and applications contribute to the growth of the subject in a symbiotic manner they are listed below without being categorized.

(i) The Matrix-Analytic Method

Starting with the introduction of phase type probability distributions, Marcel Neuts (1975) has developed an analysis technique that extends and modifies the earlier transform method to multivariables and makes it amenable for an algorithmic solution. See Neuts (1978, 1981), Sengupta (1989), and Ramaswami (1990, 2001). The use of phase type distributions in the representation of system elements and the matrix-analytic method in their analysis has significantly expanded the scope of queueing systems for which useable results can be derived. See, Chapter 8 for details.

(ii) Transform Inversion

The traditional method of analysis of queueing systems depends on inverting generating functions and/or Laplace transforms to derive useable results. The complexities of transform inversion has spurred more research on it and beginning with Abate and Dubner (1968), Dubner and Abate (1968), and Abate et al. (1968) many papers have been published on the subject. For a comprehensive survey of the state of the art of the Fourier series method of inversion see Abate and Whitt (1992).

In the inversion of Laplace transforms and probability generating functions, finding roots of characteristic equations is a key step. The celebrated Rouché's theorem only establishes the existence of the roots, not their magnitude. Pioneering and painstaking work in adapting various root finding algorithms for use in inverting transforms and generating functions is due to Professor M. L. Chaudhry (1992). Starting from the 1970s, along with his associates, he has put together a significant amount of research on various queueing systems of interest (see, Chaudhry and Templeton (1983)). For instance Chaudhry et al. (1992) provides a good illustration.

(iii) Queueing Networks

The first article on queueing networks is by J. Jackson (1957). Mathematical foundations for the analysis of queueing networks are due to Whittle (1967, 1968) and Kingman (1969), who treated them in the terminology of population processes. Complex queueing network problems have been investigated extensively since the beginning of the 1970s.

Two key concepts that advanced investigations into the properties of queueing networks are: the Poisson nature of the departure process from an M/M/s type queue (Burke 1956) and the local balance in state transitions (Whittle 1967, 1968). The $M \rightarrow M$ property, as the Poisson property has been called in computer network lierature, is a necessary condition for the limiting distribution to be in the product form. Going beyond the simple Jackson network, Baskett et al. (1975) show that the product from solutions are valid for networks more general than those with simple M/M/s type nodes, such as, with state-dependent service; heterogeneous service times; Coxian service time distributions; processor sharing discipline; and last-come, first-served discipline.

Since the publication of Baskett et al., a large body of literature has grown in the performance modeling of queueing networks. Courtois (1977), Kelley (1979), Sauer and Chandy (1981), Lavenberg (1983), Disney and Kiessler (1987), Malloy (1989), Perros (1994), Gelenbe and Pujolle (1999) and Giambene (2005) are some of the significant books that have come out on this subject.

(iv) Computer and Communication Systems

The need to analyze traffic processes in the rapidly growing computer and communication industry is the primary reason for the resurgence of queueing theory after the 1960s. Research on queueing networks (see references cited earlier) and books such as Coffman and Denning (1973) and Kleinrock (1975, 1976) laid the foundation for a vigorous growth in the application of queueing theory in computer and communication system operation.

In tracking this growth, we may cite the following survey type articles from the journal *Queueing Systems*: Denning and Buzen (1978) on the operational analysis of queueing network models; Coffman and Hoffri (1986), describing important computer devices and the queueing models used in analyzing their performance; Yashkov (1987) on analytical time-sharing models, complementary to McKinney (1969) on the same topic; three special issues of the journal edited by Mitra and Mitrani (1991), Doshi and Yao (1995), and Konstantopolous (1998); and a paper by Mitra et al. (1991) on communication systems. Research on queueing applications can also be found in various computer journals. Several books have appeared and continue to appear on the subject as well. Some of the more recent developments are discussed in Chapter 13.

(v) Manufactruring Systems

The machine interference problem analyzed by Palm (1947) and Benson and Cox (1951, 1952) was the first problem in manufacturing systems in which queueing theory methodology was used. The classical Jackson network (1957) originated out of the manufacturing setting since a job-shop is a network of machines. (Also, see Jackson (1963)). Simulation

studies reported in Conway et al. (1967) provide excellent examples of the incorporation of queueing models with job-shop scheduling. Since the 1970s, with the advent of new processes in manufacturing incorporating computers at various stages, the application of queueing theory results as well as the development of new techniques have occurred at a phenomenal rate. Three articles in Buzacott and Shanthikumar (1992) and the book Buzacott and Shanthikumar (1993) bring together most of the important developments in the application of queueing theory in manufacturing systems up to that time.

As described by Buzacott and Shanthikumar (1993) the "product-to-order" and "product-to-stock" models make direct use of queueing theory results. With demand as a customer and the manufacturing process as a server, the first model is a direct application of queueing models, while the second incorporates production–inventory system concepts, with the production system substituting for multiple or infinite number of servers. Other applications include job flow lines as tandem queues, and job-shops and flexible manufacturing systems as queueing networks. Some of the more recent applications are discussed in Chapter 12. For recent articles on the applications of queueing theory in manufacturing system modeling readers may also refer to various journals such as *Management Science, European Journal of Operational Research, IIE Transactions, Computers and Industrial Engineering*, and journals on production and manufacturing research.

(vi) Specialized Models

Specialized queueing models of the 1950s and 1960s have found broader applicability in the context of computer and communication systems. We mention below three such models that have attracted considerable attention.

Polling Models These models represent systems in which one or more servers provide service to several queues in a cyclical manner (Koenigsberg (1958)). Based on variations on the system structure and queue discipline a large number of models emerge. For research on polling models see a special issue of *Queueing Systems* edited by Boxma and Takagi (1992), as well as Takagi (1997) and Hirayama et al. (2004), all of which provide excellent bibliography on the subject.

Vacation Models Queueing systems with service breaks are not uncommon. Machine breakdowns, service disruption due to maintenance operations, cyclic server queues, and scheduled job streams are some of the examples. A key feature of the results is the ability to decompose them into results corresponding to systems without vacations and results depending on the distributions related to the vacation sequence. For bibliographies on this topic, see Doshi (1986) and Alfa (2003).

Retrial Queues In finite capacity systems, customers, denied entry to the system, trying to enter again, is quite common. Since they have already tried to get service once, they belong to a different population of customers than the original one. Problems related to this phenomenon have been extensively explored in the literature. The following papers and more recent ones appearing in journals provide bibliographies for further study: Yang and Templeton (1987), Falin (1990), and Kulkarni and Liang (1997).

(vii) Statistical Inference

In any theory of stochastic modeling statistical problems naturally arise in the applications of the models. Identification of the appropraite model, estimation of parameters from empirical data, and drawing inferences regarding future operations involve statistical procedures. These were recognized even in earlier investigations in the studies by Erlang; see Brockmeyer et al. (1960), Molina (1927), and Fry (1928).

Since elements contributing to the underlying processes in queueing systems can be modeled as random variables and their distributions, it is reasonable to assume that inference problems in queueing are not any different from such problems in statistics in general. However, often in real-world systems, sampling plans appropriate for data collection to estimate parameters of the constituent elements, may not be possible to implement. Consequently, modifications of the standard statistical procedures become necessary.

The first theoretical treatment of the estimation problem was given by Clarke (1957) who derived maximum likelihood estimates of arrival and service rates in an $M/M/1$ queueing system. Billingsley's (1961) treatment of inference in Markov processes in general and Wolff's (1965) derivation of likelihood ratio tests and maximum likelihood estimates for queues that can be modeled as birth and death processes are other significant advances that have occurred in this area. Also see Cox (1965) for a comprehensive survey of statistical problems as related to queues. Cox also provides a broad guideline for inference investigations in non-Markovian queues.

The first paper on estimating parameters in a non-Markovian system is by Goyal and Harris (1972), who used the transition probabilities of the imbedded Markov chain to set up the likelihood function. Since then, significant progress has occurred in adapting statistical procedures to various systems. Some of the examples are: Basawa and Prabhu (1981, 1988) and Acharya (1999) considered the problem of estimation of parameters in the queue $GI/G/1$; Rao et al. (1984) used a sequential probability ratio technique for the control of parameters in $M/E_k/1$ and $E_k/M/1$; Armero (1994) and Armero and Conesa (2000) used Bayesian techniques for inference in Markovian queues; Thiruvaiyaru et al. (1991) and Thiruvaiyaru and Basawa (1994) extended the maximum likelihood estimation

to include Jackson networks; Pitts (1994) considered the queue as a functional that maps the service and inter-arrival time distribution functions on to the stationary waiting time distribution function to determine its confidence bound. For a comprehensive survey of inference problems in queues see Bhat et al. (1997). More recent investigations are by Bhat and Basawa (2002) who use queue length as well as waiting time data in estimating parameters in queueing systems. A recent paper (Basawa et al. 2008) uses waiting time or system sojourn time, adjusted for idle times when necessary, to estimate parameters of inter-arrival and service times in $GI/G/1$ queues.

(viii) Design and Control

The study of real systems is motivated by the objectives of improving their design, control and effectiveness. Until the 1960s when operations researchers trained in mathematical optimization techniques got interested in queueing problems, operational problems were being handled using primarily behavioral results. It should be noted that Erlang's interest in the subject was for building better telephone systems for the company for which he was working. His paper "On the rational determination of the number of circuits" (Brockmeyer et al. (1960)) deals with the determination of the optimum number of channels so as to reduce the probability of loss in the system.

Until computers made them obsolete, graphs and tables, prepared using analytical results of measures of effectiveness, assisted the designers of communication systems such as telephones. Other examples are the papers by Bailey (1952) which looked into the appointment system in hospitals, and Edie (1956) that analyzed the traffic delays at tollbooths. From the perspective of applications of queueing results to realistic problems Morse's (1958) book has been held in high regard. This is because he presented the theoretical results available at that time in a manner appealing to the applied researchers and gave procedures for improving system design.

Hillier's (1963) paper on economic models for industrial waiting line problems is, perhaps, the first paper to introduce standard optimization techniques to queueing problems. While Hillier considered an $M/M/1$ queue, Heyman (1968) derived an optimal policy for turning the server on and off in an $M/G/1$ queue, depending on the state of the system.

Since then, operations researchers trained in mathematical optimization techniques have explored their use in much greater complexity to a large number of queueing systems. For an excellent overview, a valuable reference is a special issue of the journal *Queueing Systems* edited by Stidham (1995), which includes several review-type articles on special topics. Also see Bäuerle (2002) who considers an optimal control problem in a queueing network.

(ix) Other Topics

Even though there were a few papers on discrete time queues before the 1970s, since then, these systems have taken a larger significance because of the discreteness of time, however short the interval maybe, in computer and communication systems. It is not hard to imagine that a large portion of the results for discrete time queues are in fact derived in the same way as for continuous time queues with obvious modifications in methodology.

There have also been theoretical advances in stochastic processes with the introduction of modified processes such as Markov modulated processes, marked point processes and batch Markovian processes. These processes are used to represent various patterns such as burstiness and heterogeneity in traffic.

In the preceding paragraphs, we have outlined the growth of queueing theory identifying major developments and directions. For details of any of the facets, readers are referred to the articles and books cited above. Also see Prabhu (1987) who gives a bibliography of books and survey papers in various categories and subtopics, Adan et al. (2001) who give a broad treatment of queues with multiple waiting lines, and Dshalalow (1997) who considers systems with state-dependent parameters. The last two articles also provide extensive bibliographies. It is hoped that with the help of these references and modern Internet tools, applied researchers will be able to build on the systems covered in this text so as to establish an appropriate model to represent the system of their interest.

1.4 Modeling Exercises

These exercises are given as an introduction to modeling a random phenomenon as a queueing system. In addition to answering the questions posed in the exercises, the reader is required only to identify (i) model elements, (ii) system structure, and (iii) the assumptions one has to make in setting up the model.

1. A city bus company wants to establish a schedule for its bus fleet. In order to do this in a scientific manner, the company entrusts this job to an operations research specialist with sufficient data processing support. Describe the queueing systems involved in this process and the types of data that need to be collected in order to come up with the schedule. Identify the measures of performance for the bus system and the factors that affect these measures when the system is in operation.

2. A newly established business would like to decide on the number of telephone lines it has to install in a cost-effective manner. Identify the elements of the underlying process of the telephone answering system and indicate the specific data that need to be collected to establish the parameters of the system. Also identify the performance measures of interest.

3. In a manufacturing system, a product undergoes several stages (e.g., an automobile assembly line) and within each stage there may be several substages, including testing of components. How can such a system be modeled as a queueing system (including queueing systems for stages and substages) in order to improve the performance of the manufacturing process?

4. An airline offers three types of check-in service for the passengers: (1) First class and business class check-in, (2) regular check-in, and (3) self check-in. Describe the structure of the queueing system that can represent the check-in system and identify the data elements that need to be known to measure its performance. Also indicate the complexities that may result in improving the system by incorporating flexibilities in the system operation.

5. Several terminals used for data entry to a computer share a communication line. Terminals use the line on a first-come, first-served basis and wait in a queue when the line is busy.

   Describe the elements of this queueing system and identify the assumptions that need to be made to analyze system characteristics. (Allen (1990)).

6. In store-and-forward communication networks messages for transmission are stored in buffers of fixed size. Each message may use one or more buffers. The message is transmitted through several identical channels. Knowing the characteristics of the arrival process, transmission rate, and the message length, we are interested in the storage requirements of a network node.

   Describe the general characteristics of the approach in order to estimate the long run storage requirements for this type of a system.

7. In a warehouse, items are stacked in such a way that the most recently stacked item gets removed first. In order to use a queueing model to determine the amount of time the item is stored in the warehouse, describe the elements of such a system and say how we may characterize the time interval of interest.

8. In order to reduce the waiting time of short jobs, a round-robin (RR) service discipline is used. Under an RR queue discipline, each job gets a fixed amount of service, known as a quantum, when it is admitted to the central processing unit (CPU). If the service requirement of the job is more than the quantum, it is sent back to the end of the queue of waiting jobs. This process continues until the CPU can provide the required number of quanta of service to the job.

   Describe how the total service time of the job can be characterized in order to determine the mean amount of time the job spends in the system. (This is known as the *mean response time*.) (See Coffman and Kleinrock (1968) and Coffman and Denning (1973)).

9. A uniprogramming computer system consists of a CPU and a disk drive. After one pass at the CPU a job may need the services of the disk I/O with a certain probability, say p, and the job is complete with the probability $1-p$. There are three independent phases to disk service time: (1) seek time; (2) latency time; and (3) transfer time, each with a specified distribution. After disk service the job goes back to CPU for completing the execution. (Note that a uniprogramming system cannot start another job until the service on the one in the system is complete.)

   We are interested in determining the average response time (waiting time + service time). What type of a model is appropriate for this problem? If a queueing model is appropriate, describe the elements of the system (Trivedi (2002)).

10. In a drum storage unit a shortest-latency-time-first (SLTF) file drum is used to read or write records on files while the drum is rotating. Once a decision is made to process a particular record, the time spent waiting for the record to come under the read/write heads which are fixed is called the latency. The records are not constrained to be of any particular strength. Also, no restrictions are placed on the starting position of the records. Assume that the circumference of the drum is the unit of length and the drum rotates at a constant angular velocity, with period τ (Fuller (1980)).

    Suppose a queueing model is to be used to analyze the performance of the drum-storage unit described above. Describe the elements of such a system and the characteristics to be considered for its performance evaluation.

Chapter 2

System Element Models

2.1 Probability Distributions as Models

In building a suitable model for a queueing system, we start with its elements. Of the elements mentioned in Chapter 1, number of servers, system capacity, and discipline are normally deterministic (unless, the number of available servers becomes a random variable, which is also possible in some cases). But there are uncertainties related to arrivals and service which result in the underlying processes being stochastic.

The similarity of the arrival and service processes can be brought out by identifying similar components, such as inter-arrival times and service times; arrival epochs and departure epochs.

Of these pairs departure epochs are *almost always* from a nonempty system, whereas arrival epochs are *mostly* independent of the state of the system (exceptions are possible). Therefore, first we discuss the possibilities of using certain probability distributions to represent the process of inter-arrival times and service times. In the case of Poisson process discussed below, it is also convenient to consider the distribution of the number of events occurring in a given length of time.

To start with, we should note that depending on the properties of the basic process and convenience, we may use either continuous or discrete distributions. In many situations continuous distributions may be easier to handle analytically (algebra of discrete distributions could be cumbersome.); nevertheless, it is worthwhile to note that continuous and discrete models are mutual analogs and most of the properties carry through in both cases.

Keeping a common notation we use $Z_1, Z_2, \ldots$ as nonnegative random variables representing either inter-arrival times or service times of consecutive customers. Further, let

$$F(x) = P(Z_n \leq x), \quad n = 1, 2, \ldots$$

© Springer Science+Business Media New York 2015

U. N. Bhat, *An Introduction to Queueing Theory*, Statistics for Industry and Technology, DOI 10.1007/978-0-8176-8421-1_2

We also assume that $\{Z_n\}_{n=1}^{\infty}$ are independent and identically distributed random variables. Let

$$E[Z_n] = b, \quad n = 1, 2, \ldots$$

and define the Laplace–Stieltjes transform of $F(x)$ as

$$\psi(\theta) = \int_0^{\infty} e^{-\theta x} dF(x) \quad Re(\theta) \geq 0.$$

Clearly, we get

$$-\psi'(0) = b.$$

It should be noted that when b is the mean inter-occurrence time, $1/b$ is the rate of occurrence of the event.

In considering the suitability of a probability model for a random phenomenon, moment properties of the model distribution become useful. Many times the first two moments appear as the parameters of the model. Furthermore, the first few moments describe the shape of the density curve, thus, making them suitable measures in selecting the model (e.g., coefficient of variation (CV) = s.d./mean; coefficient of skewness = (third moment)/(s.d.)^3; coefficient of kurtosis = (fourth moment)/(s.d.)^4).

The commonly used distribution models for arrivals and service are: deterministic (when arrivals are at specified time epochs, or inter-arrival times or service times are of constant length); exponential (as distribution models for inter-arrival times or service times); Poisson (as the distribution of the number of arrivals during a specified length of time); Erlang (as distribution models for inter-arrival times or service times); and variants of these distributions. We introduce deterministic, exponential, Poisson, and Erlang distributions in the following discussion and the remainder in Appendix A.

Deterministic Distribution (D)

Let

$$\begin{aligned} F(x) &= 0 \quad x < b \\ &= 1 \quad x \geq b \end{aligned} \tag{2.1.1}$$

We get $E(Z_n) = b$ and $\psi(\theta) = e^{-\theta b}$. Also, $V(Z_n) = 0$.

This seemingly simple distribution is suitable when arrivals take place at equal intervals of time (interval length b) or service takes exactly b units of time. In practice, however, it may be hard to achieve this exactness. Early or late arrivals, early or late service completions will be closer to reality. In such cases, the assumption of a deterministic distribution should be considered a reasonable approximation of the real system.

If we are interested in an exact model for the early or late occurrence of events, we may consider the displacement from the deterministic epoch as a random variable with some distribution like the uniform or the normal. Under these conditions, it is possible to have the kth scheduled event occurring later than the occurrence of the $(k+1)$th scheduled event.

Exponential Distribution, Poisson Process (M)

Let

$$F(x) = 1 - e^{-\lambda x}, \quad x \geq 0, \quad \lambda > 0. \tag{2.1.2}$$

Then we get,

$$\begin{aligned} f(x) &= \frac{d}{dx}F(x) = \lambda e^{-\lambda x} \\ E[Z_n] &= \frac{1}{\lambda} \end{aligned}$$

and

$$\psi(\theta) = \frac{\lambda}{\theta + \lambda}.$$

Also, $V(Z_n) = \frac{1}{\lambda^2}$ and $CV(Z_n) = 1$.

Let $X(t)$ be the number of events occurring in time t such that the inter-occurrence times have the distribution given by $F(x)$. Symbolically, for the stochastic process $X(t)$ we can write

$$X(t) = \max\{n | Z_1 + Z_2 + \ldots + Z_n \leq t\}.$$

Let

$$\begin{aligned} P_n(t) &= P(X(t) = n | X(0) = 0) \\ &= P(Z_1 + Z_2 + \ldots + Z_n \leq t) \\ &\quad -P(Z_1 + Z_2 + \ldots + Z_{n+1} \leq t), \end{aligned}$$

where $F_n(t) = P(Z_1 + Z_2 + \ldots + Z_n \leq t)$ is obtained as the n-fold convolution of $F(t)$ with itself. Using the Laplace transform of $F(t)$ we find

$$\int_0^\infty e^{-\theta t} dF_n(t) = \left(\frac{\lambda}{\theta + \lambda}\right)^n.$$

On inversion this gives

$$\begin{aligned} F_n(t) &= \int_0^t e^{-\lambda y} \frac{\lambda^n y^{n-1}}{(n-1)!} dy \\ &= 1 - \sum_{r=0}^{n-1} e^{-\lambda t} \frac{(\lambda t)^r}{r!}. \end{aligned} \tag{2.1.3}$$

Thus, we get

$$\begin{aligned} P_n(t) &= F_n(t) - F_{n+1}(t) \\ &= [1 - \sum_{r=0}^{n-1} e^{-\lambda t} \frac{(\lambda t)^r}{r!} \\ &\quad -[1 - \sum_{r=0}^{n} e^{-\lambda t} \frac{(\lambda t)^r}{r!}] \\ &= e^{-\lambda t} \frac{(\lambda t)^n}{n!}, \end{aligned} \tag{2.1.4}$$

which is a Poisson distribution with mean λt. Hence, $X(t)$ is known as a *Poisson process.*

Define the probability generating function of $X(t)$ as

$$\Pi(z,t) = \sum_{n=0}^{\infty} z^n P_n(t) \quad |z| \leq 1.$$

For the Poisson process we get

$$\Pi(z,t) = e^{-\lambda(1-z)t}.$$

Also, $E[X(t)] = \lambda t$ and $V[X(t)] = \lambda t$.

The Poisson process is a special case of the Markov process which is introduced in the next chapter. It is widely used in stochastic modeling because of its properties with reference to the occurrence of events and the properties of the exponential distribution representing the corresponding inter-occurrence times of events. Two of them are given below: (a) the first describes the *memoryless property* of the exponential distribution and (b) the second generates the Erlang distribution.

(a) When $P(Z_n \leq x) = 1 - e^{-\lambda x}$ $(\lambda > 0)$

$$\begin{aligned} P(Z_n \leq t + x | Z_n > t) &= \frac{P(t < Z_n < t + x)}{P(Z_n > t)} \\ &= \frac{[1 - e^{-\lambda(t+x)}] - [1 - e^{-\lambda t}]}{e^{-\lambda t}} \\ &= 1 - e^{-\lambda x}. \end{aligned} \tag{2.1.5}$$

The implication of this property is that if an interval, such as service time, can be represented by an exponential distribution and the interval is ongoing at time t, the remaining time in the interval has the same distribution as the original one, regardless of the start of the interval. This property is commonly known as the *memoryless* property of the exponential distribution.

(b) The discussion leading to equation (2.1.3) implies that the time required for the occurrence of a given number of Poisson events has a distribution given by that expression, i.e., if Y_n is the waiting time until the nth occurrence and $\{Z_1, Z_2, \ldots\}$ are the inter-occurrence times

$$\begin{aligned} Y_n &= Z_1 + Z_2 + \ldots + Z_n \\ F_n(t) &= P(Y_n \leq t) \\ &= \int_0^t e^{-\lambda y} \frac{\lambda^n y^{n-1}}{(n-1)!} dy \end{aligned}$$

and

$$f_n(y) = e^{-\lambda y} \frac{\lambda^n y^{n-1}}{(n-1)!} dy \quad (y > 0). \tag{2.1.6}$$

The distribution given by (2.1.6) is a gamma distribution with parameters n and λ. In queueing theory it is commonly called the *Erlang distribution* with scale parameter n. It is symbolically denoted by E_n. Equation (2.1.3) also establishes a useful identity

$$\int_y^\infty e^{-\lambda x}\frac{(\lambda x)^{n-1}}{(n-1)!}\lambda dx = \sum_{r=0}^{n-1} e^{-\lambda y}\frac{(\lambda y)^r}{r!}. \tag{2.1.7}$$

For modeling purposes, Poisson process is considered an appropriate model for events occurring "at random." The reasons for such a characterization rests on its properties described in Appendix A; specifically, independence of events occurring in nonoverlapping intervals of time, the constant rate of occurrence independent of time, the independent and identically distributed nature of the inter-occurrence times, and its relationship with the uniform distribution as expressed in (A.1.4) of Appendix A. The significance of the Erlang distribution stems from the phase interpretation that can be provided for generating a suitable arrival or service process.

Consider a Poisson arrival process and suppose a queueing system admits every kth customer into the system instead of all arrivals. Now the inter-arrival time between effective arrivals to the queueing system is the sum of k exponential random variables with mean $1/\lambda$, hence, it has the distribution given by (2.1.6). Similarly, consider a service process in which a customer goes through k phases of service, each phase being exponentially distributed with mean $1/\lambda$. The total service time has the distribution (E_k), given by (2.1.6) with n replaced by k.

To facilitate comparison with the Poisson and deterministic processes consider the Erlang distribution $F(x)$ with mean $1/\lambda$. This can be accomplished by starting with an exponential distribution with parameter $k\lambda$. Then we get

$$\begin{aligned} F(x) &= \int_0^x e^{-k\lambda y}\frac{(k\lambda)^k y^{k-1}}{(k-1)!}dy \\ f(x) &= e^{-k\lambda x}\frac{(k\lambda)^k x^{k-1}}{(k-1)!}. \end{aligned} \tag{2.1.8}$$

For $k = 1$, we have the exponential distribution, which generates a Poisson process. To determine the form of $f(x)$ as $k \to \infty$, we use its transform $\psi(\theta)$. We have

$$\psi(\theta) = \left(\frac{k\lambda}{k\lambda+\theta}\right)^k = \frac{1}{(1+\theta/k\lambda)^k} \to e^{-\theta/\lambda} \text{ as } k \to \infty.$$

The resulting transform is the transform of a constant $1/\lambda$, and hence generates the deterministic distribution given in (2.1.1). Depending on the values of k, even a moderately large value of k (e.g., $k = 10$ or 15) may be sufficient for the Erlang to exhibit the property of a deterministic distribution.

2.2 Identification of Models

In the formulation of a queueing model, one starts with the identification of its elements and their properties. The system structure is easily determined. What remains is the determination of the form and properties of the input and service processes. Four major steps are essential in this analysis (i) collection of data, (ii) tests for stationarity in time, (iii) tests for independence, and (iv) distribution selection and/or estimation.

2.2.1 Collection of Data

To estimate parameters of system elements, one has to establish a sampling plan identifying the data elements to be collected with reference to specific parameters. For instance, the number of arrivals in a time period gives the arrival rate or the mean inter-arrival time, which are reciprocals of each other. Sometimes there is a tendency to use empirical performance measures to estimate parameters intrinsic to the model. For instance, in an $M/M/1$ queue, noting that the traffic intensity (which is the ratio of arrival to service rate) provides the utilization factor for the system, we may use the empirical utilization factor as its estimate. Some of the pitfalls of this approach are indicated by Cox (1965) who notes that if ρ is the traffic intensity, the efficiency of this approach is given by $1 - \rho$. Also, see the discussion by Burke following Cox's article on the bias resulting from estimating the load factor in an $M/M/$s loss system as (average number of customers in systems)/(1-probability of loss).

The length and the mode of observation are problems of interest in a sampling plan. If the arrival process is Poisson, Birnbaum (1954) has shown that observing the system until a specific number of events have occurred gives a better sample than observing for a specific amount of time. But when nothing is known regarding the processes, no such statements can be made and the efficiency of different schemes should be considered in individual cases. Another aspect of the sampling plan is the mode of observations; for discussions of what are known as the snap reading method and systematic sampling, the reader is referred to Cox (1965), and page 86 of Cox (1962), respectively.

2.2.2 Tests for Stationarity

Cox and Lewis (1966) give a comprehensive treatment of tests for stationarity in stochastic processes. In addition to the treatment of data on the occurrence of events as a time series and the determination of second-order properties of the counting processes, they consider statistical problems related to renewal processes and provide tests of significance in some general, as well as some specific cases. Lewis (1972) updates this study and considers topics such as trend analysis of nonhomogeneous Poisson processes.

In many queueing systems (such as airport and telephone traffic), the nonstationarity of the arrival process leads to a periodic behavior. Furthermore,

even though the process is nonstationary when the entire period is considered, it might be possible to consider it as a piecewise stationary process in which stationary periods can be identified (e.g., a rush hour). Under such circumstances, a procedure that can be used to test the stationarity of the process, as well as to identify stationary periods, is the Mann–Whitney–Wilcoxon test (see, for example, Conover (1971), or Randles and Wolfe (1979), or a test appropriately modified to handle ties in ranks as in Putter (1959)). The data for the test can be obtained by considering two adjacent time intervals $(0, t_1]$ and $(t_1, t_2]$ and observing the number of arrivals during such intervals for several time periods. Let $X_1, X_2, \ldots, X_n$ be the number of arrivals during the first interval for n periods, and let $Y_1, Y_2, \ldots, Y_m$ be the number of arrivals during the second interval for m periods (usually $m = n$). If F and G represent the distributions of the $X's$ and $Y's$, respectively, then the hypothesis to be tested is $F = G$ against the alternative $F \neq G$, for which the Mann–Whitney–Wilcoxon statistic can be used. Using this test, successive stationary periods can be delineated and the system can be studied in detail within such periods (see Moore (1975), who gives an algorithm for the procedure).

To analyze cyclic trends of the type discussed above, we may also use the periodogram method described by Lewis (1972) for the specific case of a nonhomogeneous Poisson process. Another test in the framework of the nonhomogeneous Poisson process is proposed by Joseph et al. (1990). They consider the output of an M/G/∞-queue where G is assumed to be known.

2.2.3 Tests for Independence

While formulating queueing models, for simplification and convenience, several assumptions of independence are made about its elements. Thus, most of the models assume that inter-arrival times and service times are independent sequences of independent and identically distributed random variables. If there are reasons to make such assumptions, statistical tests can be used for verification. Some of the tests that can be used to verify independence of a sequence of observations are tests for serial independence in point processes, described in Lewis (1972), and various tests for trend analysis and renewal processes, given by Cox and Lewis (1966). To verify the assumption of independence between inter-arrival and service times, nonparametric tests seem appropriate. Spearman's rho and Kendall's tau (Randles and Wolfe (1979), Hollander et al. (2013)) are used to test the correlation between two sequences of random variables, whereas Cramer-von Mises type statistics (see Koziol and Nemec (1979) and references cited therein) are used to test for bivariate independence directly from the definition of independence applied to random variables.

2.3 Distribution Selection

The next step in the model identification process is the determination of the best model for arrival and service processes. The distribution selection problem is based on the nature of data and availability of model distributions. For

this problem, readers are referred to books on applied statistics and data analysis (e.g., Venables and Ripley (2002)). It is advisable to start with simple distributions such as the Poisson and exponential, since analysis under such assumptions is considerably simpler. After all, a mathematical model is essentially an approximation of a real process. The simpler the model is, the easier it is to analyze and to extract information from it. Thus, the selection of a distribution should be made with due consideration to the tradeoff between the advantages of the sophistication of the model and our ability to derive useful information from it.

Distributions such as Erlang and hyperexponential, are closely related to the exponential and with an appropriate selection of parameter values, they represent a wide variety of distributions. As noted in Appendix A, Erlang with coefficient of variation ≤ 1 and hyper-exponential with coefficient of variation ≥ 1, form a family of distributions with a broad range of distribution characteristics while retaining the convenience of analysis based on Markovian properties.

Once the distribution model is chosen, the next step is the determination of parameter values that bind the model to the real system. Normally either the maximum likelihood method or the method of moments is used for parameter estimation; the former is preferred because of its desirable statistical properties whereas the latter is used for its ease of implementation. A discussion of parameter estimation and hypothesis testing in queueing theory is given in Chapter 10.

2.4 Review Exercises

1. Determine the mean, variance, and coefficient of variation (CV) for the following distributions introduced in this chapter and Appendix A.

   (i) Deterministic, (2.1.1)

   (ii) Exponential, (2.1.2)

   (iii) Hyperexponential, (A.3.1)

   (iv) Erlang, (2.1.6), (A.4.1)

   (v) Mixed Erlang, (A.5.1), (A.5.2)

   (vi) Geometric, (A.8.1)

   (vii) Binomial, (A.8.3)

   (viii) Negative binomial, (A.8.4)

2. Determine the Laplace transform or the probability-generating function, as the case may be, for the distributions listed under Ex. 1.

3. Determine the probability-generating function for

   (i) The Poisson process, (A.2.1)

   (ii) The Compound Poisson process, (A.2.2)

4. Redo Exercise 1 using the Laplace transform or probability-generating function, as the case may be.

5. Determine for a specific value of t, the mean, variance, and coefficient of variation for

   (i) Poisson process

   (ii) Compound Poisson process

6. Establish the identity (2.1.7)

7. Establish the result (A.1.3)

8. Establish the result (A.2.4)

9. Determine the maximum likelihood estimates of the mean value parameters in distributions listed under Ex. #1.

Chapter 3

Basic Concepts in Stochastic Processes

3.1 Stochastic Process

In this chapter, we introduce basic concepts used in modeling queueing systems. Analysis techniques are developed later in conjunction with the discussion of specific systems.

Uncertainties in model characteristics lead us to random variables as the basic building blocks for the queueing model. However, a random variable quantitatively represents an event in a random phenomenon. In queueing systems, and all systems that operate over time (or space or any other parameter), the model must be able to represent the system over time. That means we need a sequence or a family of random variables to represent such a phenomenon over time. Let T be the range of time of interest. Time can be continuous or discrete. We denote the time $t \epsilon T$ when it is continuous, and $n \epsilon T$ when it is discrete. Then the family of random variables $\{X(t), t \epsilon T\}$ or the sequence of random variables $\{X_n, n \epsilon T\}$ is known as a *stochastic process*. (A sample value of a random variable can be looked upon as a snapshot, whereas, a sample path of a stochastic process can be considered a video.) The space in which $X(t)$ or X_n assumes values is known as the *state space* and T is known as the *parameter space*. Another way of saying is that a stochastic process is a family or a sequence of random variables indexed by a parameter.

The underlying processes of queueing systems are the product of arrivals and service. They may be continuous or discrete. Even when we define continuous state processes such as waiting times, arrival and departure points are imbedded in them. The next two sections describe commonly occurring processes used in the analysis of queueing systems. Since general stationary and nonstationary stochastic processes are not used in the analysis of queueing models, we do not provide any information on them in our discussion.

© Springer Science+Business Media New York 2015

U. N. Bhat, *An Introduction to Queueing Theory*, Statistics for Industry and Technology, DOI 10.1007/978-0-8176-8421-1_3

3.2 Point, Regenerative, and Renewal Processes

Point Process

Consider randomly located discrete set of points in the parameter space T. These points may represent events such as arrivals in a queueing system or accidents on a stretch of road. Let $N(t), t\epsilon T$ be the number of points in $(0, t]$. Then the counting process $N(t)$ is known as a *point process* (see Lewis (1972)). There are processes in which the points may be of different types. For instance, the arrival of two types of customers. Then the process is identified as a *marked point process.*

Regenerative Process

Consider a stochastic process $\{X(t), t \epsilon T\}$ and a discrete set of points $t_1 < t_2 < \ldots < t_n \epsilon T$. Suppose the distribution properties of the process from t_i onwards is the same for all $i = 1, 2, \ldots, n$. Then we can consider the process regenerating itself at these points.

Renewal Process

Consider a discrete set of points $(t_0, t_1, t_2, \ldots)$ at which a specified event occurs and let $t_i - t_{i-1} = Z_i$ $(i = 1, 2, \ldots)$, be independent and identically distributed (i.i.d) random variables. The process of the sequence of random variables $(Z_1, Z_2, \ldots)$ is known as a *renewal process.* Let $N(t)$ be the process representing the number of events occurring in $(0, t]$. It is known as the *renewal counting process.* The periods Z_i $(i = 1, 2, \ldots)$ are *renewal periods.* Since the renewal periods are i.i.d., it is clearly seen that the renewal process is also a regenerative process.

3.3 Markov Process

Some of the simple models widely used in queueing theory are based on Markov processes. Suppose, a stochastic process $\{X(t), t\epsilon T\}$ is such, that

$$\begin{aligned} P[X(t) &\leq x|X(t_1) = x_1,\ X(t_2) = x_2, \ldots, X(t_n) = x_n] \\ &= P[X(t) \leq x|X(t_n) = x_n] \qquad (t_1 < t_2 \ldots < t_n < t) \\ &= F(x_n, x; t_n, t). \end{aligned} \tag{3.3.1}$$

Then $\{X(t)\}$ is a Markov process. When T and the state space are discrete the parallel definition is given as

$$\begin{aligned} P(X_n = j|X_{n_1} = i_1, X_{n_2} = i_2, \ldots, X_{n_k} = i_k) &= P(X_n = j|X_{n_k} = i_k) \\ &= P_{i_k,j}^{(n_k,n)}. \end{aligned} \tag{3.3.2}$$

Now the process $\{X_n, n = 1, 2, \ldots\}$ is called the *Markov chain*.

The dependence structure exhibited here is a one-step dependence, in which the state of the process is dependent only on the last parameter point at which full information of the process is available. As can be seen in the following chapters, the property of Markov dependence simplifies the analysis while retaining essential characteristics of the systems.

Since the time parameter in a Markov process has a specific range, we use transition distributions or probabilities of the process in its analysis. These are conditional statements, conditioned on the process value at the initial point t. An unconditional distribution or the probability (in the discrete case) can be obtained by the usual method of removing the condition.

For the transition probabilities of Markov processes, we use the following notations depending on the nature of state and parameter spaces.

(i) Discrete state, discrete parameter (Markov chain):

$$P_{ij}^{m,n} = P(X_n = j | X_m = i), \quad m < n, \tag{3.3.3}$$

(ii) Discrete state, continuous parameter (also called continuous time Markov chain (CTMC)):

$$P_{ij}(s,t) = P[X(t) = j | X(s) = i], \quad s < t, \tag{3.3.4}$$

(iii) Continuous state, discrete parameter:

$$F(x_m, x; m, n) = P(X_n \leq x | X_m = x_m), \quad m < n, \tag{3.3.5}$$

(iv) Continuous state, continuous parameter:

$$F(x_n, x; t_n, t) = P[X(t) \leq x | X(t_n) = x_n], \quad t_n < t. \tag{3.3.6}$$

The fundamental property of the Markov process is given by the *Chapman–Kolmogorov relation*. Corresponding to the above four cases, it can be given as follows:

(i)
$$P_{ij}^{(m,n)} = \sum_{k \epsilon S} P_{ik}^{(m,r)} P_{kj}^{(r,n)} \quad m < r < n. \tag{3.3.7}$$

(ii)
$$P_{ij}(s,t) = \sum_{k \epsilon S} P_{ik}(s,u) P_{kj}(u,t) \quad s < u < t. \tag{3.3.8}$$

(iii)
$$F(x_m, x; m, n) = \int_{y \epsilon S} d_y F(x_m, y; m, r) F(y, x; r, n) \quad m < r < n. \tag{3.3.9}$$

(iv)

$$F(x_s, x; s, t) = \int_{y \epsilon S} d_y F(x_s, y; s, u) F(y, x; u, t) \quad s < u < t. \tag{3.3.10}$$

These equations can be easily established by considering the transitions of the process in two time periods (m, r) and (r, n) when the time parameter is discrete and (s, u) and (u, t) when the time parameter is continuous, and using the basic definition of the Markov process. For instance, when both the state and parameter spaces are discrete, the probability of the transition from the initial state i to a state k $(k \epsilon S)$ in time period (m, r) is $P_{ik}^{(m,r)}$ and from state k to state j in time period (r, n) is $P_{kj}^{(r,n)}$. Equation (3.3.7) now follows by multiplying these two probabilities and summing over all values of $k \epsilon S$. Similar arguments establish (3.3.8)–(3.3.10).

The stochastic processes underlying queueing systems considered in this book primarily belong to two classes: discrete state and parameter spaces (case (i) above) and discrete state space and continuous parameter space (case (ii) above). We provide the conceptual framework for the method by which equations (3.3.7) and (3.3.8) can be used in their analysis here and in Appendix B.

Case (i): Discrete State and Parameter Space

Let $\{X_n,\ n = 0, 1, 2 \ldots\}$ be a time homogeneous Markov chain. By time homogeneous we mean that the transition probabilities $P_{ij}^{(m,n)}$ and $P_{ij}^{(m+k,n+k)}$ for $k > 0$ are the same. Without loss of generality we use $m = 0$ and write

$$P_{ij}^{(n)} = P(X_n = j | X_0 = i). \tag{3.3.11}$$

For convenience write $P_{ij}^{(1)} = P_{ij}$ as the one-step transition probability. In matrix notation, we have

$$\mathbf{P}^{(n)} = \begin{bmatrix} P_{00}^{(n)} & P_{01}^{(n)} & P_{02}^{(n)} & \cdots \\ P_{10}^{(n)} & P_{11}^{(n)} & P_{12}^{(n)} & \cdots \\ P_{20}^{(n)} & P_{21}^{(n)} & P_{22)}^{(n)} & \cdots \\ \vdots & \vdots & \vdots & \end{bmatrix}. \tag{3.3.12}$$

When $n = 1$, the matrix $\mathbf{P} \equiv \mathbf{P}^{(1)}$ is known as the *transition probability matrix.*

Note that $0 \leq P_{ij}^{(n)} \leq 1$ and the row sums of $\mathbf{P}^{(n)}$ (i.e., $\sum_{j \epsilon S} P_{ij}^{(n)}$) are equal to 1 for all values of n. With these notational simplifications, (3.3.7) can be written as

$$P_{ij}^{(n)} = \sum_{k \epsilon S} P_{ik}^{(r)} P_{kj}^{(n-r)} \quad 0 < r < n,$$

or

$$\mathbf{P}^{(n)} = \mathbf{P}^{(r)} \mathbf{P}^{(n-r)}.$$

By iterating on the value of $r = 1, 2, \ldots, n$, it follows that

$$\mathbf{P}^{(n)} = \mathbf{P}^n, \tag{3.3.13}$$

showing that the n-step transition probabilities are given by the elements of the nth power of the one-step transition probability matrix.

Case (ii): Discrete State Space and Continuous Parameter Space

As in Case (i) consider time-homogeneous Markov process in which transition probabilities $P_{ij}(s,t)$ and $P_{ij}(s+u, t+u)$ for $u > 0$ are the same. Without loss of generality, use $s = 0$ and write

$$P_{ij}(t) = P[X(t) = j | X(0) = i]. \tag{3.3.14}$$

In matrix notation the probabilities of transition among states $i, j \epsilon S$ can be given as elements of the matrix

$$\mathbf{P}(t) = ||P_{ij}(t)||.$$

Because of the continuous nature of the time parameter, we cannot get a result similar to (3.3.13) in this case. Also instead of the product representation, here we use differential equations from which $P_{ij}(t)$ can be determined. To start with note the following properties that are either obvious or assumed.

(i) $P_{ij}(t) \geq 0$

(ii) $\sum_{j \epsilon S} P_{ij}(t) = 1$

(iii) $P_{ij}(s+t) = \sum_{k \epsilon S} P_{ik}(s) P_{kj}(t)$

(iv) $P_{ij}(t)$ is continuous

(v) $\lim_{t \to 0} P_{ij}(t) = 1$ if $i = j$, and $= 0$, otherwise

Note that properties (i) and (ii) are obvious from the transition structure and (iii) is a restatement of Chapman–Kolmogorov relation. The properties (iv) and (v) are necessary (hence assumed) for deriving the differential equations.

Using Taylor series expansion, and Δt as an infinitesimal increment in t, we may write

$$P_{ij}(t, t + \Delta t) = P_{ij}(t) + \Delta t P'_{ij}(t) + \frac{\Delta t^2}{2} P''_{ij}(t) + \ldots.$$

Setting $t = 0$

$$P_{ij}(\Delta t) = P_{ij}(0) + \Delta t P'_{ij}(0) + \frac{\Delta t^2}{2} P''_{ij}(0) + \ldots.$$

Rewriting these equations and taking limits as $\Delta t \to 0$, we get

$$\lim_{\Delta t \to 0} \frac{P_{ij}(\Delta t)}{\Delta t} = P'_{ij}(0) = \lambda_{ij} \quad i \neq j \tag{3.3.15}$$

$$\lim_{\Delta t \to 0} \frac{P_{ii}(\Delta t) - 1}{\Delta t} = P'_{ii}(0) = -\lambda_{ii}, \tag{3.3.16}$$

where λ_{ij} are such that

$$\sum_{j \neq i} \lambda_{ij} = \lambda_{ii}. \tag{3.3.17}$$

Noting that λ_{ij} are *infinitesimal transition rates*, it is easy to see, that (3.3.17) is the direct consequence of the property $\sum_{j \epsilon S} P_{ij}(t) = 1$. These transition rates are also known as *generators*, displayed in a matrix as

$$\mathbf{A} = \begin{bmatrix} -\lambda_{00} & \lambda_{01} & \lambda_{02} & \dots \\ \lambda_{10} & -\lambda_{11} & \lambda_{12} & \dots \\ \lambda_{20} & \lambda_{21} & -\lambda_{22} & \dots \\ \vdots & \vdots & \vdots & \end{bmatrix}. \tag{3.3.18}$$

In continuous time Markov processes with discrete states the generator matrix $\mathbf{A}$ plays the same role in its analysis as that of the transition probability matrix $\mathbf{P}$ (matrix (3.3.12) with $n = 1$) in the analysis of a Markov chain.

The Poisson process discussed in Chapter 2 and Appendix A is a Markov process with a simple transition structure. Let $\{X(t),\ t \epsilon T\}$ be a Poisson process with parameter λ, such that

$$P_n(t) = P[X(t) = n] = e^{-\lambda t} \frac{(\lambda t)^n}{n!} \quad n = 0, 1, 2 \dots. \tag{3.3.19}$$

(See (B.2.2), Appendix B.)

Using arguments similar to those used in deriving (3.3.15)–(3.3.17), we can show that the infinitesimal transition rates $\lambda_{ij} = \lambda$ for $j = i, i+1$ and $= 0$, otherwise.

When the Poisson process and the associated exponential distribution are used to model queueing systems, their underlying processes such as the number of customers in the system are Markov, and hence require analysis techniques appropriate for Markov processes.

The differential equations used in the analysis of Markov processes are based on Chapman–Kolmogorov relations as applied to infinitesimal transitions to the process in the time interval $(t, t + \Delta t)$. These are given as

$$P'_{ij}(t) = -\lambda_{jj} P_{jj}(t) + \sum_{k \neq j} \lambda_{kj} P_{ik}(t). \tag{3.3.20}$$

This equation is known as the *Forward Kolmogorov equation* and its derivation is provided in Appendix B.

In matrix notation, we can write the equations as

$$\boldsymbol{P}'(t) = \boldsymbol{P}(t)\boldsymbol{A}. \tag{3.3.21}$$

Thus, the analysis of the behavior of a queueing system which can be modeled as a Markov process involves two key steps: the determination of the appropriate values of λ_{ij} and the solution of the resulting equation (3.3.20). The first part of this procedure is accomplished from the nature of transitions in the Markov process and the resulting differential equations are solved using standard mathematical/computational techniques. In many applications a finite time solution may not be needed. Then a limiting solution, when $t \to \infty$, is obtained to determine the limiting behavior of the system. These procedures will be introduced as and when they are needed.

3.4 Renewal Process

Consider a discrete set of points $(t_0, t_1, t_2, \ldots)$ at which a specific event occurs. Let $t_i - t_{i-1} = Z_i$ $(i = 1, 2, \ldots)$ be i.i.d. random variables with distribution

$$P(Z_i \leq x) = F(x). \tag{3.4.1}$$

The process consisting of the sequence of random variables $(Z_1, Z_2, \ldots)$ is known as a *renewal process*. Let $N(t)$ be the number of events occurring in $(0, t]$. It is known as a *renewal counting process*. The periods Z_i are known as *renewal periods*. At $t = 0$, if the renewal process is already in progress, t_0 may not be an epoch of occurrence of the renewal event. To accommodate such situations, we may define the random variable $Z_1 = t_1 - t_0$ with a distribution different from $F(x)$, say $F_1(x)$. Such a renewal process is known as a *delayed renewal process*. For our discussion we restrict ourselves to the *ordinary renewal process*, in which $F_1(x) = F(x)$; this means we assume that $t_0 = 0$ is an epoch of occurrence of the renewal event.

In the context of queueing systems, under normal conditions, the arrival process can be considered a renewal process; i.e., the inter-arrival times form a sequence of i.i.d. random variables. The service process can be a renewal process only if there are enough customers in the system to keep the server continuously busy and the queue discipline requires the server to provide a complete service to a customer once that customer's service starts. In a $G/G/1$ queue regardless of the distribution forms for G's and with a queue discipline in which the server is never idle as long as there are customers in the system, the time points at which consecutive busy periods start are renewal epochs. The renewal period in this case is made up of the combination of a busy period and an idle period, commonly known as a *busy cycle*. Thus, when we use renewal process models to analyze a queueing system, we start with a busy cycle and its distribution.

Let $S_n = Z_1+Z_2+...+Z_n$. Using $F(x)$, the distribution of S_n can be obtained as the n-fold convolution of $F(x)$ with itself, which we denote as $F_n(x)$. Define

$$\phi(\theta) = \int_0^\infty e^{-\theta x} dF(x) \quad Re(\theta) > 0 \tag{3.4.2}$$

as the Laplace–Stieltjes transform of $F(x)$. We then have

$$\int_0^\infty e^{-\theta x} dF_n(x) = [\phi(\theta)]^n . \tag{3.4.3}$$

The distribution of the renewal counting process $N(t)$ for a specific value of t can be derived as follows. Let

$$P_n(t) = P[N(t) = n] . \tag{3.4.4}$$

Consider two events

$$\{N(t) \geq n\} \text{ and } \{S_n \leq t\}.$$

These are equivalent events. By equating their probabilities, we get

$$\begin{aligned} P[N(t) \geq n] &= P[S_n \leq t] \\ &= F_n(t). \end{aligned}$$

Thus, we get

$$P_n(t) = F_n(t) - F_{n+1}(t). \tag{3.4.5}$$

The mean value function $E[N(t)]$ is called the *renewal function*, denoted by $U(t)$, and its derivative, when it exists, is called the *renewal density*, denoted by $u(t)$. From (3.4.5) it is easy to show

$$\begin{aligned} U(t) = E[N(t)] &= \sum_{n=1}^{\infty} nP_n(t) \\ &= \sum_{n=1}^{\infty} F_n(t), \end{aligned} \tag{3.4.6}$$

and

$$u(t) = \sum_{n=1}^{\infty} f_n(t),$$

where we have written $f(t)$ to denote the density function corresponding to the distribution function $F(t)$. The term "renewal density" can be intuitively

justified as follows:

$$\begin{aligned} u(t) &= \lim_{\Delta t \to 0} \frac{P(\text{renewal event occurs in } (t, t+\Delta t])}{\Delta t} \\ &= \sum_{r=1}^{\infty} \lim_{\Delta t \to 0} \frac{P(r^{th} \text{ renewal occurs in } (t, t+\Delta])}{\Delta t} \\ &= \sum_{r=1}^{\infty} \lim_{\Delta t \to 0} \frac{f_r(t)\Delta t + o(\Delta t)}{\Delta t} \\ &= \sum_{r=1}^{\infty} f_r(t) = U'(t), \end{aligned} \tag{3.4.7}$$

where we have assumed that $F(x)$ is absolutely continuous and denoted $F_r'(t) = f_r(t)$. Let

$$\begin{aligned} U^*(\theta) &= \int_0^{\infty} e^{-\theta t} U(t) dt \quad Re(\theta) > 0 \\ u^*(\theta) &= \int_0^{\infty} e^{-\theta t} u(t) dt \quad Re(\theta) > 0. \end{aligned}$$

Using the relationship between the transforms of the distribution and the density functions, we have

$$u^*(\theta) = \theta U^*(\theta). \tag{3.4.8}$$

Referring back to (3.4.2) and (3.4.3), and using (3.4.8) we get

$$U^*(\theta) = \frac{1}{\theta} \sum_{n=1}^{\infty} [\phi(\theta)]^n \tag{3.4.9}$$

$$u^*(\theta) = \sum_{n=1}^{\infty} [\phi(\theta)]^n. \tag{3.4.10}$$

From (3.4.9) and (3.4.10), we get

$$U^*(\theta) = \frac{\phi(\theta)}{\theta[1-\phi(\theta)]} \tag{3.4.11}$$

$$u^*(\theta) = \frac{\phi(\theta)}{1-\phi(\theta)}. \tag{3.4.12}$$

Rearranging terms in (3.4.11), we get

$$U^*(\theta) = \frac{\phi(\theta)}{\theta} + U^*(\theta)\phi(\theta),$$

which on inversion gives

$$U(t) = F(t) + \int_0^t U(t-\tau) dF(\tau). \tag{3.4.13}$$

Similarly from (3.4.12) we can get

$$u(t) = f(t) + \int_0^t u(t-\tau) f(\tau) d\tau. \tag{3.4.14}$$

The integral equation (3.4.13) is known as the *renewal equation*, which in its general form can be given as

$$Z(t) = h(t) + \int_0^t Z(t-\tau) dF(\tau), \tag{3.4.15}$$

where $h(t)$ is directly Riemann integrable and $F(t)$ is a distribution function. This equation can be solved to give

$$Z(t) = h(t) + \int_0^t h(t-\tau) dU(\tau). \tag{3.4.16}$$

The significance of (3.4.16) in modeling queueing systems can be described as follows.

In a stochastic process made up of renewal periods, the distribution of the state of the process at time t can be determined by convolving the renewal density of the process at time τ when the last renewal occurs before t (i.e., $dU(\tau)$) with the transition probability distribution between $t-\tau$ and t (i.e., $h(t-\tau)$). Note that the first term (i.e., $h(t)$) takes care of the possibility that no renewal has occurred during $(0, t]$. As described earlier, busy cycles are renewal periods for a queueing process. When $t \to \infty$, much simpler expressions follow, thanks to important limiting results.

1.
$$U(t+\Delta) - U(t) \to \frac{\Delta}{R} \text{ as } t \to \infty \tag{3.4.17}$$

$$u(t) \to \frac{1}{R} \text{ as } t \to \infty, \tag{3.4.18}$$

where R is the mean of the renewal period.

2. Let $h(t)$ be a nonnegative Riemann integrable function of $t > 0$, such that

$$\int_0^\infty h(t) dt < \infty.$$

Then

$$\int_0^t h(t-\tau) dU(\tau) \to \frac{1}{R} \int_0^\infty h(t) dt \quad \text{as } t \to \infty. \tag{3.4.19}$$

This result is known as the *key renewal theorem.*

For further details of the properties of renewal processes the readers are referred to Bhat and Miller (2002). Proofs for the solution of the renewal equation and the limiting behavior of the process (key renewal theorem) can be found in advanced textbooks in stochastic processes such as Karlin and Taylor (1975).

The results (3.4.18) and (3.4.19) can be easily understood if we note that $U(t)$ is the expected number of renewals in $(0, t]$. The implication of (3.4.18) is that, when $t \to \infty$, i.e., when the process is in operation for a long time, the expected number of renewals in a period of length Δ is obtained by dividing Δ by the expected length of a renewal period. Since

$$\lim_{\Delta \to 0} \frac{U(t+\Delta) - U(t)}{\Delta} = u(t)$$

the result (3.4.18) follows directly from (3.4.17) and gives the rate of occurrence of the renewal. The result (3.4.19) also follows directly by taking limits ($t \to \infty$) in (3.4.16). Since $h(t)$ is Riemann integrable it tends to 0 as $t \to \infty$ and the contribution of $dU(t)$ in the integral in (3.4.16) is $1/R$, which is the rate of occurrence of the renewal. Thus, we get

$$\lim_{t \to \infty} Z(t) = \frac{1}{R} \int_0^\infty h(t)dt. \tag{3.4.20}$$

When we observe a renewal process at an arbitrary time point t, it is unlikely that t will be a renewal epoch. Then, the time period since the last renewal epoch until t is known as the *backward recurrence time* (or *current life* in the terminology of reliability theory), and the time period until the next renewal epoch from t is known as the *forward recurrence time* (or *excess life* or *residual life*). Let $S(t)$ and $R(t)$ denote these random variables and $s_t(x)$ and $r_t(x)$ be their probability density functions, respectively. Using renewal arguments, we can write

$$s_t(x) = u(t-x)[1 - F(x)] \quad 0 < x < t \tag{3.4.21}$$

and

$$r_t(x) = f_1(t+x) + \int_{\tau=0}^{t} u(\tau) f(t - \tau + x) d\tau, \tag{3.4.22}$$

where $f_1(x)$ is the density function of the initial renewal period. As $t \to \infty$ (3.4.21) and (3.4.22) yield

$$\lim_{t \to \infty} s_t(x) = \frac{1 - F(x)}{R} \tag{3.4.23}$$

$$\lim_{t \to \infty} r_t(x) = \frac{1 - F(x)}{R}. \tag{3.4.24}$$

Taking expected values, we get

$$\int_0^\infty x \left[\frac{1 - F(x)}{R}\right] dx = \frac{E[Z^2]}{2R}, \tag{3.4.25}$$

where we have written Z as the random variable denoting the length of the renewal period.

When the renewal period has an exponential distribution with mean $1/\lambda$, $E(Z^2) = 2/\lambda^2$. As $t \to \infty$, this leads to the result, $E[\text{current life}] = E[\text{excess life}] = 1/\lambda$.

Chapter 4

Simple Markovian Queueing Systems

Poisson arrivals and exponential service make queueing models Markovian that are easy to analyze and get useable results. Historically, these are also the models used in the early stages of queueing theory to help decision-making in telephone industry. The underlying Markov process representing the number of customers in such systems is known as a *birth and death process*, which is widely used in population models. The birth–death terminology is used to represent increase and decrease in the population size. The corresponding events in queueing systems are arrivals and departures. In this chapter we present some of the important models belonging to this class.

4.1 A General Birth and Death Queueing Model

Keeping the birth (arrival)–death (departure) terminology, when population size is n, let λ_n and μ_n be the infinitesimal transition rates (generators) of birth and death respectively. When the population is the number of customers, λ_n and μ_n indicate that the arrival and service rates depend on the number in the system. Generalizing the properties of the Poisson process (see Appendix B.2), we can make the following probability statements for a transition during $(t, t + \Delta t]$.

Birth ($n \geq 0$):

$$\begin{aligned} P(\text{one birth}) &= \lambda_n \Delta t + o(\Delta t) \\ P(\text{no birth}) &= 1 - \lambda_n \Delta t + o(\Delta t) \\ P(\text{more than one birth}) &= o(\Delta t). \end{aligned}$$

© Springer Science+Business Media New York 2015

U. N. Bhat, *An Introduction to Queueing Theory*, Statistics for Industry and Technology, DOI 10.1007/978-0-8176-8421-1_4

Death ($n > 0$):

$$\begin{aligned} P(\text{one death}) &= \mu_n \Delta t + o(\Delta t) \\ P(\text{no death}) &= 1 - \mu_n \Delta t + o(\Delta t) \\ P(\text{more than one death}) &= o(\Delta t) \end{aligned}$$

where $o(\Delta t)$ is such that $\frac{o(\Delta t)}{\Delta t} \to 0$ as $\Delta t \to 0$. Note that in these statements $o(\Delta t)$ terms do not specify actual values. In each of the two cases $o(\Delta t)$ terms sum to 0 so that the total probability of the three events is equal to 1.

Let $Q(t)$ be the number of customers in the system at time t. Define

$$P_{in}(t) = P[Q(t) = n | Q(0) = i].$$

Incorporating the probabilities for transitions during $(t, t+\Delta t]$, as stated above, we get

$$\begin{aligned} P_{n,n+1}(\Delta t) &= \lambda_n \Delta t + o(\Delta t) \qquad n = 0, 1, 2, \ldots \\ P_{n,n-1}(\Delta t) &= \mu_n \Delta t + o(\Delta t) \qquad n = 1, 2, 3 \ldots \\ P_{nn}(\Delta t) &= 1 - \lambda_n \Delta t - \mu_n \Delta t + o(\Delta t) \qquad n = 1, 2, 3 \ldots \\ P_{nj}(\Delta t) &= o(\Delta t), \qquad j \neq n-1, n, n+1. \end{aligned} \tag{4.1.1}$$

In deriving terms on the right-hand side of these equations, we have made use of simplifications of the type

$$\begin{aligned} [\lambda_n \Delta t + o(\Delta t)][1 - \mu_n \Delta t + o(\Delta t)] &= \lambda_n \Delta t + o(\Delta t) \\ [1 - \lambda_n \Delta t + o(\Delta t)][1 - \mu_n \Delta t + o(\Delta t)] &= 1 - \lambda_n \Delta t - \mu_n \Delta t + o(\Delta t). \end{aligned}$$

The infinitesimal transition rates of (4.1.1) lead to the following generator matrix for the birth and death process model of the queueing system.

$$\mathbf{A} = \begin{bmatrix} -\lambda_0 & \lambda_0 & & & \\ \mu_1 & -(\lambda_1 + \mu_1) & \lambda_1 & & \\ & \mu_2 & -(\lambda_2 + \mu_2) & \lambda_2 & \\ & & \cdot & & \\ & & & \cdot & \\ & & & & \cdot \end{bmatrix}. \tag{4.1.2}$$

The generator matrix $\mathbf{A}$ of (4.1.2) results in the following forward Kolmogorov equations for $P_{in}(t)$ (See (3.3.20) and (B.1.2)). (For ease of notation, from here onwards, we write $P_{in}(t) \equiv P_n(t)$ and insert the initial state i only when needed.)

$$\begin{aligned} P_0'(t) &= -\lambda_0 P_0(t) + \mu_1 P_1(t) \\ P_n'(t) &= -(\lambda_n + \mu_n) P_n(t) + \lambda_{n-1} P_{n-1}(t) \\ &\quad + \mu_{n+1} P_{n+1}(t) \qquad n = 1, 2, \ldots \end{aligned} \tag{4.1.3}$$

As a point of digression, note that (4.1.3) can also be derived directly using (4.1.1) without going through the generator matrix as illustrated below.

Considering the transitions of the process $Q(t)$ during $(t, t+\Delta t]$, we have

$$\begin{aligned} P_0(t+\Delta t) &= P_0(t)[1-\lambda_0\Delta t + o(\Delta t)] + P_1(t)[\mu_1\Delta t + o(\Delta t)] \\ P_n(t+\Delta t) &= P_n(t)[1-\lambda_n\Delta t - \mu_n\Delta t + o(\Delta t)] \\ &\quad + P_{n-1}(t)[\lambda_{n-1}\Delta t + o(\Delta t)] \\ &\quad + P_{n+1}(t)[\mu_{n+1}\Delta t + o(\Delta t)] \\ &\quad + o(\Delta t) \qquad n = 1, 2, \ldots . \end{aligned} \tag{4.1.4}$$

Subtracting $P_n(t)$ $(n = 0, 1, 2\ldots)$ from both sides of the appropriate equation in (4.1.4) and dividing by Δt, we get

$$\begin{aligned} \frac{P_0(t+\Delta t) - P_0(t)}{\Delta t} &= -\lambda_0 P_0(t) + \mu_1 P_1(t) + \frac{o(\Delta t)}{\Delta t} \\ \frac{P_n(t+\Delta t) - P_n(t)}{\Delta t} &= -(\lambda_n + \mu_n)P_n(t) \\ &\quad + \lambda_{n-1}P_{n-1}(t) + \mu_{n+1}P_{n+1}(t) \\ &\quad + \frac{o(\Delta t)}{\Delta t}. \end{aligned}$$

Now (4.1.3) follows by letting $\Delta t \to 0$.

To determine $P_n(t)$ $[\equiv P_{in}(t)]$, (4.1.3) should be solved along with the initial conditions $P_i(0) = 1$, $P_n(0) = 0$ for $n \neq i$. (See Stewart (1994) for the numerical solution of special cases of (4.1.3).) Unfortunately, even in simple cases such as $\lambda_n = \lambda$ and $\mu_n = \mu$, $n = 0, 1, 2, 3\ldots$, that is when the arrivals are Poisson and service times are exponential ($M/M/1$ queue), deriving $P_n(t)$ explicitly is an arduous process. Furthermore, in most of the applications the need for knowing the time dependent behavior is not all that critical. The most widely used result, therefore, is the limiting result, determined from (4.1.3) by letting $t \to \infty$.

A general result on Markov processes is given below.

Theorem 4.1.1 *1) If the Markov process is irreducible (all states communicate), then the limiting distribution $lim_{t\to\infty}P_n(t) = p_n$ exists and is independent of the initial conditions of the process. The limits $\{p_n, n\epsilon\boldsymbol{S}\}$ are such that they either vanish identically (i.e., $p_n = 0$ for all $n\epsilon\boldsymbol{S}$) or are all positive and form a probability distribution (i.e., $p_n > 0$ for all $n\epsilon\boldsymbol{S}$, $\sum_{n\epsilon\boldsymbol{S}} p_n = 1$).*

(2) The limiting distribution $\{p_n, n\epsilon\boldsymbol{S}\}$ of an irreducible recurrent Markov process is given by the unique solution of the equation $\boldsymbol{pA} = 0$ and $\sum_{j\epsilon\boldsymbol{S}} p_j = 1$, where $\boldsymbol{p} = (p_0, p_1, p_2, \ldots)$.

The results presented in the theorem essentially confirm what one can think of as a state of equilibrium in a stochastic process and how that affects the Kolmogorov equations (3.3.20) in a Markov process. In a state of equilibrium,

also known as *steady state*, the behavior of the process is independent of the time parameter and the initial state; i.e.,

$$\lim_{t\to\infty} P_{in}(t) = p_n \qquad\qquad n = 0, 1, 2 \ldots$$

and therefore

$$P_n'(t) \to 0 \qquad \text{as } t \to \infty.$$

Using these results in (4.1.3), we get

$$\begin{aligned} 0 &= -\lambda_0 p_0 + \mu_1 p_1 \\ 0 &= -(\lambda_n + \mu_n) p_n + \lambda_{n-1} p_{n-1} + \mu_{n+1} p_{n+1} \qquad n = 1, 2, \ldots . \end{aligned} \tag{4.1.5}$$

These equations can be easily solved through recursion. Rearranging the first equation in (4.1.5), we have

$$p_1 = \frac{\lambda_0}{\mu_1} p_0. \tag{4.1.6}$$

For $n = 1$, the second equation gives

$$(\lambda_1 + \mu_1) p_1 = \lambda_0 p_0 + \mu_2 p_2.$$

Using (4.1.6), this equation reduces to

$$\begin{aligned} \mu_2 p_2 &= \lambda_1 p_1 \\ p_2 &= \frac{\lambda_1 \lambda_0}{\mu_2 \mu_1} p_0. \end{aligned}$$

Continuing this recursion for $n = 2, 3, \ldots$ we get

$$\mu_n p_n = \lambda_{n-1} p_{n-1} \tag{4.1.7}$$

and therefore,

$$p_n = \frac{\lambda_0 \lambda_1 \ldots \lambda_{n-1}}{\mu_1 \mu_2 \ldots \mu_n} p_0. \tag{4.1.8}$$

Theorem 4.1.1 also gives the normalizing condition $\sum_{n \epsilon \boldsymbol{S}} p_n = 1$, which when applied to (4.1.8), gives

$$p_0 = \left[1 + \sum_{n=1}^{\infty} \frac{\lambda_0 \lambda_1 \ldots \lambda_{n-1}}{\mu_1 \mu_2 \ldots \mu_n} \right]^{-1}. \tag{4.1.9}$$

The limiting distribution of the state of the birth and death queueing model is $\{p_n, n = 0, 1, 2, \ldots\}$ as given by (4.1.8) and (4.1.9). It should be noted that $\{p_n, n = 0, 1, 2, \ldots\}$ are nonzero only when

$$1 + \sum_{n=1}^{\infty} \frac{\lambda_0 \lambda_1 \ldots \lambda_{n-1}}{\mu_1 \mu_2 \ldots \mu_n} < \infty. \tag{4.1.10}$$

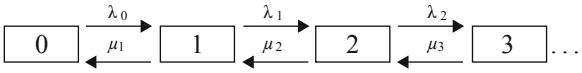


Figure 4.1 Transition diagram

In order to derive (4.1.5) deriving (4.1.3) first is not necessary. As noted in Theorem 4.1.1, with $\boldsymbol{p} = (p_0, p_1, p_2, \ldots)$ and the generator matrix $\boldsymbol{A}$, (4.1.5) can be obtained directly from

$$\boldsymbol{pA} = 0 \tag{4.1.11}$$

and

$$\sum_{n=1}^{\infty} p_n = 1.$$

For the birth and death queueing model the generator matrix $\boldsymbol{A}$ is given by (4.1.2).

Another way of looking at (4.1.5) is to consider them as representing a condition of balance among the states. Rearranging (4.1.5)

$$\begin{aligned} \lambda_0 p_0 &= \mu_1 p_1 \\ (\lambda_n + \mu_n) p_n &= \lambda_{n-1} p_{n-1} + \mu_{n+1} p_{n+1}. \end{aligned} \tag{4.1.12}$$

The transitions among the states can be pictorially represented as in Figure 4.1. Noting that the λ's and μ's represent infinitesimal transition rates in and out of the states, the equations in (4.1.12) can be interpreted as (long term probability of being in state n) $\times$ (transition rates out of state n) $= \sum_{i=n-1,n+1}$ (long term probability of being in state i) $\times$ (transition rate from state i to state n).

Such state balance equations can be easily written down using the transition diagram shown in Figure 4.1.

Given above are three ways of writing down the state balance equations:

(1) Taking the appropriate limits as $t \to \infty$ in the forward Kolmogorov equations

(2) Using the equation $\boldsymbol{pA} = 0$

(3) With the help of the transition diagram

In applications the readers may use any method with which they are comfortable. But when using the last method, care should be taken to ensure that all transitions have been accounted for. In our discussion of special models, we normally use the second method based on the generator matrix unless the transition diagram throws more light on the behavior of the system.

There are two other theorems which establish some important properties of the limiting distribution of a Markov process with an irreducible state space. The first of them addresses the concept of stationarity.

Theorem 4.1.2 *The limiting distribution of a positive recurrent irreducible Markov process is also stationary.*

A process is said to be *stationary* if the state distribution is independent of time, i.e, if

$$P_n(0) = p_n \qquad n = 0, 1, 2, \ldots$$

then

$$P_n(t) = p_n \qquad \text{for all } t.$$

Since we deal with transition distributions conditional on the initial state in stochastic processes, the stationarity means that if we use the stationary distribution as the initial state distribution, from then on all time-dependent distributions will be the same as the one we started with.

The second theorem enables us to interpret the limiting probability p_n, $n = 0, 1, 2, \ldots$ as the fraction of time the process occupies state n in the long run.

Theorem 4.1.3 *Having started from state i, let $N_{ij}(t)$ be the time spent by the Markov process in state j during $(0, t]$. Then*

$$\lim_{t\to\infty} \left[\left| \frac{N_{ij}(t)}{t} - p_j \right| > \epsilon \right] = 0.$$

The general birth and death queueing model encompasses a wide array of special cases. Some of the widely used models are discussed in the following sections.

4.2 The Queue $M/M/1$

The $M/M/1$ queue is the simplest of the queueing models used in practice. The arrivals are assumed to occur in a Poisson process with rate λ. This means that the number of customers $N(t)$ arriving during a time interval $(0, t]$ has a Poisson distribution

$$P[N(t) = j] = e^{-\lambda t} \frac{(\lambda t)}{j!} \qquad j = 0, 1, 2 \ldots.$$

It also means that the inter-arrival times have an exponential distribution with probability density

$$a(x) = \lambda e^{-\lambda x} \qquad x > 0.$$

We assume that the service times have an exponential distribution with probability density

$$b(x) = \mu e^{-\mu x} \qquad x > 0.$$

With these assumptions we have

$$\begin{aligned} E[\text{interarrival time}] &= \frac{1}{\lambda} = \frac{1}{\text{arrival rate}} \\ E[\text{serivce time}] &= \frac{1}{\mu} = \frac{1}{\text{service rate}}. \end{aligned}$$

The ratio of arrival rate to service rate plays a significant role in measuring the performance of queueing systems. Let

$$\rho = \text{Traffic intensity} = \frac{\text{Arrival rate}}{\text{Service rate}}.$$

In an $M/M/1$ queue, $\rho = \lambda/\mu$.

Clearly $M/M/1$ is a special case of the general birth and death model with $\lambda_n = \lambda$ and $\mu_n = \mu$ for $n = 0, 1, 2, \ldots$. The generator matrix is given by (state space: 0,1,2,...)

$$\mathbf{A} = \begin{bmatrix} -\lambda & \lambda & & & \\ \mu & -(\lambda+\mu) & \lambda & & \\ & \mu & -(\lambda+\mu) & \lambda & \\ & & \cdot & & \\ & & & \cdot & \\ & & & & \cdot \end{bmatrix}. \tag{4.2.1}$$

The corresponding forward Kolmogorov equations for $P_n(t)$ $(n = 0, 1, 2, \ldots)$ are

$$\begin{aligned} P_0'(t) &= -\lambda P_0(t) + \mu P_1(t) \\ P_n'(t) &= -(\lambda+\mu)P_n(t) + \lambda P_{n-1}(t) \\ &\quad +\mu P_{n+1}(t) \qquad n = 1, 2, \ldots \end{aligned} \tag{4.2.2}$$

with $P_n(0) = 1$ when $n = i$ and $= 0$ otherwise. For a complete solution of these difference–differential equations, the use of generating functions (to transform the difference equation) and Laplace transforms (to transform the differential equation) is needed. Since the resulting solution is the Laplace transform of a generating function, $P_n(t)$ can be obtained using inversion formulas. Because of the complexity of the procedure and the final result, we do not provide it in this text. Interested readers may refer to Gross et al. (2008, pp. 98–101), where the results have been derived in detail. Computational methods may also be used to solve the differential equations (4.2.2) (see Stewart 1994).

Limiting Distribution

For the limiting probabilities $\lim_{t\to\infty} P_n(t) = p_n$, we have the state balance equations (see (4.1.12))

$$\begin{aligned} \lambda p_0 &= \mu p_1 \\ (\lambda+\mu)p_n &= \lambda p_{n-1} + \mu p_{n+1} \quad n = 1, 2, 3 \ldots. \end{aligned} \tag{4.2.3}$$

Solving these equations along with $\sum_0^\infty p_n = 1$ (or specializing (4.1.8) and (4.1.9)), we get

$$p_n = (1-\rho)\rho^n \qquad n = 0, 1, 2, \ldots. \tag{4.2.4}$$

where $\rho = \lambda/\mu < 1$.

The probability that the server is busy is a performance measure for the system. Clearly, this *utilization factor* $= 1 - p_0 = \rho =$ traffic intensity in this case. Recall that we have defined $Q(t)$ as the number of customers in the system. Write $Q(\infty) = Q$ and let Q_q be the number in the queue, excluding the one in service. Now we may define two mean values, $L = E(Q)$ and $L_q = E(Q_q)$. From distribution (4.2.4) we get

$$L = \sum_{n=1}^{\infty} n(1-\rho)\rho^n = \frac{\rho}{1-\rho}$$

which can also be written as

$$= \frac{\lambda}{\mu - \lambda}. \tag{4.2.5}$$

For L_q, we get

$$\begin{aligned} L_q &= \sum_{n=1}^{\infty}(n-1)p_n \\ &= \sum_{n=1}^{\infty} np_n - \sum_{n=1}^{\infty} p_n \\ &= L - \rho = \frac{\rho^2}{1-\rho} \\ &= \frac{\lambda^2}{\mu(\mu-\lambda)}. \end{aligned} \tag{4.2.6}$$

The utilization factor ρ is the probability that the server is busy when the system is in equilibrium and therefore, it gives the expected number in service. With this interpretation, we can provide the obvious explanation for (4.2.6) as E (number in system) = E (number waiting) + E (number in service).

From (4.2.4) we obtain the variance of the number of customers in the system as

$$\begin{aligned} V(Q) &= \frac{\rho}{(1-\rho)^2} \\ &= \frac{\lambda\mu}{(\mu-\lambda)^2}. \end{aligned} \tag{4.2.7}$$

Customer Waiting Time

From a customer viewpoint, the time spent in the queue and in the system are two characteristics of importance. When the system is in equilibrium, let T_q and T be the amount of time a customer spends in queue and in the system, respectively. We assume that the system operates according to a "first-come, first-served" (FCFS) queue discipline. We note here that as long as the server remains busy when there are customers in the system, and once a service starts

it is given to its completion, the number in the system is not dependent on the order in which the customers are served. However, for waiting time the order of service is a critical factor.

With an FCFS queue discipline, the waiting time for service (T_q) of an arriving customer is the amount of time required to serve the customers already in the system. The total time in system (T) is T_q + service time. When there are n customers in the system, since service times are exponential with parameter μ, the total service time of n customers is Erlang with probability density

$$f_n(x) = e^{-\mu x} \frac{\mu^n x^{n-1}}{(n-1)!} \tag{4.2.8}$$

Let $F_q(t) = P(T_q \leq t)$, the distribution function of the waiting time T_q. Clearly

$$F_q(0) = P(T_q = 0) = P(Q = 0) = 1 - \rho. \tag{4.2.9}$$

Note that, because of the memoryless property of the exponential distribution, the remaining service time of the customer in service is also exponential with the same parameter μ. Writing $dF_q(t) = P(t < T_q \leq t + dt)$, for $t > 0$, we have

$$\begin{aligned} dF_q(t) &= \sum_{n=1}^{\infty} p_n e^{-\mu t} \frac{\mu^n t^{n-1}}{(n-1)!} dt \\ &= (1-\rho) \sum_{n=1}^{\infty} \rho^n e^{-\mu t} \frac{\mu^n t^{n-1}}{(n-1)!} dt \end{aligned}$$

which on simplification gives

$$= \lambda(1-\rho) e^{-\mu(1-\rho)t} dt. \tag{4.2.10}$$

Because of the discontinuity at 0 in the distribution of T_q, we get

$$\begin{aligned} F_q(t) &= P(T_q = 0) + \int_0^t dF_q(t) \\ &= 1 - \rho e^{-\mu(1-\rho)t} \end{aligned} \tag{4.2.11}$$

where we have combined results from (4.2.9) and (4.2.10).

Let $E(T_q) = W_q$ and $E(T) = W$. From (4.2.11) we can easily derive

$$W_q = E(T_q) = \frac{\rho}{\mu(1-\rho)} = \frac{\lambda}{\mu(\mu-\lambda)} \tag{4.2.12}$$

and

$$V(T_q) = \frac{\rho(2-\rho)}{\mu^2(1-\rho)^2}. \tag{4.2.13}$$

Recalling that the total time in the system T, is the sum of T_q and service time, we get

$$\begin{aligned} W &= E[T] = \frac{\lambda}{\mu(\mu-\lambda)} + \frac{1}{\mu} \\ &= \frac{1}{\mu-\lambda}. \end{aligned} \tag{4.2.14}$$

Comparing the result (4.2.14) with (4.2.5), we note the relationship

$$L = \lambda W. \tag{4.2.15}$$

A similar comparison between results (4.2.6) and (4.2.12) establishes

$$L_q = \lambda W_q. \tag{4.2.16}$$

The result (4.2.15) is known as *Little's law* in queueing literature. A large number of articles have been published on this result and it has been shown that it is a general property of queueing systems subject to only some restrictions on the system structure. It is discussed further in the context of queue $G/G/1$ in Chapter 9.

Busy Period

Busy period is defined as the period of time during which the server is continuously busy. When it ends, an *idle period* follows. Together they form a *busy cycle*. Since the idle period ends with an arrival, it is simply the remaining inter-arrival time after the last customer in the busy period leaves after service. With an exponential inter-arrival time, because of the memoryless property, the idle period also has the same exponential distribution.

There are several methods by which the distribution of the busy period in $M/M/1$ can be derived. None of them is simple. Here we give the outline of the method using forward Kolmogorov equations. Looking at the underlying Markov process, busy period is the duration of time, the process starting from state 1, stays continuously away from state 0. (Since the busy period starts with an arrival, it is the amount of time the process takes to get back to state 0.) Considering the transitions of the Markov process, transitions within a busy period can be brought about by converting state 0 into an absorbing state and all other states into an irreducible transient class. Then the generator matrix (4.2.7) takes the modified form

$$\mathbf{A} = \begin{bmatrix} 0 & 0 & & & \\ \mu & -(\lambda+\mu) & \lambda & & \\ & \mu & -(\lambda+\mu) & \lambda & \\ & \ddots & \ddots & \ddots & \ddots \end{bmatrix}. \tag{4.2.17}$$

The corresponding forward Kolmogorov equations for $P_n(t)$ $(n = 0, 1, 2, \ldots)$ are

$$\begin{aligned} P_0'(t) &= \mu P_1(t) \\ P_1'(t) &= -(\lambda+\mu)P_1(t) + \mu P_2(t) \\ P_n'(t) &= -(\lambda+\mu)P_n(t) + \lambda P_{n-1}(t) + \mu P_{n+1}(t) \\ & \quad n = 2, 3, \ldots . \end{aligned} \tag{4.2.18}$$

With the initial condition $P_1(0) = 1$, $P_n(0) = 0$ for $n \neq 1$. Solving these difference–differential equations require the use of probability generating functions and Laplace transforms.

Let $\pi_0(\theta)$ be the Laplace transform of the busy period defined as

$$\pi_0(\theta) = \int_0^\infty e^{-\theta t} P_0'(t) dt \qquad R(\theta) > 0.$$

After appropriate transform operations on (4.2.18) we get

$$\pi_0(\theta) = \frac{1}{2\lambda}\left[\theta + \lambda + \mu - \sqrt{(\theta+\lambda+\mu)^2 - 4\lambda\mu}\right]. \tag{4.2.19}$$

This can be inverted to give the explicit form

$$P_0'(t) = e^{-(\lambda+\mu)t}\frac{\sqrt{\mu/\lambda}}{t} I_1(2\sqrt{\lambda\mu}t) \tag{4.2.20}$$

where $I_j(x)$ is the modified Bessel function defined as

$$I_j(x) = \sum_{n=0}^{\infty} \frac{(x/2)^{2n+j}}{n!(n+j)!}.$$

Using combinatorial arguments, an alternative form for (4.2.20) can be given as

$$P_0'(t) = e^{-(\lambda+\mu)t} \sum_{n=1}^{\infty} \frac{\lambda^{n-1}\mu^n t^{2n-2}}{n!(n-1)!} \tag{4.2.21}$$

(See Prabhu 1960)

Suppose $f(x)$ is the probability density of a random variable X and, $\phi(\theta)$ its Laplace transform. (See Appendix C.2)

$$\phi(\theta) = \int_0^\infty e^{-\theta x} f(x) dx \qquad R(\theta) > 0.$$

Two easily established properties of $\phi(\theta)$ are:

$$E(X) = -\phi'(0) \tag{4.2.22}$$

$$E(X^2) = \phi''(0). \tag{4.2.23}$$

Let B represent the length of the busy period. Using (4.2.22) and (4.2.23) on the transform of B given by (4.2.19), we get

$$E[B] = \frac{1}{\mu - \lambda} \tag{4.2.24}$$

$$V[B] = \frac{1+\rho}{\mu^2(1-\rho)^3}. \tag{4.2.25}$$

There may be occasions when a busy period starts out with an initial number of i customers in the system. Because of the Markovian property of the arrival process, we can show that the transition of the underlying Markov process from i to 0 can be considered to have made up of i intervals with the same distribution representing the transitions from $i \to i-1,\ i-1 \to i-2, \ldots, 1 \to 0$. These i independent busy periods start with 1 customer in the system. Then if B_i is the random variable representing a busy period initiated by i customers, we get

$$E[B_i] = \frac{i}{\mu - \lambda} \tag{4.2.26}$$

$$V[B_i] = \frac{i(1+\rho)}{\mu^2(1-\rho)^3}. \tag{4.2.27}$$

The explicit expression of the distribution of B_i can be given as

$$P_0'(t) = e^{-\lambda+\mu)t}\frac{i\sqrt{\mu/\lambda}}{t} I_i(2\sqrt{\lambda\mu}t). \tag{4.2.28}$$

It is easy to visualize the effect of the increase in traffic intensity ρ in the range (0,1) on the length of the busy period. As ρ increases the length of the busy period should increase. This can be shown with the help of the Laplace transform (4.2.19). Consider

$$\begin{aligned}
\lim_{\theta\to 0}\pi_0(\theta) &= \lim_{\theta\to 0}\frac{(\theta+\lambda+\mu) - [(\theta+\lambda+\mu)^2 - 4\lambda\mu]^{1/2}}{2\lambda} \\
&= \frac{1}{2\lambda}\left[\lambda+\mu-\sqrt{(\lambda-\mu)^2}\right] \\
&= \begin{cases} \frac{1}{2\lambda}[\lambda+\mu-(\mu-\lambda] & \text{if } \mu \geq \lambda \\ \frac{1}{2\lambda}[\lambda+\mu-(\lambda-\mu)] & \text{if } \mu < \lambda \end{cases} \\
&= \begin{cases} 1 & \text{if } \mu > \lambda \\ \frac{\mu}{\lambda} & \text{if } \mu < \lambda. \end{cases}
\end{aligned} \tag{4.2.29}$$

But $\lim_{\theta\to 0}\pi_0(\theta) = \int_0^\infty P_0'(t)dt$ where $P_0'(t)$ is the probability density of the busy period distribution. The conclusion we can draw from (4.2.29) is, therefore, that the busy period has a proper distribution when $\rho \leq 1$ and an improper distribution when $\rho > 1$. In the latter case, the probability that it will not terminate is given by $1 - \rho^{-1}$.

4.2.1 Departure Process

The departure process is the product of processes of arrival and service. When the server is continuously busy it coincides with the service process. But when idle times intervene there is a pause in service (during idle times) there is a pause in the departures as well. Nevertheless, in equilibrium we can derive the properties of the process without reference to arrivals and service.

Let $t_1, t_2, \ldots$ be the epochs of departure from the system, and define $T_n = t_{n+1} - t_n$. When the queue is in equilibrium, i.e., when traffic intensity $\rho < 1$, denote this random variable by T. Let $Q(x)$ be the number of customers in the system x amount of time after departure and define

$$F_n(x) = P[Q(x) = n, T > x]. \tag{4.2.30}$$

We should note here, as we shall see in Chapter 5, in the $M/M/1$ queue, the limiting distribution of the process $Q(t)$ derived in (4.2.4) remains the same when t in $Q(t)$ is an arbitrary time point, an arrival point, or a departure point (see Wolff (1982)). Therefore, regardless the value of x, we have

$$P[Q(x) = n] = (1-\rho)\rho^n \qquad n = 0, 1, 2 \ldots \qquad (x \geq 0).$$

From (4.2.30), we can determine $F(x)$ as

$$F(x) = P(T > x) = \sum_0^\infty F_n(x) \tag{4.2.31}$$

For a specified n, because of the Markovian property of the underlying process, the random variable T is dependent only on n, not on the preceding inter-departure intervals. To establish the relationship between $Q(x)$ and T and to derive the distribution of T, we start by considering the transition in the interval $(x, x+\Delta x]$. In (4.2.31), $F(x)$ is the probability that T, the time interval between epochs of the last departure and the next departure, is greater than x. This means we have to consider the possibility of only arrivals during $(x, x + \Delta x)$. We have

$$\begin{aligned} F_0(x+\Delta x) &= F_0(x)\,[1 - \lambda\Delta x] + o(\Delta x) \\ F_n(x+\Delta x) &= F_n(x)\,[1 - \lambda\Delta x - \mu\Delta x] \\ &\quad + F_{n-1}(x)\lambda\Delta x + o(\Delta x) \qquad n = 1, 2, \ldots. \end{aligned} \tag{4.2.32}$$

Rearranging terms in (4.2.32), dividing by Δx, and letting $\Delta x \to 0$, we get

$$\begin{aligned} F_0'(x) &= -\lambda F_0(x) \\ F_n'(x) &= -(\lambda+\mu)F_n(x) + \lambda F_{n-1}(x) \qquad n = 1, 2, \ldots. \end{aligned} \tag{4.2.33}$$

From (4.2.30), we also have

$$F_n(0) = P\,[Q(0) = n] = p_n. \tag{4.2.34}$$

The first equation in (4.2.33) can be solved by noting

$$\frac{d}{dx}\ln F_0(x) = \frac{F_0'(x)}{F_0(x)} = -\lambda.$$

Hence

$$\ln F_0(x) = -\lambda x + C.$$

Now using the initial condition (4.2.23) to determine C, we get

$$F_0(x) = p_0 e^{-\lambda x}. \tag{4.2.35}$$

The general solution to equations (4.2.33) can be obtained by induction. For, let

$$F_{n-1}(x) = p_{n-1}e^{-\lambda x} \qquad n = 1, 2, \ldots.$$

Substituting this in the second equation of (4.2.33), we get

$$F_n'(x) + (\lambda + \mu)F_n(x) = \lambda p_{n-1}e^{-\lambda x}.$$

The general form is now confirmed by multiplying both sides by $e^{(\lambda+\mu)x}$, integrating and using the initial condition from (4.2.23). We get

$$F_n(x) = p_n e^{-\lambda x} \qquad n = 1, 2, 3, \ldots. \tag{4.2.36}$$

Thus, we get

$$\begin{aligned} F(x) &= \sum_{n=0}^{\infty} p_n e^{-\lambda x} \\ &= e^{-\lambda x} \end{aligned} \tag{4.2.37}$$

which is the same as the distribution of the inter-arrival times. Since $\{p_n\}$ is also the distribution of the number of customers in the system at departure points, equation (4.3.36) also confirms the independence of the distribution of T from the queue length distribution at departure points. Note that here we are talking about the independence of distribution of two random variables and not any relationship between their specific values. For a more exhaustive treatment of this problem see Burke (1956) who has considered this problem for the multi-server $M/M/$s queue.

The important result coming out of this analysis states that the departure process of the $M/M/1$ queue in equilibrium is the same Poisson as the arrival process. Consequently, the expected number of customers served during a length of time t when the system is in equilibrium is given by λt.

Example 4.2.1 An airport has a single runway. Airplanes have been found to arrive at the rate of 15 per hour. It is estimated that each landing takes 3 minutes. Assuming a Poisson process for arrivals and an exponential distribution for landing times, use an $M/M/1$ model to determine the following performance measures.

a. Runway utilization.

Arrival rate = 15/h (λ)

Service rate = (60/3)/h = 20/h (μ)

Utilization = $\rho = \frac{\lambda}{\mu} = \frac{3}{4}$.

Answer.

b. Expected number of airplanes waiting to land:

$$L_q = \frac{\rho^2}{1-\rho} = \frac{(.75)^2}{.25} = 2.25$$

Answer.

c. Expected waiting time:

$$E(W_q) = \frac{\lambda}{\mu(\mu-\lambda)} = \frac{15}{20(20-15)} = \frac{3}{20} \text{ h } = 9 \text{ min}$$ **Answer.**

d. Probability that the waiting will be more than 5 min? 10 min? No waiting?

$$P(\text{no waiting}) = P(T_q = 0) = 1 - \rho = .25$$ **Answer.**

$$P(T_q > t) = \rho e^{-\mu(1-\rho)t}$$

$$P(T_q > 5 \text{ min}) = \frac{3}{4} e^{-20(1-\frac{3}{4})5/60}$$

$$= \frac{3}{4} e^{-\frac{25}{60}} = 0.4944$$ **Answer.**

$$P(T_q > 10 \text{ mins}) = \frac{3}{4} e^{-\frac{50}{60}} = 0.3259$$ **Answer.**

e. Expected number of landings in a 20-minute period = $\frac{15}{60} \times 20 = 5$.

Answer.

4.3 The Queue *M/M/s*

The multi-server queue $M/M/s$ is the model used most in analyzing service stations with more than one server such as banks, checkout counters in stores, check in counters in airports and the like. The arrival of customers is assumed to follow a Poisson process, service times are assumed to have an exponential distribution and the number of servers providing service independently of each other is assumed to be s. We also assume that the arriving customers form a single queue and the one at the head of the waiting line gets into service as soon as a server is free. No server stays idle as long as there are customers to serve.

Let λ be the arrival rate and μ the service rate. (This means that the inter-arrival times and service times have exponential distributions with densities $\lambda e^{-\lambda x}$ $(x > 0)$ and $\mu e^{-\mu x}$ $(x > 0)$ respectively). Note that the service rate μ is the same for all servers. In order to use the birth and death model introduced earlier, we have to establish values for λ_n and μ_n, when there are n customers in the system. Clearly, the arrival rate does not change with the number of customers in the system (i.e., λ is the constant arrival rate). What about μ_n and how does it change?

Suppose n $(n = 1, 2, \ldots, s)$ servers are busy at time t. Then, during $(t, t+\Delta t]$, the event that a busy server will complete service has the probability $\mu\Delta t + o(\Delta t)$. Since there are n busy servers at t, the probability that any r of the n busy servers will complete service during $(t, t + \Delta t]$, can be determined using the binomial probability distribution as

$$\begin{aligned} &= \binom{n}{1} [\mu\Delta t + o(\Delta t)] [1 - \mu\Delta t + o(\Delta t)]^{n-1} \\ &= \begin{cases} n\mu\Delta t & r = 1 \\ o(\Delta t) & r = 2, 3, ..., n. \end{cases} \end{aligned} \tag{4.3.1}$$

Note that $\frac{o(\Delta t)}{\Delta t} \to 0$ as $\Delta t \to 0$.

In a similar manner, the probability that a number $r(> 1)$ of the busy servers will complete service during $(t, t + \Delta t]$ can be given as

$$\begin{aligned} &= \binom{n}{r} [\mu\Delta t + o(\Delta t)]^r [1 - \mu\Delta t + o(\Delta t)]^{n-r} \\ &= o(\Delta t). \end{aligned}$$

Therefore, when there are n busy servers at time t, the only event in $(t, t+\Delta t]$ contributing to the reduction in that number that has a nonnegligible probability, is the completion of one service, and it has the probability given in (4.3.1). Hence, the service rate at that time is $n\mu$. Then in the framework of the birth and death queueing model, we have

$$\begin{aligned} \lambda_n &= \lambda \qquad n = 0, 1, 2, \ldots \\ \mu_n &= n\mu \qquad n = 1, 2, \ldots, s-1 \\ &= s\mu \qquad n = s,\ s+1, \ldots. \end{aligned} \tag{4.3.2}$$

The generator matrix $\mathbf{A}$ for the process can be given as

$$\mathbf{A} = \begin{matrix} 0 \\ 1 \\ \vdots \\ s \\ s+1 \\ \vdots \end{matrix} \begin{bmatrix} -\lambda & \lambda & & & & & \\ \mu & -(\lambda+\mu) & \lambda & & & & \\ & \cdot & \cdot & & & & \\ & & & s\mu & -(\lambda+s\mu) & \lambda & \\ & & & & s\mu & -(\lambda+s\mu) & \lambda \\ & & & & & \ddots & \end{bmatrix}. \tag{4.3.3}$$

Let $Q(t)$ be the number of customers in the system at time t and $P_n(t) = P[Q(t) = n|Q(0) = i]$. Forward Kolmogorov equations for $P_n(t)$ can be written down specializing (4.1.3). Since solving such equations is very cumbersome we do not plan to attempt it here. For solution through transform methods interested readers may refer to Saaty (1961). For numerical solutions of the forward Kolmogorov equations see Stewart (1994). For the limiting probabilities $p_n = \lim_{t\to\infty} P_n(t)$, we have (writing out $\boldsymbol{pA} = 0$)

$$\begin{aligned} \lambda p_0 &= \mu p_1 \\ (\lambda + n\mu)p_n &= \lambda p_{n-1} + (n+1)\mu p_{n+1} \qquad 0 < n < s \\ (\lambda + s\mu)p_n &= \lambda p_{n-1} + s\mu p_{n+1} \qquad s \le n < \infty. \end{aligned} \tag{4.3.4}$$

A recursive procedure on the lines of that used in the case of $M/M/1$ provides the following solution

$$\begin{aligned} n\mu p_n &= \lambda p_{n-1} \qquad n = 1, 2, \ldots, s \\ s\mu p_n &= \lambda p_{n-1} \qquad n = s+1, s+2, \ldots. \end{aligned}$$

(Also see (4.1.7)). Therefore,

$$\begin{aligned} p_n &= \frac{1}{n!}\left(\frac{\lambda}{\mu}\right)^n p_0 \qquad 0 \le n < s \\ p_{s+n} &= \left(\frac{\lambda}{s\mu}\right)^n p_s \qquad n = 0, 1, 2, \ldots \\ p_n &= \left(\frac{\lambda}{s\mu}\right)^{n-s} p_s \qquad n = s, s+1, \ldots. \end{aligned} \tag{4.3.5}$$

Writing $\frac{\lambda}{s\mu} = \rho$ and simplifying, we get

$$\begin{aligned} p_n &= \frac{1}{n!}(s\rho)^n p_0 \qquad 0 \le n < s \\ &= \frac{1}{s!}(s\rho)^s \rho^{n-s} p_0 \qquad s \le n < \infty. \end{aligned} \tag{4.3.6}$$

Using the condition $\sum_0^\infty p_n = 1$, (4.3.6) gives

$$\begin{aligned} p_0 &= \left[\sum_{r=0}^{s-1} \frac{(s\rho)^r}{r!} + \frac{(s\rho)^s}{s!(1-\rho)}\right]^{-1} \\ p_n &= \frac{(s\rho)^n}{n!} p_0 \qquad 0 \le n < s \\ &= \frac{s^s \rho^n}{s!} p_0 \qquad s \le n < \infty \end{aligned} \tag{4.3.7}$$

provided $\frac{\lambda}{s\mu} = \rho < 1$. Since $s\mu$ is the maximum service rate, we may consider ρ as defined above as the traffic intensity for the system. Writing the last equation in (4.3.5) as

$$p_n = \rho^{n-s} p_s \qquad n \ge s \tag{4.3.8}$$

we may say that when the number of customers in the system is $\geq s$, the system behaves like an $M/M/1$ with service rate $s\mu$. For convenience we may also write $\alpha = \frac{\lambda}{\mu}$, so that $\alpha/s = \rho$. An alternative form of (4.3.7) using α, can be given as

$$\begin{aligned} p_0 &= \left[\sum_{r=0}^{s-1} \frac{\alpha^r}{r!} + \frac{\alpha^s}{s!}(1-\alpha/s)^{-1}\right]^{-1} \\ p_n &= \frac{\alpha^n}{n!} p_0 \qquad 0 \leq n < s \\ &= \frac{\alpha^s}{s!}\left(\frac{\alpha}{s}\right)^{n-s} p_0 \qquad s \leq n < \infty. \end{aligned} \tag{4.3.9}$$

It should be noted that customers will have to wait for service only if the number in the system is $\geq s$. The probability of this event is given by $\sum_{n=s}^{\infty} p_n$ and hence

$$\begin{aligned} P(\text{customer delay}) &= C(s,\alpha) \\ &= \frac{\alpha^s}{s!}\left(1-\frac{\alpha}{s}\right)^{-1}\left[\sum_{r=0}^{s-1}\frac{\alpha^r}{s!} + \frac{\alpha^s}{s!}\left(1-\frac{\alpha}{s}\right)^{-1}\right]^{-1}. \end{aligned} \tag{4.3.10}$$

The formula for $C(s,\alpha)$ is known in the literature as *Erlang's Delay Formula* or *Erlang's Second Formula* and it is also denoted as $E_{2,s}(\alpha)$. (This result was first published by Erlang in 1917.) Before the advent of computers, the telephone industry used $C(s,\alpha)$ charts plotted for different combinations of s and α in the determination of the optimum number of lines.

Writing L and L_q as the mean number of customers in the system and the number in the queue respectively, we may derive them as follows. Using expressions from (4.3.7), we get (writing $s\rho = \alpha$ when convenient)

$$\begin{aligned} \sum_{n=1}^{\infty} n p_n &= p_0\left[\sum_{n=1}^{s} n\frac{\alpha^n}{n!} + \sum_{n=s+1}^{\infty} n\rho^{n-s}\frac{\alpha^s}{s!}\right] \\ &= p_0\left[\alpha\sum_{n=1}^{s-1}\frac{\alpha^{n-1}}{(n-1)!} + \frac{\alpha^s}{s!}\sum_{n-s=1}^{\infty} n\rho^{n-s}\right] \\ &= p_0\left[\alpha\sum_{r=0}^{s-1}\frac{\alpha^r}{r!} + \frac{\alpha^s}{s!}\sum_{r=1}^{\infty}(r+s)\rho^r\right] \\ &= p_0\left[\alpha\sum_{r=0}^{s-1}\frac{\alpha^r}{r!} + \frac{\alpha^s}{s!}\left(\frac{\rho}{(1-\rho)^2} + \frac{s\rho}{(1-\rho)}\right)\right] \\ &= \frac{\rho\alpha^s p_0}{s!(1-\rho)^2} + \alpha p_0\left[\sum_{r=0}^{s-1}\frac{\alpha^r}{r!} + \frac{\alpha^s}{s!(1-\rho)}\right]. \end{aligned}$$

Note that the terms inside [ ] above is $= p_0^{-1}$ (See (4.3.7)) Thus we get

$$L = \alpha + \frac{\rho\alpha^s p_0}{s!(1-\rho)^2} \tag{4.3.11}$$

which can also be written as

$$L = \alpha + \frac{\rho p_s}{(1-\rho)^2}. \tag{4.3.12}$$

To derive L_q, we write

$$\begin{aligned} L_q &= \sum_{n=s+1}^{\infty} (n-s)\frac{\alpha^s}{s!}\rho^{n-s}p_0 \\ &= \frac{\alpha^s}{s!}p_0 \sum_{n-s=1}^{\infty} (n-s)\rho^{n-s} \\ &= \frac{\alpha^s}{s!}p_0 \sum_{r=1}^{\infty} r\rho^r \\ &= \frac{\rho\alpha^s p_0}{s!(1-\rho)^2} \end{aligned} \tag{4.3.13}$$

which can also be written as

$$L_q = \frac{\rho p_s}{(1-\rho)^2}. \tag{4.3.14}$$

The expression for the variance of the number in the system gets cumbersome, so we do not present it here.

Comparing expressions for L and L_q, we can surmise that $s\rho(=\alpha)$ represents the expected number of busy servers. This can also be determined as the contribution of the utilization factor corresponding to s servers. For, we may write

$$\text{Individual Server Utilization} = \sum_{n=1}^{s-1} \frac{n}{s} p_n + \sum_{n=s}^{\infty} p_n. \tag{4.3.15}$$

Using expressions for p_n from (4.3.7) in (4.3.15) and simplifying, we find the individual server utilization factor to be ρ, i.e., in the long run the probability (or the fraction of time) a server will be busy is ρ.

Waiting Time

For the discussion on waiting times of customers, we assume that they are served with an FCFS queue discipline. When the number of customers in the system is

$\geq s$, the inter-departure times are exponential with rate parameter $s\mu$. Let T_q be the waiting time of the customer as $t \to \infty$ and $F_q(t) = P[T_q \leq t]$. Clearly

$$\begin{aligned} F_q(0) &= P[T_q = 0] = P(Q < s) \\ &= \sum_{n=0}^{s-1} p_n \\ &= p_0 \sum_{n=0}^{s-1} \frac{\alpha^n}{n!}. \end{aligned}$$

From the first equation in (4.3.9), we have

$$\sum_{n=0}^{s-1} \frac{\alpha^n}{n!} = \frac{1}{p_0} - \frac{\alpha^s}{s!}(1-\rho)^{-1}$$

giving

$$F_q(0) = 1 - \frac{\alpha^s p_0}{s!(1-\rho)}. \tag{4.3.16}$$

Also following the arguments leading to (4.2.10) for the queue $M/M/1$, in the multi-server case we have, (using (4.3.8) in the simplification)

$$\begin{aligned} dF_q(t) &= \sum_{n=s}^{\infty} p_n e^{-s\mu t} \frac{(s\mu t)^{(n-s)!}}{(n-s)!} s\mu dt \\ &= p_s e^{-s\mu t} \sum_{n=s}^{\infty} \rho^{n-s} \frac{(s\mu t)^{n-s}}{(n-s)!} s\mu dt \\ &= s\mu p_s e^{-s\mu(1-\rho)t} dt \end{aligned} \tag{4.3.17}$$

which can also be written as

$$= \frac{s\mu\alpha^s}{s!} p_0 e^{-s\mu(1-\rho)t} dt. \tag{4.3.18}$$

Noting that $F_q(0)$ does not contribute any term for the expected value of T_q, from (4.3.17), we have

$$\begin{aligned} W_q &= \int_0^{\infty} t dF_q(t) = \int_0^{\infty} s\mu p_s t e^{-s\mu(1-\rho)t} dt \\ &= \frac{p_s}{s\mu(1-\rho)^2}. \end{aligned} \tag{4.3.19}$$

Using p_0 instead of p_s, we may also write

$$W_q = \frac{\alpha^s p_0}{s! s\mu(1-\rho)^2}. \tag{4.3.20}$$

Comparing (4.3.14) with (4.3.19) or (4.3.13) with (4.3.20), we can again verify Little's formula $L_q = \lambda W_q$.

The distribution function $F_q(t)$ of the waiting time can now be obtained from (4.3.16) and (4.3.18).

$$\begin{aligned} F_q(t) &= F_q(0) + \int_0^t \frac{s\mu\alpha^s}{s!} p_0 e^{-s\mu(1-\rho)x} dx \\ &= 1 - \frac{\alpha^s p_0}{s!(1-\rho)} + \frac{\alpha^s p_0}{s!(1-\rho)} \int_0^t s\mu(1-\rho) e^{-s\mu(1-\rho)x} dx \\ &= 1 - \frac{\alpha^s p_0}{s!(1-\rho)} e^{-s\mu(1-\rho)t}. \end{aligned} \tag{4.3.21}$$

Busy Period

The meaning of the busy period in a multi-server queue requires further elaboration. If the busy period is the time when arriving customers have to wait for service, in a multi-server queue it is the time when all servers are busy. In $M/M/s$ this period has the same characteristics as a busy period in an $M/M/1$ queue, with the same arrival rate λ but with a service rate $s\mu$. But if it has to include periods during which at least one of the servers is busy, we need new results, which are beyond the scope of this discussion. The theoretical construct for the equations remains the same as (4.2.18), but because of the varying service rates the equations get much harder to simplify.

Departure Process

As mentioned during the discussion of the departure process of the queue $M/M/1$, the procedure outlined there applies to $M/M/s$ as well. In fact, the differential equations (4.2.33) can be extended to include varying service rates and the inductive procedure adopted in their solution applies in this case as well. Using the same notations as before, we get

$$F_n(x) = p_n e^{-\lambda x}, \qquad n = 0, 1, 2, \ldots$$

and

$$\begin{aligned} F(x) &= \sum_{n=0}^{\infty} p_n e^{-\lambda x} \\ &= e^{-\lambda x} \end{aligned} \tag{4.3.22}$$

(See Burke (1956) for details. Also Reich (1965)).

Example 4.3.1 In the airport problem of Example 4.2.1, how would the performance measures change if there are two runways while assuming the same arrival and service rates?

a. Runway utilization.

Arrival rate = 15/h (λ)
Service rate = 20/h (μ)
No. of servers = 2 (s)
Utilization of each runway = $\rho = \frac{\lambda}{s\mu} = \frac{3}{8}$ **Answer**.

b. Expected number of airplanes waiting to land:

$$L_q = \frac{\rho\alpha^s p_0}{s!(1-\rho)^2}$$

Note: $\alpha = s\rho = \frac{3}{4}$

$$\begin{aligned} p_0 &= \left[\sum_{r=0}^{1}\frac{\alpha^r}{r!} + \frac{\alpha^s}{s!(1-\rho)}\right]^{-1} \\ &= \left[1 + \frac{3}{4} + \frac{(3/4)^2}{2}(1-\frac{3}{8})^{-1}\right]^{-1} \\ &= 0.4545 \\ L_q &= \left[(\frac{3}{8})(\frac{3}{4})^2(0.4545)\right] \Big/ 2(\frac{5}{8})^2 \\ &= 0.1227 \quad \textbf{Answer}. \end{aligned}$$

c. Expected waiting time

$$\begin{aligned} W_q &= \frac{\alpha^s p_0}{s!s\mu(1-\rho)^2} \\ &= \left[(\frac{3}{4})^2(0.4545)\right] \Big/ 2\times 2\times 20(1-\frac{3}{8})^2 \\ &= 0.00818 \text{ h} = 0.49 \text{ min} \quad \textbf{Answer}. \end{aligned}$$

d . Probability that the waiting will be more than 5 minutes? 10 minutes? no waiting?

$$\begin{aligned} P(\text{no waiting}) &= F_q(0) = 1 - \frac{\alpha^s p_0}{s!(1-\rho)} \\ &= 1 - \frac{(\frac{3}{4})^2(0.4545)}{2(1-3/8)} \\ &= 0.7955 \quad \textbf{Answer}. \\ P(T_q > t) &= \frac{\alpha^s p_0}{s!(1-\rho)} e^{-s\mu(1-\rho)t} \\ P(T_q > 5 \text{ min}) &= \frac{(3/4)^2(0.4545)}{2(5/8)} e^{-2(\frac{1}{3})(\frac{5}{8})5} \\ &= 0.1245 \quad \textbf{Answer}. \\ P(T_q > 10 \text{ min}) &= 0.0155 \quad \textbf{Answer}. \end{aligned}$$

e. Expected number of landings in a 20-minute period $=\frac{15}{60}\times 20 = 5$. **Answer.** (The departure process is Poisson with parameter λ).

Example 4.3.2 A bank has established two counters, one for commercial banking and the second for personal banking. Arrival and service rates at the commercial counter are 6 and 12 per hours, respectively. The corresponding numbers at the personal banking counter are 12 and 24, respectively. Assume that arrivals occur in Poisson processes and service times have exponential distributions.

(i) Assuming that the two counters operate independently of each other determine the expected number of waiting customers and their mean waiting time at each counter.

	Commercial	Personal	
λ	6/h	12/h	
μ	12/h	24/h	
$\rho = \frac{\lambda}{\mu}$	0.5	0.5	
$L_q = \frac{\rho^2}{1-\rho}$	0.5	0.5	**Answer.**
$W_q = \frac{\rho}{\mu(1-\rho)}$	5 min	2.5 min	**Answer.**

(ii) What is the effect of operating the two queues as a two-server queue with arrival rate 18/h and service rate 18/h? What conclusion can you draw from this operation?

	Two-server queue	
λ	18/h	
μ	18/h	
Number of servers (s)	2	
$\rho = \frac{\lambda}{s\mu}$	0.5	
$\alpha = \frac{\lambda}{\mu}$	1	
$p_0 = [\sum_0^1 \frac{\alpha^r}{r!} + \frac{\alpha^2}{2(1-\rho)}]^{-1}$	0.33	
$L_q = \frac{\rho\alpha^2 p_0}{2(1-\rho)^2}$	0.33	**Answer.**
$W_q = \frac{\alpha^2 p_0}{(2)\ 2\mu(1-\rho)^2}$	1.33 min	**Answer.**

Conclusion: The two-server queue operation is more efficient than the two single-server operations.

Incidentally, the efficiency of multi-server over single-server systems is the reason that multi-server service systems, whenever possible, use single waiting lines feeding multiple counters for service, e.g., airline checkin counters, checkout counters in stores effectively operate this way because of jockeying among the waiting lines (see Smith and Whitt (1981)).

4.4 The Finite Queue $M/M/s/K$

When the waiting room in a queueing system has a capacity limit we get a finite queue. In most situations, a finite queue occurs more naturally than a queue with waiting room of infinite size. However, as the capacity limit gets larger, the behavior of the system approximates that of an infinite capacity system and in such cases we are justified in ignoring the size limit. A communication system with a finite buffer and several service channels is a good example of a finite queue.

Consider an s-server queueing system with Poisson arrivals, exponential service and a capacity limit of K for the number in the system. Clearly $K \geq s$. Assume that λ and μ are the arrival and service rates respectively. These assumptions result in the following infinitesimal transition rates in the generalized birth and death queueing model.

$$\begin{aligned} \lambda_n &= \lambda \qquad n = 0, 1, 2, \ldots, K-1 \\ \mu_n &= n\mu \qquad n = 1, 2, \ldots, s-1 \\ &= s\mu \qquad n = s, s+1, \ldots, K. \end{aligned} \tag{4.4.1}$$

Note that we assume the arrivals to be denied entry to the system (or the arrival process stops) once the number in the system reaches K. The generator matrix $\mathbf{A}$ is essentially the same as (4.3.3), in the first K rows,

$$\mathbf{A} = \begin{array}{c} 0 \\ 1 \\ \vdots \\ \vdots \\ K-1 \\ K \end{array} \begin{bmatrix} -\lambda & \lambda & & & & & \\ \mu & -(\lambda+\mu) & \lambda & & & & \\ & \ddots & & & & & \\ & & \ddots & & & & \\ & & & s\mu & -(\mu+s\mu) & \lambda & \\ & & & & s\mu & & -s\mu \end{bmatrix}. \tag{4.4.2}$$

For the limiting probabilities $\{p_n\}$ $n = 0, 1, 2, \ldots, K$, the state balance equations can be written down in a manner similar to (4.3.4). The solution corresponding to (4.3.6) can be given as

$$\begin{aligned} p_n &= \frac{1}{n!}\left(\frac{\lambda}{\mu}\right)^n p_0 \qquad 0 \leq n \leq s \\ &= \frac{1}{s!}\left(\frac{\lambda}{\mu}\right)^s \left(\frac{\lambda}{s\mu}\right)^{n-s} p_0 \qquad s \leq n \leq K. \end{aligned}$$

Writing $\frac{\lambda}{s\mu} = \rho$ and $\frac{\lambda}{\mu} = \alpha$, p_0 can be obtained using the condition $\sum_{n=0}^{K} p_n = 1$.

$$p_0 = \left[\sum_{r=0}^{s-1} \frac{\alpha^r}{r!} + \frac{\alpha^s}{s!} \sum_{n=s}^{K} \rho^{n-s}\right]^{-1}.$$

Since the second sum on the right-hand side of the expression for p_0 is a finite sum, we need not impose the condition $\rho < 1$ for a solution with $p_0 > 0$. Thus we have

$$\begin{aligned} p_0 &= \left[\sum_{r=0}^{s-1} \frac{\alpha^r}{r!} + \frac{\alpha^s}{s!}\frac{1-\rho^{K-s+1}}{1-\rho}\right] && \rho \neq 1 \\ &= \left[\sum_{r=0}^{s-1} \frac{\alpha^r}{r!} + \frac{\alpha^s}{s!}(K-s+1)\right]^{-1} && \rho = 1 \\ p_n &= \frac{\alpha^n}{n!}p_0 && 0 \leq n < s \\ &= \frac{\alpha^s}{s!}\rho^{n-s}p_0 && s \leq n \leq K. \end{aligned} \tag{4.4.3}$$

Because of the unwieldy nature of the expressions for the mean number in the system (L) and in the queue (L_q) we do not present them here. The procedure for deriving them starts with the limiting distribution given by (4.4.3).

In discussing the characteristics of the waiting time of customers in a finite queue, we need to allow for the possibility of an arriving customer not joining the system. When the system is in equilibrium, the probability that the arriving customer will not join the system is p_K. Hence, when there are n ($n < K$) customers in the system, the joint probability that there are n customers and an arrival will join the system is given by $p_n/(1-p_K)$. Thus, using the notations used earlier for the distribution for the waiting time, we have

$$F_q(t) = F_q(0) + P(0 < W_q \leq t)$$

where

$$F_q(0) = \sum_{n=0}^{s-1} p_n/(1-p_K).$$

Also,

$$dF_q(t) = \sum_{n=s}^{K-1} \frac{p_n}{1-p_K} e^{-s\mu t}\frac{(s\mu t)^{n-s}}{(n-s)!}s\mu dt \tag{4.4.4}$$

$$\begin{aligned} F_q(t) &= F_q(0) + \frac{1}{1-p_K}\sum_{n=s}^{K-1} p_n \int_0^t e^{-s\mu t}\frac{(s\mu t)^{n-s}}{(n-s)!}s\mu dt \\ &= F_q(0) + \frac{1}{1-p_K}\sum_{n=s}^{K-1} p_n \left(1 - \int_t^\infty e^{-s\mu t}\frac{(s\mu t)^{n-s}}{(n-s)!}s\mu dt\right). \end{aligned}$$

In simplifying this expression, we note that

$$F_q(0) + \frac{1}{1-p_K}\sum_{n=s}^{K-1} p_n = 1$$

and

$$\int_t^\infty e^{-s\mu t}\frac{(s\mu t)^{n-s}}{(n-s)!}s\mu dt = \sum_{r=0}^{n-s} e^{-s\mu t}\frac{(s\mu t)^r}{r!}$$

(see (2.1.3)). Then we get

$$F_q(t) = 1 - \frac{1}{1-p_K}\sum_{n=s}^{K-1} p_n \sum_{r=0}^{n-s} e^{-s\mu t}\frac{(s\mu t)^r}{r!}. \tag{4.4.5}$$

Taking expectations, we get

$$\begin{aligned} W_q = \int_0^\infty t dF_q(t) &= \sum_{n=s}^{K-1}\frac{p_n}{1-p_K}\int_0^\infty e^{-s\mu t}\frac{(s\mu t)^{n-s}}{(n-s)!}s\mu t dt \\ &= \frac{1}{s\mu(1-p_K)}\sum_{n=s}^{K-1}(n-s+1)p_n. \end{aligned} \tag{4.4.6}$$

The expected time in the system can be obtained as

$$W = W_q + \frac{1}{\mu}. \tag{4.4.7}$$

The expected number of customers in the queue and in the system are obtained by noting that the effective arrival rate is $\lambda(1-p_K)$.

$$L = \lambda(1-p_K)W \tag{4.4.8}$$

$$L_q = \lambda(1-p_K)W_q \tag{4.4.9}$$

Two special cases of this system have been used widely in applications:

(i) $M/M/1/K$

(ii) $M/M/\mathrm{s}/\mathrm{s}$

The Finite Queue *M/M/1/K*

For single-server systems with limited waiting room, $M/M/1/K$ is a better model than the infinite waiting room queue $M/M/1$. A direct specialization of results (4.4.3)–(4.4.7) yields the following results. Note that $s = 1$ and $\alpha = \rho = \frac{\lambda}{\mu}$.

$$\begin{aligned} p_0 &= \frac{1-\rho}{1-\rho^{K+1}} && \rho \neq 1 \\ &= \frac{1}{K+1} && \rho = 1 \end{aligned} \tag{4.4.10}$$

$$\begin{aligned} p_n &= \frac{(1-\rho)\rho^n}{1-\rho^{K+1}} && \rho \neq 1 \\ &= \frac{1}{K+1} && \rho = 1. \end{aligned} \tag{4.4.11}$$

Also

$$\begin{aligned} 1-p_K &= \frac{1-\rho^K}{1-\rho^{K+1}} && \rho \neq 1 \\ &= \frac{K}{K+1} && \rho = 1 \\ F_q(t) &= 1-\frac{1-\rho}{1-\rho^K}\sum_{n=1}^{K-1}\rho^n\sum_{r=0}^{n-1}e^{-\mu t}\frac{(\mu t)^r}{r!} && \rho \neq 1 \\ &= 1-\frac{1}{K}\sum_{n=1}^{K-1}\sum_{r=0}^{n-1}e^{-\mu t}\frac{(\mu t)^r}{r!} && \rho = 1 \end{aligned} \tag{4.4.12}$$

$$\begin{aligned} W_q &= \frac{1}{\mu}\left[\frac{\rho}{1-\rho}-\frac{K\rho^K}{1-\rho^K}\right] && \rho \neq 1 \\ &= \frac{1}{2\mu}(K-1) && \rho = 1 \end{aligned} \tag{4.4.13}$$

$$\begin{aligned} W &= \frac{1}{\mu}\left[\frac{1}{1-\rho}-\frac{K\rho^K}{1-\rho^K}\right] && \rho \neq 1 \\ &= \frac{K+1}{2\mu} && \rho = 1 \end{aligned} \tag{4.4.14}$$

$$\begin{aligned} L_q &= \frac{\rho}{1-\rho}-\frac{\rho(1+K\rho^K)}{1-\rho^{K+1}} && \rho \neq 1 \\ &= \frac{K(K-1)}{2(K+1)} && \rho = 1 \end{aligned} \tag{4.4.15}$$

$$\begin{aligned} L &= \frac{\rho(1-\rho^K)}{(1-\rho)(1-\rho^{K+1})}-\frac{K\rho^{K+1}}{1-\rho^{K+1}} && \rho \neq 1 \\ &= \frac{K}{2} && \rho = 1. \end{aligned} \tag{4.4.16}$$

The readers should note that in the simplifications leading to some of the results given above we have used the formula

$$\sum_{n=1}^{K-1} n\rho^{n-1} = \frac{d}{d\rho}\left(\frac{1-\rho^K}{1-\rho}\right).$$

Example 4.4.1 A small mail order business has one telephone line and a facility for call waiting for two additional customers. Orders arrive at the rate of 1 per

minute and each order requires 2 minutes and 30 seconds to take down the particulars. Model this system as an $M/M/1/3$ queue and answer the following questions.

(a) What is the expected number of calls waiting in the queue? What is the mean wait in queue?

Assuming that the arrivals are in a Poisson process with rate 1 per minute (λ) and the service times are exponential with mean 2.5 minute ($1/\mu$). We have $\rho = 2.5$. Also $K = 3$. Using the first result from (4.4.15) we get

$$\begin{aligned} L_q &= \frac{2.5}{1-2.5} - \frac{(2.5)[1+3(2.5)^3]}{1-(2.5)^4} \\ &= 1.4778 \end{aligned}$$ **Answer.**

Since $\lambda = 1$, the mean waiting time in queue

$$W_q = 3.7271 \text{ min.}$$ **Answer.**

(b) What is the probability that the call has to wait for more than 1.5 minute before getting served?

We use the formula for $1 - F_q(t)$ from (4.4.12) with $t = 1.5$, $1/\mu = 2.5$ and $\rho = 2.5$. We get,

$$\begin{aligned} &\text{P(Wait in queue} > 1.5 \text{ min)} \\ &= \frac{1-2.5}{1-(2.5)^3} \sum_{n=1}^{3-1} (2.5)^n \sum_{r=0}^{n-1} e^{-\frac{1.5}{2.5}} \frac{(1.5/2.5)^r}{r!} \\ &= 0.7036 \end{aligned}$$ **Answer.**

(c) Because of the excessive waiting time of customers, the business decides to use two telephone lines instead of one, keeping the same total capacity for the number in the system, namely 3. What improvements result in the performance measures considered under (a) and (b)?

With two lines, now $s = 2$ and we have an $M/M/2/3$ system. Accordingly in (4.4.3), we have $\alpha = 2.5$, $\rho = 1.25$, $s = 2$ and $K = 3$. We get

$$\begin{aligned} p_0 &= 0.0950, \qquad p_1 = 0.2374 \\ p_2 &= 0.2969, \qquad p_3 = 0.3711 \end{aligned}$$

Using these results in (4.4.6), (4.4.9), and (4.4.5), we get

$$W_q = 0.5902 \text{ min}$$ **Answer.**

$$L_q = \lambda(1-p_3)W_q = 0.3712$$ **Answer.**

P(wait in queue > 1.5 min):

$$1 - F_q(1.5) = 0.1422.$$ **Answer.**

(d) What is the impact of increasing the capacity to four customers in the system? Now we have an $M/M/2/4$ queue. Using the formulas as in (c) we get,

$$\begin{aligned} &p_0 = 0.0649, \qquad p_1 = 0.1622 \\ &p_2 = 0.2028, \qquad p_3 = 0.2535 \\ &p_4 = 0.3169 \\ &W_q = 1.2989 \text{ min.} && \textbf{Answer}. \\ &L_q = 0.8873 && \textbf{Answer}. \end{aligned}$$

P (Wait in queue > 1.5 min):

$$1 - F_q(1.5) = 0.3353 \qquad \textbf{Answer}.$$

It is instructive to note that the performance has not improved from the viewpoint of the customer, because the system now accepts more customers than before. But from the management perspective, fewer customers are being denied access to the system ($p_4 = 0.3169$ vs. $p_3 = 0.3711$).

The Loss System $M/M/$s/s

The queue $M/M/$s/s in which customers arriving when all servers are busy, are not allowed entry to the system is one of the earliest systems considered by A. K. Erlang (1917). Before the introduction of call waiting buffers, telephone systems operated strictly as loss systems.

Let customer arrivals be Poisson with parameter λ and service times be exponential with mean $1/\mu$. There are s servers and all customers arriving when all servers are busy are lost to the system. Thus, the state space for the number of customers in the system is $\{0, 1, 2, \ldots, s\}$. The generator matrix for the birth and death model is a modified version of (4.3.3):

$$\mathbf{A} = \begin{array}{c} 0 \\ 1 \\ \vdots \\ s-1 \\ s \end{array} \begin{bmatrix} -\lambda & \lambda & & & & \\ \mu & -(\lambda+\mu) & \lambda & & & \\ & \vdots & & & & \\ & & (s-1)\mu & -[\lambda+(s-1)\mu] & \lambda & \\ & & & s\mu & & -s\mu \end{bmatrix}. \tag{4.4.17}$$

Accordingly, the limiting probabilities are obtained using the state balance equations,

$$\begin{aligned} \lambda p_0 &= \mu p_1 \\ (\lambda + n\mu)p_n &= \lambda p_{n-1} + (n+1)\mu p_{n+1} \qquad 1 \le n < s \\ s\mu p_s &= \lambda p_{s-1}. \end{aligned} \tag{4.4.18}$$

Writing $\frac{\lambda}{\mu} = \alpha$, (4.4.18) can be solved recursively to give

$$p_0 = \left[1 + \alpha + \alpha^2/2! + \ldots + \alpha^s/s!\right]^{-1}$$

$$p_n = \frac{\alpha^n}{n!} p_0 \qquad n = 0, 1, \ldots, s. \tag{4.4.19}$$

This gives

$$p_s = \frac{\alpha^s \Big/ s!}{1 + \alpha + \frac{\alpha^2}{2!} + \ldots + \frac{\alpha^s}{s!}} \tag{4.4.20}$$

which is the probability that a customer is blocked from entering the system. (Telephone calls are lost.) Also λp_s gives the expected number of customers who will be blocked from entering the system in unit time. Equation (4.4.20) is commonly known as *Erlang's Loss Formula* or *Erlang's First Formula* and denoted as $\mathrm{E}_{1,s}(\alpha)$ or $B(s, \alpha)$. This formula has been extensively used in designing telephone systems by traffic engineers.

For ready reference p_s values are plotted for different values of s, against varying values of the *offered load* α. In the teletraffic literature, it is common to measure the offered load (the ratio of arrival rate to the service rate) in *Erlangs* for convenience. It should be noted that in the telephone industry parlance the *carried load* is given by $\alpha(1-p_s)$, since a proportion p_s of the arriving customers are lost to the system.

The right-hand side expression in the formula (4.4.20) has been shown to be a convex function of s in $[0, \infty)$ for $\alpha > 0$. (See Smith and Whitt (1981) and Jagers and van Doorn (1986)). Another characteristic of this formula is its validity even when service times have a general distribution.

4.5 The Infinite Server Queue $M/M/\infty$

Even though calling a system with an infinite number of servers, consequently with no waiting line, *an infinite server queue*, is a misnomer, the system $M/M/\infty$ is being identified as such because of its structure. The customers arrive in a Poisson process and the service times have an exponential distribution. Let λ and μ be the arrival and service rates. We assume that the system is able to provide service as soon as the customer arrives. A simple example is a large grocery store or a supermarket where customers serve themselves while picking up merchandise. The checkout counters will then have to be modeled as an $M/M/$s system. Another example is a large parking lot.

When there are n customers in the system the service rate is $n\mu$ ($n = 1, 2, \ldots$). For the birth and death parameters of the queueing model, we have

$$\begin{aligned} \lambda_n &= \lambda \qquad n = 0, 1, 2, \ldots \\ \mu_n &= n\mu \qquad n = 1, 2, 3 \ldots . \end{aligned} \tag{4.5.1}$$

The generator matrix is obtained by extending the first part of the matrix (4.3.3) of the multi-server queue $M/M/s$, for $n = s+1, s+2, \ldots$. We get

$$\mathbf{A} = \begin{array}{c} 0 \\ 1 \\ 2 \\ \vdots \end{array} \begin{bmatrix} -\lambda & \lambda & & & \\ \mu & -(\lambda+\mu) & \lambda & & \\ & 2\mu & -(\lambda+2\mu) & \lambda & \cdots \\ & & \vdots & \vdots & \end{bmatrix}. \tag{4.5.2}$$

The state balance equations for the limiting probabilities $\{p_n,\ n = 0, 1, 2 \ldots\}$ take the form

$$\begin{aligned} \lambda p_0 &= \mu p_1 \\ (\lambda + n\mu)p_n &= \lambda p_{n-1} + (n+1)\mu p_{n+1} \qquad n = 1, 2, \ldots. \end{aligned} \tag{4.5.3}$$

These equations along with $\sum_0^\infty p_n = 1$ give the solution

$$\begin{aligned} p_0 &= e^{-\lambda/\mu} \\ p_n &= e^{-\lambda/\mu}(\frac{\lambda}{\mu})^n \qquad n = 0, 1, 2, \ldots \end{aligned} \tag{4.5.4}$$

which is Poisson with parameter λ/μ.

Because of the structure of the birth and death parameters (4.5.1), the system can also be considered a queueing system with arrivals in a Poisson process and exponential service times with a linearly dependent arrival rate $n\mu$ when there are $n\mu$ customers in the system. It should be noted that the departure process properties established in Section 4.2 apply to this system as well and therefore, the departure process of customers in equilibrium, has the same Poisson distribution, as the arrival process. This property justifies the use of an $M/M/s$ model for the checkout counters in the supermarket example cited above.

4.6 Finite Source Queues

The source of customers is an important element of a queueing system. In the models discussed so far, we have assumed that the source is infinite. This assumption is essential in characterizing the arrivals to the system as being Poisson. If the source of customers is finite, a prespecified number in the population, even though we cannot assume a Poisson process for arrivals, we can generate a Markovian arrival process with the following arrival scheme.

Suppose there are M customers in the population. Each customer goes through two alternating phases, not needing service and being in need of service. An example is a population of machines that require service when they become inoperative. Another example is a subscriber group in an information exchange. Assume that the phase during which the customer does not require service is

exponentially distributed with mean $1/\lambda$. This implies that if at time t, a customer is in this phase, during $(t, t+\Delta t]$, the event that it will need service has the probability $\lambda t + o(\Delta t)$. Thus, if there are k customers in that phase requiring no service at time t, the probability that one of them will call for service is $k\lambda + o(\Delta t)$ (see discussion preceeding (4.3.1)). When the service times are exponential, we are able to use a birth and death queueing model for such a system.

For convenience, let us define the state of the process as the number of customers requiring service. It takes values in $S : \{0, 1, 2, \ldots, M\}$. Note that if n is the number requiring service, the leftover population size that can generate customers for service is $M - n$. Also assume that there are s $(s \leq M)$ servers. These assumptions lead to the birth and death parameters as:

$$\begin{aligned} \lambda_n &= (M-n)\lambda && n = 0, 1, \ldots, M \\ &= 0 && n > M \\ \mu_n &= n\mu && n = 1, 2, \ldots, s-1 \\ &= s\mu && n = s, s+1, \ldots, M \\ &= 0 && n > M. \end{aligned} \tag{4.6.1}$$

There are two classical examples with these characteristics treated in the queueing literature. The *machine interference problem* has M machines and s repairmen. Naturally, inoperative machines wait for their turn when all repairmen are busy. The second problem is similar to the $M/M/\mathrm{s}/\mathrm{s}$ loss system in which, customers arriving when all servers are busy are lost and the lost customers have to reinitiate the request for service.

The Machine Interference Problem

Let $Q(t)$ be the number of inoperative machines at time t out of a total number M. Assume that the call for service and the completion of service have the characteristics leading to the birth and death parameters as described in (4.6.1). Define

$$P_n(t) = P[Q(t) = n | Q(0) = i]$$

and $p_n = \lim_{t\to\infty} P_n(t)$.

The generator matrix has a structure similar to that of (4.4.2) with obvious modifications to the birth rate. We give below the state balance equations.

$$\begin{aligned} M\lambda p_0 &= \mu p_1 \\ [(M-n)\lambda + n\mu]\, p_n &= (M-n+1)\lambda p_{n-1} + (n+1)\mu p_{n+1} && 1 \leq n < s \\ [(M-n)\lambda + s\mu]\, p_n &= (M-n+1)\lambda p_{n-1} + s\mu p_{n+1} && s \leq n < M \\ s\mu p_M &= \lambda p_{M-1}. \end{aligned} \tag{4.6.2}$$

Solving these equations recursively,

$$p_1 = M\left(\frac{\lambda}{\mu}\right) p_0$$

$$[(M-1)\lambda + \mu] p_1 = M\lambda p_0 + 2\mu p_2$$

giving

$$\begin{aligned}
p_2 &= \frac{M(M-1)}{2}\left(\frac{\lambda}{\mu}\right)^2 p_0 \\
&\ddots \\
p_n &= \binom{M}{n}\left(\frac{\lambda}{\mu}^n\right)p_0 \qquad 0 \le n \le s \\
[(M-s)\lambda + s\mu]\, p_s &= (M-s+1)\lambda p_{s-1} + s\mu p_{s+1} \\
p_{s+1} &= \frac{(M-s)\lambda}{s\mu} p_s \\
&= \binom{M}{s+1}\frac{(s+1)!}{s!s}\left(\frac{\lambda}{\mu}\right)^{s+1} p_0 \\
&\ddots \\
p_n &= \binom{M}{n}\frac{n!}{s!s^{n-s}}\left(\frac{\lambda}{\mu}\right)^n p_0 \qquad s \le n \le M. \qquad (4.6.3)
\end{aligned}$$

The limiting probabilities $\{p_n, n = 0, 1, \ldots, M\}$ are now determined in the usual manner using the condition $\sum_0^M p_n = 1$. In particular, when $s = 1$, writing $\frac{\lambda}{\mu} = \alpha$, we get

$$p_n = \frac{M!}{(M-n)!}\alpha^n p_0$$

and

$$p_0 = \left[1 + \frac{M!}{(M-1)!}\alpha + \ldots + M!\alpha^M\right]^{-1}. \qquad (4.6.4)$$

In the context of machines and repairmen, two measures of effectiveness can be defined as (using L to stand for the mean number of inoperative machines):

$$\begin{aligned}
\text{Machine availability} &= 1 - \frac{L}{M} \\
\text{Operative utilization} &= \sum_{n=0}^{s-1}\frac{np_n}{s} + \sum_{n=s}^{M} p_n.
\end{aligned}$$

Illustrative values of operative utilization for different values of α, M, and s can be easily determined from (4.6.3) as shown in the table below (Table 4.1).

Table 4.1 Illustrative values of operative utilization

α	M	s	Operative utilization
0.45	4	1	0.881
	8	2	0.934
	16	4	0.994
0.05	15	1	0.656
	30	2	0.682
	60	4	0.705

An obvious conclusion we can draw from the numbers in the table is that it is better to use repairmen in a pool, rather than assigning a certain number of machines to each of them as long as characteristics of service are the same among repairmen.

The number of machines actually waiting for service could also be of interest. Because of the permutations occurring in the expressions, unfortunately, we are unable to give closed form expressions for the mean number of inoperative machines in the system directly from the equation. To determine the number actually waiting, L_q, we may use the relation obtained for the $M/M/$s queue in (4.3.11) and (4.3.13). However, the arrival rate in this case is dependent on the remaining number of operative machines in the population. Let λ' be the effective arrival rate. Then we have

$$\begin{aligned} \lambda' &= \sum_{n=0}^{M-1} (M-n)\lambda p_n \\ &= \lambda(M-L). \end{aligned} \tag{4.6.5}$$

We get

$$L_q = L - \frac{\lambda'}{\mu} = L - \alpha(M-L). \tag{4.6.6}$$

The expressions for waiting time for service can be obtained using Little's law with λ' as the arrival rate instead of λ.

The Finite Source Loss System

Consider an information exchange with M subscribers. The exchange has s servers and there is no facility for call waiting when all servers are busy. Assume that the call arrivals are initiated in the same manner as in the machine interference problem with parameter λ, and the service times are exponential with rate μ. The state of the system is the number of calls being serviced and state space is therefore $S:\ 0, 1, 2, \ldots, s$. The state balance equations for the limiting

probabilities p_n, $n = 0, 1, 2, \ldots, s$ can be written down as

$$\begin{aligned} M\lambda p_0 &= \mu p_1 \\ [(M-n)\lambda + n\mu]\, p_n &= (M-n+1)\lambda p_{n-1} + (n+1)\mu p_{n+1} \quad 1 \le n < s \\ s\mu p_s &= (M-s+1)\lambda p_{s-1}. \end{aligned} \tag{4.6.7}$$

Solving these equations recursively we get

$$\begin{aligned} p_n &= \binom{M}{n}\alpha^n p_0 \qquad 0 \le n \le s \\ p_0 &= [1 + \binom{M}{1}\alpha + \binom{M}{2}\alpha^2 + \ldots + \binom{M}{s}\alpha^s]^{-1} \end{aligned} \tag{4.6.8}$$

and therefore

$$p_n = \frac{\binom{M}{n}\alpha^n}{\sum_{k=0}^{s}\binom{M}{r}\alpha^r} \qquad n = 0, 1, 2, \ldots, s. \tag{4.6.9}$$

To determine the probability that one of the M sources will find the system busy while initiating a call, we have to consider the probability that all servers are busy serving calls from the remaining $M-1$ sources. Let this probability be b_s. Then we have

$$b_s = \frac{\binom{M-1}{s}\alpha^s}{\sum_{k=0}^{s}\binom{M-1}{k}\alpha^k}. \tag{4.6.10}$$

This result is often called the *Engset formula* in the literature. Clearly, this is the proportion of calls lost to the system. The distribution (4.6.9) is known as the *Engset distribution*.

In the discussion of $M/M/s/s$ system, we mentioned that the distribution (4.4.19) was valid even when the service time is general. In a similar manner, it has been shown that the distribution (4.6.9) holds even when the arrival process has a more general structure.

4.7 Other Models

In this section, we present additional models that may be considered as specializations of the general birth and death queueing model.

4.7.1 The $M/M/1/1$ System

Even though this system can be considered a specialization of the finite queue $M/M/1/K$ of Section 4.4, the $M/M/1/1$ system is significant in its own right because it corresponds to the two-state Markov process useful in a large number of applications.

Let customers arrive in a Poisson process with parameter λ and get served by a single server. The service time distribution is exponential with mean $1/\mu$. The

system can accommodate only one customer who is being served, and customers arriving when the server is busy leave the system without service. Let $Q(t)$ be the number of customers in the system at time t and $lim_{t\to\infty}Q(t) = Q$. The random variable Q can assume two values (0,1) and let $P(Q = n) = p_n$ $(n = 0, 1)$. Clearly, $\{Q(t), t\epsilon T\}$ is a Markov process with the generator matrix

$$\boldsymbol{A} = \begin{matrix} \\ 0 \\ 1 \end{matrix} \begin{matrix} \begin{matrix} 0 & \quad 1 \end{matrix} \\ \begin{bmatrix} -\lambda & \lambda \\ \mu & -\mu \end{bmatrix} \end{matrix}. \tag{4.7.1}$$

For the transition probability $P_{ij}(t) = P[Q(t) = j|Q(0) = i]$ $(i, j = 0, 1)$ we get the forward Kolmogorov equations

$$\begin{aligned} P'_{i0}(t) &= -\lambda P_{i0}(t) + \mu P_{i1}(t) \\ P'_{i1}(t) &= -\mu P_{i1}(t) + \lambda P_{i0}(t). \end{aligned} \tag{4.7.2}$$

If we note that $P_{i0}(t) + P_{i1}(t) = 1$, the two equations in (4.7.2) give a single linear first order differential equation

$$P'_{i0}(t) = \mu - (\lambda + \mu)P_{i0}(t). \tag{4.7.3}$$

Using the initial condition

$$P_{i0}(0) = \begin{cases} 1 & \text{if } i = 0 \\ 0 & \text{if } i = 1 \end{cases},$$

(4.7.3) can be solved through standard techniques to give

$$\begin{aligned} P_{00}(t) &= \frac{\mu}{\lambda+\mu} + \frac{\lambda}{\lambda+\mu}e^{-(\lambda+\mu)t} \\ P_{10}(t) &= \frac{\mu}{\lambda+\mu} - \frac{\mu}{\lambda+\mu}e^{-(\lambda+\mu)t}. \end{aligned} \tag{4.7.4}$$

Also, we have

$$\begin{aligned} P_{01}(t) = 1 - P_{00}(t) &= \frac{\lambda}{\lambda+\mu} - \frac{\lambda}{\lambda+\mu}e^{-(\lambda+\mu)t} \\ P_{11}(t) = 1 - P_{10}(t) &= \frac{\lambda}{\lambda+\mu} + \frac{\mu}{\lambda+\mu}e^{-(\lambda+\mu)t}. \end{aligned}$$

The limiting probabilities p_n, $n = 0, 1$, can be determined either by letting $t \to \infty$ in (4.7.4) or by solving the state balance equation

$$\lambda p_0 = \mu p_1 \tag{4.7.5}$$

along with the normalizing condition $p_0 + p_1 = 1$. We get

$$p_0 = \frac{\mu}{\lambda+\mu}; \quad p_1 = \frac{\lambda}{\lambda+\mu}. \tag{4.7.6}$$

These probabilities can be expressed in terms of the mean busy and idle periods. Dividing the numerator and denominator of the expressions for p_0 and p_1 by $\lambda\mu$, we get

$$p_0 = \frac{1/\lambda}{1/\mu + 1/\lambda}; \quad p_1 = \frac{1/\mu}{1/\mu + 1/\lambda}. \tag{4.7.7}$$

Note that $1/\lambda$ is the mean idle period and $1/\mu$ is the mean busy period. Generalizing this concept to a process that occupies two alternate states 0 and 1, represented by two independent random variables X and Y, it can be shown, with the help of renewal theory, that in the long run the probabilities that the process can be found in the states 0 and 1, are given by

$$p_0 = \frac{E(X)}{E(X) + E(Y)}; \quad p_1 = \frac{E(Y)}{E(X) + E(Y)}. \tag{4.7.8}$$

The breadth of applicability of this model can be easily seen if we look at the process alternating between two states: busy or idle in the context of a service system, working or under repair in the context of a machine in operation, "locked" or "ready to register signals" in a Type I counter, etc.

Suppose there are N such multiple processes undergoing transitions between alternate states independent of each other with the transition structure as described above. The probability distribution of the number of processes in states (0,1) is then given by the binomial distribution:

$$P(N = k) = \binom{N}{k} p_0^k p_1^{N-k} \quad k = 0, 1, 2, \ldots, N. \tag{4.7.9}$$

4.7.2 Markovian Queues with Balking

Balking is a phenomenon in which an arriving customer decides not to join the queue. The reason for balking could be external or internal to the queue; in the latter case, normally, it depends on the number in the systems.

As a general model, consider a single server queueing system with Poisson arrivals and exponential service, the rates of arrival and service being λ_n and μ_n respectively, when there are n customers in the system. In order to incorporate balking in the arrival process, we consider serveral special forms for the arrival rate λ_n. The limiting probability p_n, $n = 0, 1, 2, \ldots$ for the number of customers in the system is given by (4.1.8) and (4.1.9) of the general birth and death model

(i)

$$\begin{aligned} \lambda_n &= \lambda\alpha \qquad n = 0, 1, 2, \ldots; \quad 0 \le \alpha \le 1 \\ \mu_n &= \mu \qquad\quad n = 1, 2, \ldots. \end{aligned} \tag{4.7.10}$$

This case assumes that only a certain portion α of the arriving customers decide to join the queue. Substituting in (4.1.8) and (4.1.9), we get

$$p_n = (1 - \rho)\rho^n \quad n = 0, 1, 2 \ldots \tag{4.7.11}$$

where $\rho = \frac{\lambda\alpha}{\mu}$.

(ii)

$$\begin{aligned} \lambda_n &= \frac{\lambda}{n+1} \quad n = 0, 1, 2, \ldots \\ \mu_n &= \mu \quad n = 1, 2, 3, \ldots . \end{aligned} \tag{4.7.12}$$

The arrival rate here is inversely proportional to the number of customers in the system (Haight 1957). Substituting in (4.1.9) and (4.1.8), we get

$$\begin{aligned} p_0 &= \left[1 + \frac{\lambda}{\mu} + \frac{1}{2}\left(\frac{\lambda}{\mu}\right)^2 + \frac{1}{3!}\left(\frac{\lambda}{\mu}\right)^3 + \ldots\right]^{-1} \\ &= e^{-\rho} \\ p_n &= \frac{1}{n!}\left(\frac{\lambda}{\mu}\right)^n p_0 \\ &= e^{-\rho}\frac{\rho^n}{n!} \qquad n = 0, 1, 2, \ldots \end{aligned} \tag{4.7.13}$$

where $\rho = \frac{\lambda}{\mu}$.

(iii)

$$\begin{aligned} \lambda_n &= \frac{N-n}{N(n+1)} \quad n = 0, 1, 2, \ldots, N \\ &= 0 \quad n > N \\ \mu_n &= \mu \quad n = 1, 2, \ldots, N \end{aligned} \tag{4.7.14}$$

In this case the blocking phenomenon also includes the factor that the customers do not join the queue when its size reaches N (Haight 1957). Substituting in (4.1.8), we get

$$\begin{aligned} p_n &= \frac{N(N-1)\ldots(N-n+1)}{n!}\left(\frac{1}{N\mu}\right)^n \\ &= \binom{N}{n}\left(\frac{1}{N\mu}\right)^n . \end{aligned}$$

Using (4.1.9)

$$\begin{aligned} p_0 &= \left[\sum_{n=0}^{N}\binom{N}{n}\left(\frac{1}{N\mu}\right)^n\right]^{-1} \\ &= \left(1 + \frac{1}{N\mu}\right)^{-N} \end{aligned}$$

Thus we get

$$\begin{aligned} p_n &= \binom{N}{n}\left(\frac{1}{N\mu}\right)^n\left(1+\frac{1}{N\mu}\right)^{-N} \\ &= \binom{N}{n}\left(\frac{1}{1+N\mu}\right)^n\left(\frac{N\mu}{1+N\mu}\right)^{N-n} \\ & \quad n=0,1,2,...,N. \end{aligned} \tag{4.7.15}$$

(iv)

$$\begin{aligned} \lambda_n &= \lambda e^{-n\alpha/\mu} \quad n=0,1,2,\ldots;\ \alpha>0 \\ \mu_n &= \mu \qquad\qquad\quad n=1,2,3,\ldots. \end{aligned} \tag{4.7.16}$$

The arrival rate here, incorporates a fraction that reflects an estimate of waiting time t and an impatience factor α in the customer's decision to join the queue (Morse 1958, p. 24). Substituting in (4.1.9) and (4.1.8), we get

$$\begin{aligned} p_0 &= \left[\sum_{n=0}^{\infty}\rho^n\Pi_{i=1}^{n-1}e^{-i\alpha/\mu}\right]^{-1} \\ &= \left[\sum_{n=0}^{\infty}\rho^n e^{\frac{-n(n-1)\alpha}{2\mu}}\right]^{-1} \\ p_n &= \left[\rho^n e^{\frac{-n(n-1)\alpha}{2\mu}}\right]p_0 \qquad\qquad n=1,2,\ldots \end{aligned} \tag{4.7.17}$$

where $p=\frac{\lambda}{\mu}$.

4.7.3 Markovian Queues with Reneging

After joining the queue, if a customer abandons its desire to get served and leaves the system, the customer is said to have *reneged*. One way to incorporate this factor in modeling is to assume a distribution, normally an exponential distribution in between successive customer reneging. Let β be the rate, independent of the number in the system, at which reneging occurs. Then assuming a constant arrival rate λ and service rate μ, we can give the birth and death parameters for the model as

$$\begin{aligned} \lambda_n &= \lambda \qquad\quad n=0,1,2,\ldots \\ \mu_n &= \mu+\beta \qquad n=1,2,3\ldots. \end{aligned} \tag{4.7.18}$$

Writing $\mu+\beta=\gamma$ and $\rho=\frac{\lambda}{\gamma}$, for the limiting probabilities, we have (with $\rho<1$)

$$p_n=(1-\rho)\rho^n \quad n=0,1,2,\ldots. \tag{4.7.19}$$

4.7.4 Phase Type Machine Repair

The $M/M/1/1$ system discussed in Section 4.7.1 can be generalized to consider a machine repair requiring k phases. Suppose, a machine requires service after it has been in operation for a length of time exponentially distributed with mean $1/\lambda$. Let the repair require k phases of service, where the ith phase ($i = 1, 2, \ldots, k$) is exponentially distributed with mean $1/\mu_i$. The operating and repair states of the machine are 0 (operating), and i (representing phase i, $i = 1, 2, \ldots, k$). Because of the exponential distributions involved in the process, the machine can be considered to undergo transitions in a Markov process with the following generator matrix

$$\mathbf{A} = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ \vdots \\ k \end{array} \begin{array}{c} \begin{array}{ccccc} 0 & 1 & 2 & \ldots & k \end{array} \\ \begin{bmatrix} -\lambda & \lambda & & & \\ & -\mu_1 & \mu_1 & & \\ & & -\mu_2 & \mu_2 & \\ & & & \ddots & \\ \mu_k & & & & -\mu_k \end{bmatrix} \end{array}. \tag{4.7.20}$$

Let $p_n = (p_0, p_1, p_2, \ldots, p_k)$ be the limiting probabilities for the state of the machine. For state balance equations, we have

$$\begin{aligned} \lambda p_0 &= \mu_k p_k \\ \mu_1 p_1 &= \lambda p_0 \\ &\vdots \\ \mu_k p_k &= \mu_{k-1} p_{k-1}. \end{aligned} \tag{4.7.21}$$

Solving recursively, we get

$$\begin{aligned} p_1 &= \frac{\lambda}{\mu_1} p_0 \\ p_2 &= \frac{\lambda}{\mu_2} p_0 \\ \cdots p_k &= \frac{\lambda}{\mu_k} p_0 \end{aligned}$$

Using the normalizing condition $\sum_0^k p_n = 1$, we get

$$\begin{aligned} p_0 &= \left[1 + \lambda \sum_1^k \frac{1}{\mu_i}\right]^{-1} \\ p_n &= \left(\frac{\lambda}{\mu_n}\right) \left[1 + \lambda \sum_1^k \frac{1}{\mu_i}\right]^{-1} \\ & \qquad n = 1, 2, \ldots, k. \end{aligned} \tag{4.7.22}$$

We may note that the overall repair time has a generalized Erlang distribution of (A.14) with the transform

$$\psi(\theta) = \sum_{i=1}^{k} \left(\frac{\mu_i}{\theta + \mu_i} \right). \tag{4.7.23}$$

4.8 Remarks

In this chapter, we have discussed only a few queueing systems for which generalized birth and death process models are suitable. We shall discuss a few more extended models in Chapters 6 and 7. There are many more examples in the queueing literature where such models have been effectively used. For instance, Syski (1960) has provided a large number of models for queueing systems applicable to telephone industry. Further perusal of the telecommunication systems literature would reveal models developed since 1960.

There are other application areas, such as computer and manufacturing systems, where investigators use birth and death process models as a first line of attack in solving problems. The major advantages of these models are their Markovian structure often leading to useable explicit results, and the ability to use numerical investigations without complex computational problems when explicit results are not forthcoming. After all, queueing models are approximate representation of real systems, and starting with a Markovian model provides a good starting point for the understanding of their approximate behavior.

4.9 Exercises

1. Compare the system idle time probability (p_0) in the three systems (1) $M/M/s/s$, (2) $M/M/s$, and (3) $M/M/\infty$ and show that

$$p_0^{(1)} > p_0^{(2)} \text{and} p_0^{(3)} > p_0^{(2)}. \tag{4.9.1}$$

2. An airline employs two counters, one exclusively for first class and business class passengers and the other for coach class passengers. The service times at both counters have been found to be exponential with mean 3 minute. The coach class passengers arrive at the rate of 18 per hour and upper class passengers arrive at the rate of 15 per hour. Is there any advantage in keeping the exclusivity of service in the counters? Answer this question using server utilization, mean number of customers in the system, and the mean waiting time, all in steady state.

3. A customer service counter has s telephone lines. Service requests arrive in a Poisson process with rate λ and the length of service is exponentially distributed with mean $1/\mu$. What is the probability that a request will encounter a busy system? What is the probability that a service request will arrive when the service center is busy?

4. Customer arrivals at a 7-Eleven is Poisson at the rate of 20 per hour. They can be assumed to spend an average of 12 minutes picking up merchandise, with the length of time having an exponential distribution. Two checkout counters provide service with a service rate of 15 per hour at each counter. We may also assume that the service times have an exponential distribution. Determine the limiting results for the following.
   (1) The distribution of the number of customers picking up merchandise and its mean.
   (2) The mean length of time the customers wait at the counter for service.
   (3) The mean total amount of time the customers spend in the store.

5. In a taxi stand there is space for only five taxicabs. Taxis arrive in a Poisson process with rate 12 per hour. If there is no waiting room, arriving taxis leave without passengers. Customers arrive at the taxi stand in a Poisson process once every 6 minutes on the average.
   (1) Determine the limiting distribution of the number of customers waiting for taxis.
   (2) What is the probability that there are taxis waiting for customers?
   (3) Determine the mean waiting time for a customer.

6. An automobile service station has one station for oil and filter change. On an average the oil and filter change takes 7 minutes, the amount of time having an exponential distribution. Cars arrive in a Poisson process at the rate of 6 per hour. What is the probability that an arriving car has to wait more than 10 minutes to get served?

   What is the effect to the waiting time of adding another station with identical service characteristics? Determine the probability that the waiting time will be more than 5 minutes with two stations for oil and filter change.

7. Customer arrivals to a service counter is in a Poisson process at the rate of 10 per hour. The service time distribution can be assumed to be exponential. Determine the minimum rate of service that would result in the customer waiting time being greater than 5 minutes. with a probability 0.10 or less.

8. In a manufacturing process production machines breakdown at the rate of 3 per hour. We may assume that the process of breakdowns is Poisson. The repair times of machines can be assumed to have an exponential distribution. The repairs can be run at two rates: 4 per hour at a cost of \$20/h and 5 per hour at a cost of \$30/h. Considering the loss of productivity of the machines while they are either waiting for service or being in service, what rate of repair would make it beneficial to provide service at the faster rate? You may assume an 8-hour workday in your calculations.

9. Customer arrivals to a store are in a Poisson process with a rate of 50 per hour. On an average, each customer spends 15 minutes in the store and we may assume the times the customer spends in the store to have an exponential distribution. Currently the store provides parking space for 15 cars. What is the probability that no parking space will be available, if a customer were to arrive at some time? How many more spaces will be needed to make sure that the arriving customer will find parking space 99 % of the time?

10. Suppose the arrival and the service rates in Ex.#9 are changed to: arrivals, 100 per hour and mean service time 30 minutes. How many parking spaces should be provided to make sure that the arriving customers will find parking space 99 % of the time?

11. A single switchboard is used to direct calls coming to a doctor's office. The calls arrive in Poisson process at a rate of 15 per hour. Call holding times can be assumed to be exponential with a mean of 2 minute. What is the probability that the calls will not have to wait for more than 2 minutes before getting to the receptionist? Assume unlimited room for all waiting.

    Suppose it is decided to establish an upper limit K for the number of calls waiting, such that the waiting time will be less than 2 minute with a 90 % probability. Determine K by successively increasing its value.

12. In the $M/M/s/s$ (loss system) show that in the long run

$$L = \rho[1 - P_B] \tag{4.9.2}$$

    where L = long run expected number of customers in the system;

$$\begin{aligned} \rho &= \frac{\text{arrival rate}}{\text{service rate}} \\ P_B &= \text{Prob. that an arriving customer is blocked from entering the system.} \end{aligned}$$

13. In a drug store customers arrive at the counter (with one server per counter) in a Poisson process at the rate of 48 per hour. The service time can be assumed to be exponential with an average of 1 minute and the service is provided at a counter attended by a server. The number of counters is varied depending on the number of customers waiting or being served as follows:

0–4 customers	1 counter
5–9 customers	2 counters
10–14 customers	3 counters
15 or more	4 counters

Assume that this policy is used to increase or decrease the number of servers.

Determine the following:

a. Probability of system idleness.

b. How often would the store need more than one counter?

c. What is the average number of customers either waiting for service or being served?

d. What is the average waiting time in the queue?

14. The atmopsheric quality at time t—denoted $A(t)$—is measured by the number of pollutant units residing in the airshed at that time. These units are emitted from pollutant sources one unit at a time with rate α. The emission process can be assumed to be poisson. Each unit thus emitted gets diffused in an average time of length β. Also assume that the diffusion times are exponential random variables which are independent and identically distributed. Obtain the mean and variance of $A(t)$ as $t \to \infty$.

15. (1) Writing $\beta = \frac{1}{\alpha} = \frac{\mu}{\lambda}$ in (4.6.4), using s in place of M, show that p_0 of (4.6.4) can be expressed as

$$p_0 = (\beta^s/s!) \Big/ \left(\sum_{n=0}^{s} \frac{\beta^n}{n!} \right)$$

which is the probability of blocking in an $M/M/s/s$ system (see (4.4.20)).

(2) Let λ^* be the effective arrival rate of machines for repair. Noting that λ^* can be expressed also as

$$\lambda^* = \frac{M}{(1/\lambda) + W_q + (1/\mu))}$$

show that the mean waiting time of a machine repair (waiting + service) is given by

$$W = \frac{M}{\lambda^*} - \frac{1}{\lambda}.$$

16. Ten terminals used for data entry in a hospital share a communication line. Terminals use the line on a FCFS basis and wait in a queue when the line is busy. It has been observed that the data entry job takes, on an average 100 seconds and once the terminal is free, it is ready for the job in 5 seconds, on the average. Determine the throughput rate (effective arrival rate λ^* of Exercise 15) and the mean response time W. (Total time for job completion = waiting + service).

17. A computer system has s servers. Since each server can be accessed separately, each of the s servers can be considered a separate subsystem as well. The arrival of jobs to each server is Poisson with rate λ and the service time is exponential with mean $1/\mu$. The main system operator would like to find out whether pooling resources would be advantageous in terms of response time (the amount of time the job spends in the system). With this objective consider the following three setups when $s = 3$.

    (1) Three separate systems

    (2) Arrivals are pooled into a single queue and processed separately as a multi-server queue.

    (3) The arrivals are pooled as in (2). In addition the servers are connected, such that together they process jobs as a single server with rate 3μ.

    Let W_i be the mean response time with the ith setup ($i = 1, 2, 3$). Show that

    $$W_1 > W_2 > W_3.$$

18. In a cyclic queue model of a single CPU and an I/O processor, the number of jobs in the system remains a constant N. After receiving service at the CPU, the job leaves the system with probability α and joins the I/O queue with probability $1 - \alpha$. Soon after a job leaves the system a new job is admitted to the CPU queue. The service times at the CPU and the I/O are exponential with mean $\frac{1}{\mu_1}$ and $\frac{1}{\mu_2}$ respectively. Determine the limiting distribution of the number of jobs waiting and being served at the CPU queue. Also determine the mean time in system for a job (Coffman and Denning 1973).

19. In a communication system, messages are transmitted through M identical channels. Messages are segmented for storage in fixed size buffers (bins). An individual message may require several buffers, but no buffer contains data from more than one message. When messages release the buffers from which they are transmitted, the buffers are ready for reuse.

    Assume that messages arrive in a Poisson process with rate λ. The messages are of length L that is exponentially distributed with mean $1/\mu_L$. The transmission rate for the messages is R, so that the transmission time is exponential with mean $\frac{1}{R\mu_L}$.

    The messages are stored in buffers. The data field size per buffer is b. Let N be the random variable representing the number of buffers used by a message.

    (a) Obtain the distribution of N.

    (b) Obtain the limiting probability that no message is present in the system.

(c) Determine the distribution of the number of occupied buffers under statistical equilibrium and its mean and variance in terms of the limiting probability of no messages present in the system. (Pederson and Shah 1972).

20. The following model describes a simplified representation of a multiprogramming system. Let the drum storage unit with a shortest-latency-time-first (SLTF) file drum described in Ex.10 of Chapter 1 be connected to a central processor unit with a fixed number of m tasks circulating in a closed system, alternately requesting service at the processor and the drum. Let μ_n be the service rate at the file drum unit as described in Ex.10 of Chapter 1, and λ be the service rate at the central processor. Let $p_n,\ n = 0, 1, 2, \ldots, m$, be the limiting distribution of the queue length (including the one in service) at the file drum unit.

    Determine $\{p_n\}$ and the expected processor utilization for various values of m (which is known as the degree of multiprogramming) (Fuller 1980).

21. A simplified model of the drum storage unit described in Ex.20, assumes a Poisson arrival of requests for files with rate λ. Let the service rate μ_n be determined by the formula

$$\frac{1}{\mu_n} = \frac{\tau}{n+1} + \frac{1}{\mu}$$

    where τ is the period of rotation and n is the number of requests in the system. Determine the mean waiting time of a request (Fuller (1980)).

22. In a time-shared computer system M terminals share a central processor. Let μ be the processing rate at the CPU, with the processing time having an exponential distribution. If a terminal is free at time t, the probability that it will initiate a job in the infinitesimal interval $(t, t + \Delta t]$ is $\lambda\Delta t + o(\Delta t)$ and it will continue to be free at $t + \Delta t$ has the probability $1 - [\lambda\Delta t + o(\Delta t)]$.

    (a) Let $\{p_n\}$ be the probability distribution of the number of busy terminals as $t \to \infty$. Determine p_n, $n = 0, 1, 2, \ldots, M$.

    (b) Show that in the long run, the arrival rate at the CPU is given by

$$\frac{M\lambda}{1 + \lambda W}$$

    where W is the mean response time (= mean waiting time of a job arriving at the terminal.)

    (c) Equating the arrival rate with the departure rate from the processor show that the mean response time can be obtained as

$$\frac{M}{\mu(1 - p_0)} - \frac{1}{\lambda}.$$

(Fuller 1980).

23. Consider a two server Markovian queue $M/M_i/2$, in which customer arrivals are in a Poisson process with parameter λ and the service times of the two servers are distributed exponentially with rates $\mu_1 > \mu_2$. An arriving customer finding both servers free, always chooses the faster server. But if there is only one server free when an arrival occurs, it enters service with the free server regardless of the service rate. If both servers are busy, the arriving customer waits in line for service in the order of arrival.

    Determine the limiting distribution of the number of customers in the system.

    Compare numerically the mean number of customers in the heterogeneous system $M/M_i/2$ with the corresponding homogeneous system $M/M/2$, when the service rate in the latter system is $(\mu_1 + \mu_2)/2$.

    (Singh 1970).

24. Extend Ex.23, above to an $M/M_i/3$ heterogeneous queue and determine the limiting distribution of the number of customers in it. Also carry out a numerical comparison of the mean number of customers in the systems between $M/M_i/3$ and $M/M/3$, when the service rate in the latter system is the average of the three heterogeneous rates.

    (Singh 1971).

Chapter 5

Imbedded Markov Chain Models

In the last chapter, we used Markov process models for queueing systems with Poisson arrivals and exponential service times. To model a system as a Markov process, we should be able to give complete distribution characteristics of the process beyond time t, using what we know about the process at t and changes that may occur after t, without referring back to the events before t. When arrivals are Poisson and service times are exponential, because of the memoryless property of the exponential distribution we are able to use the Markov process as a model. If the arrival rate is λ and service rate is μ, at any time point t, time to next arrival has the exponential distribution with rate λ, and if a service is in progress, the remaining service time has the exponential distribution with rate μ. If one or both of the arrival and service distributions are non-exponential, the memoryless property does not hold and a Markov model of the type discussed in the last chapter does not work. In this chapter, we discuss a method by which a Markov model can be constructed, not for all t, but for specific time points on the time axis.

5.1 Imbedded Markov Chains

In an $M/G/1$ queueing system, customers arrive in a Poisson process and get served by a single server. We assume that service times of customers are independent and identically distributed (i.i.d) with an unspecified (general) distribution. Let $Q(t)$ be the number of customers in the system at time t. For the complete description of the state of the system at time t, we need the value of $Q(t)$ as well as information on the remaining service time of the customer in service, if there is one being served at that time. Let $R(t)$ be the remaining service time of such a customer. Now the vector $[Q(t), R(t)]$ is a vector Markov process since both of its components, viz., the number in the system and the

© Springer Science+Business Media New York 2015

U. N. Bhat, *An Introduction to Queueing Theory*, Statistics for Industry and Technology, DOI 10.1007/978-0-8176-8421-1_5

remaining service time, are now completely specified. The earliest investigation to analyze this vector process by itself was by Cox (1955), who used information on $R(t)$ as a supplementary variable in constructing the forward Kolmogorov equations given in Chapter 3. Since it employs analysis techniques beyond the scope set for this text, we shall not cover it here.

In two papers in the 1950s D. G. Kendall (1951, 1953) developed a procedure to convert the queue length processes in $M/G/1$ and $G/M/s$ into Markov chains. (In the queue $G/M/s$, the service time has the memoryless property. Therefore, in the vector process $[Q(t), R(t)]$, $R(t)$ now represents the time until a new arrival.) The strategy is to consider departure epochs in the queue $M/G/1$ and arrival epochs in the queue $G/M/s$. Let $t_0 = 0, t_1, t_2, \ldots$ be the points of departure of customers in the $M/G/1$ queue and define $Q(t_n + 0) = Q_n$. Thus, Q_n is defined as the value of $Q(t)$ soon after departure. At the points $\{t_n,\ n = 0, 1, 2, \ldots\}$, $R(t)$ is equal to zero; hence, Q_n can be studied without reference to the random variable $R(t)$. Because of the Markov property of the Poisson distribution the process $\{Q_n, n = 0, 1, 2, \ldots\}$ is a Markov chain with discrete parameter and state spaces. Due to the imbedded nature of the process it is known as an imbedded Markov chain. In the queue $G/M/s$, arrival points generate the imbedded Markov chain. We discuss these two systems in the next two sections.

Imbedded Markov chains can also be used to analyze waiting times in the queue $G/G/1$. A limited exploration of that technique will be given in Chapter 9. The matrix-analytic method described in Chapter 8 is entirely based on imbedded Markov chains as well.

5.2 The Queue $M/G/1$

Let customers arrive in a Poisson process with parameter λ and get served by a single server. Let the service times of these customers be i.i.d. random variables $\{S_n,\ n = 1, 2, 3, \ldots\}$ with $P(S_n \le x) = B(x)$, $x \ge 0$; $E(S_n) = b$; $V(S_n) = \sigma_s^2$. We assume that S_n is the service time of the nth customer. Let $Q(t)$ be the number of customers in the system at time t and identify $t_0 = 0, t_1, t_2, \ldots$ as the departure epochs of customers. As described above, at these points the remaining service times of customers are zero. Let $Q_n = Q(t_n + 0)$ be the number of customers in the system soon after the nth departure. We can show that $\{Q_n,\ n = 0, 1, 2, \ldots\}$ is a Markov chain as follows.

Let X_n be the number of customers arriving during S_n. With the Poisson assumption for the arrival process, we have

$$\begin{aligned} k_j = P(X_n = j) &= \int_0^\infty P(X_n = j | S_n) P(t < S_n \le t + dt) \\ &= \int_0^\infty e^{-\lambda t} \frac{(\lambda t)^j}{j!} dB(t) \quad j = 0, 1, 2, \ldots . \end{aligned} \tag{5.2.1}$$

In writing $dB(t)$ in (5.2.1), we use the Stieltjes notation in order to accommodate discrete, continuous, and mixed distributions. (See Appendix C.)

Consider the relationship between Q_n and Q_{n+1}. We have

$$Q_{n+1} = \begin{cases} Q_n + X_{n+1} - 1 & \text{if } Q_n > 0 \\ X_{n+1} & \text{if } Q_n = 0 \end{cases}. \tag{5.2.2}$$

The first expression for Q_{n+1} is obvious. The second expression (i.e., X_{n+1} if $Q_n = 0$) results from the fact that T_{n+1} is the departure point of the customer who arrives after t_n. It is in fact $= 1 - 1 + X_{n+1}$.

As can be seen from (5.2.2), Q_{n+1} can be expressed in terms of Q_n and a random variable X_{n+1}, which does not depend on any event before t_n. Since X_{n+1} is i.i.d., it does not depend on Q_n either. The one-step dependence of a Markov chain holds. Hence, $\{Q_n,\ n = 0, 1, 2, \ldots\}$ is a Markov chain. Its parameter space is made up of departure points, and the state space S is the number of customers in the system; $S = \{0, 1, 2, \ldots\}$. Because of the imbedded nature of the parameter space, it is known as an *imbedded Markov chain.*

Let

$$P_{ij}^{(n)} = P(Q_n = j | Q_0 = i), \quad i, j \epsilon S \tag{5.2.3}$$

and write $P_{ij}^{(1)} \equiv P_{ij}$.

From the relationship (5.2.2) and the definition of k_j in (5.2.1), we can write

$$\begin{aligned} P_{ij} &= P(Q_{n+1} = j | Q_n = i) \\ &= \begin{cases} P(i + X_{n+1} - 1 = j) & \text{if } i > 0 \\ P(X_{n+1} = j) & \text{if } i = 0 \end{cases} \\ &= \begin{cases} k_{j-i+1} & \text{if } i > 0 \\ k_j & \text{if } i = 0 \end{cases}. \end{aligned} \tag{5.2.4}$$

The transition probability matrix $\mathbf{P}$ for the Markov chain is

$$\mathbf{P} = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ 3 \\ \ddots \end{array} \begin{array}{c} \begin{array}{cccc} 0 & 1 & 2 & \ldots \end{array} \\ \begin{bmatrix} k_0 & k_1 & k_2 & \ldots \\ k_0 & k_1 & k_2 & \ldots \\ & k_0 & k_1 & \ldots \\ & & k_0 & \ldots \\ & & & \ldots \end{bmatrix} \end{array}. \tag{5.2.5}$$

For the Markov chain to be irreducible the state space should have a single equivalence class. (See Appendix B.3 for the definition of the term *irreducible* and the procedure of classification of states.) For this to happen in this context the following two conditions must hold: $k_0 > 0$ and $k_0 + k_1 < 1$. It is easy to see that if $k_0 = 0$, with one or more customer arrivals for each departure, there is no possibility for the system to attain stability and the number in the system will only increase with time. If $k_0 + k_1 = 1$, the only two states $\{0, 1\}$ are possible in the system eventually. (If the system starts with $i > 1$ customers, once it gets 0 or 1 it will remain in $\{0, 1\}$).

Further classification of states depends on $E(X_n)$, the expected number of customers arriving during a service time.

Define the Laplace–Stieltjes transform of the service time distribution

$$\psi(\theta) = \int_0^\infty e^{-\theta t} dB(t) \quad \text{Re}(\theta) > 0 \tag{5.2.6}$$

and the probability generating function (PGF) of the number of customers arriving during a service time

$$K(z) = \sum_{j=0}^{\infty} k_j z^j \quad |z| \leq 1. \tag{5.2.7}$$

The following results follow from well-known properties of Laplace–Stieltjes transforms and PGFs:

$$\begin{aligned} E(S_n) &= b = -\psi'(0) \\ E(S_n^2) &= \psi''(0) \\ E(X_n) &= K'(1) \\ E(X_n^2) &= K''(1) + K'(1). \end{aligned} \tag{5.2.8}$$

From (5.2.1), we get

$$\begin{aligned} K(z) &= \int_0^\infty e^{-\lambda t} \sum_{j=0}^{\infty} \frac{(\lambda t z)^j}{j!} dB(t) \\ &= \int_0^\infty e^{-(\lambda - \lambda z)t} dB(t) \\ &= \psi(\lambda - \lambda z). \end{aligned}$$

Hence

$$\begin{aligned} K'(z) &= -\lambda \psi'(\lambda - \lambda z) \\ K'(1) &= -\lambda \psi'(0) \\ &= \lambda b. \end{aligned} \tag{5.2.9}$$

Note that λb =(arrival rate) $\times$ (mean service time). This quantity is called the *traffic intensity* of the queueing system denoted by ρ. The value of ρ determines whether the system is in equilibrium (attains steady state) when the time parameter n (of t_n) $\to \infty$. It can be shown that when $\rho < 1$, the Markov chain is positive recurrent (i.e., the process returns to any state with probability one and the mean time for the eventual return $< \infty$); when $\rho = 1$, the chain is null recurrent (i.e., the process returns to any state with probability one, but the mean time for the eventual return $= \infty$); and when $\rho > 1$, the chain is transient (i.e., the process may not return to the finite states at all. Then the probability that the process will be found in one of the finite states is zero.) These derivations are beyond the scope of this text. Nevertheless, these properties are easy

to comprehend if we understand the real significance of the value of the traffic intensity.

Recalling the result derived in (3.3.13), the n-step transition probabilities $P_{ij}^{(n)}$ $(i, j = 0, 1, 2, \ldots)$ of the Markov chain $\{Q_n\}$ are obtained as elements of the nth power of the (one-step) transition probability matrix $\mathbf{P}$. In considering $\mathbf{P}^n$ in real systems the following three observations will be useful.

1. The result (3.3.13) holds regardless of the structure of the matrix.
2. As n in $\mathbf{P}^n$ increases, the nonzero elements cluster within submatrices representing recurrent equivalence classes.
3. In an aperiodic irreducible positive recurrent Markov chain, as n in $\mathbf{P}^n$ increases the elements in each column tend toward an intermediate value.

The probability $P_{ij}^{(n)}$ for $j = 0, 1, 2 \ldots$ and finite n gives the time-dependent behavior of the queue length process $\{Q_n\}$. There are analytical techniques for deriving these probabilities. However, they are beyond the scope of this text. For example see Takács (1962), who uses PGFs to simplify recursive relations generated by the Chapman–Kolmogorov relations for $P_{ij}^{(n)}$. Prabhu and Bhat (1963) look at the transitions of Q_n as some first passage problems and use combinatorial methods in solving them (also see Prabhu (1965a)). In practice, however, with the increasing computer power for matrix operations, simple multiplications of $\mathbf{P}$ to get its nth power seem to be the best course of action. When the state space is not finite, the observations given above can be used to limit it without losing significant amount of information.

Limiting Distribution

The third observation given above, stems from the property of aperiodic positive recurrent irreducible Markov chains which results in $\lim_{n\to\infty} \mathbf{P}^n$ becoming a matrix with identical rows. Computationally, this property can be validated by getting successive powers of $\mathbf{P}^n$; as n increases the elements in the columns of the matrix tend to a constant intermediate value. This behavior of the Markov chain is codified in the following theorem and the corollary, given without proof.

Theorem 5.2.1 *(1) Let i be a state belonging to an aperiodic recurrent equivalence class. Let $P_{ii}^{(n)}$ be the probability of the n-step transition $i \to i$, and μ_i be its mean recurrence time. Then $\lim_{n\to\infty} P_{ii}^{(n)}$ exists and is given by*

$$\lim_{n\to\infty} P_{ii}^{(n)} = \frac{1}{\mu_i} = \pi_i, \text{ say.}$$

(2) Let j be another state belonging to the same equivalence class and $P_{ji}^{(n)}$ be the probability of the n-step transition $j \to i$. Then

$$\lim_{n\to\infty} P_{ji}^{(n)} = \lim_{n\to\infty} P_{ii}^{(n)} = \pi_i.$$

Corollary 5.2.1 *If i is positive recurrent, $\pi_i > 0$ and if i is null recurrent, $\pi_i = 0$.*

See Karlin and Taylor (1975) for a proof of this theorem.

Note that the term *recurrence time* in the theorem signifies the number of steps the Markov chain takes to return to the starting state for the first time. See Appendix B for other definitions.

Theorem 5.2.1 applies to Markov chains whether their state space is finite or countably infinite.

For a state space $S : \{0, 1, 2, \ldots\}$ let $(\pi_0, \pi_1, \pi_2, \ldots)$ be the limiting probability vector where $\pi_i = \lim_{n\to\infty} P_{ji}^{(n)}$, $i, j \epsilon S$. Let $\mathbf{\Pi}$ be the matrix with identical rows $\boldsymbol{\pi} = (\pi_0, \pi_1, \pi_2, \ldots)$. Now, using Chapman–Kolmogorov relations we may write

$$\mathbf{P}^{(n)} = \mathbf{P}^{n-1}\mathbf{P}$$

(see discussion leading up to (3.3.13)).

Applying Theorem 5.2.1 to $\mathbf{P}^{(n)}$ and $\mathbf{P}^{(n-1)}$, it is easy to write

$$\mathbf{\Pi} = \mathbf{\Pi}\,\mathbf{P}$$

or

$$\boldsymbol{\pi} = \boldsymbol{\pi}\mathbf{P}. \tag{5.2.10}$$

Furthermore, multiplying both sides of (5.2.10) repeatedly by $\mathbf{P}$, we can also establish that

$$\begin{aligned} \boldsymbol{\pi}\mathbf{P} &= \boldsymbol{\pi} = \boldsymbol{\pi}\mathbf{P}^2 \\ \boldsymbol{\pi}\mathbf{P} &= \boldsymbol{\pi} = \boldsymbol{\pi}\mathbf{P}^n. \end{aligned} \tag{5.2.11}$$

The last equation shows that if we use the limiting distribution as the initial distribution of the state of an irreducible, aperiodic, and positive recurrent Markov chain, the state distribution after n transitions ($n = 1, 2, 3 \ldots$) is also given by the same limiting distribution. Such a property is known as the *stationarity* of the distribution. The following theorem summarizes these results and provides a procedure by which the limiting distribution can be determined.

Theorem 5.2.2 *(1) In an irreducible, aperiodic, and positive recurrent Markov chain, the limiting probabilities $\pi_i (i = 0, 1, 2, \ldots)$ satisfy the equations*

$$\begin{aligned} \pi_j &= \sum_{i=0}^{\infty} \pi_i P_{ij} \quad j = 0, 1, 2, \ldots \\ \sum_{j=0}^{\infty} \pi_j &= 1. \end{aligned} \tag{5.2.12}$$

The limiting distribution is stationary.

(2) Any solution of the equations

$$\sum_{i=0}^{\infty} x_i P_{ij} = x_j \quad j = 0, 1, 2, \ldots \tag{5.2.13}$$

is a scalar multiple of $\{\pi_i,\ i = 0, 1, 2, \ldots\}$ *provided* $\sum |x_i| < \infty$.

Thus, the limiting distribution of the Markov chain can be obtained by solving the set of simultaneous equations (5.2.12) and normalizing the solution using the second equation $\sum_0^\infty \pi_j = 1$. Note that because the row sums of the Markov chain are equal to 1, (5.2.12) by itself yields a solution only up to a multiplicative constant. The normalizing condition is, therefore, essential in the determination of the limiting distribution.

With this background on the general theory of Markov chains, we are now in a position to determine the limiting distribution of the imbedded Markov chain of the $M/G/1$ queue.

Let $\boldsymbol{\pi} = (\pi_0, \pi_1, \pi_2, \ldots)$ be the limiting distribution of the imbedded chain. Using the transition probability matrix (5.2.5) in the equation $\boldsymbol{\pi}\mathbf{P} = \boldsymbol{\pi}$ (which is (5.2.12)), we have

$$\begin{aligned} k_0\pi_0 + k_0\pi_1 &= \pi_0 \\ k_1\pi_0 + k_1\pi_1 + k_0\pi_2 &= \pi_1 \\ k_2\pi_0 + k_2\pi_1 + k_1\pi_2 + k_0\pi_3 &= \pi_2 \\ &\ddots \end{aligned} \tag{5.2.14}$$

A convenient way of solving these equations computationally is to define

$$\nu_0 \equiv 1 \text{ and } \nu_i = \pi_i \Big/ \pi_0$$

and rewrite (5.2.14) in terms of ν_i $(i = 1, 2, \ldots)$ as

$$\begin{aligned} \nu_1 &= \frac{1 - k_0}{k_0} \\ \nu_2 &= \frac{1 - k_1}{k_0}\nu_1 - \frac{k_1}{k_0} \\ &\vdots \\ \nu_j &= \frac{1 - k_1}{k_0}\nu_{j-1} - \frac{k_2}{k_0}\nu_{j-2} - \ldots \frac{k_{j-1}}{k_0}\nu_1 - \frac{k_{j-1}}{k_0} \\ &\vdots \end{aligned} \tag{5.2.15}$$

These equations can be solved recursively to determine ν_i $(i = 1, 2, \ldots)$. The limiting probabilities $(\pi_0, \pi_1, \pi_2, \ldots)$ are known to be monotonic and concave, and therefore, for larger values of n they become extremely small. Clearly

$\nu_i = \pi_i/\pi_0$ will also have the same properties and for computational purposes it is easy to establish a cutoff value for the size of the state space.

In order to recover π_i's from ν_i's, we note that

$$\sum_{i=0}^{\infty} \nu_i = 1 + \sum_{i=1}^{\infty} \frac{\pi_i}{\pi_0} = \frac{\sum_{i=0}^{\infty} \pi_i}{\pi_0} = \frac{1}{\pi_0}.$$

Here we have incoporated the normalizing condition $\sum_0^\infty \pi_i = 1$. Thus, we get

$$\pi_0 = \left(1 + \sum_{i=1}^{\infty} \nu_i\right)^{-1}$$

and

$$\pi_i = \frac{\nu_i}{1 + \sum_{i=1}^{\infty} \nu_i}. \tag{5.2.16}$$

Analytically, the limiting distribution $(\pi_0, \pi_1, \pi_2, \ldots)$ can be determined by solving equations (5.2.14) using generating functions. Unfortunately, deriving the explicit expressions for the probabilities require inverting the resulting PGF. However, we can obtain the mean and variance of the distribution using standard techniques. We give below the procedure for the determination of the mean and variance of $lim_{n\to\infty} Q_n$ using the PGF of its distribution. If one is interested only in the results they are given in (5.2.27)–(5.2.29). Define

$$\Pi(z) = \sum_{j=0}^{\infty} \pi_j z^j \quad |z| \leq 1$$

and

$$K(z) = \sum_{j=0}^{\infty} k_j z^j \quad |z| \leq 1.$$

Multiplying equations (5.2.14) with appropriate powers of z and summing, we get

$$\begin{aligned} \Pi(z) &= \pi_0 K(z) + \pi_1 K(z) + \pi_2 z K(z) + \ldots \\ &= \pi_0 K(z) + \frac{K(z)}{z}(\pi_1 z + \pi_2 z^2 + \ldots) \\ &= \pi_0 K(z) + \frac{K(z)}{z}[\Pi(z) - \pi_0]. \end{aligned}$$

Rearranging terms,

$$\begin{aligned} \Pi(z)[1 - \frac{K(z)}{z}] &= \pi_0 K(z)[1 - \frac{1}{z}] \\ \Pi(z) &= \frac{\pi_0 K(z)(z-1)}{z - K(z)}. \end{aligned} \tag{5.2.17}$$

The unknown quantity π_0 on the right-hand side expression for $\Pi(z)$ in (5.2.17) can be determined using the normalizing condition $\sum_{j=0}^{\infty} \pi_j = 0$. We must have

$$\Pi(1) = \sum_{j=0}^{\infty} \pi_j = 1.$$

Letting $z \to 1$ in (5.2.17), we get (applying l'Hôpital's rule)

$$1 = \frac{\lim_{z\to 1} \pi_0[K(z) - (z-1)K'(z)]}{\lim_{z\to 1}[1 - K'(z)]}.$$

Recalling that $K(1) = 1$ and $K'(1) = \rho$, (from (5.2.9)) we have

$$\begin{aligned} 1 &= \frac{\pi_0}{1-\rho} \\ \pi_0 &= 1 - \rho. \end{aligned} \tag{5.2.18}$$

Thus, we get

$$\Pi(z) = \frac{(1-\rho)(z-1)K(z)}{z - K(z)}. \tag{5.2.19}$$

Explicit expressions for probabilities $\{\pi_j,\ j = 0, 1, 2, \ldots\}$ can be obtained by expanding $\Pi(z)$ in special cases. An alternative form of $\Pi(z)$ works out to be easier for this expansion. We may write

$$\begin{aligned} \Pi(z) &= \frac{(1-\rho)K(z)}{(z - K(z))/(z-1)} \\ &= \frac{(1-\rho)K(z)}{1 - [1 - K(z)]/(1-z)}. \end{aligned} \tag{5.2.20}$$

Note that $\sum_{j=0}^{\infty} z^j(k_{j+1} + k_{j+2} + \ldots)$ can be simplified to write as

$$\frac{1 - K(z)}{1 - z} = C(z), \quad \text{say.}$$

(Also see algebraic simplifications leading to (5.2.17)).

For $|z| \le 1$

$$|C(z)| = \left|\frac{1 - K(z)}{1 - z}\right| < 1 \quad \text{if } \rho < 1. \tag{5.2.21}$$

Now using a geometric series expansion, we may write

$$\Pi(z) = (1-\rho)K(z)\sum_{j=0}^{\infty}[C(z)]^j. \tag{5.2.22}$$

The explicit expression for π_j is obtained by expanding the right-hand side of (5.2.22) as a power series in z and picking the coefficient of z^j in it.

In a queueing system, the queue length process $Q(t)$ may be considered with three different time points: (1) when t is just before an arrival epoch, (2) when t is soon after a departure epoch, and (3) when t is an arbitrary point in time. In general, the distribution of $Q(t)$ with reference to these three time points may not be the same. However, when the arrival process is Poisson, it can be shown that the limiting distributions of $Q(t)$ in all three cases are the same. The property of the Poisson process that makes this happen is its relationship with the uniform distribution mentioned in Appendix B. See Wolff (1982) who coined the acronym PASTA (Poisson Arrivals See Time Averages). For proofs of this property also see Cooper (1981) and Gross et al. (2008).

The PGF $\Pi(z)$ derived in (5.2.19), therefore, also gives the limiting distribution $\lim_{t\to\infty} Q(t)$. There are several papers in the literature deriving the transition distribution of $Q(t)$ for finite t. Among them are Prabhu and Bhat (1963b) and Bhat (1968) who obtain the transition distribution using recursive methods and renewal theory arguments. The explicit expression for the limiting distribution of $Q(t)$ (and the limiting distribution of Q_n in the imbedded chain case) derived in these papers is given by

$$\begin{aligned} \pi_0 &= 1-\rho \\ \pi_j &= (1-\rho)\int_0^\infty e^{-\lambda t}\sum_{n=0}^{\infty}\left[\frac{(\lambda t)^{n+j-1}}{(n+j-1)!}-\frac{(\lambda t)^{n+j}}{(n+j)!}\right]dB_n(t) \end{aligned} \quad (5.2.23)$$

for $\rho<1$, where $B_n(t)$ is the n-fold convolution of $B(t)$ with itself (Prabhu and Bhat 1963a, b; Bhat 1968).

The mean and variance of $\lim_{n\to\infty} Q_n$ can be determined from the PGF (5.2.19) through standard techniques. Writing $Q^* = \lim_{n\to\infty} Q_n$, we have

$$\begin{aligned} L &= E(Q^*) = \Pi'(1) \\ V(Q^*) &= \Pi''(1)+\Pi'(1)-[\Pi'(1)]^2. \end{aligned} \quad (5.2.24)$$

Differentiating $\Pi(z)$ w.r.t. z, we get

$$\begin{aligned} \Pi'(z) = \frac{1-\rho}{[z-K(z)]^2}\{&[z-K(z)][(K(z)+(z-1)K'(z)] \\ &-(z-1)[1-K'(z)]K(z)\}. \end{aligned}$$

Using l'Hôpital's rule twice while taking limits $z\to 1$, we get

$$\Pi'(1) = \frac{2K'(1)[1-K'(1)]+K''(1)}{2[1-K'(1)]}. \quad (5.2.25)$$

But note from (5.2.9), $K'(1)=\rho$ and

$$\begin{aligned} K''(1) &= \lambda^2\psi''(0) \\ &= \lambda^2 E(S^2) \end{aligned} \quad (5.2.26)$$

where we have used a generic notation for the service time. Substituting from (5.2.26) in (5.2.25), we get, after simplifications,

$$L = E(Q^*) = \rho + \frac{\lambda^2 E(S^2)}{2(1-\rho)} \tag{5.2.27}$$

which is often referred to as *Pollaczek–Khintchine formula*.

Noting that ρ is the expected number in service (which is the same as the probability the server is busy in a single server queue), L_q, the mean number in the queue is obtained as

$$L_q = \frac{\lambda^2 E(S^2)}{2(1-\rho)}. \tag{5.2.28}$$

Extending the differentiation to get $\Pi''(z)$, and taking limits as $z \to 1$ with the multiple use of l'Hôpital's rule to get $\Pi''(1)$ we obtain

$$\begin{aligned} V(Q^*) &= \rho(1-\rho) + \frac{\lambda^2 E(S^2)}{2(1-\rho)}\left[3 - 2\rho + \frac{\lambda^2 E(S^2)}{2(1-\rho)}\right] \\ &\quad + \frac{\lambda^3 E(S^3)}{3(1-\rho)}. \end{aligned} \tag{5.2.29}$$

Recall that σ_S^2 is the variance of the service time distribution. Hence $\sigma_S^2 = E(S^2) - [E(S)]^2$. Using this expression in (5.2.27) and noting that $\lambda E(S) = \rho$, we get an alternative form for $E(Q^*)$.

$$E(Q^*) = \rho + \frac{\rho^2}{2(1-\rho)} + \frac{\lambda^2 \sigma_S^2}{2(1-\rho)} \tag{5.2.30}$$

which clearly shows that the mean queue length increases with the variance of the service time distribution. For instance, when $\sigma_S^2 = 0$, i.e., when the service time is a constant (in the queue $M/D/1$)

$$E(Q^*) = \rho + \frac{\rho^2}{2(1-\rho)} = \frac{\rho}{1-\rho}(1 - \frac{\rho}{2}). \tag{5.2.31}$$

On the other hand, when the service time distribution is Erlang with mean $\frac{1}{\mu}$ and scale parameter k (i.e., by writing $\lambda = \mu$ in (2.1.8)), we get $\sigma_S^2 = \frac{1}{k\mu^2}$ and

$$\begin{aligned} E(Q^*) &= \rho + \frac{\rho^2}{2(1-\rho)} + \frac{\rho^2}{2k(1-\rho)} \\ &= \rho + \frac{\rho^2(1+k)}{2k(1-\rho)}. \end{aligned} \tag{5.2.32}$$

When $k = 1$, we get $E(Q^*)$ in $M/M/1$ as

$$E(Q^*) = \frac{\rho}{1-\rho}.$$

Waiting Time

The concept of waiting time has been used earlier in the context of the $M/M/1$ queue. Since we had the distribution of the queue length explicitly, we were then able to determine the distribution of the waiting time. But in the $M/G/1$ case the explicit expression for the limiting distribution of the queue length, viz., equation (5.2.23), is not easy to handle, even for computations. Consequently, we approach this problem indirectly using the PGF $\Pi(z)$ of the queue length.

Assume that the queue discipline is first-come, first-served (FCFS). Let T be the total time spent by the customer in the system in waiting and service which we may call system time or time in system, and T_q the actual waiting time, both as $t \to \infty$. Let $E(T) = W$ and $E(T_q) = W_q$. The determination of W and W_q requires the use of PGFs and Laplace–Steiltjes transforms (LSTs) which may be skipped in first reading, moving directly to the results (5.2.42) and (5.2.43).

Let $F(\cdot)$ be the distribution function of T with a Laplace–Stieltjes transform

$$\Phi(\theta) = \int_0^\infty e^{-\theta t} dF(t) \quad \text{Re}(\theta) > 0.$$

Consider a customer departing from the system. It has spent a total time of T, in waiting and service. Suppose the departing customer leaves n customers behind; clearly, these customers have arrived during its time in system T. Then we have

$$P(Q^* = n) = \int_0^\infty e^{-\lambda t} \frac{(\lambda t)^n}{n!} dF(t) \quad n \geq 0. \tag{5.2.33}$$

Using generating functions,

$$\begin{aligned} \Pi(z) &= \sum_{n=0}^\infty P(Q^* = n) z^n = \sum_{n=0}^\infty z^n \int_0^\infty e^{-\lambda t} \frac{(\lambda t)^n}{n!} dF(t) \\ &= \int_0^\infty e^{-\lambda t} \sum_{n=0}^\infty \frac{(\lambda t z)^n}{n!} dF(t) \\ &= \int_0^\infty e^{-(\lambda - \lambda z)t} dF(t) \\ &= \Phi(\lambda - \lambda z). \end{aligned} \tag{5.2.34}$$

Comparing (5.2.19) with (5.2.34), we have

$$\frac{(1-\rho)(z-1)K(z)}{z - K(z)} = \Phi(\lambda - \lambda z). \tag{5.2.35}$$

Recall that

$$K(z) = \psi(\lambda - \lambda z).$$

Substituting in (5.2.35),

$$\Phi(\lambda - \lambda z) = \frac{(1-\rho)(z-1)\psi(\lambda - \lambda z)}{z - \psi(\lambda - \lambda z)}.$$

Writing $\lambda - \lambda z = \theta$, we get $z = 1 - \frac{\theta}{\lambda}$; hence

$$\begin{aligned}\Phi(\theta) &= \frac{(1-\rho)\frac{\theta}{\lambda}\psi(\theta)}{\psi(\theta) - (\lambda-\theta)/\lambda} \\ &= \frac{(1-\rho)\theta\psi(\theta)}{\theta - \lambda[1-\psi(\theta)]}. \end{aligned} \tag{5.2.36}$$

Since the system time T is the sum of the actual waiting time T_q and service time S, defining the Laplace–Stieltjes transform of the distribution of T_q as $\Phi_q(\theta)$, we have

$$\Phi(\theta) = \Phi_q(\theta)\psi(\theta). \tag{5.2.37}$$

Comparing (5.2.36) and (5.2.37), we write

$$\Phi_q(\theta) = \frac{(1-\rho)\theta}{\theta - \lambda[1-\psi(\theta)]} \tag{5.2.38}$$

which can be expressed as

$$\begin{aligned}\Phi_q(\theta) &= \frac{1-\rho}{1-\frac{\lambda}{\theta}[1-\psi(\theta)]} \\ &= (1-\rho)\sum_{n=0}^{\infty}\left[\frac{\lambda}{\theta}[1-\psi(\theta)]\right]^n. \end{aligned} \tag{5.2.39}$$

In using the geometric series for (5.2.39), we can show that $|\frac{\lambda}{\theta}[1-\psi(\theta)]| < 1$ for $\rho < 1$.

In Chapter 3, we have introduced a renewal process as a sequence of independent and identically distributed random variables. Suppose $t_{n+1} - t_n = Z_n$ is the nth member of such a sequence. Let t be a time point such that $t_n < t \le t_{n+1}$. Then $t_{n+1} - t = R(t)$ is the *forward recurrence time* the density function of which was introduced in the functional form as in equation (3.4.22). If $B(\cdot)$ is the distribution function of Z_n (Note that in Chapter 3 we have used $F(\cdot)$ for this distribution and R for its mean.) (3.4.24) can be rewritten as

$$\lim_{t\to\infty} r_t(x) = \frac{1}{E[Z_n]}[1 - B(x)]. \tag{5.2.40}$$

Using this concept, we can invert (5.2.39) to give the distribution function of T_q as

$$F_q(t) = (1-\rho)\sum_{n=0}^{\infty}\rho^n R^{(n)}(t) \tag{5.2.41}$$

where $R^{(n)}(t)$ is the n-fold convolution of the distribution of the remaining service time $R(t)$ with itself.

As stated in Chapter 4, Little's law ($L = \lambda W$) applies broadly to queueing systems with only some restrictions on structure and discipline. (See Section 9.2 for details.) Hence, using the law on (5.2.27) and (5.2.28), we get

$$W = E(S) + \frac{\lambda E(S^2)}{2(1-\rho)} \tag{5.2.42}$$

$$W_q = \frac{\lambda E(S^2)}{2(1-\rho)}. \tag{5.2.43}$$

These means can also be determined from the transforms $\Phi(\theta)$ and $\Phi_q(\theta)$. For, we have

$$\begin{aligned} W &= E(T) = \Phi'(0) \\ \sigma_T^2 &= V(T) = \Phi''(0) - [\Phi'(0)]^2 \end{aligned}$$

and similar expressions for W_q and $\sigma_{T_q}^2$. The following result, derived in this manner, might be useful in some applications.

$$\sigma_{T_q}^2 = V(T_q) = \frac{\lambda E(S^3)}{3(1-\rho)} + \frac{\lambda^2 [E(S^2)]^2}{4(1-\rho)^2}. \tag{5.2.44}$$

The Busy Period

In the context of an imbedded Markov chain, the length of the busy period is measured in terms of the number of transitions of the chain without visiting the state 0. Let B_i be the number of transitions of the Markov chain before it enters state 0 for the first time, having initially started from state i.

A key property of B_i is that it can be thought of as the sum of i random variables each with the distribution of B_1. This is equivalent to saying that the transition $i \to 0$ can be considered to be occurring in i segments, $i \to i-1$, $i-1 \to i-2, \ldots, 1 \to 0$. This is justified by the fact that the downward transition can occur only one step at a time. Since all these transitions are structurally similar to each other we can consider B_i as the sum of i random variables each with the distribution of B_1. To derive $E(B_i)$ we first use an indirect method involving the PGF of B_1. The reader may go directly to the result (5.2.50) in first reading.

Let

$$g_i^{(n)} = P[B_i = n] \quad n = 1, 2, \ldots. \tag{5.2.45}$$

As a consequence of the property stated above $g_i^{(n)}$ is the i-fold convolution of $g_1^{(n)}$ with itself. Thus, for the PGF of $g_i^{(n)}$

$$G_i(z) = \sum_{n=i}^{\infty} g_i^{(n)} z^n = [G_1(z)]^i. \tag{5.2.46}$$

Noting that the busy period cannot end before the ith transition of the Markov chain, we have

$$\begin{aligned} g_i^{(i)} &= k_0^{(i)} \\ g_i^{(n)} &= \sum_{r=1}^{n-1} k_r^{(i)} g_r^{(n-i)} \end{aligned} \tag{5.2.47}$$

where $k_r^{(i)}$ is the i-fold convolution of the probability k_r that r customers arrive during a service period (see (5.2.1)).

For $i = 1$

$$g_1^{(1)} = k_0$$

and

$$g_1^{(n)} = \sum_{r=1}^{n-1} k_r g_r^{(n-1)} \quad n \geq 2.$$

Multiplying by appropriate powers of z to both sides of these equations, we get

$$\begin{aligned} g_1^{(1)} z &= k_0 z \\ \sum_{n=2}^{\infty} g_1^{(n)} z^n &= \sum_{n=2}^{\infty} z^n \sum_{r=1}^{n-1} g_r^{(n-1)}. \end{aligned}$$

Hence

$$\begin{aligned} G_1(z) &= z k_0 + z \sum_{r=1}^{\infty} k_r \sum_{n=r+1}^{\infty} z^{n-1} g_r^{(n-1)} \\ &= z \left[k_0 + \sum_{r=1}^{\infty} k_r [G_1(z)]^r \right] \\ &= z K[G_1(z)]. \end{aligned} \tag{5.2.48}$$

From the definition of $K(z)$ earlier, we have

$$K(z) = \psi(\lambda - \lambda z)$$

where $\psi(\theta)$ is the Laplace–Stieltjes transform of the service time distribution. Thus, the PGF $G_1(z) \equiv G(z)$ is such that it satisfies the functional equation

$$\omega = z\psi(\lambda - \lambda\omega). \tag{5.2.49}$$

It is possible to show that $G(z)$ is the least positive root (≤ 1) of the functional equation (5.2.49) when $\rho < 1$ and determine explicit expressions for specific distributions. (For other ways of deriving the busy period distribution in explicit forms see Prabhu and Bhat (1963a).)

We can easily obtain the mean length of the busy period by implicit differentiation of (5.2.49). We have

$$G(z) = z\psi(\lambda - \lambda G(z)).$$

On differentiation,

$$G'(z) = \psi[\lambda - \lambda G(z)] + z\psi'[\lambda - \lambda G(z)][-\lambda G'(z)].$$

As $z \to 1$,

$$G'(1) = \psi(0) + \psi'(0)[-\lambda G'(1)].$$

Rearranging terms

$$\begin{aligned} G'(1)[1 + \lambda\psi'(0)] &= 1 \\ G'(1) &= \frac{1}{1 + \lambda\psi'(0)}. \end{aligned}$$

Referring back to the definitions given earlier

$$E[B_1] = G'(1) = \frac{1}{1 - \rho}.$$

Following the arguments leading to (5.2.46), for the busy period B_i initiated by i customers, we get

$$E(B_i) = \frac{i}{1 - \rho}. \tag{5.2.50}$$

Since we are counting the number of transitions, to get the exact mean length of a busy period we multiply it by the mean length of time taken for each transition, viz., the service period. Hence

$$\text{Mean length of the busy period } = \frac{E(S)}{(1 - \rho)}. \tag{5.2.51}$$

Noting that a busy cycle is made up of a busy period and an idle period and that the mean length of the idle period in $M/G/1$ with arrival rate λ is $1/\lambda$, we get

$$\text{Mean length of the busy cycle } = \frac{E(S)}{1 - \rho} + \frac{1}{\lambda} = \frac{1}{\lambda(1 - \rho)}. \tag{5.2.52}$$

The Queue $M/G/1/K$

Consider the $M/G/1$ queue described earlier, with the restriction that the capacity for the number of customers in the system is K. Since the state space for the imbedded Markov chain is the number in the system soon after departure,

K will not be included in the state space. $S = \{0, 1, 2, \ldots, K-1\}$. Thus, corresponding to (5.2.2) we have the relation

$$Q_{n+1} = \begin{cases} \min\,(Q_n + X_{n+1} - 1, K-1) & \text{if } Q_n > 0 \\ \min\,(X_{n+1}, K-1) & \text{if } Q_n = 0 \end{cases}. \tag{5.2.53}$$

Using the probability distribution $\{k_j, j = 0, 1, 2, \ldots\}$ defined in (5.2.1), we get the transition probability matrix

$$\mathbf{P} = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ \vdots \\ K-1 \end{array} \begin{array}{c} \begin{array}{ccccc} 0 & 1 & 2 & \ldots & K-1 \end{array} \\ \begin{bmatrix} k_0 & k_1 & k_2 & \ldots & 1 - \sum_0^{k-2} k_j \\ k_0 & k_1 & k_2 & \ldots & 1 - \sum_0^{k-2} k_j \\ & k_0 & k_1 & \ldots & 1 - \sum_0^{k-3} k_j \\ & & & & \vdots \\ & & & & 1 - k_0 \end{bmatrix} \end{array}. \tag{5.2.54}$$

Let $\boldsymbol{\pi} = (\pi_0, \pi_1, \ldots, \pi_{k-1})$ be the limiting distribution for the state of the Markov chain. These probabilities are determined by solving the equations

$$\begin{aligned} \pi_j &= \sum_i \pi_i P_{ij} \quad j = 0, 1, 2, \ldots, K-1 \\ \sum_0^{K-1} \pi_j &= 1. \end{aligned} \tag{5.2.55}$$

The first $K-1$ equations are identical to those for $M/G/1$ with no capacity restriction. Therefore, we can use the computational method outlined in (5.2.15) for the solution of (5.2.55). We may note here that one of the k simultaneous equations in (5.2.55) is redundant because of the Markov chain structure of the coefficients. In its place, we use the normalizing condition $\sum_0^{K-1} \pi_j = 1$ for the solution. We may also note from the computational solution technique that the finite case solution is obtained by using the same ν_i's as in the infinite case for $i = 0, 1, 2, \ldots, K-1$ and determining

$$\begin{aligned} \pi_0 &= \left[\sum_0^{K-1} \nu_i\right]^{-1} \\ \pi_i &= \pi_0 \nu_i. \end{aligned} \tag{5.2.56}$$

The discussion of the waiting time distribution is a bit complicated in $M/G/1/K$, since for the Markov chain the state space is only $\{0, 1, 2, \ldots, K-1\}$ while our arrival may find K customers in the system (before a departure). Thus, from the viewpoint of an arrival, we need the limiting distribution at an arbitrary point in time. For further details, the readers are referred to more advanced texts on the subject.

The busy period analysis given earlier for the queue $M/G/1$ cannot be easily modified for the finite capacity case. Computationally, the best approach seems

to be to consider the busy period as a first passage problem in the irreducible Markov chain from state 1 to state 0. This can be done by converting state 0 into an absorbing state and using the concept of the fundamental matrix in the determination of the expected number of transitions required for the first passage transition. For details, the readers are referred to Bhat and Miller (2002), Chapter 2. A further discussion of this method is also given in Section 7.2 of Chapter 7.

Example 5.2.1 Consider a computer network node in which requests for data arrive in a Poisson process at the rate of 0.5 per unit time. Assume that the data retrieval (service) takes a constant amount of one unit of time.

We can model this system as an $M/D/1$ queue and use the techniques developed in this section for its analysis. We have

$$k_j = e^{-0.5}\frac{(0.5)^j}{j!} \quad j = 0, 1, 2, \ldots$$

which on evaluation gives

$$k_0 = 0.607;\ k_1 = 0.303;\ k_2 = 0.076;\ k_3 = 0.012;\ k_4 = 0.002.$$

The Laplace–Stieltjes transform of the service time distribution is

$$\psi(\theta) = e^{-(0.5)\theta}$$

and the PGF of k_j, $j = 0, 1, 2 \ldots$ is

$$K(z) = e^{-0.5(1-z)}.$$

These give, the PGF $\Pi(z)$ of the limiting distribution as

$$\Pi(z) = \frac{(1-0.5)(z-1)e^{-0.5(1-z)}}{z - e^{-0.5(1-z)}}.$$

Now $\Pi(z)$ can be inverted to determine the distribution explicitly. However, computationally it is easier to use the method described in (5.2.56). We have

$$\begin{aligned}
\nu_1 &= \frac{1-k_0}{k_0} = 0.647; \nu_2 = \frac{1-k_1}{k_0}\nu_1 - \frac{k_1}{k_0} = 0.244; \nu_3 = 0.074; \\
\nu_4 &= 0.022; \nu_5 = 0.006; \nu_6 = 0.002; \nu_7 = 0.001; \\
\sum_{i=0}^{7} \nu_i &= 1.996; \\
\pi_0 &= (\sum_{0}^{7} \nu_i)^{-1} = 0.501; \\
\pi_i &= \nu_i \pi_0 \quad i = 1, 2, \ldots, 7.
\end{aligned}$$

Thus, we get

$$\begin{aligned} \pi_0 &= 0.501; \quad \pi_1 = 0.324; \quad \pi_2 = 0.122; \quad \pi_3 = 0.037; \\ \pi_4 &= 0.011; \quad \pi_5 = 0.003; \quad \pi_6 = 0.001; \quad \pi_7 = 0.000. \end{aligned}$$

The mean number of customers in the system as $t \to \infty$ can be determined either from the formula (5.2.31) or from the distribution $\boldsymbol{\pi}$ determined above. We get

$$L = E(Q) = 0.75.$$

Using Little's law, $L = \lambda W$, the mean system time can be obtained as

$$W = 0.75/0.5 = 1.5 \text{ time units.}$$

Also for the mean length of a busy period B, we have

$$E(B_1) = \frac{1}{1 - 0.5} = 2 \text{ time units.} \qquad \textbf{Answer}$$

Example 5.2.2 In an automobile garage with a single mechanic, from the records kept by the owner, the distribution of the number of vehicles arriving during the service time of a vehicle is obtained as follows:

$$\begin{aligned} P \text{ (no new arrivals)} &= 0.5 \\ P \text{ (one new arrival)} &= 0.3 \\ P \text{ (two new arrivals)} &= 0.2. \end{aligned}$$

If we assume that the arrival of vehicles for service follow a Poisson distribution, we can model this system as an $M/G/1$ queue, even when we do not have a distribution form for the service times. With this assumption, we get

$$k_0 = 0.5; \quad k_1 = 0.3; \quad k_2 = 0.2$$

with E (# of arrivals during one service period) $= 0.7 =$ traffic intensity ρ. The computational method for the determination of the limiting distribution is the most appropriate, since no distribution form is available for the service time. Using equations (5.2.15), we get

$$\begin{aligned} &\nu_0 = 1; \quad \nu_1 = 1; \quad \nu_2 = 0.8; \quad \nu_3 = 0.32; \\ &\nu_4 = 0.128; \quad \nu_5 = 0.051; \quad \nu_6 = 0.021 \quad \nu_7 = 0.008; \\ &\nu_8 = 0.003; \quad \nu_9 = 0.001; \quad \nu_{10} = 0.001. \end{aligned}$$

Hence $\sum_{i=0}^{10} \nu_i = 3.333$. Since $\pi_0 = \left(\sum_{i=0}^{10} \nu_i\right)^{-1}$ and $\pi_i = \nu_i \pi_0$, we get

$$\begin{aligned} &\pi_0 = 0.300; \quad \pi_1 = 0.300; \quad \pi_2 = 0.240; \quad \pi_3 = 0.096; \\ &\pi_4 = 0.038; \quad \pi_5 = 0.015; \quad \pi_6 = 0.006; \quad \pi_7 = 0.002; \\ &\pi_8 = 0.001; \quad \pi_9 = 0.000. \end{aligned}$$

The mean of this distribution is obtained as

$$L = E(Q^*) = 1.353.$$

Using Little's law, for the mean system time we get

$$W = 1.353/0.7 = 1.933 \text{ service time units.}$$

Note that we use the mean service time as the unit of time for the purpose of determining the mean waiting time. **Answer**

Example 5.2.3 The following example (Coffman and Kleinrock (1968)) illustrates the procedure for the determination of the response time which is the mean total time spent in a queueing system by a customer in receiving service with a round robin (RR) service discipline. In RR service discipline, service is provided for a fixed amount of time, called quantum (Q), with every visit to the server, and if the service time of a job is longer than Q, the customer is sent back to the end of the queue to wait for its turn again. Clearly, such a discipline favors customers with short service times.

The readers should note that the derivations in this example are rather advanced in nature.

Consider a single server queueing system with arrivals in a Poisson process with parameter λ. Let the service times have a geometric distribution

$$g_i = (1-\sigma)\sigma^{i-1} \quad i = 1, 2, \ldots \quad 0 < \sigma < 1 \tag{5.2.57}$$

where g_i is the probability that the service time consists of i quanta, each of length Q.

We are interested in the determination of the conditional mean response time (W_k) of a customer requiring k quanta of service. The response time is made up of three components: the service time of the customer in service at the time of the arrival, total service time of the customers waiting in queue at the time of the arrival, and the service time of the arriving customer. The mean number of customers waiting or in service at the time of the arrival is given by (5.2.27) as

$$L = \rho + \frac{\lambda^2 E(S^2)}{2(1-\rho)} \tag{5.2.58}$$

where $\rho = \lambda E(S)$ is the traffic intensity with S denoting service time and $E(S^2)$, the second moment of the service time distribution. From the distribution (5.2.57), we have

$$\begin{aligned} E(S) &= \left(\sum_{i=1}^{\infty} i g_i)\right) Q \\ &= \frac{Q}{1-\sigma}. \end{aligned} \tag{5.2.59}$$

$$\begin{aligned} E(S^2) &= Q^2 \sum_{i=1}^{\infty} i^2 g_i \\ &= Q^2 \sum_{i=1}^{\infty} [i(i-1)+i] g_i \\ &= (1-\sigma) Q^2 \sum_{i=1}^{\infty} [i(i-1)+i] \sigma^{i-1} \\ &= (1-\sigma) Q^2 [2\sigma(1-\sigma)^{-3} + (1-\sigma)^{-2}] \\ &= \frac{1+\sigma}{(1-\sigma)^2} Q^2. \end{aligned} \tag{5.2.60}$$

Substituting from (5.2.59) and (5.2.60) in (5.2.58), we get

$$L = \rho + \frac{\rho^2(1+\sigma)}{2(1-\rho)} \tag{5.2.61}$$

where $\rho = \lambda Q/(1-\sigma)$.

Let $W_k(j)$ be the conditional expectation of the time spent in the system by an arriving customer requiring k quanta of service when there are j customers in the system. Then for the conditional mean response time, we have

$$W_k = \sum_{j=0}^{\infty} p_j W_k(j) \tag{5.2.62}$$

where $\{p_j,\ j = 0, 1, 2 \ldots\}$ is the limiting distribution of the number of customers in the system at an arrival epoch. Note that, because of the PASTA property discussed earlier, the distribution $\{p_j\}_{j=0}^{\infty}$ is identical to the limiting distribution $\{\pi_j\}_{j=0}^{\infty}$ of customers in the system soon after a departure epoch and its mean value is given by (5.2.61).

Because of the RR nature of service, we have to breakdown $W_k(j)$ in terms of the number of times the customer passes through service, which is k. Let $U_i(j)$, $i = 1, 2, \ldots, k$ be the random variable denoting the time required for the ith pass, assuming that the customer arrives when there are j customers in the system. For the sake of simplicity we shall suppress the argument j until, its inclusion is necessary.

Suppose $U_i = x$, $i \geq 2$. Then U_{i+1} is made up three components: (i) The amount of time required to serve those who are ahead of the customer at the ith pass. This is $\sigma\left[\left(\frac{x}{Q}\right) - 1\right] Q$, where σ is the probability of a customer returning for another quantum of service, and x/Q is the number of quanta of service ahead of the customer at the ith pass, (ii) the amount of time needed to provide one quantum of service for those who arrive during U_i, and (iii) the customer's quantum of service. Thus, we get

$$E[U_{i+1} \Big| U_i = x] = \sigma(\frac{x}{Q} - 1)Q + \lambda x Q + Q \tag{5.2.63}$$

giving

$$E[U_{i+1}] = (\lambda Q + \sigma)E(U_i) + Q(1-\sigma) \quad i = 2, 3, \ldots, k. \tag{5.2.64}$$

By successive iteration, from (5.2.64) we get

$$E[U_i] = \alpha^{i-2}E(U_2) + Q(1-\sigma)\frac{1-\alpha^{i-2}}{1-\alpha} \quad i = 2, 3, \ldots, k \tag{5.2.65}$$

where we have written $\lambda Q + \sigma = \alpha$. Also, in the first pass for $U_1(j)$, we have

$$\begin{aligned} U_1(j) &= \text{(time to complete the service in progress)} \\ &\quad +\text{(total time to serve } j-1 \text{ customers)} \\ &\quad +\text{(one quantum of service for the arriving customer)} \\ &= \rho(\frac{Q}{2}) + (j-\rho)Q + Q \\ &= (1-\frac{\rho}{2})Q + jQ. \end{aligned} \tag{5.2.66}$$

The term $\rho(\frac{Q}{2})$ leading to (5.2.66) represents the mean of a uniform distribution in $(0, Q)$ with ρ as the probability of finding a customer in service. Using similar arguments

$$E(U_2(j)) = \lambda Q E(U_1) + \sigma(jQ) + Q. \tag{5.2.67}$$

Now

$$W_k(j) = \sum_{i=1}^{k} E(U_i(j)). \tag{5.2.68}$$

Substituting from (5.2.65) and (5.2.67), we get

$$\begin{aligned} W_k(j) &= E(U_1) \\ &\quad + \sum_{i=2}^{k}\left[\alpha^{i-2}(\lambda Q E(U_1) + Q(\sigma j + 1)) + Q(1-\sigma)\frac{1-\alpha^{i-2}}{1-\alpha}\right] \\ &= E(U_1) + [\lambda Q E(U_1) + Q(\sigma j + 1)]\frac{1-\alpha^{k-1}}{1-\alpha} \\ &\quad + \frac{Q(1-\sigma)}{1-\alpha}[(k-1) - \frac{1-\alpha^{k-1}}{1-\alpha}] \\ &= E(U_1) + \frac{(k-1)Q}{1-\rho} + Q[\lambda E(U_1) + \sigma j - \frac{\rho}{1-\rho}]\frac{1-\alpha^{k-1}}{1-\alpha}. \end{aligned} \tag{5.2.69}$$

In deriving (5.2.69), we have used the following simplifications:

$$\alpha = \lambda Q + \sigma; \quad \rho = \frac{\lambda Q}{1-\sigma}; \quad 1-\rho = \frac{1-\sigma-\lambda Q}{1-\sigma}.$$

Taking expectations as in (5.2.62), for the conditional mean response time we get

$$W_k = W_1 + \frac{(k-1)Q}{1-\rho} + Q[\lambda W_1 + \sigma L - \frac{\rho}{1-\rho}]\frac{1-\alpha^{k-1}}{1-\alpha} \quad k \geq 1 \qquad (5.2.70)$$

with $W_1 = (1-\rho/2)Q + LQ$, and L given by (5.2.61).

When the service time distribution is exponential, $\mu e^{-\mu x}(x > 0)$, we get $\sigma = e^{-\lambda Q}$ and

$$g_i = (1 - e^{-\lambda Q})e^{-(i-1)\lambda Q} \quad i = 1, 2, \ldots. \qquad (5.2.71)$$

For the ramifications of making Q very small, and other variations readers may refer to Coffman and Denning (1973).

Example 5.2.4 The storage in a warehouse is such that the most recent item stored gets to be taken out first. This is the example of a last-come, first-served (LCFS) service discipline if the process of replenishment of the item and its disposal is looked upon as a queueing process. Let the replenishment process be Poisson with parameter λ and let $B(\cdot)$ be the distribution function of the inter-arrival times of demands.

We want to determine the average time an item stays in the warehouse before it is disposed of.

Considering the inter-demand times as the service times of the queueing system, we have here an $M/G/1$ queue with LCFS service discipline.

Since, the customer arriving last gets served first in an LCFS queueing system, the amount of time the customer spends while waiting is the sum of the remainder of the service that is in progress and the length of the busy period initiated by the number of customers who arrive during that period.

When the service time distribution has a general form $B(\cdot)$, as $t \to \infty$ the remainder of the service time at the time of the customer arrival can be considered to be the forward recurrence time of a renewal process, the density function of which was briefly introduced in (5.2.40). We have

$$r(x) = \lim_{t\to\infty} r_t(x) = \frac{1}{E(S)}[1 - B(x)] \qquad (5.2.72)$$

where we have used S to denote the service time random variable. The mean of this distribution can be obtained as follows:

$$\begin{aligned}\int_0^\infty xr(x)dx &= \frac{1}{E(S)}\int_0^\infty x[1-B(x)]dx \\ &= \frac{1}{E(S)}\int_{x=0}^\infty x\left[\int_{y=x}^\infty dB(y)\right]dx \\ &= \frac{1}{E(S)}\int_{y=0}^\infty \left[\int_{x=0}^y xdx\right]dB(y) \\ &= \frac{1}{E(S)}\int_{y=0}^\infty \frac{y^2}{2}dB(y) \\ &= \frac{E(S^2)}{2E(S)}. \end{aligned} \tag{5.2.73}$$

With the Poisson arrival rate λ, the expected number of customers arriving during the remainder of the service time can be given as

$$\frac{\lambda E(S^2)}{2E(S)}. \tag{5.2.74}$$

As described earlier, the customer's mean waiting time W_q is the sum of the mean length of the remainder of the service time and the mean length of the busy period initiated by the number of customers who arrive during that period. Using (5.2.51), we get

$$\begin{aligned} W_q &= \frac{\lambda E(S^2)}{2E(S)} \times \frac{1}{1-\rho}\cdot E(S) \\ &= \frac{\lambda E(S^2)}{2(1-\rho)}. \end{aligned} \tag{5.2.75}$$

The average time the item stays in the warehouse is the time spent in the system $W = W_q + E(S)$. We have

$$W = E(S) + \frac{\lambda E(S^2)}{2(1-\rho)}. \tag{5.2.76}$$

Comparing these results with (5.2.42) and (5.2.43), we note that in this example we have shown that the mean waiting time (and also the mean queue length) of the customer in the system is the same whether the queue discipline is FCFS or LCFS.

5.3 The Queue $G/M/1$

Let customers arrive at time points $t_0 = 0,\ t_1,\ t_2, \ldots$ and get served by a single server. Let $Z_n = t_{n+1} - t_n$, $n = 1, 2, 3, \ldots$, be independent and identically distributed random variables with distribution function $A(\cdot)$ with mean a. Also let

the service time distribution be exponential with mean $1/\mu$. It should be noted that this system has been traditionally represented by the symbol $GI/M/1$ (GI—General Independent). We use the symbolic representation $G/M/1$ for symmetry with the system $M/G/1$. Also I in GI does not really add any additional information.

Let $Q(t)$ be the number of customers in the system at time t and define $Q(t_n - 0) = Q_n$, $n = 1, 2, \ldots$. Thus, Q_n is the number in the system just before the nth arrival. Define X_n as the number of potential service completions during the inter-arrival period Z_n. (Note that we use the word "potential" to indicate that there may not be X_n actual service completions, if the number of customers in the system soon after t_n is less than that number.) Let $\{b_j,\ j = 0, 1, 2, \ldots\}$ be the distribution of X_n. We have

$$b_j = P(X_n = j) = \int_0^\infty e^{-\mu t} \frac{(\mu t)^j}{j!} dA(t). \tag{5.3.1}$$

Now consider the relationship between Q_n and Q_{n+1}. We have

$$Q_{n+1} = \begin{cases} Q_n + 1 - X_{n+1} & \text{if } Q_n + 1 - X_{n+1} > 0 \\ 0 & \text{if } Q_n + 1 - X_{n+1} \leq 0 \end{cases}. \tag{5.3.2}$$

We should note that since X_{n+1} is defined as the potential number of departures, $Q_n + 1 - X_{n+1}$ can be < 0. Clearly Q_{n+1} does not depend on any random variable with an earlier index parameter than n; hence $\{Q_n,\ n = 0, 1, 2, \ldots\}$ is a Markov chain imbedded in the queue length process. From (5.3.2) we get the transition probability

$$\begin{aligned} P_{ij} &= P(Q_{n+1} = j | Q_n = i) \\ &= \begin{cases} P(X_{n+1} = i - j + 1) & \text{if } j > 0 \\ P(X_{n+1} \geq i + 1) & \text{if } j = 0 \end{cases} \end{aligned}$$

giving

$$\begin{aligned} P_{ij} &= b_{i-j+1} \quad j > 0 \\ P_{i0} &= \sum_{r=i+1}^{\infty} b_r. \end{aligned} \tag{5.3.3}$$

The transition probability matrix takes the form

$$\mathbf{P} = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ \vdots \end{array} \begin{array}{c} \begin{array}{ccccc} 0 & 1 & 2 & 3 & \ldots \end{array} \\ \begin{bmatrix} \sum_1^\infty b_r & b_0 & & & \\ \sum_2^\infty b_r & b_1 & b_0 & & \\ \sum_3^\infty b_r & b_2 & b_1 & b_0 & \ldots \\ \vdots & \vdots & \vdots & & \end{bmatrix} \end{array}. \tag{5.3.4}$$

For the Markov chain to be irreducible $b_0 > 0$ and $b_0 + b_1 < 1$. (These two conditions can be justified in much the same way as for the queue $M/G/1$.) We can easily determine that the Markov chain is aperiodic. Let

$$\phi(\theta) = \int_0^\infty e^{-\theta t} dA(t) \quad Re(\theta) > 0$$

be the Laplace–Stieltjes transform of $A(\cdot)$. Using $\phi(\theta)$, the PGF of $\{b_j\}$ is obtained as

$$\begin{aligned} \beta(z) &= \sum_{j=0}^{\infty} b_j z^j \quad |z| \le 1 \\ &= \int_0^\infty e^{-(\mu - \mu z)t} dA(t) \\ &= \phi(\mu - \mu z). \end{aligned}$$

Using definitions similar to those given in (5.2.8), we get (using generic symbols X and Z for X_n and Z_n)

$$E(Z) = \beta'(1) = -\mu\phi'(0) = a\mu. \tag{5.3.5}$$

We define the traffic intensity $\rho =$ (arrival rate)/(service rate). From (5.3.5) we get

$$\rho = \frac{1}{a\mu}. \tag{5.3.6}$$

It can be shown that the Markov chain is positive recurrent when $\rho < 1$, null recurrent when $\rho = 1$, and transient when $\rho > 1$. (See discussion under $M/G/1$ for the implications of these properties. Also a proof is provided later in (5.3.31) and the remarks following that equation.)

The n-step transition probabilities $P_{ij}^{(n)}$ $(i, j = 0, 1, 2, \ldots)$ of the Markov chain $\{Q_n\}$ are obtained as elements of the nth power of $\mathbf{P}$. The observations made under $M/G/1$ regarding the behavior of $\mathbf{P}^n$ hold in the $G/M/1$ case as well. For analytical expressions for $P_{ij}^{(n)}$ the readers may refer to the same references, Takács (1962), Prabhu and Bhat (1963a), and Prabhu (1965a). In practice however, if the state space can be restricted to a manageable size depending on the computer power, successive multiplication of $\mathbf{P}$ to get its power $\mathbf{P}^n$ is likely to turn out to be the best course of action.

Limiting Distribution

Let $\boldsymbol{\pi} = (\pi_0, \pi_1, \pi_2, \ldots)$ be the limiting probabilities defined as $\pi_j = \lim_{n\to\infty} P_{ij}^{(n)}$. Based on Theorem 5.2.1, this limiting distribution exists when the Markov chain is irreducible, aperiodic, and positive recurrent, i.e., when $\rho < 1$. Theorem 5.2.2

provides the method to determine the limiting distribution. Thus, from (5.2.12) we have the equations

$$\begin{aligned} \pi_j &= \sum_{i=0}^{\infty} \pi_i P_{ij} \quad j = 0, 1, 2 \ldots \\ \sum_{0}^{\infty} \pi_j &= 1. \end{aligned}$$

Using P_{ij}'s from (5.3.4), we get

$$\begin{aligned} \pi_0 &= \sum_{i=0}^{\infty} \pi_i \Big(\sum_{r=i+1}^{\infty} b_r \Big) \\ \pi_1 &= \pi_0 b_0 + \pi_1 b_1 + \pi_2 b_2 + \ldots \\ \pi_2 &= \pi_1 b_0 + \pi_2 b_1 + \pi_3 b_2 + \ldots \\ &\vdots \\ \pi_j &= \sum_{r=0}^{\infty} \pi_{j+r-1} b_r \quad (j \geq 1). \end{aligned} \tag{5.3.7}$$

The best computational method for the determination of the limiting distribution seems to be the direct matrix multiplication to get $\mathbf{P}^n$ for increasing values of n until the rows can be considered to be reasonably identical. The computational technique suggested for $M/G/1$ (see (5.2.15)) does not work because of the lower triangular structure of $\mathbf{P}$. As we will see later in the discussion of the finite queue $G/M/1/K$, unless we start with a large enough K, restricting the state space to a finite value alters the last row of the matrix on which the technique has to be anchored.

Given below is an analytical method using finite differences, which may be skipped in first reading. The solution is given in (5.3.13).

This procedure is mathematically simple as well as elegant. (For background in techniques for solving finite difference equations, see standard texts on the subject, e.g., Hildebrand (1968) and Boole (1970).)

Define the finite difference operator D as

$$D\pi_i = \pi_{i+1}. \tag{5.3.8}$$

Using (5.3.8), equation (5.3.7) can be written as

$$\pi_{j-1}(D - b_0 - Db_1 - D^2 b_2 - D^3 b_3 - \ldots) = 0. \tag{5.3.9}$$

Appealing to finite difference methods, a nontrivial solution to the equation (5.3.9) is obtained by solving its characteristic equation

$$\begin{aligned} D - b_0 - Db_1 - D^2 b_2 - \ldots &= 0 \\ D &= \sum_{j=0}^{\infty} b_j D^j \\ D &= \beta(D). \end{aligned} \tag{5.3.10}$$

Hence, the solution to (5.3.10) should satisfy the functional equation

$$z = \beta(z). \tag{5.3.11}$$

In (5.3.10) and (5.3.11), we have used the fact that $\beta(z)$ is the PGF of $\{b_j, j = 0, 1, 2, \ldots\}$.

To obtain roots of (5.3.11), consider two equations $y = z$ and $y = \beta(z)$. The intersections of these two equations give the required roots.

We also have the following properties.

- $\beta(0) = b_0 > 0$; $\beta(1) = \sum_0^\infty b_j = 1$; $\beta'(1) = \rho^{-1}$
- $\beta''(z) = 2b_2 + 6b_3 z + \ldots > 0$ for $z > 0$.

Hence, $\beta'(z)$ is monotone increasing and therefore $\beta(z)$ is convex.

Of the two equations, $y = z$ is a straight line passing through 0, and since $\beta(0) = b_0 > 0$, $\beta(1) = 1$, and $\beta(z)$ is convex, equation $y = z$ and $y = \beta(z)$ intersect at most twice, once at $z = 1$. Let ζ_s be the second root. Whether ζ_s lies to the left or to the right of 1 is dependent on the value of the traffic intensity ρ.

Case 1: $\rho < 1$**.**

When $\rho < 1$, $\beta'(1) > 1$; then $y = \beta(z)$ intersects $y = z$ approaching from below at $z = 1$. But $b_0 > 0$. Hence $\zeta_s < 1$.

Case 2: $\rho > 1$**.**

When $\rho > 1$, $\beta'(1) < 1$. Then $y = \beta(z)$ intersects $y = z$ approaching from above at $z = 1$. Hence $\zeta_s > 1$.

Case 3: $\rho = 1$**.**

In this case $\beta'(1) = 1$ and $y = z$ is a tangent to $y = \beta(z)$ at $z = 1$. This means ζ_s and 1 coincide.

Let ζ be the least positive root. We have $\zeta < 1$ if $\rho < 1$ and $\zeta = 1$ is $\rho \geq 1$. This root is used in the solution of the finite difference equation (5.3.9). (Note that our solution is in terms of probabilities which are ≤ 1, so the root we use must be ≤ 1 as well.)

Going back to the difference equation (5.3.9), we can say that

$$\pi_j = c\zeta^j \quad (j > 0) \tag{5.3.12}$$

is a solution. Since $\zeta < 1$, $\sum_0^\infty \pi_j = 1$, we get

$$\sum_j \pi_j = c\sum_{j=0}^{\infty} \zeta^j = \frac{c}{1-\zeta} = 1$$

giving

$$c = 1 - \zeta.$$

Substituting this back into (5.3.12), we get

$$\pi_j = (1-\zeta)\zeta^j \quad j = 0, 1, 2, \ldots \tag{5.3.13}$$

as the limiting distribution of the state of the system in the queue $G/M/1$.

Referring back to the definition of $\beta(z)$, we note that ζ is the root of the equation

$$z = \phi(\mu - \mu z). \tag{5.3.14}$$

In most cases the root ζ of (5.3.14) has to be determined using numerical techniques. For efficient root-finding algorithms the readers may refer to Chaudhry (1992) and references cited therein.

With the geometric structure for the limiting distribution (5.3.13), the mean and variance of the number in the system, say Q^A, are easily obtained. We have (the superscript A denotes arrival point restriction)

$$\begin{aligned} L^A &= E(Q^A) = \frac{\zeta}{1-\zeta}; \quad L_q^A = \frac{\zeta^2}{1-\zeta} \\ V(Q^A) &= \frac{\zeta}{(1-\zeta)^2}. \end{aligned} \tag{5.3.15}$$

It is important to note that the imbedded Markov chain analysis gives the properties of the number in the system at arrival epochs. (For convenience we have used the number before the arrivals). As pointed out under the discussion on the $M/G/1$ queue the limiting distributions of the number of customers in the system at arrival epochs, departure epochs, and at arbitrary points in time are the same only when the arrivals occur as a Poisson process. Otherwise we have to make appropriate adjustments to the distribution derived above. In this context, the results derived in Prabhu (1965) and Bhat (1968) are worth mentioning. Writing $p_j = \lim_{t\to\infty} P[Q(t) = j]$, where $Q(t)$ is the number at an arbitrary time t, these authors arrive at the following expression for the limiting distribution $\{p_j,\ j = 0, 1, 2 \ldots\}$, when $\rho < 1$.

$$\begin{aligned} p_0 &= 1 - \rho \\ p_j &= \rho(1-\zeta)\zeta^{j-1} \quad j \geq 1. \end{aligned} \tag{5.3.16}$$

From (5.3.16) these results follow on mean queue length:

$$L = \frac{\rho}{1-\zeta}; \quad L_q = \frac{\rho\zeta}{1-\zeta}. \tag{5.3.17}$$

As an example, consider the queue $M/M/1$. Let $A(t) = 1 - e^{-\lambda t}$ $(t \geq 0)$. Then we have

$$\begin{aligned} \phi(\theta) &= \frac{\lambda}{\lambda+\theta} \\ \phi(\mu - \mu z) &= \frac{\lambda}{\lambda+\mu-\mu z}. \end{aligned}$$

Now the functional equation (5.3.11) takes the form

$$\begin{aligned} z &= \frac{\lambda}{\lambda+\mu-\mu z} \\ -\mu z^2 + (\lambda+\mu)z - \lambda &= 0. \end{aligned} \tag{5.3.18}$$

This quadratic equation has two roots 1 and $\frac{\lambda}{\mu} = \rho$. Substituting ρ in place of ζ in (5.3.16)–(5.3.18) we have the limiting distribution and mean values for the queue $M/M/1$, which match with the results derived in Chapter 4.

Waiting Time

To determine the distribution of the waiting time of a customer we need the distribution of the number of customers in the system at the time of its arrival. The limiting distribution derived in (5.3.13) is in fact an arrival point distribution in $G/M/1$. Furthermore, its structure is the same as the geometric distribution we had for $M/M/1$, with ζ taking the place of ρ of the $M/M/1$ result. The service times of customers in the system are exponential with rate μ, also as in $M/M/1$. Hence, the waiting time results for $G/M/1$ have the same forms as those for $M/M/1$ with ζ replacing ρ. Without going into the details of their derivation, we can write,

$$\begin{aligned} F_q(t) &= P(T_q \leq t) = 1 - \zeta e^{-\mu(1-\zeta)t} \\ W_q &= E[T_q] = \frac{\zeta}{\mu(1-\zeta)} \\ V[T_q] &= \frac{\zeta(2-\zeta)}{\mu^2(1-\zeta)^2}. \end{aligned} \tag{5.3.19}$$

The time T spent by the customer in the system is obtained by adding service time to T_q. We get

$$\begin{aligned} W &= E(T) = E(T_q) + \frac{1}{\mu} = \frac{1}{\mu(1-\zeta)} \\ V(T) &= V[T_q + S] = \frac{1}{\mu^2(1-\zeta)^2}. \end{aligned} \tag{5.3.20}$$

Busy Cycle

A busy cycle of a $G/M/1$ queue, when modeled as an imbedded Markov chain is the number of transitions the process takes to go from state 0 to state 0 for the first time. This interval is also known as the *recurrence time* of state 0. The busy cycle includes the busy period when the server is continuously busy, and the idle period, when there is no customer in the system. Let R denote the number of transitions in a busy cycle. (Note that we are using a generic symbol R, with the assumption that all such busy cycles have the same distribution.) Unfortunately, the derivation of the distribution and its mean is not very simple. We give below the procedure for advanced readers. The mean values of R and the length of the busy cycle are given in (5.3.31) and (5.3.32).

Let $h_j^{(n)}$ be the probability that the number of customers, just before the nth arrival in a busy cycle is j. Working backward from n, considering the arrival time of the first of those j customers we can write, for $j \geq 1$,

$$\begin{aligned} h_j^{(j)} &= b_0^{(j)} \\ h_j^{(n)} &= \sum_r h_r^{(n-j)} b_r^{(j)} \quad n \geq j \end{aligned} \tag{5.3.21}$$

where $b_r^{(j)}$ is the j-fold convolution of b_r with itself. Looking back to relations (5.2.47), we see that (5.3.21) is structurally similar to (5.2.47) with $h_j^{(n)}$ replacing $g_i^{(n)}$ and $b_r^{(i)}$ replacing $k_r^{(i)}$. Define

$$H_j(z) = \sum_{n=j}^{\infty} h_j^{(n)} z^n \quad |z| \leq 1. \tag{5.3.22}$$

Using arguments similar to those used in determining $G(z)$, we can show that

$$H_j(z) = [\eta(z)]^j, \quad j \geq 1 \tag{5.3.23}$$

where $\eta(z)$ is the unique root in the unit circle of the equation

$$\omega = z\beta(\omega). \tag{5.3.24}$$

The distribution of R is given by $h_0^{(n)}$. Considering the transitions during the nth transition interval, we have

$$\begin{aligned} h_0^{(n)} &= \sum_{r=1}^{n-1} h_r^{(n-1)} \left(\sum_{k=r+1}^{\infty} b_k \right) \\ h_0^{(1)} &= \sum_1^{\infty} b_k. \end{aligned} \tag{5.3.25}$$

Taking generating functions

$$\begin{aligned} H_0(z) &= \sum_{n=1}^{\infty} h_0^{(n)} z^n \\ &= \left(\sum_{1}^{\infty} b_k\right) z + \sum_{n=2}^{\infty} \sum_{r=1}^{n-1} h_r^{(n-1)} z^n \left(\sum_{k=r+1}^{\infty} b_k\right). \end{aligned} \tag{5.3.26}$$

The right-hand side of (5.3.26) can be simplified as follows. For ease of notation write $\sum_{r+1}^{\infty} b_k = \beta_r$. The right-hand side of (5.3.26) simplifies to

$$\begin{aligned} \beta_0 z &+ z \sum_{n=2}^{\infty} z^{n-1} \sum_{r=1}^{n-1} \beta_r h_r^{(n-1)} \\ &= z\left[\beta_0 + \sum_{r=1}^{\infty} \beta_r \sum_{n=r+1}^{\infty} h_r^{(n-1)} z^{n-1}\right] \\ &= z\left[\sum_{r=0}^{\infty} \beta_r [\eta(z)]^r\right] \end{aligned}$$

where we have used (5.3.22) and (5.3.23). But

$$\sum_{r=0}^{\infty} \beta_r z^r = \frac{1-\beta(z)}{1-z}$$

since

$$\begin{aligned} \sum_{r=0}^{\infty} z^r \sum_{j=r+1}^{\infty} b_j &= \sum_{j=1}^{\infty} b_j \sum_{r=0}^{j-1} z^r \\ &= \sum_{j=1}^{\infty} b_j \left(\sum_{r=0}^{\infty} - \sum_{r=j}^{\infty}\right) z^r \\ &= \sum_{j=1}^{\infty} b_j \left[\frac{1}{1-z} - z^j \sum_{r=j}^{\infty} z^{r-j}\right] \\ &= \frac{1-\beta(z)}{1-z}. \end{aligned}$$

Thus, we get

$$H_0(z) = \frac{z - z\beta[\eta(z)]}{1-\eta(z)}.$$

But $\eta(z)$ is such that

$$\eta(z) = z\beta[\eta(z)].$$

Hence

$$H_0(z) = \frac{z - \eta(z)}{1 - \eta(z)}. \tag{5.3.27}$$

Letting $z \to 1$ in $H_0(z)$, we can show that R is a proper random variable (i.e., $P(R < \infty)$) when $\rho \leq 1$. The expected length of the busy cycle (recurrence time of state 0) is obtained as $\lim_{z\to 1} H_0'(z)$. We have

$$\begin{aligned} H_0'(z) &= \frac{[1-\eta(z)][1-\eta'(z)] + [z-\eta(z)]\eta'(z)}{[1-\eta(z)]^2} \\ &= \frac{1-\eta'(z)}{1-\eta(z)} + \frac{[z-\eta(z)]\eta'(z)}{[1-\eta(z)]^2}. \end{aligned} \tag{5.3.28}$$

To simplify (5.3.28) further, we need values for $\eta(1)$ and $\eta'(1)$. Referring back to the functional equation (5.3.24), we find that for $z = 1$ it is the solution of the functional equation (5.3.11) which we have found to be the least positive root ζ (< 1 or $= 1$). Hence

$$\eta(1) = \zeta \text{ if } \rho < 1 \text{ and } = 1, \text{ if } \rho \geq 1. \tag{5.3.29}$$

Consider $\eta(z) = z\beta[\eta(z)]$. We get

$$\begin{aligned} \eta'(z) &= z\beta'[\eta(z)]\,\eta'(z) + \beta[\eta(z)] \\ \eta'(z)\,[1 - z\beta'[\eta(z)]] &= \beta[\eta(z)]. \end{aligned}$$

Letting $z \to 1$, and using (5.3.29)

$$\begin{aligned} \eta'(1)\,[1 - \beta'(\zeta)] &= \beta(\zeta) \\ \eta'(1) &= \frac{\beta(\zeta)}{1 - \beta'(\zeta)}. \end{aligned} \tag{5.3.30}$$

Substituting from (5.3.29) and (5.3.30) in (5.3.28)

$$\begin{aligned} \lim_{z\to 1} H_0'(z) &= \left[1 - \frac{\beta(\zeta)}{1-\beta'(\zeta)}\right]\frac{1}{1-\zeta} + \left[(1-\zeta)\frac{\beta(\zeta)}{1-\beta'(\zeta)}\right] \times \frac{1}{(1-\zeta)^2} \\ &= \frac{1}{1-\zeta} < \infty \text{ if } \rho < 1. \end{aligned} \tag{5.3.31}$$

Similarly, we can also show that $H'_{\lim z\to 1}(z) = \infty$ when $\rho = 1$.

We may note that these results establish the classification properties of positive recurrence, null recurrence, and transience of the imbedded Markov chain.

The mean length of the busy cycle is obtained as the product (expected number of transitions) $\times$ (mean inter-arrival time).

$$E[\text{busy cycle}] = \frac{E(Z)}{1-\zeta}. \tag{5.3.32}$$

Since, the busy period terminates during the last transition of the Markov chain and the transition interval (inter-arrival time) has a general distribution, the determination of the mean busy period is too complicated to be covered at this stage.

The Queue $G/M/s$

The imbedded Markov chain analysis of the queue $G/M/1$ can be easily extended to the multi-server queue $G/M/s$. Since the Markov chain is defined at arrival points, the structure of the process is similar to that of $G/M/1$, except for the transition probabilities. Retaining the same notations, for the relationship between Q_n and Q_{n+1}, we get

$$Q_{n+1} = \begin{cases} Q_n + 1 - X_{n+1} & \text{if } Q_n + 1 - X_{n+1} > 0 \\ 0 & \text{if } Q_n + 1 - X_{n+1} \leq 0, \end{cases}$$

where X_{n+1} is the total number of potential customers who can be served by s servers during an inter-arrival time with distribution $A(\cdot)$.

To determine transition probabilities P_{ij} $(i, j = 0, 1, 2 \ldots)$ we have to consider three cases for the initial value i and the final value j: $i + 1 \geq j \geq s$; $i + 1 \leq s$ and $j \leq s$; and $i + 1 > s$ but $j < s$. Note that when $Q_n = i$, the transition starts with $i + 1$, and j is always $\leq i + 1$. Since the service times are exponential with density $\mu e^{-\mu x}$ $(x > 0)$, the probability that a server will complete service during $(0, t]$ is $1 - e^{-\mu t}$ and the probability that the service will continue beyond t is $e^{-\mu t}$. Incorporating these concepts along with the assumptions that the servers work independently of each other, we get the following expressions for P_{ij}.

Case 1: $i + 1 \geq j \geq s$

$$P_{ij} = \int_0^\infty e^{-s\mu t} \frac{(s\mu t)^{i+1-j}}{(i+1-j)!} dA(t). \tag{5.3.33}$$

This represents $i+1-j$ service completions during an inter-arrival period, when all s servers are busy. See discussion under $M/M/s$ to justify the service rate $s\mu$ when all servers are busy.

Case 2:

$i + 1 \leq s$ and $j \leq s$

$$P_{ij} = \binom{i+1}{i+1-j} \int_0^\infty (1 - e^{-\mu t})^{i+1-j} e^{-j\mu t} dA(t). \tag{5.3.34}$$

This expression takes into account the event in which $i + 1 - j$ out of $i + 1$ customers complete service during $(0, t]$ while j customers are still being served. Because of the independence of servers among one another, each service can be considered a Bernoulli trial and the outcome has a binomial distribution with success probability $1 - e^{-\mu t}$.

Case 3:

$i+1 > s$ but $j < s$

$$P_{ij} = \int_{t=0}^{\infty} \int_{\tau=0}^{t} e^{-s\mu\tau} \frac{(s\mu\tau)^{i-s}}{(i-s)!} s\mu \binom{s}{s-j} [1 - e^{-\mu(t-\tau)}]^{s-j} e^{-j\mu(t-\tau)} d\tau dA(t). \tag{5.3.35}$$

Initially, $i+1-s$ customers complete service with rate $s\mu$, and then $s-j$ out of the remaining s complete their service independently of each other.

The transition probability matrix of the imbedded chain has a structure similar to the one displayed in (5.3.4). Because of the structure of P_{ij} values under cases 2 and 3, the finite difference solution given earlier for the limiting distribution need major modifications. Interested readers are referred to advanced texts on the subject, e.g., Gross et al.(2008). Taking into consideration the complexities of these procedures, the computational method developed below for $G/M/1/K$ could turn out to be advantageous in this case, if it is possible to work with a finite limit for the number of customers in the system.

5.3.1 The Queue $G/M/1/K$

Consider the $G/M/1$ queue described earlier with the restrictions that the system can accommodate only K customers at a time. Since the imbedded chain is defined just before an arrival epoch, the number of customers in the system soon after the arrival epoch is K, whether it is K or $K-1$ before that time point. If it is K before, the arriving customer does not get admitted to the system. Thus, in place of (5.3.2) we have the relation

$$Q_{n+1} = \begin{cases} \min\,(Q_n + 1 - X_{n+1}, K) & \text{if } Q_n + 1 - X_{n+1} > 0 \\ 0 & \text{if } Q_n + 1 - X_{n+1} \le 0. \end{cases} \tag{5.3.36}$$

Using probabilities b_j, $j = 0, 1, 2 \ldots$ defined in (5.3.1) the transition probability matrix $\mathbf{P}$ can be displayed as follows:

$$\mathbf{P} = \begin{array}{c} \\ 0 \\ 1 \\ \vdots \\ K-1 \\ K \end{array} \begin{array}{c} \begin{array}{ccccccc} 0 & 1 & 2 & \ldots & K-1 & K \end{array} \\ \begin{bmatrix} \sum_1^{\infty} b_r & b_0 & 0 & & & \\ \sum_2^{\infty} b_r & b_1 & b_0 & & & \\ \vdots & \vdots & & & & \\ \sum_K^{\infty} b_r & b_{K-1} & b_{K-2} & \ldots & b_1 & b_0 \\ \sum_K^{\infty} b_r & b_{K-1} & b_{K-2} & \ldots & b_1 & b_0 \end{bmatrix} \end{array} \tag{5.3.37}$$

Note that the last two rows of the matrix $\mathbf{P}$ are identical because the Markov chain effectively starts off with K customers from either of the states $K-1$ and K.

Let $\boldsymbol{\pi} = (\pi_0, \pi_i, \ldots, \pi_K)$ be the limiting distribution of the imbedded chain. Writing out the equation $\pi_j = \sum_{i=0}^{K} \pi_i P_{ij}$, we have

$$\begin{aligned}
\pi_0 &= \sum_{i=0}^{K-1} \pi_i \left(\sum_{r=i+1}^{\infty} b_r \right) + \left(\sum_{r=k}^{\infty} b_r \right) \pi_k \\
\pi_1 &= \pi_0 b_0 + \pi_1 b_1 + \ldots + \pi_{K-1} b_{K-1} + \pi_K b_{K-1} \\
\pi_2 &= \pi_1 b_0 + \pi_2 b_1 + \ldots + \pi_{K-1} b_{K-2} + \pi_K b_{K-2} \\
&\vdots \\
\pi_{K-1} &= \pi_{K-2} b_0 + \pi_{K-1} b_1 + \pi_K b_1 \\
\pi_K &= \pi_{K-1} b_0 + \pi_K b_0. \qquad (5.3.38)
\end{aligned}$$

If the value of K is not too large, solving these simultaneous equations in π_j, $j = 0, 1, 2, \ldots, K$, along with the normalizing condition $\sum_0^K \pi_j = 1$ directly could be computationally practical. Or for that matter getting $\mathbf{P}^n$ for increasing values of n until the row elements are close to being identical will also give the limiting distribution under these circumstances. An alternative procedure is to develop a computational recursion as done in the case of the $M/G/1$ queue (see (5.2.15)).

To do so we start with the last equation of (5.3.38) and define

$$\nu_i = \frac{\pi_i}{\pi_{i-1}}, \quad i = 1, 2, \ldots, K.$$

We have

$$\begin{aligned}
\pi_i &= \nu_i \pi_{i-1} \\
&= \nu_i \nu_{i-1} \pi_{i-2} \\
&= \nu_i \nu_{i-1} \ldots \nu_1 \pi_0. \qquad (5.3.39)
\end{aligned}$$

From the last equation in (5.3.38), we get

$$\begin{aligned}
\nu_K &= b_0 + \nu_K b_0 \\
\nu_K &= \frac{b_0}{1 - b_0}.
\end{aligned}$$

From the next to the last equation in (5.3.38), we get

$$\begin{aligned}
\nu_{K-1} &= b_0 + \nu_{K-1} b_1 + \nu_K \nu_{K-1} b_1 \\
\nu_{K-1} &= \frac{b_0}{1 - b_1 - \nu_K b_1}
\end{aligned}$$

and so on.

Since $\sum_0^K \pi_j = 1$, from (5.3.39) we get

$$(1 + \nu_1 + \nu_1 \nu_2 + \ldots + \nu_1 \nu_2 \ldots \nu_K) \pi_0 = 1$$

and hence

$$\pi_0 = \frac{1}{1 + \sum_{i=1}^{K} \Pi_{r=1}^{i} \nu_r}. \tag{5.3.40}$$

Substituting these back in (5.3.39), we get π_j, $j = 0, 1, 2, \ldots K$.

Note that in developing the recursion we have defined ν_i's as ratio of consecutive π_i's, unlike in the case of (5.2.15). We do so for the reason that π_j's decrease in value as j increases and dividing by a very small π_j is likely to result in large computational errors. Looking at the structure of the limiting distribution of $G/M/1$, the ratio of consecutive terms of $\boldsymbol{\pi}$, are likely to be close to the constant ζ.

Example 5.3.1 In a service center job arrivals occur in a deterministic process, one job per one unit of time. Service is provided by a single server with an exponential service time distribution with rate 1.5 jobs per unit time.

In order to determine the limiting distribution, using a $D/M/1$ model, we note that

$$\begin{aligned} \phi(\theta) &= \int_0^\infty e^{-\theta t} dA(t) \\ &= e^{-\theta}. \end{aligned}$$

With an exponential service time distribution we have $\mu = 1.5$. Hence

$$\begin{aligned} \beta(z) &= \phi(\mu - \mu z) \\ &= e^{-1.5(1-z)}. \end{aligned}$$

The limiting distribution is expressed in terms of ζ which is the unique root in the unit circle, of the functional equation

$$z = e^{-1.5(1-z)}.$$

We can easily solve this equation by successive substitution starting wtih $z = 0.4$. We get

z	$\beta(z)$
0.400	0.407
0.407	0.411
0.411	0.413
0.413	0.415
0.415	0.416
0.416	0.416

We use $\zeta = 0.416$ in the limiting distribution $\boldsymbol{\pi} = (\pi_0, \pi_1, \pi_2, \ldots)$ given by (5.3.13). We get

$\pi_0 = 0.584$; $\pi_1 = 0.243$; $\pi_2 = 0.101$; $\pi_3 = 0.042$;
$\pi_4 = 0.017$; $\pi_5 = 0.007$; $\pi_6 = 0.003$; $\pi_7 = 0.001$.

Answer

Example 5.3.2 Consider the service center Example 5.3.1 above with a capacity restriction of K customers in the system.

In this case, we use the computational recursion developed in (5.3.39) for two values of $K = 4$ and 7.

The distribution of the potential number of customers served during an inter-arrival period is Poisson with mean 1.5. We have

$$\begin{array}{llll} b_0 = 0.223; & b_1 = 0.335; & b_2 = 0.251; & b_3 = 0.125 \\ b_4 = 0.047; & b_5 = 0.015; & b_6 = 0.003; & b_7 = 0.001\ . \end{array}$$

$\underline{K = 4}$

Using $\nu_i = \pi_i/\pi_{i-1}$ in (5.3.38) with $K = 4$, we get

$$\begin{aligned} \nu_4 &= b_0\,[1 - b_0]^{-1} \\ \nu_3 &= b_0\,[1 - b_1 - \nu_4 b_1]^{-1} \\ \nu_2 &= b_0\,[1 - b_1 - \nu_3 b_2 - \nu_4\nu_3 b_2]^{-1} \\ \nu_1 &= b_1\,[1 - b_1 - \nu_2 b_2 - \nu_3\nu_2 b_3 - \nu_4\nu_3\nu_2 b_3]^{-1}\,. \end{aligned}$$

Substituting appropriate values of $b_j = 0, 1, 2, 3$, we get

$$\nu_1 = 0.413; \quad \nu_2 = 0.398; \quad \nu_3 = 0.392; \quad \nu_4 = 0.287.$$

But we have

$$\begin{aligned} \pi_4 &= \nu_4\nu_3\nu_2\nu_1\pi_0 \\ \pi_3 &= \nu_3\nu_2\nu_1\pi_0 \\ \pi_2 &= \nu_2\nu_1\pi_0 \\ \pi_1 &= \nu_1\pi_0. \end{aligned}$$

Using $\sum_0^4 \pi_j = 1$, we get

$$\begin{aligned} \pi_0 &= [1 + \nu_1 + \nu_1\nu_2 + \nu_1\nu_2\nu_3 + \nu_1\nu_2\nu_3\nu_4]^{-1} \\ &= 0.602. \end{aligned}$$

Thus, we have the limiting distribution

$$\pi_0 = 0.602; \quad \pi_1 = 0.249; \quad \pi_2 = 0.099; \quad \pi_3 = 0.039; \quad \pi_4 = 0.001.$$

$\underline{K = 7}$

Looking at the structure of ν_i's, it is clear that $\nu_4, \ldots, \nu_1$ determined above, in fact, yield $\nu_7, \ldots, \nu_4$, when $K = 7$. Extending the equations to determine the remaining ν's, viz., ν_3, ν_2, and ν_1, we get the following set of values:

$$\nu_7 = 0.287;\ \nu_6 = 0.392;\ \nu_5 = 0.398;\ \nu_4 = 0.413;$$
$$\nu_3 = 0.415;\ \nu_2 = 0.416;\ \nu_1 = 0.417.$$

Converting these back to π's, we get

$\pi_0 = 0.585; \quad \pi_1 = 0.244; \quad \pi_2 = 0.101$;
$\pi_3 = 0.042; \quad \pi_4 = 0.017; \quad \pi_5 = 0.007$;
$\pi_6 = 0.003; \quad \pi_7 = 0.001.$ **Answer**

A comparison of these values with those obtained under Example 5.3.1 shows that when $K = 7$ the effect of the capacity limit is negligible for the long run distribution of the process.

5.4 Exercises

1. Specialize the mean waiting time results (5.2.42) and (5.2.43) when the service time distribution is (a) deterministic (constant service time) and (b) Erlang E_k.

2. Obtain the transition probability matrices $\mathbf{P}$ of the Markov chains representing the number of customers in the system at epochs at which service completion occurs in (a) Example 5.2.1 and (b) Example 5.2.2.

   Determine the limiting distributions of the queueing systems in (a) and (b) by obtaining $\mathbf{P}^n$ for large enough n.

3. Obtain the transition probability matrix $\mathbf{P}$ of the Markov chain representing the number of customers at arrival epochs in the problem described in Example 5.3.1. Determine the limiting distribution of the queueing system by obtaining $\mathbf{P}^n$ for large enough n.

4. A mail order business receives orders for various items of merchandize in a Poisson process at the rate of 15 per hour. The amount of time required to fill an order has a mean of 3.5 minutes and variance 2.5 minutes^2. Determine the expected number of orders waiting to be filled. Also, determine the mean length of the period from the time the order is received until it is filled.

   How much of an improvement in service can be accomplished if (a) the length of service time is shortened to 3 minutes, without changing the variance? (b) if the variance of service time is reduced to 2 minutes^2, without altering the mean?

5. (a) Cars arrive at a single station carwash at the rate of 15 per hour. The automatic carwash is set to take up exactly 3 minutes. Assuming the arrivals are in a Poisson process determine (1) the expected number of cars waiting for wash at any time, and (2) the expected waiting time of each automobile.

   (b) The owner of the carwash wants to reduce the waiting time by shortening the amount of time taken for each wash. However, a quick survey of his customers reveals that a third of his customers would like to have a longer wash. To satisfy their need he sets up two wash times, 5 and

2.5 minutes for the two groups. Both these groups will pass through the same station. With this change has he improved the situation or worsened it? Determine the expected number of cars waiting and the mean waiting time in the long run.

6. In a doctor's office, appointments to see a doctor are made at 15 minute intervals. But the amount of time the doctor spends with her patients are mostly less than 15 minute, but some of them may take much longer. It has been found that these times can be represented by an exponential distribution with mean 12 minutes. Assuming steady state, determine the expected number of patients in the waiting room at any time. Also, determine the mean amount of time a patient waits at the doctor's office per visit. Suppose the office personnel and the doctor decide to take breaks when there are no customers in the system. During a 7 hour work day how often will they be able to take such breaks?

7. At a taxi-stop on a busy street, the inter-arrival times of taxis dropping off passengers (and then being ready for the pick up) have a mean 5 minutes with a standard deviation 1 minute. The taxis are not allowed to wait for customers at the stop. Customers arrive at the stop in a Poisson process once every 6 minutes on the average.

   (a) Determine the expected number of customers waiting, when a taxi leaves the stop.

   (b) What is the probability that there would be no waiting customer, when a taxi leaves the stop?

   (c) Determine the mean waiting time of a customer.

8. In Example 5.2.3, let the time be discretized into segments, each of Q units of time in length. Assume that the arrivals occur at the end of each such interval with probability λQ. The service times have a geometric distribution as described in (5.2.57) and the queue discipline is RR as described in the example. Following the same arguments as in the example derive the conditional mean response time W_k corresponding to the result (5.2.70).

9. In the computer system model of Exercise 9 of Chapter 1, the following numerical value and distributional assumptions are made. Determine the average response time for the system.

   (a) Arrivals are in a Poisson process with rate 1 per second.

   (b) CPU time of the job (for the first time or after I/0 use) is exponential with mean 0.1 second.

   (c) The three phases of disk service have the following characteristics.

   Seek time: Exponentially distributed with mean 0.03 seconds.

   Latency time: Uniformly distributed with mean 0.01 second.

Transfer time: Constant = 0.01 second.

(d) P (a job will need disk service) = 0.8.

10. The memory disk in a computer is organized into tracks and sectors with a read–write head per track. Requests for the use of the memory disk arrive in a Poisson process with rate λ. Let the disk rotation time be R units and the number of sectors per track be K. Determine the mean response time of a request under the following two alternatives.

    (a) The requests follow a single queue and are handled on a FCFS basis. Assume that the duration of service can take any of the K equally probable values, $\frac{iR}{K}$, $(i = 1, 2, \ldots, K)$.

    (b) Each sector has its own queue and the requests select the K sectors with equal probability. Once a request has been handled, the next in queue must wait for the sector to come around again during the next rotation (Krakowiak (1988)).

11. In the RR discipline model of Example 5.2.3, if we let the quantum $Q \to 0$, the resulting discipline is a *Processor Sharing* (PS) service discipline. This is because when $Q \to 0$, it is as if, service is provided simultaneously to all customers in the system.

    Letting $Q \to 0$ in Example 5.2.3, show that the mean response time in a processor sharing system with Poisson arrivals and any arbitrary service time distribution, is given by

$$W = \frac{1}{\mu(1-\rho)}$$

    where $1/\mu$ is the mean service time and $\rho = \frac{\lambda}{\mu}$ is the traffic intensity. (Hint: When $Q \to 0$, α and $\sigma \to 1$ and $L \to \rho/(1-\rho)$. Set $\frac{S}{Q} = k$ where S is the service time requested by the customer.)

Chapter 6

Extended Markov and Renewal Models

The queueing systems discussed in the last two chapters were devoid of any features such as group arrivals, group service, priority service, etc., that would make modeling difficult. In this chapter we introduce them in a limited sense, so that Markov process modeling is still possible by extending the models as well as procedures for analyzing them. However, the readers should be warned that the extension of the procedure comes at the cost of employing a higher level of mathematics in the use of probability generating functions (PGFs).

6.1 The Bulk Queue $M^{(X)}/M/1$

The queueing systems discussed in the previous chapters assume that the customers arrive one at a time. There are many situations where customers arrive in groups, e.g., customer arrivals in restaurants, voice or data traffic segmented as packets in a communication system. Queueing systems in which customer arrivals and/or service occur in groups are known as *bulk queues* in the literature.

Let customers arrive in groups of size X, where, in general, X is a random variable assuming integer values greater than zero. Let the groups arrive in a Poisson process with rate λ and the customer service be provided one at a time with an exponential service time distribution with rate μ. For simplicity, we use the symbolic notation $M^{(X)}/M/1$, to signify this system.

Let $d_r = P(X = r)$, $r = 1, 2, \ldots$ be the distribution of the size of the arriving group of customers. We assume that the group size is independent of other characteristics of the system. Thus whenever an arrival occurs, the number of customers in the system increases by the size of the group.

Let $Q(t)$ be the number of customers in the system at time t, and let Q represent $Q(t)$ as $t \to \infty$. Because $Q(t)$ increases by the arriving group size at arrival points, $\{Q(t)\}$ is a modified birth and death process in which increases in

© Springer Science+Business Media New York 2015

U. N. Bhat, *An Introduction to Queueing Theory*, Statistics for Industry and Technology, DOI 10.1007/978-0-8176-8421-1_6

the state space can occur by more than 1. Let $p_n = P(Q = n)$, $n = 0, 1, 2, \ldots$. Making appropriate modifications to the state balance equations for $M/M/1$ given in (4.2.3), we have

$$\begin{aligned} \lambda p_0 &= \mu p_1 \\ (\lambda + \mu) p_n &= \lambda \sum_{r}^{n} d_r p_{n-r} + \mu p_{n+1} \quad n = 1, 2, \ldots . \end{aligned} \tag{6.1.1}$$

It should be noted that the first term in the right-hand side of the second equation in (6.1.1) exists only if $n - r \geq 0$.

Unfortunately, (6.1.1) cannot be solved using recursive methods, as done in the $M/M/1$ case. Instead, we use PGFs to simplify the equations. Let

$$P(z) = \sum_{n=0}^{\infty} p_n z^n; \quad \delta(z) = \sum_{r=1}^{\infty} d_r z^r \quad |z| \leq 1.$$

Multiplying equations in (6.1.1) with appropriate powers of z, we have

$$\begin{aligned} \lambda p_0 &= \mu p_1 \\ (\lambda + \mu) \sum_{n=1}^{\infty} p_n z^n &= \lambda \sum_{n=1}^{\infty} z^n \sum_{r=1}^{n} d_r p_{n-r} + \mu \sum_{n=1}^{\infty} p_{n+1} z^n. \end{aligned} \tag{6.1.2}$$

Interchanging summations on the right-hand side of (6.1.2) and simplifying, we get

$$\begin{aligned} (\lambda + \mu) P(z) - \mu p_0 &= \lambda \sum_{r=1}^{\infty} d_r z^r \sum_{n=r}^{\infty} z^{n-r} p_{n-r} \\ &\quad + \mu \sum_{n=0}^{\infty} z^n p_{n+1} \\ &= \lambda \delta(z) P(z) + \frac{\mu}{z} \sum_{m=1}^{\infty} z^m p_m \\ &= \lambda \delta(z) P(z) + \frac{\mu}{z} [P(z) - p_0]. \end{aligned}$$

Rearranging terms and simplifying

$$P(z) = \frac{\mu p_0 (1 - z)}{\mu(1 - z) - \lambda z [1 - \delta(z)]}. \tag{6.1.3}$$

To determine p_0, we use the normalizing condition $\sum_n p_n = 1$ and note $\lim_{z \to 1} P(z) = 1$. Taking limits on the right-hand side of (6.1.3) using l'Hôpital's rule, we get

$$\lim_{z \to 1} P(z) = \frac{\lim_{z \to 1} \mu p_0 (1 - z)}{\lim_{z \to 1} [\mu(1 - z) - \lambda z (1 - \delta(z))]}$$

giving

$$\begin{aligned} 1 &= \frac{\mu p_0}{\mu + \lambda(1-\delta'(1))} \\ p_0 &= 1 - \frac{\lambda\delta'(1)}{\mu}. \end{aligned} \tag{6.1.4}$$

Note that

$$\begin{aligned} \delta'(1) &= \lim_{z\to 1} \sum_{r=1}^{\infty} r d_r z^{r-1} \\ &= E(X) = d, \quad \text{say.} \end{aligned}$$

which is the average group size. We also note that $\frac{\lambda d}{\mu} = \rho$ is the traffic intensity. This leads us to the result

$$p_0 = 1 - \rho \tag{6.1.5}$$

and

$$P(z) = \frac{\mu(1-z)(1-\rho)}{\mu(1-z) - \lambda z(1-\delta(z))}. \tag{6.1.6}$$

Unfortunately, even with simple forms of the distribution $\{d_r\}$, inverting the PGF (6.1.6) is not simple. See discussion following the PGF (5.2.19) of the limiting distribution of the imbedded chain in the queue $M/G/1$. Nevertheless, (6.1.6) can be used easily for the determination of the mean value of Q as $t \to \infty$, by noting that $E(Q) = \lim_{z\to 1} P'(z)$. Because of the term $(1-z)$ in the numerator and $(1-\delta(z))$ in the denominator of (6.1.6), we use l'Hôpital's rule in taking limits in $P'(z)$. After simplifications we get

$$E(Q) = \lim_{z\to 1} P'(z) = \frac{2\rho + \frac{\lambda}{\mu}\delta''(1)}{2(1-\rho)} \tag{6.1.7}$$

But $\delta''(1) = E(X^2) - E(X)$. With d as the mean group size, (6.1.7) simplifies to

$$L = E(Q) = \frac{\rho + \frac{\lambda}{\mu}E(X^2)}{2(1-\rho)}. \tag{6.1.8}$$

The variance of Q can be determined by letting $z \to 1$ in $P''(z)$ and noting that $V(Q) = P''(1) + P'(1) - [P'(1)]^2$. The algebra in the determination of $V(Q)$ involves the use of l'Hôpital's rule multiple times.

When the group sizes are a constant K, the mean number of customers in the system, given by (6.1.8), simplifies to

$$\begin{aligned} L = E(Q) &= \frac{\rho + \frac{\lambda}{\mu}K^2}{2(1-\rho)} \\ &= \left(\frac{K+1}{2}\right)\frac{\rho}{1-\rho}, \end{aligned} \tag{6.1.9}$$

since $\rho = \frac{\lambda K}{\mu}$.

6.2 The Bulk Queue $M/M^{(X)}/1$

In the queueing model $M/M^{(X)}/1$, we assume that the customers arrive one at a time, but get served in groups of size X. For simplicity we also assume that X is a constant K. When the service is in groups there are two other factors of queue discipline that can complicate the analysis: (1) whether the server waits for customer arrivals when there are fewer than K customers in the queue at the time of a service completion and (2) if the server starts service with less than K customers in the group, whether the new arrivals are allowed to join the ongoing service or are required to wait for the next batch. To keep the algebra simple, we make the assumption that the server starts service only when the batch is full. For an analysis of the system under queue discipline in which service starts even with a single customer and the arriving customers are allowed to join the batch in service to fill the vacancies, see Gross et al. (2008).

Thus, the customers arrive one at a time in a Poisson process with parameter λ and get served in groups of size K if there are K or more customers in the queue at the completion of a service. If there are less than K customers waiting at the completion of a service, the server waits until the service batch of K is full. The service time distribution is exponential with parameter μ. With these assumptions, for the limiting distribution $\{p_n,\, n = 0, 1, 2, \ldots\}$ of the number of customers in the system as $t \to \infty$, the state balance equations can be presented as follows:

$$\begin{aligned} \lambda p_0 &= \mu p_K \\ \lambda p_n &= \lambda p_{n-1} + \mu p_{n+K} \quad n = 1, 2, \ldots, K-1 \\ (\lambda+\mu) p_n &= \lambda p_{n-1} + \mu p_{n+K} \quad n = K, K+1, \ldots. \end{aligned} \tag{6.2.1}$$

The method we use to solve these equations makes use of PGFs. Multiply both sides of (6.2.1) by appropriate powers of z and add. We get

$$\begin{aligned} &\lambda p_0 + \lambda \sum_{n=1}^{K-1} p_n z^n + (\lambda+\mu) \sum_{n=K}^{\infty} p_n z^n \\ &= \mu p_K + \lambda \sum_{n=1}^{K-1} p_{n-1} z^n + \mu \sum_{n=1}^{K-1} p_{n+K} z^n \\ &\quad + \lambda \sum_{n=K}^{\infty} p_{n-1} z^n + \mu \sum_{n=K}^{\infty} p_{n+K} z^n. \end{aligned}$$

Noting that $\sum_{n=0}^{\infty} p_n z^n = P(z)$ and making appropriate simplifications, we can write

$$\begin{aligned}
&(\lambda+\mu)P(z) - \mu \sum_{n=0}^{K-1} p_n z^n \\
&= \lambda z \sum_{n=1}^{\infty} p_{n-1} z^{n-1} + \frac{\mu}{z^K} \sum_{n=0}^{\infty} p_{n+K} z^{n+K} \\
&= \lambda z P(z) + \frac{\mu}{z^K} \left[P(z) - \sum_{m=0}^{K-1} p_m z^m \right].
\end{aligned}$$

Rearranging terms and multiplying by z^K, we get

$$\left[(\lambda+\mu)z^K - \lambda z^{K+1} - \mu\right] P(z) = \mu(z^K - 1) \sum_{n=0}^{K-1} p_n z^n$$

$$P(z) = \frac{(1-z^K)\sum_{n=0}^{K-1} p_n z^n}{(\lambda/\mu)z^{K+1} - (\frac{\lambda}{\mu}+1)z^K + 1}. \tag{6.2.2}$$

For the complete determination of the PGF $P(z)$, we need to determine $\sum_{n=0}^{K-1} p_n z^n$ in the numerator. For this we have to make use of Rouché's theorem from the theory of complex variables. Being a PGF, $P(z)$ must converge inside the unit circle. The denominator of (6.2.2) has $K+1$ zeros. Thus, for $P(z)$ to be a proper PGF, the numerator of (6.2.2) must vanish at these $K+1$ zeros. It is easily seen that $z = 1$ is a zero of the numerator as well as the denominator. Appealing to Rouché's theorem (we leave out the details of using the theorem here because its theory is beyond the scope of this text; interested readers may refer to more advanced books on queueing theory), we can show that exactly $K-1$ zeros of the denominator are within the unit circle, leaving one zero lying outside. Let z_0 (> 1) be the root of the equation

$$\left(\frac{\lambda}{\mu}\right) z^{K+1} - \left(\frac{\lambda}{\mu}+1\right) z^K + 1 = 0. \tag{6.2.3}$$

Clearly if we divide the denominator of (6.2.2) by $(z-1)(z-z_0)$, we are left with a polynomial with $K-1$ roots within the unit circle. The portion of the numerator with zeros within the unit circle is $\sum_{n=0}^{K-1} p_n z^n$; therefore, this function and the leftover of the denominator can differ by at most a multiplicative constant. We get

$$\sum_{n=0}^{K-1} p_n z^n = C \frac{(\lambda/\mu)\, z^{K+1} - \left(\frac{\lambda}{\mu}+1\right) z^K + 1}{(z-1)(z-z_0)}. \tag{6.2.4}$$

Substituting this result in (6.2.2), we have

$$\begin{aligned} P(z) &= \frac{C(1-z^K)}{(z-1)(z-z_0)} \\ &= \frac{C}{z_0-z}\sum_{n=0}^{K-1} z^n. \end{aligned} \tag{6.2.5}$$

Since $P(1)=1$, setting $z=1$ in (6.2.5), we get

$$C = \frac{z_0-1}{K}$$

and

$$P(z) = \frac{(z_0-1)\sum_{n=0}^{K-1} z^n}{K(z_0-z)}. \tag{6.2.6}$$

The right-hand side of (6.2.6) can be expanded as a power series in z to determine $p_n,\ n=0,1,2,\ldots$ explicitly as follows:

$$P(z) = \frac{z_0-1}{Kz_0}\left(\sum_{s=0}^{K-1} z^s\right)\left(\sum_{r=0}^{\infty}\left(\frac{z}{z_0}\right)^r\right) \tag{6.2.7}$$

$$\begin{aligned} p_n &= \frac{z_0-1}{Kz_0}\sum_{r=0}^{n}\left(\frac{1}{z_0}\right)^r && n<K \\ &= \frac{z_0-1}{Kz_0^{n-K+2}}\sum_{r=0}^{K-1}\left(\frac{1}{z_0}\right)^r && n\geq K. \end{aligned} \tag{6.2.8}$$

Noting that

$$\sum_{r=0}^{n}\left(\frac{1}{z_0}\right)^r = \frac{1-(\frac{1}{z_0})^{n+1}}{1-(\frac{1}{z_0})}$$

(6.2.8) can be presented as

$$\begin{aligned} p_n &= \frac{z_0^{n+1}-1}{Kz_0^{n+1}} && n<K \\ &= \frac{z_0^{K-1}}{Kz_0^{n+1}} && n\geq K. \end{aligned} \tag{6.2.9}$$

As mentioned above, finding the root z_0 lying outside the unit circle of the equation (6.2.3) is essential in the determination of the limiting distribution. This is a common problem in the analysis of systems of this type and there are root-finding algorithms available specifically applicable in such cases. For elaboration on the appropriate root finding algorithms, the readers may refer to Chaudhry and Templeton (1983) and journal articles by M. L. Chaudhry and his associates (see for instance Chaudhry et al. (1992)).

6.3 Imbedded Markov Chain Analysis of Bulk Queues $M^{(Z)}/G/1$ and $G/M^{(Z)}/1$

In the queue $M^{(Z)}/G/1$, let customers arrive in groups of size Z, an integer valued r.v. (> 0), with the distribution

$$P(Z = r) = d_r \quad r = 1, 2, \ldots. \tag{6.3.1}$$

Except for this factor, all other characteristics of the system are the same as the $M/G/1$ queue analyzed in Section 5.2. Consequently, the system $M^{(Z)}/G/1$ can be analyzed on the same lines as described in Section 5.2 with the necessary modification for the distribution $\{k_j\}$ of the number of customers arriving during a service interval. It can be given as (see 5.2.1)

$$k_j = \int_0^\infty e^{-\lambda t} \sum_{r=0}^{j} \frac{(\lambda t)^r}{r!} d_j^{(r)} dB(t) \quad j = 0, 1, 2, \ldots, \tag{6.3.2}$$

where $d_j^{(r)}$ denotes the r-fold convolution of d_j with itself (meaning, that it is the probability of r groups bringing in j customers).

Since service is provided for one customer at a time, (5.2.2) holds as it is; so does the transition probability matrix. The only change that needs to be made is in the determination of $K(z)$ of (5.2.7). We get

$$K(z) = \int_0^\infty e^{-\lambda t} \sum_{j=0}^{\infty} z^j \sum_{r=0}^{j} \frac{(\lambda t)^r}{r!} d_j^{(r)} dB(t). \tag{6.3.3}$$

Interchanging the summations in (6.3.3)

$$K(z) = \int_0^\infty e^{-\lambda t} \sum_{r=0}^{\infty} \frac{(\lambda t)^r}{r!} \sum_{j=r}^{\infty} d_j^{(r)} z^j dB(t). \tag{6.3.4}$$

Let

$$D(z) = \sum_{k=1}^{\infty} d_k z^k \quad |z| \leq 1.$$

A well-known property of PGF is that the PGF of the sum of independent random variables is the product of PGFs of the random variables. Using this property to $d_j^{(r)}$, which is the probability distribution of the sum of r random variables representing group sizes, we get

$$\sum_{j=r}^{\infty} d_j^{(r)} z^j = [D(z)]^r. \tag{6.3.5}$$

Using this result in (6.3.4), we get

$$\begin{aligned} K(z) &= \int_0^\infty e^{-\lambda t} \sum_{r=0}^{\infty} \frac{\lambda t)^r}{r!} [D(z)]^r dB(t) \\ &= \int_0^\infty e^{-\lambda t} e^{\lambda t D(z)} dB(t) \\ &= \psi(\lambda - \lambda D(z)). \end{aligned} \tag{6.3.6}$$

Accordingly

$$\begin{aligned} K'(z) &= -\lambda D'(z) \psi'(\lambda - \lambda D(z)) \\ K'(1) &= -\lambda D'(1) \psi'(0). \end{aligned}$$

Note that $D'(1) = \sum_k k d_k = E(X) = d$, say. Also, $-\lambda \psi'(0) = b$, mean service time. Thus, we have

$$K'(1) = \lambda d b = \rho \tag{6.3.7}$$

showing that the traffic intensity now has the value $\rho = \lambda d b$.

Another term that needs modification if $K''(1)$, that occurs in (5.2.25) and therefore in $E(Q^*)$. We have

$$\begin{aligned} K''(z) &= [-\lambda D'(z)]^2 \psi''(\lambda - \lambda D(z)) \\ &\quad + \psi'(\lambda - \lambda D(z))[-\lambda D''(z)] \\ K''(1) &= \lambda^2 d^2 \psi''(0) - \lambda D''(1) \psi'(0) \\ &= \lambda^2 d^2 E(S^2) + b D''(1). \end{aligned} \tag{6.3.8}$$

But

$$\begin{aligned} D''(z) &= \sum_k k(k-1) d_k z^{k-2} \\ D''(1) &= E(Z^2) - E(Z) \end{aligned}$$

giving

$$K''(1) = \lambda^2 d^2 E(S^2) + b E(Z^2) - d$$

and

$$L = E(Q^*) = \rho + \frac{\lambda^2 d^2 E(S^2) + b E(Z^2) - d}{2(1-\rho)}. \tag{6.3.9}$$

An expression for $V(Q^*)$ can be obtained in a similar manner.

Modifications to be made in the analysis of $G/M/1$ to obtain results for the queue $G/M^{(Z)}/1$ are similar. The PGF $\beta(z)$ of the number of potential service completions in one inter-arrival period has the form

$$\beta(z) = \phi(\mu - \mu D(z)) \tag{6.3.10}$$

and the traffic intensity ρ is given by

$$\rho = \frac{1}{ad\mu}. \tag{6.3.11}$$

The limiting distribution of the number of customers in the system just before a customer arrival is the same as (5.3.13) where, ζ is now the least positive root of the equation $z = \beta(z)$, where $\beta(z)$ has the form given in (6.3.10).

6.4 The Queues $M/E_k/1$ and $E_k/M/1$

In Section 2.1, we defined the Erlang distribution, E_k, as the distribution of the sum of k independent and identically distributed (i.i.d.) exponential random variables. From Chapter 8, we note that Erlang E_k is the simplest phase-type distribution and it represents the distribution of the time taken by a Markov process to traverse k phases of exponential service. We may use this representation to provide a Markov model for the number of customers in the system in queues $M/E_k/1$ and $E_k/M/1$.

Let us first consider the queue $M/E_k/1$. Suppose arrivals occur in a Poisson process with rate λ. Let service be provided by a single server with service time distribution

$$f(x) = e^{-k\mu x}\frac{(k\mu x)^{k-1}k\mu}{(k-1)!} \quad x > 0 \tag{6.4.1}$$

which has the mean $= 1/\mu$. Using the observations made in Section 2.1 and Appendix A, we may note that when the service time has the Erlang distribution (6.4.1), it can be considered as made up of k phases, each with an exponential distribution with density $k\mu e^{-k\mu x}$ $(x > 0)$, which has a mean $1/k\mu$. Thus, if we associate a number representing the number of phases of service yet to be used (we use the unexpended number for convenience) for the customer being served along with the number of customers in the system, we have a representation of the state of the process that can be considered Markovian. Using {(number of customers in the system, number of phases of service yet to be used)} as the bivariate process, the state space can be represented as $\{(0,0);(1,0),(1,1),\ldots,(1,k);(2,0),(2,1),\ldots,(2,k);...\}$. Defining the limiting distribution of this process $\{p_{n_1,n_2},\ n_1 = 0,1,2,\ldots;n_2 = 0,1,\ldots,k\}$ appropriately, we may write down the state balance equations and solve them using PGFs. (See Prabhu (1997) for details.)

An alternative method is to count the number of exponential phases waiting to be served or in service. When there are n customers in the system and the number of phases of service yet to be used for the customer in service is r, the total count for the number of phases is $(n-1)k + r$. In order to be able to use this approach, each arriving customer should be thought of as bringing k phases of service to the system. Accordingly consider an $M^k/M/1$ queue, in which customers arrive in a Poisson process in groups of size k. The rate of arrival for groups is λ. Each customer demands service that has an exponential

distribution with mean $1/k\mu$. The total number of phases waiting for or in service in this system is the same as the total number of phases waiting for or in service in an $M/E_k/1$ system. The limiting distribution of the state of the system is given in Section 6.1. Since these results are in terms of the corresponding number phases, all we need now is a procedure to convert the phases into the corresponding number of customers.

As described above, when there are n customers and r service phases yet to be used in the system, the total phase count is $(n-1)k+r$. Reversing this procedure, when there are a total number of n phases in the system, the number of customers in the system can be obtained as $\left[\frac{n}{k}\right]+1$, where $[\ ]$ signifies the largest integer contained in $\frac{n}{k}$, when n is not a multiple of k and $\frac{n}{k}$ when n is a multiple of k.

Let $\{p_n,\ n=0,1,2,\ldots\}$ be the limiting distribution of the number of customers in an $M/E_k/1$ system and let $\{p_n^{(b)},\ n=0,1,2,\ldots\}$ be the limiting distribution of the corresponding group arrival queue $M^k/M/1$ for the phases. We then have

$$\begin{aligned} p_n &= \sum_{j=(n-1)k+1}^{nk} p_j^{(b)} \qquad n \geq 1 \\ p_0 &= p_0^{(b)}. \end{aligned} \tag{6.4.2}$$

Another alternative analysis of $M/E_k/1$ is via the imbedded Markov chain approach of Section 5.2. The Laplace transform (5.2.6) of the service time distribution takes the form

$$\psi(\theta) = \left(\frac{k\mu}{k\mu+\theta}\right)^k \tag{6.4.3}$$

and the PGF (5.2.7) of the number of customers arriving during a service period is now given by

$$K(z) = \left(\frac{k\mu}{k\mu+\lambda-\lambda z}\right)^k. \tag{6.4.4}$$

The resulting PGF of the limiting distribution $\{\pi_j,\ j=0,1,2,\ldots\}$ has the form

$$\Pi(z) = \frac{(1-\rho)(z-1)(k\mu)^k}{z(k\mu+\lambda-\lambda z)^k-(k\mu)^k}. \tag{6.4.5}$$

Since the arrival process is Poisson, the limiting distribution $\{\pi_j\}$ of the imbedded Markov chain $\{Q_n\}$ and the limiting distribution $\{p_n\}$ of its continuous time analog $\{Q(t)\}$, are the same in the queue $M/E_k/1$. Hence, as a practical matter, any of the alternative procedures suggested above should lead to the same result.

Similar alternative procedures for analysis can be suggested for the queue $E_k/M/1$.

(i) *The use of a bivariate Markov process*

When the inter-arrival times are distributed as Erlang E_k, each of them may be considered to have made up of k exponentially distributed phases. Now keeping track of the number of elapsed phases in an inter-arrival time helps in defining the Markov process with the number of customers in the system as the first variable and the number of elapsed exponential inter-arrival phases as the second variable. State balance equations may be written down for the bivariate process and solved using PGF. For details, see Prabhu (1997).

(ii) *Using the $M/M^k/1$ model*

In addition to a real customer arriving at the end of the kth exponential phase of an Erlangian inter-arrival time, we may assume that $k-1$ virtual customers arrive at the end of the preceding $k-1$ phases. Since these virtual customers are associated with a real customer, all k customers (one real and $k-1$ virtual) will have to be served as a group. The modified system now has a single server, customer arrivals in a Poisson process in such a way that every kth customer is real which is preceded by $k-1$ virtual customers, and all these k customers are served in a group. Then the number of "customers" (which include real and virtual customers) can be modeled as an $M/M^k/1$ queue. The limiting distribution of the number of customers in the model as given in Section 6.2 gives the limiting distribution of the number of "customers" in the $E_k/M/1$ queue. Let $\{p_n,\ n = 0,1,2\ldots\}$ be the limiting distribution of the number of customers in the queue $E_k/M/1$ and $\{p_n^{(b)},\ n = 0,1,2,\ldots\}$ be the limiting distribution of the number of customers in the $M/M^k/1$ queue. Then $\{p_n,\ n = 0,1,2,\ldots\}$ can be determined using the relation

$$\begin{aligned} p_n &= \sum_{j=nk}^{nk+k-1} p_j^{(b)} \quad n = 1,2,\ldots \\ p_0 &= \sum_{j=0}^{k-1} p_j^{(b)}. \end{aligned} \tag{6.4.6}$$

(iii) *Using the imbedded Markov chain of $E_k/M/1$ as a special case of $G/M/1$ described in Section 5.3*

Let the inter-arrival time distribution be given as

$$f(x) = e^{-k\lambda x}\frac{(k\lambda x)^{k-1}k\lambda}{(k-1)!} \quad x > 0. \tag{6.4.7}$$

The Laplace transform of (6.4.7) takes the form

$$\phi(\theta) = \left(\frac{k\lambda}{k\lambda+\theta}\right)^k \tag{6.4.8}$$

and the PGF $\beta(z)$ of the number of potential services during an inter-arrival period can be given as

$$\beta(z) = \left(\frac{k\lambda}{k\lambda + \mu - \mu z}\right)^k. \tag{6.4.9}$$

The limiting distribution of the number of customers in the system just before an arrival is obtained as

$$\pi_j = (1-\zeta)\zeta^j \qquad j = 0, 1, 2 \ldots, \tag{6.4.10}$$

where ζ is the least positive root of the equation

$$z = \beta(z). \tag{6.4.11}$$

6.5 The Bulk Queues $M/G^K/1$ and $G^K/M/1$

In the last three sections, we have assumed that both inter-arrival times and service times of customers are exponential or Erlangian, thus making it easy to use Markov processes (although in an extended sense) in the analysis. Here we briefly describe a practical approach based on imbedded Markov chains that can be used when one of the element distributions does not have the nice properties of the exponential distribution even when arrival or service occurs in groups. For readers interested in the continuous time analog of the results that can be derived using imbedded Markov chains, the best comprehensive reference seems to be Chaudhry and Templeton (1983).

Let us first consider the queue $M/G^K/1$ with the following description. Customers arrive one at a time in a Poisson process with rate λ. There is a single server providing service to groups of exactly K customers at a time. The service times have a general distribution $B(\cdot)$. If there are less than K customers waiting in the queue at the completion of a service, server waits until the number K is reached to start the service. It should be noted that we have made this policy assumption for convenience. Modifications to this policy, such as starting service with at least a specified number of customers less than K, require making appropriate changes to the expressions.

Let $\{Q_n, n = 0, 1, 2, \ldots\}$ be the number of customers in the system soon after the nth group departure. Let X_n be the number of customers arriving during the nth service. Following the arguments used in Section 5.2, for the distribution $\{k_j,\ j = 0, 1, 2, \ldots\}$ of X_n we have

$$k_j = P(X_n = j) = \int_0^\infty e^{-\lambda t}\frac{(\lambda t)^j}{j!}dB(t) \quad j = 0, 1, 2, \ldots. \tag{6.5.1}$$

We also have the random variable relationship between Q_n and Q_{n+1}

$$Q_{n+1} = \begin{cases} Q_n + X_{n+1} - K & \text{if } Q_n > K \\ X_{n+1} & \text{if } Q_n \le K. \end{cases} \tag{6.5.2}$$

(The justification for this relationship is exactly the same as given following (5.2.2) except that now we need K customers to start the service instead of 1.)

For $P_{ij} = P(Q_{n+1} = j | Q_n = i)$, (6.5.2) gives

$$\begin{aligned} P_{ij} &= \begin{cases} P(i + X_{n+1} - K = j) & \text{if } i > K \\ P(X_{n+1} = j) & \text{if } i \le K \end{cases} \\ &= \begin{cases} k_{j-1+K} & \text{if } i > K \\ k_j & \text{if } i \le K. \end{cases} \end{aligned} \tag{6.5.3}$$

Displaying these probabilities in a matrix form, we get the transition probability matrix $\mathbf{P}$ of the imbedded Markov chain

$$\mathbf{P} = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ \vdots \\ K \\ K+1 \\ \vdots \end{array} \begin{array}{c} \begin{array}{ccc} 0 & 1 & \dots \end{array} \\ \begin{bmatrix} k_0 & k_1 & \dots \\ k_0 & k_1 & \dots \\ k_0 & k_1 & \dots \\ \vdots & \vdots & \\ k_0 & k_1 & \dots \\ & k_0 & \dots \\ & \vdots & \end{bmatrix} \end{array}. \tag{6.5.4}$$

Comparing (6.5.4) with (5.2.5), we note that (6.5.4) has $K + 1$ identical rows, instead of 2 as in (5.2.5). If we are interested in a mathematical expression for the limiting distribution of the Markov chain, we proceed the same way as in Section 5.2. In order to completely specify the PGF of the distribution, it would be necessary to specify the zeros of its denominator using Rouché's theorem. However as a practical approach, in this age of computers, we may obtain $\mathbf{P}^n$ by matrix multiplication as suggested in Section 5.2. Note that the elements k_j's of the matrix are known (can be determined numerically) and the limiting distribution is given by $\lim_{n\to\infty} \mathbf{P}^n$. Also recall that the limiting matrix has identical rows.

The imbedded Markov chain analysis of the queue length process in queue $G^K/M/1$ follows the method outlined in Section 5.3 by considering the number of customers in the system just before arrival points. Let $A(\cdot)$ be the distribution function of the inter-arrival times and $f(x) = \mu e^{-\mu x}$ $(x > 0)$ be the service time distribution. We assume that customers arrive in groups of constant size K. (If we assume variable group sizes, we have to incorporate the group size distribution in our analysis.) For reasons explained following equation (5.3.2), we define X_{n+1} as the number of potential departures during the $(n + 1)$th inter-arrival period. Let $\{Q_n, n = 0, 1, 2, \dots\}$ be the number of customers in the system just before the nth group arrival. Analogous to (5.3.2) we have

$$Q_{n+1} = \begin{cases} Q_n + K - X_{n+1} & \text{if } Q_n + K - X_{n+1} > 0 \\ 0 & \text{if } Q_n + K - X_{n+1} \le 0. \end{cases} \tag{6.5.5}$$

Let $P(X_n = j) = b_j$, $j = 0, 1, 2, \ldots$ as given in (5.3.1). Following the steps used in Section 5.3, for the transition probability $P_{ij} = P(Q_{n+1} = j | Q_n = i)$, from (6.5.5) we get

$$P_{ij} = \begin{cases} P(i + K - X_{n+1} = j) & \text{if } j > 0 \\ P(i + K - X_{n+1} \leq 0) & \text{if } j = 0 \end{cases} \tag{6.5.6}$$

$$= \begin{cases} b_{i+K-j} & j > 0 \\ \sum_{i+K}^{\infty} b_r & j = 0. \end{cases} \tag{6.5.7}$$

The transition probability matrix has the form

$$\mathbf{P} = \begin{array}{c} \\ 0 \\ 1 \\ \\ K \\ \\ \end{array} \begin{array}{c} \begin{array}{cccc} 0 & 1 & 2 & \ldots \end{array} \\ \begin{bmatrix} \sum_{K}^{\infty} b_r & b_{K-1} & b_{K-2} & \ldots \\ \sum_{K+1}^{\infty} b_r & b_K & b_{K-1} & \ldots \\ \vdots & & & \\ \sum_{2K}^{\infty} b_r & b_{2K-1} & b_{2K-2} \ldots & \\ \vdots & & \vdots & \end{bmatrix} \end{array}. \tag{6.5.8}$$

For the limiting distribution $\boldsymbol{\pi} = (\pi_0, \pi_1, \pi_2, \ldots)$ we may use the standard procedure of solving equation $\boldsymbol{\pi}\mathbf{P} = \boldsymbol{\pi}$ and $\sum_0^{\infty} \pi_j = 1$ on the same lines as illustrated in Section 5.3. But as a practical matter, since the elements of $\mathbf{P}$ can be determined numerically from (5.3.1), obtaining a close approximation to $\lim_{n\to\infty} \mathbf{P}^n$ by matrix multiplication is likely to be simpler.

For the determination of analytical and numerical solutions in bulk queueing systems that lead to algorithmic procedures, the matrix-analytic solution techniques developed by M. F. Neuts and his associates and as detailed in Chapter 8 are highly recommended. A thorough knowledge of matrix analysis is essential in understanding them. See also Chaudhry and Templeton (1983) and subsequent articles by Professor Chaudhry and his associates as an alternative set of references on bulk queueing systems.

6.6 The Queues $E_k/G/1$ and $G/E_k/1$

The queueing systems $E_k/G/1$ and $G/E_k/1$ can also be analyzed as bulk queueing systems $M/G^k/1$ and $G^k/M/1$, respectively on the same lines as described in Section 6.4. Since the results of the bulk queueing analysis are in terms of the number of phases in the system, appropriate conversion has to be made to give results in terms of the number of customers in the system.

A better alternative is the use of Neuts' matrix-analytic solution technique on the bivariate Markov chain. In the queue $E_k/G/1$, the state space of the imbedded Markov chain is given by two variables {number of customers in the

system; number of elapsed exponential phases since the last arrival}. In the queue $G/E_k/1$, the state space is characterized by two variables {number of customers in the system, number of exponential service phases yet to be used for customer in service}. For more details, see Chapter 8 and references provided in it.

6.7 The Queue *M/D/s*

When jobs are mechanized in manufacturing systems constant service times are common. Also jobshops employ multiple machines in parallel to maintain job flow. Under these circumstances a Poisson arrival, constant service time, and multiple server queueing system is a natural model. Let customers arrive in a Poisson process with rate λ, the service time be a constant b, and let the number of servers be s. Even though this does not look like a Markovian system, an imbedded Markov chain can be identified in the queue length process of this system.

Let $(0, b, 2b, 3b, \ldots)$ be the epochs of observation on the time axis. Define $Q(t)$ as the number of customers in the system at time t and $Q_n = Q(nb)$. Let $\{X_n,\ n = 1, 2, \ldots\}$ be the number of customers arriving during $[(n-1)b, nb]$. We have

$$k_j = P(X_n = j) = e^{-\lambda b}\frac{(\lambda b)^j}{j!} \quad j = 0, 1, 2 \ldots . \tag{6.7.1}$$

Considering the number of arrivals and service completions during $[nb, (n+1)b]$, we may write

$$Q_{n+1} = \begin{cases} Q_n + X_{n+1} - s & Q_n > s \\ X_{n+1} & Q_n \leq s. \end{cases} \tag{6.7.2}$$

This relationship can be justified by noting that if there are $> s$ customers in the system at time nb, s of them will be in service and will depart before $(n+1)b$. If the number of customers in the system at time nb is $\leq s$, all of them will depart before $(n+1)b$, and X_{n+1}, the number of customers arriving during $[nb, (n+1)b]$ will be left in the system at $(n+1)b$.

The relationship (6.7.2) is exactly the same as (6.4.2) obtained for the bulk queue $M/G^K/1$ (with s replacing K). Thus $\{Q_n,\ n = 0, 1, \ldots\}$ is a Markov chain imbedded in the queue length process. Its transition probability matrix is given by (6.5.4) with K replaced by s, and its analysis follows on similar lines.

6.8 The Queue *M/M/*1 with Priority Disciplines

Queue disciplines which assign priority service to some customers is common in service systems. The priority can be based on factors such as customer class, the type of service, and even the length of service. With the advent of computers, a wide variety of priority disciplines have been introduced for

improving system performance. The round robin, shortest processing time, and earliest due date first disciplines are some of the examples. Since the analysis of most of the variants involve underlying processes much more complex than those we consider in this text, we shall not delve into them here. However, we introduce the simple two-class priority model under the $M/M/1$ setting and discuss the fundamental issues to be confronted in its analysis.

To begin with, when we consider priority queues the following factors need attention:

1. There are more than one class of customers based on their needs or importance to the system.

2. The customers in one class have a higher priority for service than others. When there are more than two classes, we can arrange them in a hierarchy of priorities with regard to service.

3. The priority accorded to a class of customers can be *preemptive* or *non-preemptive*. If a customer has preemptive priority over another customer, the priority customer will preempt the non-priority customer for service. If the priority is non-preemptive, the priority customer will enter service on the completion of the ongoing service at the time of its arrival. (Such a priority discipline is also known as *head of the line priority discpline.*)

4. When preemption of service is allowed, the service to the preempted customer can be resumed after the priority customers are served, from the point the service was preempted or starting from the beginning all over again. These two alternatives are known as *preemptive resume* and *preemptive repeat* disciplines, respectively. For the purpose of analysis, preemptive repeat discipline can be further divided into *different* and *identical*, depending on the service time selected while resuming service. Under *preemptive repeat different* discipline, the sample realization of service is different from the one originally chosen, while under *preemptive repeat identical* discipline, the same sample realization is used.

5. As long as the priority assignment is not based on the length of service and there is no preemption, the total number of customers in the priority system has the same distribution as the number of customers in the system without priorities. Thus in this case, the priority discipline affects only the waiting time of the class of customers entering the system.

Consider an $M/M/1$ queue with two priority classes. Let class 1 customers have higher priority for service over class 2 customers. Also, let the Poisson arrival and exponential service rates of the customers of the two classes be as follows:

Class 1—arrival rate λ_1; service rate μ_1

Class 2—arrival rate λ_2; service rate μ_2.

Let us first assume a non-preemptive priority system with service provided by a single server. Because of the Poisson arrival and exponential service time assumption, we may use a generalized birth and death process model for this system. The state space of the underlying Markov process has to be represented with three components: (number of class 1 customers; number of class 2 customers; class of customer in service). Let $\{p_{mnr},\ m, n = 0, 1, 2, \ldots; r = 1, 2\}$ be the limiting distribution of the state space for the process. When $m = n = 0$, for convenience we denote the probability by p_0. In general, the elements of the state space are

$$\{0, 101, 012, m01, 0n2, mn1, mn2; m, n \geq 1\}.$$

The following state balance equations, written down using the principles described in Section 4.1, determine the limiting distribution of the state of the process $(\lambda = \lambda_1 + \lambda_2)$

$$\begin{aligned}
\lambda p_0 &= \mu_1 p_{101} + \mu_2 p_{012} \\
(\lambda + \mu_1) p_{101} &= \lambda_1 p_0 + \mu_1 p_{201} + \mu_2 p_{112} \\
(\lambda + \mu_2) p_{012} &= \lambda_2 p_0 + \mu_1 p_{111} + \mu_2 p_{022} \\
(\lambda + \mu_1) p_{m01} &= \lambda_1 p_{m-1,01} + \mu_1 p_{m+1,01} + \mu_2 p_{m12} && m \geq 1 \\
(\lambda + \mu_2) p_{0n2} &= \lambda_2 p_{0,n-1,2} + \mu_1 p_{1n1} + \mu_2 p_{0,n+1,2} && n \geq 1 \\
(\lambda + \mu_2) p_{m12} &= \lambda_1 p_{m-1,12} && m \geq 1 \\
(\lambda + \mu_1) p_{1n1} &= \lambda_2 p_{1,n-1,1} + \mu_1 p_{2n1} + \mu_2 p_{1,n+1,2} && n \geq 1 \\
(\lambda + \mu_1) p_{mn1} &= \lambda_1 p_{m-1,n1} + \lambda_2 p_{m,n-1,1} \\
&\quad + \mu_1 p_{m+1,n1} + \mu_2 p_{m,n+1,2} && m \geq 2,\ n \geq 1 \\
(\lambda + \mu_2) p_{mn2} &= \lambda_1 p_{m-1,n2} + \lambda_2 p_{m,n-1,2} && m \geq 1,\ n \geq 2
\end{aligned} \tag{6.8.1}$$

and $\sum p_{mnr} = 1$.

For a complete analytical solution of these equations the best approach is to use PGFs. See Morse (1958), Miller (1981), and Gross et al. (2008) for details. However, when the number of customers allowed in the system is small, these equations can be solved numerically. Before embarking on a numerical solution, one should note that the number of equations (accordingly the number of unknowns) can get very large even with low capacity limits. If K $(K > 2)$ is the number allowed in the system, the number of equations turns out to be $K^2 + K + 1$. Thus if $K = 10$, one should be ready to deal with 111 equations. Example 6.8.1 is illustrative of the procedure when $K = 2$.

Without going through the analysis, we state the following conclusions which are useful in understanding the effect of priority assignment on customers as well as the system.

1. When the service rates of the two classes of customers are different, the mean number of low-priority customers waiting is larger and the mean

number of high-priority customers waiting is smaller than the corresponding means with a two-class system with no priorities.

2. For the queueing system, the priority scheme is useful only if the service rate of the higher priority customers is larger than the service rate of the lower priority customers.

3. If the reduction in the number of customers waiting in the system is a design criterion, extending the conclusion in 2 above, we can infer that the priority scheme known as the *shortest processing time* discipline is optimal. In this queue discipline the customer with the shortest service time gets the highest priority.

4. Because of the non-preemptive nature of the priority discipline, the mean waiting time of the high-priority customer will be larger than its mean waiting time under a preemptive priority. This difference is equal to the mean service time of the low-priority customer conditioned on the arrival of the high-priority customer when there are no high-priority customers in the system. This can be obtained as

$$\frac{1}{\mu_2} \cdot \frac{\lambda_1}{\lambda} \left(\sum_{n=1}^{\infty} p_{0n2} \right) = \frac{\lambda_1 \lambda_2}{\lambda \mu_2^2}. \tag{6.8.2}$$

5. The derivation of the mean waiting time of the low-priority customer is much more complex. Suppose when a low-priority customer arrives there are m high-priority (class 1) and n low-priority (class 2) customers in the system. Then the components of its waiting time are the following:

   (a) The remaining service time of the customer in service

   (b) Total length of the busy periods of class 1 customers who arrive during the remaining service time in (a)

   (c) Total length of the m class 1 busy periods

   (d) Total length of the busy periods initiated by class 1 customers who arrive during the service times of n class 2 customers

For details see Cobham (1954). The key point to note here is that for every class 1 customer in the system the class 2 customer will be delayed by an amount of time equivalent to the length of the busy period initiated by that class 1 customer.

Let us now consider the effect of preemption on the service of the low-priority customer due to the arrival of a high-priority customer. Consider the $M/M/1$ system with two priority classes as described above with incorporation of preemption in the priority queue discipline. Suppose when a lower priority (class 2) customer is in service, on the arrival of a high-priority customer (class 1), its service is terminated and the new arrival is taken into service right away.

Since the service time of the low-priority customer is exponential, the distribution of the remaining service time of the low-priority customer has the same distribution as the original one.

Due to the preemptive nature of priority, when there is a high-priority customer in the system, it will be in service. Hence the state space can be identified with only two components: (number of class 1 customers in the system; number of class 2 customers in the system). Let $\{p_{mn}, m, n > 0\}$ and p_0 be the limiting probabilities for the number of customers in the system. For state balance equations we have ($\lambda = \lambda_1 + \lambda_2$)

$$\begin{aligned} \lambda p_0 &= \mu_1 p_{10} + \mu_2 p_{01} \\ (\lambda + \mu_1) p_{m0} &= \lambda_1 p_{m-1,0} + \mu_1 p_{m+1,0} && m \geq 1 \\ (\lambda + \mu_2) p_{0n} &= \lambda_2 p_{0,n-1} + \mu_1 p_{1n} + \mu_2 p_{0,n+1} && n \geq 1 \\ (\lambda + \mu_1) p_{mn} &= \lambda_1 p_{m-1,n} + \lambda_2 p_{m,n-1} + \mu_1 p_{m+1,n} && m, n > 0 \end{aligned} \quad (6.8.3)$$

and $\sum p_{mn} = 1$.

Again for the complete analytical solution of these equations, the use of PGFs is essential. As in the case of non-preemptive priority, if we think of numerical solutions of these equations when the capacity limit of the system is K, number of equations to be solved will be $2K + \frac{K(K-1)}{2} + 1$. Thus if $K = 10$, the number of equations in this case is 66 as compared to 111 for the non-preemptive case.

When there are only two priority classes and the priority is preemptive, as far as the higher priority customer is concerned, the system performs just like a regular $M/M/1$ system. But the effect on the low-priority customer is two fold: On the waiting time as well as the service time. Here we define waiting time as the amount of time between the customer's arrival epoch and the time point it is taken into service for the first time. (Note that because of preemption, the customer's service could be interrupted repeatedly.)

Let us first consider the time between the moment a low-priority customer enters service for the first time and the time point at which it departs after completion of service. When it is interrupted because of the arrival of a high-priority customer, it can get back into service only after the service of the high-priority customer and the corresponding busy period initiated by that service. This happens with every high-priority arrival. Hence, the amount of time between the moment low-priority customer enters service for the first time and the time it departs from the system is made up of its service time and r busy periods of high-priority customers, where r is the number of high-priority customers arriving during the low-priority customer's service time. This time period is known as *completion time* in the literature (Jaiswal (1968)). Other terms used to identify this time period are *server sojourn time* and *residence time*.

Suppose there are m high-priority customers and n low-priority customers at the time of the arrival of the low-priority customer. Then the waiting time of this customer is made up m high-priority busy periods and n completion times.

In the foregoing discussion, for the sake of simplicity, we have used only two priority classes. If there are more classes of customers, normally there would be a hierarchy of priorities, say, $1, 2, \ldots$. Then for any class i, $i = 1, 2, \ldots, r$, its performance would be affected by the performance of classes $1, 2 \ldots, i-2$, through the performance of class $i-1$. For instance class $i-1$ completion time is the high-priority service time as seen by the class i customer. We shall not go into a discussion of such systems because of its complexity. Interested readers may refer to Jaiswal (1968) and journal articles that have appeared since then.

The example given below illustrates the complexities involved in solving a problem with the simplest priority discipline.

Example 6.8.1 A service center is set up primarily to provide one type of service, which we identify as class 1. However, in order to ensure that the service personnel stay busy as much as possible, it accepts another type of service, called class 2, on a low-priority basis. At any time only two customers are allowed to be present in the system. Let λ_1 and λ_2 be the Poisson arrival rates of these two classes of customers and μ_1 and μ_2 be their service rates on the assumption that the service times are exponential.

Let us determine the limiting distribution of the number of customers in the system under non-preemptive and preemptive priority disciplines.

Non-preemptive Priority This discipline assumes that once a non-priority customer starts service, it is carried out to conclusion even if a priority customer arrives in the mean time. The states representing the number of customers in the system and the class of customer in service are: $(0, 101, 012, 201, 022, 111, 112)$. The state balance equations are given below ($\lambda = \lambda_1 + \lambda_2$).

$$\begin{aligned}
\lambda p_0 &= \mu_1 p_{101} + \mu_2 p_{012} && (6.8.4)\\
(\lambda + \mu_1) p_{101} &= \lambda_1 p_0 + \mu_1 p_{201} + \mu_2 p_{112} && (6.8.5)\\
(\lambda + \mu_2) p_{012} &= \lambda_2 p_0 + \mu_1 p_{111} + \mu_2 p_{022} && (6.8.6)\\
\mu_1 p_{201} &= \lambda_1 p_{101} && (6.8.7)\\
\mu_2 p_{022} &= \lambda_2 p_{012} && (6.8.8)\\
\mu_1 p_{111} &= \lambda_2 p_{101} && (6.8.9)\\
\mu_2 p_{112} &= \lambda_1 p_{012} && (6.8.10)\\
\sum p_{mnr} &= 1.
\end{aligned}$$

Substituting from (6.8.7) and (6.8.10) in (6.8.5),

$$\begin{aligned}
(\lambda + \mu_1) p_{101} &= \lambda_1 p_0 + \lambda_1 p_{101} + \lambda_1 p_{012}\\
(\lambda_2 + \mu_1) p_{101} - \lambda_1 p_{012} &= \lambda_1 p_0. && (6.8.11)
\end{aligned}$$

But from (6.8.4),

$$\mu_1 p_{101} + \mu_2 p_{012} = \lambda p_0.$$

Solving for p_{101} and p_{012},

$$\begin{aligned} p_{101} &= \frac{\lambda_1(\lambda+\mu_2)}{\lambda_1\mu_1+\lambda_2\mu_2+\mu_1\mu_2}p_0 \\ &= Ap_0, \text{ say} \end{aligned} \tag{6.8.12}$$

$$\begin{aligned} p_{012} &= \frac{1}{\mu_2}\left[\lambda - \frac{\lambda_1\mu_1(\lambda+\mu_2)}{\lambda_1\mu_1+\lambda_2\mu_2+\mu_1\mu_2}\right]p_0 \\ &= Bp_0, \text{ say.} \end{aligned} \tag{6.8.13}$$

Substituting these values in (6.8.7)–(6.8.10), we get

$$\begin{aligned} p_{201} &= \frac{\lambda_1}{\mu_1}Ap_0 \quad p_{111} = \frac{\lambda_2}{\mu_1}Ap_0 \\ p_{022} &= \frac{\lambda_2}{\mu_2}Bp_0 \quad p_{112} = \frac{\lambda_1}{\mu_2}Bp_0. \end{aligned} \tag{6.8.14}$$

Now p_0 is obtained from the normalizing condition $\sum p_{mnr} = 1$.

$$p_0 = \left[1 + (1+\frac{\lambda}{\mu_1})A + (1+\frac{\lambda}{\mu_2})B\right]^{-1}. \tag{6.8.15}$$

Summarizing, we get

$$\begin{aligned} P(0) = P(\text{service counter idle}) &= p_0 = \left[1 + (1+\frac{\lambda}{\mu_1})A + (1+\frac{\lambda}{\mu_2})B\right]^{-1} \\ P(1) = P(\text{class 1 in service}) &= p_{101} + p_{201} + p_{111} \\ &= \left(1+\frac{\lambda}{\mu_1}\right)Ap_0 \\ P(2) = P(\text{class 2 in service}) &= p_{012} + p_{022} + p_{112} \\ &= \left(1+\frac{\lambda}{\mu_2}\right)Bp_0. \end{aligned} \tag{6.8.16}$$

For the sake of comparing the two disciplines, let $\lambda_1 = 3/\text{h}$, $\lambda_2 = 2/\text{h}$, and the mean service times be 30 minutes each ($\mu_1 = \mu_2 = 2/\text{h}$). Then we get the following results.

$$\begin{aligned} A &= \frac{3}{2}; \quad B = 1 \\ p_0 &= \frac{4}{39}; \\ p_{101} &= \frac{6}{39}; \qquad p_{201} = \frac{9}{39}; \qquad p_{111} = \frac{6}{39}; \\ p_{012} &= \frac{4}{39}; \qquad p_{022} = \frac{4}{39}; \qquad p_{112} = \frac{6}{39}; \\ P(0) &= 0.103; \quad P(1) = 0.538; \quad P(2) = 0.359. \end{aligned}$$

Answer

Preemptive Priority Due to the preemptive nature of the discipline, we further assume that when there are two class 2 customers in the system and one of them in service, an arriving class 1 customer will displace the one in service. This means one class 2 customer will be removed from the center. The states representing the number of customers in the system are: (0, 10, 20, 01, 02, 11). The state balance equations are given below ($\lambda = \lambda_1 + \lambda_2$).

$$\begin{aligned}
\lambda p_0 &= \mu_1 p_{10} + \mu_2 p_{01} \\
(\lambda + \mu_1) p_{10} &= \lambda_1 p_0 + \mu_1 p_{20} \\
\mu_1 p_{20} &= \lambda_1 (p_{10} + p_{11}) \\
(\lambda + \mu_2) p_{01} &= \lambda_2 p_0 + \mu_1 p_{11} + \mu_2 p_{02} \\
(\lambda_1 + \mu_2) p_{02} &= \lambda_2 p_{01} \\
(\lambda_1 + \mu_1) p_{11} &= \lambda_2 p_{10} + \lambda_1 (p_{01} + p_{02}) \\
\sum p_{mn} &= 1
\end{aligned}$$

This is an example in which the help of a computer in solving the simultaneous equations is likely to work out better. Below we give the standard elimination technique for the solution. For convenience write

$$\begin{aligned}
\lambda &= a; \quad \mu_1 = b; \quad \mu_2 = c; \quad \lambda + \mu_1 = d; \quad \lambda_1 = e; \\
\lambda + \mu_2 &= g; \quad \lambda_2 = h; \quad \lambda_1 + \mu_2 = j; \quad \lambda_1 + \mu_1 = k.
\end{aligned}$$

The second to the last equation allows us to write

$$p_{02} = \frac{h}{j} p_{01}. \tag{6.8.17}$$

Eliminating p_{02} from the set of equations, we get

$$\begin{aligned}
bp_{10} + cp_{01} &= ap_0 && (6.8.18) \\
dp_{10} - bp_{20} &= ep_0 && (6.8.19) \\
ep_{10} - bp_{20} + ep_{11} &= 0 && (6.8.20) \\
-mp_{01} - bp_{11} &= hp_0 && (6.8.21) \\
hp_{10} + np_{01} - kp_{11} &= 0 && (6.8.22)
\end{aligned}$$

where we have written $c(\frac{h}{j}) - g = m$ and $e(1 + \frac{h}{j}) = n$.

Eliminating p_{11} from (6.8.20) and (6.8.21),

$$bep_{10} - b^2 p_{20} - emp_{01} = hep_0. \tag{6.8.23}$$

Eliminating p_{01} from (6.8.18) and (6.8.23),

$$(bem + bce)p_{10} - b^2 cp_{20} = (aem + che)p_0. \tag{6.8.24}$$

Eliminating p_{20} from (6.8.19) and (6.8.24),

$$b(em + ce - cd)p_{10} = e(am + ch - bc)p_0$$

giving

$$p_{10} = \frac{e(am + ch - bc)}{b(em + ce - cd)} p_0 = Ap_0, \quad say. \tag{6.8.25}$$

Using this result in (6.8.19), we get

$$p_{20} = \frac{1}{b}(dA - e)p_0 = Bp_0, \quad say. \tag{6.8.26}$$

Substituting the values of p_{10} and p_{20} in (6.8.23),

$$\begin{aligned} p_{01} &= \frac{1}{em}[beA - b^2B - eh]p_0 \\ &= Cp_0, \quad say. \end{aligned} \tag{6.8.27}$$

Substituting this result in (6.8.21),

$$\begin{aligned} p_{11} &= \frac{-1}{b}[mC + h]p_0 \\ &= Dp_0, \quad say. \end{aligned} \tag{6.8.28}$$

Finally, going back to (6.8.17),

$$p_{02} = \frac{h}{j} Cp_0 = Ep_0, \quad say. \tag{6.8.29}$$

Using the results from (6.8.25)–(6.8.29) in the normalizing condition $\sum p_{ij} = 1$, we get

$$p_0 = (1 + A + B + C + D + E)^{-1}. \tag{6.8.30}$$

With the numerical values used in the non-preemptive case we get the following results:

$$\begin{aligned} A &= 1.748; \quad B = 4.618; \quad C = 0.752; \quad D = 5.2; \quad E = 0.301. \\ p_0 &= 0.073; \quad p_{10} = 0.128; \quad p_{20} = 0.339 \\ p_{01} &= 0.055; \quad p_{02} = 0.022; \quad p_{11} = 0.382 \end{aligned}$$

Hence

$P(0) = 0.073$, $P(1) = 0.849$, $P(2) = 0.077$. **Answer**

In the foregoing discussion, we described how we can determine the limiting distribution of the number of customers in the various priority classes in the system. As illustrated above, even when the number of classes is relatively small, the problem becomes exceedingly difficult to solve. However, if we are interested only in the mean values of the queue lengths and waiting times in steady state, a method given by Cobham (1954) can be used to determine them with relative ease. We illustrate the procedure below when the service times are exponential for each of the priority classes, even though it is valid for arbitrary service time distributions. The changes to be made to the expressions in the

latter case will be indicated at the end of the discussion. Unless faced with a priority queueing system of this type in an application, beginning readers may skip the following analysis because of its intricate details.

Consider a non-preemptive priority queueing system with k priority classes. Customers of class i arrive in a Poisson process with rate λ_i $(i = 1, 2, \ldots, k)$; their service time distribution is exponential with mean $1/\mu_i$. Service is provided by one server. We wish to determine the mean waiting time $W_q^{(i)}$ of a customer belonging to the ith priority class.

Let $\rho_i = \frac{\lambda_i}{\mu_i}$ and $\sigma_i = \sum_{j=1}^{i} \rho_j$.

Clearly, the limiting distributions of the queue lengths and waiting times exist only if $\sigma_k = \sum_{j=1}^{k} \rho_j = \rho < 1$.

Let $T_q^{(i)}$ be the waiting time of an arriving customer of class i. (This is the time the customer waits in line before entering service.) Suppose there are n_r $(r = 1, 2, \ldots, i)$ customers ahead of the arriving customer. The time it has to wait has three components.

(1) The remaining service time of the customer in service at the time of arrival, say S_0.

(2) The total service time of the customers who are ahead of it. Let S_r be the total service time of n_r customers of rth priority class $(r = 1, 2, \ldots, i)$.

(3) While waiting for service, the arriving customer must also wait for the completion of service of arriving customers belonging to a higher priority class. Let n'_r $(r = 1, 2, \ldots, i-1)$ be the number of higher priority customers arriving during $T_q^{(i)}$. The S'_r be the total service time of these arrivals.

Combining these three components, we have

$$T_q^{(i)} = S_0 + \sum_{r=1}^{i} S_r + \sum_{r=1}^{i-1} S'_r \tag{6.8.31}$$

taking expectations

$$W_q^{(i)} = E(S_0) + \sum_{r=1}^{i} E(S_r) + \sum_{r=1}^{i-1} E(S'_r). \tag{6.8.32}$$

Since all service times are exponential, the remaining service time of the customer in service is also exponential with the same rate, appropriate for its priority class. Because the class of the customer is not known, we may use ρ_r/ρ as the probability that it belongs to class r. Furthermore, we must also account for the probability of the system being busy at the time of arrival. This probability is $= \rho$. Combining these terms, we get

$$E(S_0) = \sum_{r=1}^{k} \frac{1}{\mu_r} \left(\frac{\rho_r}{\rho} \right) \rho = \sum_{r=1}^{k} \frac{\rho_r}{\mu_r}. \tag{6.8.33}$$

Let S_r be the service time of customers in the rth priority class. When there are n_r customers of rth priority class ahead of the arriving customer, the total expected service time of customers in that class is given by

$$\begin{aligned} E(S_r) &= E(n_r)E(S_r) \\ &= E(n_r) \cdot \frac{1}{\mu_r}. \end{aligned} \tag{6.8.34}$$

Here we have used the property that the number of customers and the service times are independent of each other, and the service time distribution is exponential with mean $1/\mu_r$. Now using Little's formula ($L_q = \lambda W_q$), we get

$$E(S_r) = \frac{\lambda_r W_q^{(r)}}{\mu_r} = \rho_r W_q^{(r)}. \tag{6.8.35}$$

We may use similar arguments, for the total service time of the higher priority customers arriving during $T_q^{(i)}$. However, we should note that the number of customers to be included here are new arrivals during $T_q^{(i)}$ and therefore the expected number of such new customers of class r is $\lambda_r W_q^{(i)}$. Thus we have

$$\begin{aligned} E(S_r') &= \frac{\lambda_r W_q^{(i)}}{\mu_r} \\ &= \rho_r W_q^{(i)}. \end{aligned} \tag{6.8.36}$$

Combining the three expressions from (6.8.33), (6.8.35), and (6.8.36), we get

$$W_q^{(i)} = E(S_0) + \sum_{r=1}^{i} \rho_r W_q^{(r)} + \sum_{r=1}^{i-1} \rho_r W_q^{(i)} \tag{6.8.37}$$

$$\begin{aligned} [1 - \rho_i - \sigma_{i-1}] W_q^{(i)} &= E(S_0) + \sum_{r=1}^{i-1} \rho_r W_q^{(r)} \\ W_q^{(i)} &= \frac{1}{1 - \sigma_i} \left[\sum_{r=1}^{i-1} \rho_r W_q^{(r)} + E(S_0) \right]. \end{aligned} \tag{6.8.38}$$

From (6.8.37), noting that $\rho_0 = 0$, we get

$$\begin{aligned} W_q^{(1)} &= \frac{E(S_0)}{1 - \sigma_1} \\ W_q^{(2)} &= \frac{1}{1 - \sigma_2} [\sigma_1 W_q^{(1)} + E(S_0)] \\ &= \frac{E(S_0)}{(1 - \sigma_2)(1 - \sigma_1)}. \end{aligned}$$

For induction, assume that

$$W_q^{(i)} = \frac{E(S_0)}{(1-\sigma_{i-1})(1-\sigma_i)}. \tag{6.8.39}$$

From (6.8.37) we get

$$\begin{aligned}(1-\sigma_{i-1})W_q^{(i)} &= \sum_{r=1}^{i} \rho_r W_q^{(r)} + E(S_0) \\ W_q^{(i)} &= \frac{\sum_{r=1}^{i} \rho_r W_q^{(r)} + E(S_0)}{1-\sigma_{i-1}}. \end{aligned} \tag{6.8.40}$$

Equating the right hand sides of (6.8.39) and (6.8.40), we get

$$\sum_{r=1}^{i} \rho_r W_q^{(r)} = \frac{\sigma_i E(S_0)}{1-\sigma_i}. \tag{6.8.41}$$

Now using the assumed form of $W_q^{(i)}$ from (6.8.39) in (6.8.38) and using (6.8.41),

$$\begin{aligned} W_q^{(i+1)} &= \frac{1}{(1-\sigma_{i+1})}\left[\sum_{r=1}^{i} \rho_r W_q^{(r)} + E(S_0)\right] \\ &= \frac{1}{(1-\sigma_{i+1})}\left[\frac{\sigma_i E(S_0)}{1-\sigma_i} + E(S_0)\right] \\ &= \frac{E(S_0)}{(1-\sigma_i)(1-\sigma_{i+1})} \end{aligned} \tag{6.8.42}$$

which shows by induction, that the general form of $W_q^{(i)}$ is given by

$$W_q^{(i)} = \frac{E(S_0)}{(1-\sigma_{i-1})(1-\sigma_i)}. \tag{6.8.43}$$

Substituting from (6.8.33)

$$W_q^{(i)} = \frac{\sum_{r=1}^{k} (\rho_r/\mu_r)}{(1-\sigma_{i-1})(1-\sigma_i)}. \tag{6.8.44}$$

When the service times have arbitrary distributions (i.e., the system is $M/G/1$ with k priority classes) independent of other characteristics of the system, the result for $W_q^{(i)}$ remains the same as (6.8.43), except $E(S_0)$ is determined as follows.

Using $S_r^{(j)}$ to denote the service time of a customer in the rth priority class, the expected value of the remaining service time of a customer in the priority class r is given by (5.2.73)

$$R = \frac{E[(S_r^{(j)})^2]}{2E(S_r^{(j)})}$$

and

$$\rho_r = \lambda_r E(S_r^{(j)}).$$

Now from the arguments used in deriving (6.8.33), we get

$$\begin{aligned} E(S_0) &= \sum_{r=1}^{k} \frac{E[(S_r^{(j)})^2]}{2E(S_r^{(j)})} \cdot \frac{\lambda_r E(S_r^{(j)})}{\rho} \cdot \rho \\ &= \sum_{r=1}^{k} \frac{1}{2} \lambda_r E[(S_r^{(j)})^2] \end{aligned} \tag{6.8.45}$$

giving

$$W_q^{(i)} = \frac{\sum_{r=1}^{k} \lambda_r E[(S_r^{(j)})^2]}{2(1-\sigma_{i-1})(1-\sigma_i)}. \tag{6.8.46}$$

When the mean waiting times are known, the mean time in the system is obtained by adding the mean service time. The corresponding mean queue lengths are obtained by using Little's formula.

Analogous results for the queue with shortest processing time discipline can be easily derived from (6.8.46), first by discretizing the service times to a geometric distribution with a quantum Q as the time unit and then making $Q \to 0$ to get the continuous time analog. Details of the procedure can be found in Coffman and Denning (1973).

6.9 Renewal Process Models

In the preceding pages, we have used only Markov models for analyzing queueing systems. As indicated in several places, Markov models are not general enough to provide the complete analysis of the systems discussed. For instance, the queue length process $\{Q(t), t \in T\}$ is Markovian only at departure points in the queue $M/G/1$, and at arrival points in the queue $G/M/1$. Although we have described the imbedded Markov chain technique to analyze such systems, it does not give us the complete answer. We can use the Markov process analysis technique only by defining supplementary variables to represent the remaining service time at time t in $M/G/1$, and the time since the last arrival in $G/M/1$. However, explicit results are difficult to obtain. Readers may consult advanced texts on queueing theory for information on these techniques. In this section, we provide an alternative approach based on renewal processes to overcome this difficulty. (See Section 3.4 for definitions and properties of the renewal process.)

Assuming the independence of arrival processes, normally the inter-arrival times are i.i.d., and they form a renewal process. But since the departure cannot take place when there are no customers in the system, the departure process is not renewal even when the service times are i.i.d. random variables. (Here we ignore the possibility that may occur in some applications, in which the arrival and service processes operate independently of each other.) They form a renewal

process only during the period when servers are continuous busy, i.e., during a busy period. Since a busy period is followed by an idle period, when the queue discipline dictates that the server does not stay idle when there are customers in the system, the starting points of busy periods form a set of renewal points for the queue length process, with the sequence of busy period—idle period pairs forming the renewal periods. The bare essentials of how such a framework can be used to analyze queues $M/G/1$ and $G/M/1$ is given below with necessary references for advanced reading.

For the number of customers in the system $Q(t)$ let

$$P_{ij}(t) = P[Q(t) = j | Q(0) = i]$$

be its transition probability distribution for the period $(0, t]$. Let ${}^0P_{ij}(t)$ be the probability that the transition of $Q(\cdot)$ from i to j, occurs in $(0, t]$ avoiding state zero, during that period. Probabilistically this can be defined as

$${}^0P_{ij}(t) = P\left[Q(t) = j, Q(\tau) \neq 0,\ 0 < \tau < t | Q(0) = i\right]. \tag{6.9.1}$$

We should note that ${}^0P_{ij}(t)$ is the probability of transition of the process $\{Q(t)\ t\epsilon T\}$ within a busy period. Further noting that the busy cycle is the renewal period for the queue length process, we can write

$$P_{ij}(t) = {}^0P_{ij}(t) + \int_0^t {}^0P_{0j}(t-\tau)dU(\tau). \tag{6.9.2}$$

The two terms on the right-hand side of this equation give the probabilities of two mutually exclusive and collectively exhaustive events in the transition: (1) the process does not visit state 0 and (2) the process visits state 0 at τ $(0 < \tau < t)$ for the last time, and between τ and t the transition is zero-avoiding. The equation now is in the form of (3.4.15) and therefore using the key renewal theorem (3.4.19), we get

$$\lim_{t\to\infty} P_{ij}(t) = \frac{1}{R}\int_0^\infty {}^0P_{0j}(t)dt \tag{6.9.3}$$

where $R = E$ [busy cycle].

In the case of the queue $M/G/1$, R has been obtained in (5.2.52) as

$$R = \frac{1}{\lambda(1-\rho)}. \tag{6.9.4}$$

In the case of queue $G/M/1$ from (5.3.32), we have

$$R = \frac{1}{\lambda(1-\zeta)} \tag{6.9.5}$$

where $1/\lambda$ is the mean inter-arrival time. Note that ζ is the least positive root of the functional equation (5.3.14).

The determination of the transition probabilities ${}^0P_{0j}(t)$ is complicated in both of these systems and we do not present it here. Interested readers may refer to Bhat (1968) for the results as well as a complete analysis of the queue length processes in the two-queueing systems. (Also see Takács (1962)).

A *semi-Markov process* or a *Markov renewal process* is defined by incorporating the concepts of the renewal process into the structure of discrete time Markov chains. Using a semi-Markovian model, we can extend the imbedded Markov chain analysis in systems such as $M/G/1$ and $G/M/1$ to derive performance characteristics in continuous time. Again, these analyses are beyond the scope of this text. Interested readers may refer to Takács (1962), who does not explicitly use a semi-Markov model in his investigations, and Neuts (1966, 1967).

6.10 Exercises

Note: Exercises in this chapter may require the use of computational tools.

1. In a service system, groups of customers arrive and get served individually by a single server. The customer groups arrive in a Poisson process with rate 5 per hour and the group size has a distribution with mean 2.5 and variance 2. The service times are exponential with mean 4 minutes. Determine the expected number of customers in the system in the long run.

2. In an emergency medical clinic, patients arrive for treatment at the rate of 5 per hour and their inter-arrival times can be assumed to have an Erlang E_k distribution with $k = 3$. Assume that each patient requires the services of the doctor for an amount of time that has an exponential distribution with mean 10 minutes. Determine the average time a patient has to spend in the clinic.

3. An automobile service station has one station for general checkups such as oil and filter change, tire rotation, checking fluid levels, etc. On the average the checkup takes 15 minutes, the amount of time having an Erlang E_k distribution with $k = 4$. Cars arrive in a Poisson process at the rate of 3 per hour. Determine the average number of cars in the system in the long run.

4. In an assembly line, by the time a product reaches the assembly station, it would have passed through three earlier stations, at each of which it would have spent an amount of time exponentially distributed with mean 3 minutes. The assembly time is exponential with mean 5 minutes. Determine the expected number of products waiting at the assembly station, assuming that an unlimited number of products are available at the first station and they pass through the first three stations without delay.

5. In Exercise 4, assume that the assembly time at the specific station is exactly 5 minutes. Using an $E_k/D/1$ model, determine the transition probability matrix of the imbedded Markov chain of the number of "products" waiting to be assembled at the specific station. Note that "products" represent the total number of phases making up the inter-arrival time.

   Using an appropriate finite capacity for the waiting room for products, determine the limiting distribution of the number of phases as well as the expected number of products waiting to be assembled.

6. In Exercise 6 of Chapter 5, suppose the amount of time the doctor spends with a patient has the Erlang distribution with mean 10 minutes and $k = 3$. Using a $D/E_k/1$ model and the phase interpretation of the patient's time with the doctor, determine the transition probability matrix of the imbedded Markov chain of the number of outstanding phases of service waiting to be performed at the time of a patient's arrival. Obtain the limiting distribution of this process and the expected number of patients waiting using an appropriate capacity limit.

7. In an airport, check-in counters for an airline are supplemented with four additional self-service counters for ticketed passengers. The passengers arrive at the self-service counters in a Poisson process at the rate of 80 per hour and take exactly 2 minutes to get their boarding passes. Using an $M/D/s$ model obtain the transition probability matrix of the imbedded Markov chain of the process representing the number of customers in the self-service system and determine its limiting distribution. Use an appropriate capacity limit to make the computations feasible.

8. Extend the numerical portion of Example 6.8.1 to allow four customers to be present in the system, and determine the limiting distribution and the three probabilities P (service counter idle), P (class 1 in service), and P (class 2 in service). Compare the results in the non-preemptive and preemptive priority cases.

9. A computer technician has maintenance contracts with three customers. The three customers C_1, C_2, and C_3 have different preemptive priority assignment for service. Under this scheme, C_1 has preemptive priority over C_2 and C_3, and C_2 has preemptive priority over C_3. The customers C_1, C_2, and C_3 call for service with rates λ_1, λ_2, and λ_3, and get service with rates μ_1, μ_2, and μ_3, respectively, in such a way that each such process can be modeled as a two state Markov process described in Section 4.7.1.

   Obtain the state space for the underlying process and determine its limiting distribution. Also determine the long run probabilities P (technician is idle), $P(C_1$ is being served), $P(C_2$ is being served), $P(C_2$ is waiting for service), $P(C_3$ is being served), and $P(C_3$ is waiting for service).

10. Four classes of customers arrive at a counter for service and get served based on a preemptive priority discipline, with class 1 having the highest

priority and class 4, the lowest. The arrivals of the four classes of customers are in Poisson processes with rates 3, 6, 6, and 9 per hour, respectively. Service times of all customers have exponential distributions with mean 2 minutes. Determine the mean waiting time for customers in each class.

Chapter 7

Queueing Networks

7.1 Introduction

The queueing systems considered in the preceding chapters had customers demanding service from a single facility. But there are many real-world systems in which customers get served in more than one station arranged in a network structure, which is a collection of nodes connected by a set of paths. In a queueing network, a group of servers operating from the same facility are identified as a node. As described in Chapter 1, under the historical perspective, a large portion of the advances occurring in queueing theory after the 1960s is connected to networks of queues one way or the other. Computer, communication, and manufacturing systems where queueing theory has found major application areas abound with such networks.

In a queueing network, customers demand service from more than one server. All customers may not require service from the same set of servers. Also, often they may have to go back to the same server more than once. Figure 7.1 is a simple illustration of a network in which the sequencing of the service is shown by directional arrows between the nodes. Figure 7.1 also shows that customers arrive at Nodes 1 and 4 and depart from Nodes 3 and 5. A queueing network with this feature is known as an *open network*. All nodes of the network represent queues and let $Q_i(t)$ be the number of customers at node i at time t. The total number of customers in the network is $\sum_i Q_i(t)$. When no new customers are allowed to enter the network and no customers in the network exits from it, i.e., when $\sum_i Q_i(t) = Q$, a constant, we have a *closed network*. A service center supporting a fixed number of machines is an example of a closed network. When the arrival rate into and the departure rate out of the network are the same (or approximately the same) it can be modeled as a closed network without sacrificing too much accuracy. (With a finite set of equations a closed network sometimes is easier to analyze, depending on its structure.)

© Springer Science+Business Media New York 2015

U. N. Bhat, *An Introduction to Queueing Theory*, Statistics for Industry and Technology, DOI 10.1007/978-0-8176-8421-1_7

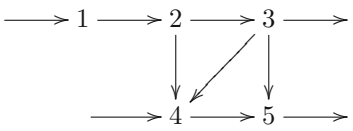


Figure 7.1 Open network

As we point out later, queueing networks in their generality, present formidable problems in their analysis. What we intend to cover here are *Markovian networks* in which the queueing systems at nodes are Markovian and the nodes themselves have a Markovian structure. We start with analyzing the node networks. A Markovian node network is often called the *routing chain.*

7.2 The Markovian Node Network

Consider a network of nodes $\{1, 2, \ldots, k\}$. After getting served at node i, suppose a customer moves to node j with probability P_{ij} $(i, j = 1, 2, \ldots, k)$. Customer opting for a repeat service at node i is represented by probability P_{ii}. In the context of an open queueing network, to account for the outside world from which the customers arrive and to which they go after departing from any state in the network, we have to define an extra state 0, with transition probabilities $P_{00} = 0$; $P_{0j} \geq 0$, $j = 1, 2, \ldots, k$, and $P_{i0} \geq 0$, $i = 1, 2, \ldots, k$. The transition probability matrix $\mathbf{P}$, also known as the *routing matrix* can be represented as

$$\mathbf{P} = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ \vdots \\ k-1 \\ k \end{array} \begin{array}{c} \begin{array}{ccccc} 0 & 1 & 2 & \ldots & k \end{array} \\ \begin{bmatrix} 0 & P_{01} & P_{02} & \ldots & P_{0k} \\ P_{10} & P_{11} & P_{12} & \ldots & P_{1k} \\ P_{20} & P_{21} & P_{22} & \ldots & P_{2k} \\ \vdots & \vdots & & & \\ P_{k-1,0} & P_{k-1,1} & P_{k-1,2} & \ldots & P_{k-1,k} \\ P_{k0} & P_{k1} & P_{k2} & \ldots & P_{k,k} \end{bmatrix} \end{array}. \tag{7.2.1}$$

For the closed network, the node 0 is unnecessary.

A surprising amount of information, exclusive of the queue phenomenon, can be derived from the transition probability matrix $\mathbf{P}$ of the node network. It should be noted that the Markov chain is irreducible (all states communicate with each other), and in general it is also aperiodic unless a special structure is imposed on it.

(i) *Relative throughput*: The rate of customers passing through each node is known as the throughput of that node. Under stable conditions rates of

customer arrivals at each node must attain input–output parity. Let λ_i be the arrival rate at node i, and let $\boldsymbol{\lambda} = (\lambda_0, \lambda_1, \ldots, \lambda_k)$. Under steady state, therefore, we have $\boldsymbol{\lambda}\mathbf{P} = \boldsymbol{\lambda}$, which is the same basic equation we solve for obtaining the limiting distribution of the Markov chain. In the absence of the normalizing condition $\sum_0^k \pi_i = 1$ as in the case of the limiting distribution, the solution of the equation $\boldsymbol{\lambda}\mathbf{P} = \boldsymbol{\lambda}$, gives us only the relative throughput in the network. When the arrival rate from outside the network is known, one can obtain the actual throughputs. In a closed network, however, the actual values depend on the traffic circulating in the system.

(ii) *Throughput time exclusive of waiting*: Consider a customer passing through the nodes of the network with a given transition structure. Let $1/\mu_j$ be the mean time the customer spends at node j and ν_{ij} be the expected number of visits the customer makes to node j having started initially from node i. Then the mean throughput time exclusive of waiting $= \sum_{j=1}^k \nu_{ij} \left(\frac{1}{\mu_j}\right)$. The expected number of visits ν_{ij}, $j = 1, 2, \ldots, k$ can be determined as elements of the *fundamental matrix* of the finite Markov chain $\mathbf{P}$, after converting state 0 to be an absorbing state.

The transition probability matrix (7.2.1) has the following structure when state 0 is made absorbing.

$$\mathbf{P} = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ \vdots \\ k \end{array} \begin{array}{c} \begin{array}{ccccc} 0 & 1 & 2 & \ldots & k \end{array} \\ \begin{bmatrix} 1 & 0 & 0 & \ldots & 0 \\ P_{10} & P_{11} & P_{12} & \ldots & P_{1k} \\ P_{20} & P_{21} & P_{22} & \ldots & P_{2k} \\ \vdots & \vdots & & & \\ P_{k0} & P_{k1} & P_{k2} & \ldots & P_{k,k} \end{bmatrix} \end{array} \tag{7.2.2}$$

$$= \begin{bmatrix} 1 & 0 \\ \boldsymbol{R} & \boldsymbol{Q} \end{bmatrix}. \tag{7.2.3}$$

Matrix (7.2.2) is partitioned and denoted as shown in (7.2.3). Based on the theory of finite Markov chains (see Bhat and Miller (2002)), the elements of the matrix $(\mathbf{I} - \mathbf{Q})^{-1}$ which is known as the *fundamental matrix*, give the expected numbers of visits of the Markov chain to the various states before it ultimately visits the absorbing state 0. Let

$$(\mathbf{I} - \mathbf{Q})^{-1} = \begin{bmatrix} \nu_{11} & \nu_{12} & \ldots & \nu_{1k} \\ \nu_{21} & \nu_{22} & \ldots & \nu_{2k} \\ \vdots & \vdots & & \vdots \\ \nu_{k1} & \nu_{k2} & \ldots & \nu_{kk} \end{bmatrix}. \tag{7.2.4}$$

Suppose, the Markov chain is initially in state i. The expected numbers of visits of the process to states $1, 2, \ldots, k$ before it gets absorbed in 0

are $\nu_{i1}, \nu_{i2}, \ldots, \nu_{ik}$, respectively. The expected total number of visits is, therefore, given by $\sum_{j=1}^{k} \nu_{ij}$.

In the node network, let the initial state be i with probability α_i. With the assumption that the process spends an average of $1/\mu_j$ time units in state j, the total throughput time of a customer exclusive of waiting is given by $\sum_{i=1}^{k} \alpha_i \sum_{j=1}^{k} \nu_{ij}(\frac{1}{\mu_j})$ time units. For an elaboration of this procedure, including expressions for the variance of the throughput time, the readers may refer to Bhat and Miller (2002).

In the formulation given above, we have used $1/\mu_j$ as the mean service time of a customer in node j per visit. It can also be the average total amount of time a customer spends in that node including waiting and service, as long as the queueing systems in the various nodes are independent of each other.

(iii) *Reliability of the network.* The fundamental matrix approach can also be used to determine the reliability of the node network. This is done by introducing a failure node, say $k+1$, which is also absorbing. The new transition probability matrix will have the structure as in (7.2.5):

$$\mathbf{P} = \begin{array}{c} \\ 0 \\ k+1 \\ 1 \\ 2 \\ \vdots \\ k \end{array} \begin{array}{c} \begin{array}{cccccc} 0 & k+1 & 1 & 2 & \ldots & k \end{array} \\ \begin{bmatrix} 1 & & & & & \\ & 1 & & & & \\ P_{10} & P_{1,k+1} & P_{11} & P_{12} & \ldots & P_{1k} \\ P_{20} & P_{2,k+1} & P_{21} & P_{22} & \ldots & P_{2k} \\ \vdots & \vdots & \vdots & \vdots & & \vdots \\ P_{k0} & P_{k,k+1} & P_{k1} & P_{k2} & \ldots & P_{kk} \end{bmatrix} \end{array} \tag{7.2.5}$$

$$= \begin{bmatrix} \boldsymbol{I} & 0 \\ \boldsymbol{R} & \boldsymbol{Q} \end{bmatrix}. \tag{7.2.6}$$

Assume that the process starts in state 1. Let f_{ij} be the probability that starting from state i $(i = 1, 2, \ldots, k)$, the Markov chain ultimately gets absorbed in j $(j = 0, k+1)$. Define

$$\mathbf{F} = \begin{bmatrix} f_{10} & f_{20} & \cdots & f_{k0} \\ f_{1,k+1} & f_{2,k+1} & \cdots & f_{k,k+1} \end{bmatrix}. \tag{7.2.7}$$

Again appealing to the theory of finite Markov chains, we have

$$\mathbf{F} = (\mathbf{I} - \mathbf{Q})^{-1}\mathbf{R}, \tag{7.2.8}$$

where $\mathbf{R}$ is the submatrix as defined in (7.2.5) and (7.2.6). Thus for a customer starting from state i,

$$P(\text{customer will pass through the network without system failure}) = \sum_{j=1}^{k} \nu_{ij} P_{j,0}$$

$$P(\text{system failure}) = \sum_{i=1}^{k} \sum_{j=1}^{k} \nu_{ij} P_{j,k+1}.$$

Assuming that a customer starts from state i with probability α_i, the reliability R of the system is obtained as $R = 1 - P$ (system failure) $= \sum_{i=1}^{k} \alpha_i \sum_{j=1}^{k} \nu_{ij} P_{j0}$.

For a finite time reliability analysis of a Markovian network the readers are referred to Bhat and Kavi (1987).

7.3 Queues in Series

The simplest open queueing network structure is when service facilities are located in series and customers pass through them sequentially. Such systems are also known as *tandem queues*. Examples of queues in series abound in most of the application areas. Work done in assembly lines, traffic signals in a road network, and sequential computations to be carried out in a computer system are some obvious instances. Assume that at each node, the system operates as a Markovian queue ($M/M/$s) with one or more servers. Customers from outside the network always start at the first facility. There is no blocking between successive service stations; this means the waiting room feeding customers to each of these stations has infinite capacity.

To illustrate the behavior of a series of queues, consider two ($M/M/1$) queues in series. Assume that there is a waiting room of infinite size in front of each server. Let customers arrive at the first queue in a Poisson process with rate λ, and assume that the service times are exponential with means $1/\mu_1$ and $1/\mu_2$, respectively. Let $Q_1(t)$ and $Q_2(t)$ be the number of customers at time t in the two queues. As a consequence of the assumptions made on the arrival process and service time distributions, $\{Q_1(t), Q_2(t)\}$ is a vector Markov process, with states (n_1, n_2), $n_1, n_2 = 0, 1, 2 \ldots$. Also, the arrival process to the second queue, as $t \to \infty$, is also Poisson with rate λ. The transition probabilities of the process $\{Q_1(t), Q_2(t)\}$ for finite t can be derived theoretically starting with forward Kolmogorov equations and using transforms. However, the solutions will turn out to be much more complex than the transition probabilities for a single ($M/M/1$) queue, and furthermore, the transition probabilities of the two systems in series will not be independent of each other. The situation is much different when $t \to \infty$. Let Q_1 and Q_2 be the limiting queue lengths in the two

queues. Define

$$p_{n_1,n_2} = P(Q_1 = n_1, Q_2 = n_2), \quad n_1, n_2 = 0, 1, 2, \ldots \tag{7.3.1}$$

which exists when $\rho_1 = \frac{\lambda}{\mu_1} < 1$ and $\rho_2 = \frac{\lambda}{\mu_2} < 1$. The transition diagram with reference to (n_1, n_2) and its neighboring states can be shown as in Figure 7.2.

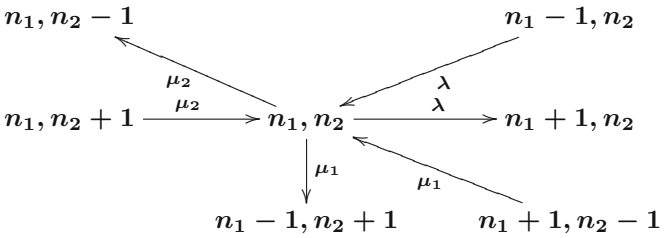


Figure 7.2 Transition diagram

Writing down the corresponding state balance equations, we get

$$\begin{aligned}
\lambda p_{00} &= \mu_2 p_{01} \\
(\lambda + \mu_2) p_{0n_2} &= \mu_1 p_{1,n_2-1} + \mu_2 p_{0,n_2+1} \\
(\lambda + \mu_1) p_{n_1,0} &= \mu_2 p_{n_1,1} + \lambda p_{n_1-1,0} \\
(\lambda + \mu_1 + \mu_2) p_{n_1 n_2} &= \mu_1 p_{n_1+1,n_2-1} + \mu_2 p_{n_1,n_2+1} \\
&\quad + \lambda p_{n_1-1,n_2} \qquad n_1, n_2 > 0.
\end{aligned} \tag{7.3.2}$$

Because of the bivariate structure of states, the recursive solution technique employed in solving such equations in the univariate case does not work in this case. Instead, we appeal to the uniqueness property of the solution to (7.3.2) and start with a trial solution

$$p_{n_1,n_2} = \rho_1^{n_1} \rho_2^{n_2} p_{00} \tag{7.3.3}$$

(see R. R. P. Jackson (1954)). It is easy to see that (7.3.3) satisfies the equations (7.3.2). Now using the normalizing condition

$$\sum_{n_1=0}^{\infty} \sum_{n_2=0}^{\infty} p_{n_1 n_2} = 1,$$

we get

$$p_{00} = (1 - \rho_1)(1 - \rho_2).$$

Hence

$$p_{n_1 n_2} = (1 - \rho_1)(1 - \rho_2) \rho_1^{n_1} \rho_2^{n_2}. \tag{7.3.4}$$

Extending this approach to a series of k queues, each an $(M/M/1)$ system, with μ_i as the parameter of the exponential service time distribution of the ith system, and $\rho_i = \frac{\lambda}{\mu_i} < 1$, $i = 1, 2, \ldots, k$, we get

$$p_{n_1 n_2 \ldots n_k} = \Pi_{i=1}^{k} (1 - \rho_i) \rho_i^{n_i} \tag{7.3.5}$$

(See R. R. P. Jackson (1956)).

Observing the solution (7.3.5), it is clear that we would have obtained the same solution if we had considered the k systems operating independently of each other. But in fact they operate in series and in finite time their behaviors depend on each other. This is the consequence of the departure process result we established in Section 4.2.1 where we found that the departure process of an $(M/M/\text{s})$ type queue was also Poisson with the same rate as the arrival process, as $t \to \infty$. In the queueing network literature, this property has been sometimes denoted as the $M \to M$ *property*. The significance of this property is that it is a necessary condition for the limiting distribution to be in the product form as shown in (7.3.5). In the case of the series of queues, we may conclude that even though in finite time the individual queues are not independent, in the long term they behave as if they are independent.

Another concept closely related to the $M \to M$ property is *local balance*. While discussing the general birth and death queueing model, we established the balance relation (4.1.7) for transitions between two neighboring states. In the broader context of networks of queues, that property is known as local balance and because of the bivariate nature of states, there may be more than one way of identifying neighboring states. The necessary underlying assumption is the Markovian property of the arrival and service processes. (See Chandy (1972) and Muntz (1973).)

In the two-queue series described in this section, consider the state balance equation (7.3.2). They can be broken up into the following local balance equations

$$\begin{aligned}
\lambda p_{n_1 n_2} &= \mu_2 p_{n_1, n_2+1} && n_1 = 0, 1, 2, \ldots; && n_2 = 0, 1, 2, \ldots \\
\mu_1 p_{n_1 n_2} &= \lambda p_{n_1 - 1, n_2} && n_1 = 1, 2, \ldots; && n_2 = 0, 1, 2, \ldots \\
\mu_2 p_{n_1, n_2} &= \mu_1 p_{n_1+1, n_2-1} && n_1 = 0, 1, 2 \ldots; && n_2 = 1, 2, \ldots
\end{aligned} \tag{7.3.6}$$

The validity of such local balance equations is established by back-substitution in the global state balance equations (7.3.2). For the rationale behind the local balance equations the readers are referred to Bhat (1984), Chapters 7 and 12.

The structure of local balance equations leads directly to a product form solution for the limiting distribution of the bivariate process. We illustrate this property using equations (7.3.6). From the second equation in (7.3.6), we get (writing $\frac{\lambda}{\mu_1} = \rho_1$, and $\frac{\lambda}{\mu_2} = \rho_2$)

$$p_{n_1 n_2} = \frac{\lambda}{\mu_1} p_{n_1 - 1, n_2}.$$

Using this equation recursively, we obtain

$$p_{n_1 n_2} = \rho_1^{n_1} p_{0 n_2}. \tag{7.3.7}$$

From the first equation in (7.3.6), we get

$$p_{n_1, n_2+1} = \frac{\lambda}{\mu_2} p_{n_1 n_2}$$

giving

$$p_{n_1 n_2} = \rho_2 p_{n_1, n_2 - 1}.$$

This yields

$$p_{n_1 n_2} = \rho_2^{n_2} p_{n_1 0}. \tag{7.3.8}$$

Inserting the value of $p_{0 n_2}$ from (7.3.8) in (7.3.7), we get

$$p_{n_1 n_2} = \rho_1^{n_1} \rho_2^{n_2} p_{00}. \tag{7.3.9}$$

The result (7.3.4) now follows on the application of the normalizing condition

$$\sum_{n_1=0}^{\infty} \sum_{n_2=0}^{\infty} p_{n_1 n_2} = 1.$$

As seen in these derivations identifying local balance equations is not as simple as it seems to be. The preceding derivation has been provided to illustrate the close connection among the three properties: (1) the $M \to M$, (2) local balance, and (3) the product form solution. A general approach to the analysis of these systems is provided in the Section 7.5.

7.4 Queues with Blocking

The analysis gets complicated if blocking is introduced when customers move from one station to the next. This occurs when there is a waiting room of finite size in between two stations and the customer completing service from the first has to wait until the completion of the ongoing service at the second station when the waiting room between them is full. Thus, any specification of the state space must include information on the numbers of customers in all the stations adding an extra measure of complication. We illustrate these factors below with an example.

Example 7.4.1 A machine repair has two stages, and there are two repairmen working sequentially one for each stage. The system is set up in such a way that there can be a maximum number of three machines waiting for repair or being repaired at any time, two with the first mechanic and one with the second mechanic. In case the first mechanic completes his work while the second

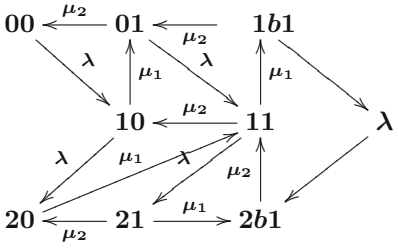
mechanic is still working on the second stage on the previous machine, the first mechanic stops working until the second mechanic is ready to work on it. Repair requests arriving when there are two jobs with the first mechanic (one waiting and one being worked on or one waiting and one blocked from entering into the second stage) are not allowed into the system. Repair requests arrive in a Poisson process with rate λ and repair times at the two stages have exponential distributions with rates μ_1 and μ_2, respectively.

Let the numbers of machines in the two stages represent a bivariate Markov process (the process is Markovian because of the Poisson and exponential assumptions). The state space of the process can be identified as:

stage 1	0	0	1	1	1b	2	2	2b
stage 2	0	1	0	1	1	0	1	1

.

Note that because of blocking, we had to increase the state space to include two points (1b, 1) and (2b, 1) representing blocked machines at stage 1. We present the corresponding transition diagram for the bivariate Markov process in Figure 7.3.



□

Figure 7.3 Transition diagram

Let $p_{ij} = P(Q_1 = i; Q_2 = j)$. Using the state balance principle of the equality of the input and the output with reference to each state, we have the following eight state balance equations.

$$\begin{aligned}
\lambda p_{00} &= \mu_2 p_{01} \\
(\lambda + \mu_2) p_{01} &= \mu_1 p_{10} + \mu_2 p_{1b,1} \\
(\lambda + \mu_1) p_{10} &= \lambda p_{00} + \mu_2 p_{11} \\
(\lambda + \mu_1 + \mu_2) p_{11} &= \lambda p_{01} + \mu_2 p_{2b,1} + \mu_1 p_{20} \\
(\lambda + \mu_2) p_{1b,1} &= \mu_1 p_{11} \\
\mu_1 p_{20} &= \lambda p_{10} + \mu_2 p_{21} \\
(\mu_1 + \mu_2) p_{21} &= \lambda p_{11} \\
\mu_2 p_{2b,1} &= \lambda p_{1b,1} + \mu_1 p_{21}.
\end{aligned}$$

Solving these equations with appropriate substitutions, we get

$$\begin{aligned} p_{01} &= \frac{\lambda}{\mu_2}p_{00}; \quad p_{10} = Ap_{00}; \quad p_{11} = Bp_{00} \\ p_{1b,1} &= Cp_{00}; \quad p_{21} = Dp_{00}; \quad p_{20} = Ep_{00} \\ p_{2b,1} &= Fp_{00}, \end{aligned}$$

where

$$\begin{aligned} A &= \frac{\lambda(\lambda+\mu_2)^2 + \lambda\mu_1\mu_2}{\mu_1\mu_2(2\lambda+\mu_1+\mu_2)} \\ B &= [(\lambda+\mu_1)A - \lambda]\frac{1}{\mu_2} \\ C &= \frac{\mu_1}{\lambda+\mu_2}B \\ D &= \frac{\lambda}{\mu_1+\mu_2}B \\ E &= \frac{\lambda}{\mu_1}A + \frac{\mu_2}{\mu_1}D \\ F &= \frac{\mu_1}{\mu_2}D + \frac{\lambda}{\mu_2}C. \end{aligned}$$

Using the normalizing condition that requires these probabilities sum to 1, we get

$$p_{00} = [1 + \frac{\lambda}{\mu_2} + A + B + C + D + E + F]^{-1}.$$

Answer

The solution to Example 7.4.1 illustrates the magnitude of the problem in dealing with blocking in queues in series. For instance, even a minor change in the blocking rule such as allowing the first mechanic to repair the waiting machine while the machine that has received the first stage repair is made to wait for the second stage repair, instead of the one used in the example above—will change the transition diagram significantly, thus requiring the need for rewriting the equations and reworking the solution. This modification is left as an exercise to the reader. As a reference for dealing with the blocking phenomenon in queueing networks, we may cite Perros (1994).

7.5 Open Jackson Networks

Suppose in the Markovian node network of Section 7.2, each node represents an $(M/M/\mathrm{s})$ queue, with s_i servers at node i, $(i = 1, 2, \ldots, k)$ and there is no blocking for transitions among the nodes. This means each of the queues is an $(M/M/\mathrm{s})$ system with a waiting room of infinite size. Also, assume that

customers arrive at node i from outside the network in a Poisson process with rate λ_i and service times at node i are exponential with mean $1/\mu_i$. Let α_{ij} be the probability that a customer completing service at node i, requests service from node j, $j \neq i$, and α_{i0} be the probability that it will leave the network after service at node i. Let $Q_1, Q_2, \ldots, Q_k$ be the number of customers in the k nodes respectively as $t \to \infty$, and define

$$p_{n_1 n_2 \ldots n_k} = P(Q_1 = n_1, Q_2 = n_2, \ldots, Q_k = n_k). \tag{7.5.1}$$

This is an example of what is commonly known as an open Jackson network after J. R. Jackson (1957) who analyzed it for the first time. For the limiting distribution $p_{n_1 n_2 \ldots n_k}$ of (7.5.1) Jackson has shown that

$$p_{n_1 n_2 \ldots n_k} = p_1(n_1) p_2(n_2) \ldots p_k(n_k), \tag{7.5.2}$$

where

$$p_i(r) = \begin{cases} p_i(0) \frac{(\gamma_i/\mu_i)^r}{r!} & r = 0, 1, 2, \ldots, s_i \\ p_i(0) \frac{(\gamma_i/\mu_i)^r}{s_i! s_i^{r-s_i}} & r = s_i, s_i + 1, \ldots \end{cases} \tag{7.5.3}$$

and

$$\gamma_i = \lambda_i + \sum_{j \neq i} \alpha_{ji} \gamma_j \quad i = 1, 2, \ldots, k. \tag{7.5.4}$$

Given λ_i and α_{ij} $(i, j = 1, 2, \ldots, k)$, the quantity γ_i can be determined from (7.5.4). It should be noted that γ_i is the effective arrival rate at node i after taking into account the traffic from outside the network and the $k - 1$ other nodes within the network. Thus, if $\rho_i = \gamma_i/\mu_i$ is the effective traffic intensity at each node, clearly $\rho_i < 1$ for $i = 1, 2, \ldots, k$ for the limiting distribution to exist. Now $p_i(0)$ for $i = 1, 2, \ldots, k$ can be determined using the normalizing condition

$$\sum_{n_1} \sum_{n_2} \cdots \sum_{n_k} p_{n_1 n_2 \ldots n_k} = 1.$$

The structure of the distribution $p_i(r)$ in (7.5.3) is similar to the limiting distribution of the queue $(M/M/\mathrm{s})_i$ with arrival rate γ_i and service rate μ_i. Does this mean that the arrival process at the ith node is Poisson? In reality it is not true even when $t \to \infty$. This is because of the feedback feature of transitions between nodes. In a series of queues with only feed-forward transitions, we could apply Burke's (1956) result on the departure process (see, Section 4.2.1) and conclude that as $t \to \infty$, the feed-forward transition generates a Poisson process. On the other hand if the transition includes the feedback feature, the resulting arrival process is not Poisson. In fact Burke (1976) has shown that in an $(M/M/1)$ queue with feedback, the effective inter-arrival time distribution is a mixture of exponentials (i.e., hyper-exponential). Thus, from the limiting distribution $p_{n_1 n_2 \ldots n_k}$ of (7.5.3), which is the product of limiting distributions of $(M/M/\mathrm{s})_i$ queueing systems, the only conclusion we can draw is that in the

limit, the Jackson network behaves as if it is a cluster of $(M/M/\mathrm{s})_i$ queues, without being so in actuality. For a discussion of these features of queueing networks readers are referred to Disney and Kiessler (1987).

Markovian network models used in queueing are also known as *Markov population processes*. A systematic procedure for the analysis of such processes with particular reference to their limiting distributions has been given by Kingman (1969). Kingman's results verify the results derived by Jackson, who also generalizes his earlier result to incorporate production systems composed of special purpose service centers in Jackson (1963), and Whittle (1967, 1968) who has derived limiting distributions for migration processes. See Bhat (1984) for details.

The derivation of the limiting distribution (7.5.3) is complex and cumbersome, even when there are only two nodes in the system as can be seen from the following outline. Assume that $k = 2$, and $s_1 = s_2 = 1$. Using properties of state transitions, we can write down the state balance equations as follows:

$$\begin{aligned}
(\lambda_1+\lambda_2)p_{00} &= \mu_1\alpha_{10}p_{10} + \mu_2\alpha_{20}p_{01} \\
(\lambda_1+\lambda_2+\mu_1)p_{10} &= \lambda_1 p_{00} + \mu_2\alpha_{21}p_{01} + \mu_1\alpha_{10}p_{20} \\
(\lambda_1+\lambda_2+\mu_2)p_{01} &= \lambda_2 p_{00} + \mu_1\alpha_{12}p_{10} + \mu_2\alpha_{20}p_{02} \\
(\lambda_1+\lambda_2+\mu_1+\mu_2)p_{11} &= \lambda_1 p_{01} + \lambda_2 p_{10} \\
&\quad +\mu_1\alpha_{10}p_{21} + \mu_2\alpha_{20}p_{12} \\
&\quad +\mu_1\alpha_{12}p_{20} + \mu_2\alpha_{21}p_{02} \\
&\quad \vdots \\
(\lambda_1+\lambda_2+\mu_1+\mu_2)p_{n_1n_2} &= \lambda_1 p_{n_1-1,n_2} + \lambda_2 p_{n_1,n_2-1} \\
&\quad +\mu_1\alpha_{10}p_{n_1+1,n_2} + \mu_2\alpha_{20}p_{n_1,n_2+1} \\
&\quad +\mu_1\alpha_{12}p_{n_1+1,n_2-1} + \mu_2\alpha_{21}p_{n_1-1,n_2+1}, \\
&\quad n_1, n_2 > 0. \qquad (7.5.5)
\end{aligned}$$

Calculating the effective arrival rates to each of the two nodes, we get

$$\begin{aligned}
\gamma_1 &= \lambda_1 + \alpha_{21}\gamma_2 \\
\gamma_2 &= \lambda_2 + \alpha_{12}\gamma_1. \qquad (7.5.6)
\end{aligned}$$

Solving for γ_1 and γ_2 in (7.5.6)

$$\begin{aligned}
\gamma_1 &= \frac{\lambda_1 + \lambda_2\alpha_{21}}{1-\alpha_{12}\alpha_{21}} \\
\gamma_2 &= \frac{\lambda_2 + \lambda_1\alpha_{12}}{1-\alpha_{12}\alpha_{21}}. \qquad (7.5.7)
\end{aligned}$$

Write $\rho_i = \frac{\gamma_i}{\mu_i}$, $i = 1, 2$. Suppose a trial solution is

$$p_{n_1n_2} = C\rho_1^{n_1}\rho_2^{n_2}. \qquad (7.5.8)$$

Verifying that (7.5.8) is, in fact, the correct solution to the state balance equations (7.5.5) with a normalizing condition $\sum_{n_1}\sum_{n_1} p_{n_1 n_2} = 1$, is not a simple task. For details of such a procedure in the general case, with k nodes and multiple servers in each node, the readers are referred to Gross et al. (2008).

7.6 Closed Jackson Networks

Suppose $\lambda_i = 0$ and $\alpha_{i0} = 0$ in the assumptions made while defining the open Jackson network. Let $Q = \sum_{i=1}^{k} Q_i$ be the total number of customers in the network. Now we have a closed Jackson network, which can be used to model a network of queues with a fixed number of customers.

Following the same reasoning as in open networks with k nodes, and the ith node supporting s_i servers $(i = 1, 2, ..., k)$, the limiting distribution $p_{n_1 n_2 \ldots n_k} = P(Q_1 = n_1, Q_2 = n_2, \ldots, Q_k = n_k)$ can be obtained in the product form as

$$p_{n_1 n_2 \ldots n_k} = C\Pi_{i=1}^{k} \frac{\rho_i^{n_i}}{a_i(n_i)} \tag{7.6.1}$$

where

$$a_i(n_i) = \begin{cases} n_i! & n_i < s_i \\ s_i! s_i^{n_i - s_i} & n_i \geq s_i \end{cases} \tag{7.6.2}$$

and $\rho_i = \frac{\gamma_i}{\mu_i}$ with γ_i satisfying the relation

$$\gamma_i = \sum_{j=1}^{k} \gamma_j \alpha_{ji}.$$

This relation can be written as

$$\mu_i \rho_i = \sum_{j=1}^{k} \mu_j \rho_j \alpha_{ji}. \tag{7.6.3}$$

The constant term C in (7.6.1) is determined using the normalizing condition $\sum_{n_1 n_2 \ldots n_k} p_{n_1 n_2} \ldots n_k = 1$. We note here that the term "product form" is used only to the portion of the result involving $n_1, n_2, \ldots, n_k$. In this case constant C does not factor out corresponding to the nodes as it did in the open network. In solving (7.6.3) to determine $\rho_i,\ i = 1, 2, \ldots, k$, we should note that since the total traffic is known, only $k - 1$ equations are independent. Hence, we start by setting one of the ρ_i's as equal to 1.

The determination of $C \equiv C(Q)$ is not a simple problem. We have

$$C^{-1}(Q) \equiv [C(Q)]^{-1} = \sum_{n_1 + n_2 + \ldots + n_k = Q} \Pi_{i=1}^{k} \frac{\rho_i^{n_i}}{a_i(n_i)}, \tag{7.6.4}$$

where the sum extends over all possible ways of choosing $n_1, n_2, \ldots, n_k$ such that $\sum_1^k n_i = Q$. The number of ways this can be done is given by the combinatorial term $\binom{Q+k-1}{Q}$. (The equivalent combinatorial problem is that of distributing Q balls in k cells, which in turn is equivalent to randomly assigning positions to $k-1$ bars among $Q+k-1$ positions arranged in a row.) One of the earliest algorithms to compute $G(Q) = C^{-1}(Q)$ systematically has been given by Buzen (1973). He defines

$$f_i(n_i) = \frac{\rho_i^{n_i}}{a_i(n_i)} \tag{7.6.5}$$

so that

$$G(Q) = \sum_{\Sigma n_r = Q} \Pi_{i=1}^k f_i(n_i).$$

Consider

$$g_m(n) = \sum_{n_1+n_2+\ldots n_k=n} \sum_{i=1}^{m} f_i(n_i) \tag{7.6.6}$$

and $g_k(Q) = G(Q)$ ($m = k$ and $n = Q$). We may write

$$\begin{aligned} g_m(n) &= \sum_{r=0}^{n} \left[\sum_{n_1+n_2+\ldots+n_{m-1}+r=n} \Pi_{i=1}^m f_i(n_i) \right] \\ &= \sum_{r=0}^{n} f_m(r) \left[\sum_{n_1+n_2+\ldots+n_{m-1}=n-r} \Pi_{i=1}^{m-1} f_i(n_i) \right] \\ &= \sum_{r=0}^{n} f_m(r) g_{m-1}(n-r) \quad n = 0, 1, 2, \ldots, Q. \end{aligned} \tag{7.6.7}$$

Also $g_1(n) = f_1(n)$ and $g_m(0) = 1$. Equation (7.6.7) gives a recursive structure for the determination of $G(Q)$. The algorithm used in calculating $G(Q)$ is known as the *convolution algorithm* and it will be illustrated numerically in Chapter 13, Section 13.3.

There are several computational algorithms in the literature, some of which are improvements over Buzen's algorithm, for the calculation of $G(Q)$ and the marginal distributions $p_i(n)$. (See, for instance, Gelenbe and Pujolle (1998).) For a discussion of their relative advantages, the readers may refer to books on the performance modeling of computer networks, such as Sauer and Chandy (1981). For an illustration of the use of recursive solutions see Gross et al. (2008).

7.7 Cyclic Queues

Consider the special case of the closed queueing network in which

$$\alpha_{ij} = \begin{cases} 1 & j = i+1, \quad 1 \le i \le k-1 \\ 1 & i = k, \qquad j = 1 \\ 0 & \text{otherwise .} \end{cases} \tag{7.7.1}$$

This is a cyclic queue (Koenigsberg 1958) where service is provided cyclically by one or more servers. Cyclic queue models are forerunners of polling models that have been mentioned in Chapter 1. For simplicity we assume that there is only one server at each station. Using the same notations as in the last section, corresponding to (7.6.3) we have the following equations

$$\begin{aligned} \mu_1\rho_1 &= \mu_k\rho_k \\ \mu_2\rho_2 &= \mu_1\rho_1 \\ &\vdots \\ \mu_k\rho_k &= \mu_{k-1}\rho_{k-1}. \end{aligned} \tag{7.7.2}$$

From these we get

$$\begin{aligned} \rho_2 &= \frac{\mu_1}{\mu_2}\rho_1 \\ \rho_3 &= \frac{\mu_1}{\mu_3}\rho_1 \\ &\vdots \\ \rho_k &= \frac{\mu_1}{\mu_k}\rho_1. \end{aligned} \tag{7.7.3}$$

Without loss of generality we set $\rho_1 = 1$. For the limiting distribution, we get

$$p_{n_1,n_2\ldots,n_k} = \frac{1}{G(Q)} \frac{\mu_1^{Q-n_1}}{\mu_2^{n_2}\mu_3^{n_3}\ldots\mu_k^{n_k}}. \tag{7.7.4}$$

The factor $G(Q)$ in (7.7.4) is determined using Buzen's algorithm as described in the last section.

Example 7.7.1 Suppose there are only two stations in a closed cyclic network. Service times at the two stations have exponential distributions with rates μ_1 and μ_2. Following the arguments used in deriving (7.7.2)–(7.7.4), we have

$$\rho_2 = \frac{\mu_1}{\mu_2}\rho_1.$$

Setting $\rho_1 = 1$ we get $\rho_2 = \frac{\mu_1}{\mu_2}$.

$$p_{n_1,Q-n_1} = \frac{1}{G(Q)} \left(\frac{\mu_1}{\mu_2}\right)^{Q-n_1}.$$

Using the normalizing condition $\sum_{n_1} p_{n_1,Q-n_1} = 1$,

$$\begin{aligned} \frac{1}{G(Q)} \sum_{n_1=0}^{Q} \left(\frac{\mu_1}{\mu_2}\right)^{Q-n_1} &= 1 \\ G(Q) &= \frac{1 - \left(\frac{\mu_1}{\mu_2}\right)^{Q+1}}{1 - \frac{\mu_1}{\mu_2}}. \end{aligned}$$

The limiting distribution is now obtained as

$$p_{n_1,Q-n_1} = \frac{1 - (\mu_1/\mu_2)}{1 - (\mu_1/\mu_2)^{Q+1}} \left(\frac{\mu_1}{\mu_2}\right)^{Q-n_1}.$$

Answer

7.8 Remarks

We have described in the preceding sections some of the fundamental models for queueing networks. In practice, however, networks of the real world are normally much more complex. Since the 1970s, with increased attention to models necessary to analyze traffic in computer and communication systems, researchers have developed other indirect or approximate techniques of analysis. The *mean value analysis* (see Section 13.3) is one such example. In the category of approximations is the *method of isolation and aggregation* in which systems are analyzed through loosely dependent subsystems. For these and other approximate methods readers are referred to books such as Gelenbe and Pujolle (1999).

The paper by Baskett et al. (1975) on the open, closed, and mixed queueing networks with different classes of customers was one of the earliest attempts to go beyond the Jackson network. Their extensions include the use of distributions other than the exponential (e.g., Coxian), and service disciplines other than the FCFS (processor sharing, no queueing, and LCFS). Since the publication of this article, the literature on the performance modeling of queueing networks has greatly increased. What we have provided here is an introduction to the topic. Interested readers and researchers may consult books such as Courtois (1977), Kelley (1979), Sauer and Chandy (1981), Lavenberg (1983), Molloy (1989), Perros (1994), and Gelenbe and Pujolle (1998) and articles in various journals dealing with computer and communication networks.

7.9 Exercises

Note: Exercises in this chapter may need the use of computational tools.

1. Solve Example 7.4.1 with the change in the service discipline such that the first mechanic starts working on a waiting machine, if there is one, on the completion of service to a machine even when it is blocked by the on going service at the second stage.

2. Solve Example 7.4.1 when the total number of machines allowed in the system is four, with two with each mechanic.

   Solve the problem under service disciplines in the following two cases when a machine is blocked from starting service at the second stage.

   (a) The first mechanic stops work as in Example 7.4.1
   (b) The first mechanic starts work on a waiting machine, if there is one, as in Exercise 1 above.

3. In a two node open queueing network with blocking, let the number of waiting spaces in front of the second server be m. Let $m+1$ represent the blocked state. If n_1 and n_2 are the numbers of customers in the two nodes, respectively (including those in service), we have $n_1 = 0, 1, 2, \ldots$ and $n_2 = 0, 1, 2, \ldots, m, m+1$. Let λ be the Poisson arrival rate and μ_1 and μ_2 be the service rates at the two nodes with exponential service time distributions.

   Examine the impact of two special cases:

   (a) $\mu_1 \to \infty$, when the customer at the first node receives an infinitesimal amount of service
   (b) The first node is saturated, meaning that there is always at least one customer waiting for service.
   (See Perros (1994)).

4. In Exercise 3(b) above, what is the percentage of time the first server is providing service? Specialize the result for $\mu_1 = \mu_2$.

5. In a computer center jobs are submitted from N terminals in Poisson processes, each with rate λ. Each job requests service from a processor followed by one of the two I/O devices. The I/O devices are chosen with probability β_1 and β_2, respectively. The job then exits the system with probability β_3 or proceeds for consultation of a file on another server with probability β_4. (Note that $\beta_1+\beta_2 = 1$ and $\beta_3+\beta_4 = 1$.) After consulting the file, the job joins the first queue for another round. Assume that the services provided at the various locations are all distributed exponentially with the following rates: CPU - μ_1; I/O(1) $-\mu_2$; I/O (2) $-\mu_3$ and, file server $-\mu_4$. Determine the limiting distribution of jobs at each station and the mean response time for a job in the entire system given the following values.

   $N = 40$ $\quad \lambda = 0.01/\text{sec}; \quad \beta_1 = 0.6; \quad \beta_3 = 0.4$

   $\frac{1}{\mu_1} = 0.8$ sec; $\quad \frac{1}{\mu_2} = 0.3$ sec; $\quad \frac{1}{\mu_3} = 0.6$ sec; $\quad \frac{1}{\mu_4} = 1$ sec.

   (Krakowiak (1988)).

6. Consider the following motor vehicle registration process with four stations.

   (a) *Reception.* Customers arrive in a Poisson process with rate 12 per hour. The receptionist takes an amount of time that is exponentially distributed with 20 seconds to direct each customer to either one of two processing clerks with probabilities 0.3 (Clerk 1) and 0.7 (Clerk 2).

   (b) *Clerk 1* handles out-of-state and new licenses and takes on the average 10 minutes, the amount of time having an exponential distribution.

   (c) *Clerk 2* handles standard in-state renewal applications and takes on the average 5 minutes. An exponential distribution assumption is appropriate for this time as well.

   (d) 20 % of applications processed by Clerk 1 go to Clerk 2 and 10 % of applications processed by Clerk 2 go to Clerk 1 for further processing. When the processing is completed by the two clerks (80 % by Clerk 1 and 90 % by Clerk 2), the customers move to the cashier for paying the fees.

   (e) The amount of time spent by the cashier with a customer is exponential with mean 1 minutes.

   Model this system as an open network and obtain the limiting distribution of the number of customers at each station. Also determine (i) the average total amount of time a customer spends in the system and (ii) the average total amount of time a customer with an in-state license spends in the system. (Molloy (1989)).

7. An information network has N centers $C_1, C_2, \ldots, C_N$ and a message arrival center C_0. If a message cannot be satisfied completely in center C_i, it is sent to one of the remaining centers C_j. Consider a strictly hierarchical message transfer network in which the message referral occurs in a path $C_N \to C_{N-1} \to ... \to C_2 \to C_1$. In addition to the originating center C_0, include a center C_R that deals with all rejected messages. Let P_{ij} $(i, j = 1, 2, \ldots, N)$ be the probabilities of the referral path and include P_{i0} and P_{iR} $(i = 1, 2, \ldots, N)$ as probabilities of satisfaction and rejection at center C_i $(i = 1, 2, \ldots, N)$. Let c_{ij} be the cost associated with the referral path $i \to j$ and let γ_{ij} and η_{ij} be its mean and variance. Assume that n_i messages originate at center C_0 during a given length of time and let K be the total cost associated with these messages. Determine the mean and variance of K. (Bhat et al. (1975); also see Bhat (1984), Section 13.5.)

Chapter 8

Matrix-Analytic Queueing Models

Contributing Author: Professor Srinivas R. Chakravarthy[1]

8.1 Introduction

Successive inter-arrival times in the queueing systems considered in the previous chapters form a renewal process. However, in many practical situations, especially in production and manufacturing systems, the inter-arrival times between jobs may not form a renewal process. In this chapter, we present versatile Markovian point processes (VMPP) introduced by Neuts (1974) that enable us to model arrival processes that may not be renewal.

A brief introduction to phase type distributions (PH) is given in Section 8.2. In stochastic modeling, PH-distributions lend themselves naturally to algorithmic implementation. They have closure properties and a related matrix formalism that make them convenient for use in practice (see Chapter 2 of Neuts (1981)). In order to facilitate computational (or algorithmic) analysis of stochastic models Neuts developed the PH-distributions, introduced matrix-geometric solutions, VMPP, and matrix-analytic methods (MAMs). The VMPP was initially referred to as N-process by Ramaswami (1980). Lucantoni et al. (1990) introduced the terms Markovian arrival process (MAP) and batch MAP (BMAP) to describe the VMPP with much simpler notation. The BMAP not only generalizes the phase type renewal process but also provides a way to model correlated arrivals.

This chapter is dedicated to the memory of Professor Marcel Neuts, pioneer of MAM and algorithmic probability, who died on March 9, 2014.

[1]Department of Industrial and Manufacturing Engineering, Kettering University, Flint MI 48504

© Springer Science+Business Media New York 2015

U. N. Bhat, *An Introduction to Queueing Theory*, Statistics for Industry and Technology, DOI 10.1007/978-0-8176-8421-1_8

The basic structures in MAMs can be classified into one of three types: (a) The $M/G/1$ type wherein any one-step transition from a level, say i, can occur to level $i-1$ on the left or to any level i or higher. That is, the structure is *skip-free to the left*; (b) the $G/M/1$ type wherein any one-step transition from a level, say i, can occur to any level up to level $i+1$. That is, the structure is *skip-free to the right*; (c) the quasi-birth-and-death (QBD) type wherein any one-step transition from a level, say i, can occur only to levels $i-1, i, i+1$. Thus, this structure can be thought of as the intersection of the first two types. In all of these types, the transition structure is assumed to possess spatial homogeneity. That is, except for the boundary levels (which in most cases could be simply level 0), the probabilities governing the transitions from level i to level j depend only on the difference $i-j$ and not on the values of i and j.

Using the fact that for the classical $M/G/1$ queue the busy period plays an important role (see Takacs (1962)). Neuts (1989) generalized the busy period analysis for the $M/G/1$ type queues using matrix formalisms. The key concept that plays a role in the $M/G/1$ type queues is the first passage time, also referred to as the "fundamental period". The matrix, G, of the conditional probabilities of the first passage time is obtained as the minimal nonnegative solution to a nonlinear matrix equation. In the classical $G/M/1$ queue, it is known that the steady-state distribution of the number of customers in the system has a geometric solution (5.3.13). Neuts (1978) generalized this result to the $G/M/1$ type queues where the key role in these types of queues is played by the rate matrix, R, which has a probabilistic interpretation. Again, this matrix is obtained as the minimal nonnegative solution to a nonlinear matrix equation (Neuts 1981) (see (5.3.14) in the $G/M/1$ case).

In the rest of this chapter, the readers should note that we have used some notations that are not the same as in previous chapters; for example the generator matrix is identified as $\mathbf{Q}$ while in earlier chapters it is identified as $\mathbf{A}$. Also what we have called earlier as continuous time Markov process we call it here as continuous-time Markov chain (CTMC).

8.2 Phase Type Distributions

PH-distributions are natural generalizations of exponential distributions (see A.6) and have found applications in areas such as reliability, inventory, and warranty models in addition to queueing. Furthermore, these distributions require simple matrix formalisms not only in the analysis but also in algorithmic implementation.

For use in sequel, let $\boldsymbol{e}(r), e_j(r)$, and I_r denote, respectively, the (column) vector of dimension r consisting of 1's, column vector of dimension r with 1 in the jth position and 0 elsewhere, and an identity matrix of dimension r. When there is no need to emphasize the dimension of these vectors we will suppress the suffix. Thus, $\boldsymbol{e}$ will denote a column vector of 1's of appropriate dimension. The notation $\otimes$ and $\oplus$, respectively, will stand for the Kronecker product and

Kronecker sum of two matrices. Thus, if $\mathbf{A}$ is a matrix of order $m \times n$ and if $\mathbf{B}$ is a matrix of order $p \times q$, then $\mathbf{A} \otimes \mathbf{B}$ will denote a matrix of order $mp \times nq$ whose (i, j)th block matrix is given by $a_{ij}\mathbf{B}$; the Kronecker sum of two square matrices, say, $\mathbf{G}$ of order g and $\mathbf{H}$ of order h, denoted by $\mathbf{G} \oplus \mathbf{H}$ is given by $\mathbf{G} \otimes \mathbf{I} + \mathbf{I} \otimes \mathbf{H}$, a square matrix of dimension gh. For more details on Kronecker products and sums, we refer the reader to Steeb and Hardy (2011).

Suppose that $\{Y_t\}$ is CTMC on $\{1, 2, ..., m, m+1\}$ with m transient states $1, 2, ..., m$ and an absorbing state $m+1$. For details on CTMC we refer the reader to Appendix B. The generator of the MC is of the form

$$\tilde{\mathbf{Q}} = \begin{bmatrix} \mathbf{S} & \boldsymbol{S}^0 \\ \mathbf{0} & 0 \end{bmatrix}, \tag{8.2.1}$$

where $\mathbf{S}$ is an $m \times m$ matrix, $\boldsymbol{S}^0$ is a column vector of order m such that $\boldsymbol{S}\boldsymbol{e} + \boldsymbol{S}^0 = \mathbf{0}$. Assume that the matrix $\mathbf{S} + \mathbf{S}^0\boldsymbol{\beta}$ is irreducible. Let $(\beta_1, \ldots, \beta_m, \beta_{m+1}) = (\boldsymbol{\beta}, \beta_{m+1})$ be a probability vector giving the initial probabilities. That is, $P(Y_0 = i) = \beta_i, 1 \le i \le m+1$. Suppose we start the CTMC in one of the m transient states. We are interested in finding the time until absorption into the absorbing state.

Let X denote the time until absorption into state $m+1$. Then X is a continuous random variable on $[0, \infty)$ and its probability density and cumulative probability distribution functions are given by (we assume that $\beta_{m+1} = 0$ which is the case in most applications.)

$$f(t) = \boldsymbol{\beta} e^{\boldsymbol{S}t}\mathbf{S}^0, \quad t \ge 0, \quad F(t) = P(X \le t) = 1 - \boldsymbol{\beta} e^{\boldsymbol{S}t}\boldsymbol{e}, \quad t \ge 0. \tag{8.2.2}$$

In this case, we say that X follows a PH-distribution with representation $(\boldsymbol{\beta}, \boldsymbol{S})$ of order m and denote this by $X \equiv \text{PH}\ (\boldsymbol{\beta}, \boldsymbol{S})$ of order m. The mean and variance of X are given by

$$\mu_X = \boldsymbol{\beta}(-\boldsymbol{S})^{-1}\boldsymbol{e} \quad \text{and} \quad \sigma_X^2 = 2\boldsymbol{\beta}\boldsymbol{S}^{-2}\boldsymbol{e} - \mu_X^2. \tag{8.2.3}$$

In equation (8.2.2) and also in the sequel we will be using the exponential matrix frequently. Exponential matrix is defined as $e^{\mathbf{A}} = \sum_{n=0}^{\infty} \mathbf{A}^n/n! = \mathbf{I} + \mathbf{A} + \mathbf{A}^2/2! + \cdots$.

Some well-known distributions are special cases of PH-distributions. These are given below. (Also see Appendix A.)

Exponential: By taking $m = 1, \boldsymbol{\beta} = 1, \boldsymbol{S} = (-\lambda)$, we get the exponential distribution with parameter λ.

Erlang: Erlang of order m with parameter λ is a PH-distribution with representation $(\boldsymbol{\beta}, \boldsymbol{S})$ of order m given by

$$\boldsymbol{\beta} = (1, 0, ..., 0), \quad \mathbf{S} = \begin{bmatrix} -\lambda & \lambda & 0 & \ldots & 0 \\ 0 & -\lambda & \lambda & \ldots & 0 \\ 0 & 0 & -\lambda & \ldots & 0 \\ \vdots & \vdots & \vdots & \ldots & \vdots \\ 0 & 0 & 0 & \ldots & -\lambda \end{bmatrix}.$$

Generalized Erlang: Generalized Erlang of order m with parameters $\lambda_1, ..., \lambda_m$ is a PH-distribution with representation $(\boldsymbol{\beta}, \boldsymbol{S})$ of order m given by $\boldsymbol{\beta} = (1, 0, ..., 0)$ and

$$\mathbf{S} = \begin{bmatrix} -\lambda_1 & \lambda_1 & 0 & \dots & 0 \\ 0 & -\lambda_2 & \lambda_2 & \dots & 0 \\ 0 & 0 & -\lambda_3 & \dots & 0 \\ \vdots & \vdots & \vdots & \dots & \vdots \\ 0 & 0 & 0 & \dots & -\lambda_m \end{bmatrix}.$$

Hyperexponential: A hyperexponential is a mixture of m exponentials with parameters $\lambda_1, ..., \lambda_m$. The mixing probabilities are $p_1, ..., p_m$. This is a PH-distribution with representation $(\boldsymbol{\beta}, \boldsymbol{S})$ of order m given by $\boldsymbol{\beta} = (p_1, ..., p_m)$ and

$$\mathbf{S} = \begin{bmatrix} -\lambda_1 & 0 & 0 & \dots & 0 \\ 0 & -\lambda_2 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \dots & \vdots \\ 0 & 0 & 0 & \dots & -\lambda_m \end{bmatrix}.$$

Example 8.2.1 Suppose that the time (in minutes) between two phone calls arriving is modeled as a continuous time PH-distribution with

$$\boldsymbol{\beta} = (0.4, 0.3, 0.3), \quad \mathbf{S} = \begin{bmatrix} -5 & 2 & 2 \\ 0 & -6 & 6 \\ 1 & 2 & -4 \end{bmatrix}.$$

Find the mean and the variance of the time between two phone calls.

First note that $(-\mathbf{S})^{-1} = \begin{bmatrix} \frac{1}{3} & \frac{1}{3} & \frac{2}{3} \\ \frac{1}{6} & \frac{1}{2} & \frac{5}{6} \\ \frac{1}{6} & \frac{1}{3} & \frac{5}{6} \end{bmatrix}, (-\mathbf{S})^{-1}\boldsymbol{e} = \begin{bmatrix} \frac{4}{3} \\ \frac{3}{2} \\ \frac{4}{3} \end{bmatrix}$, and hence using expressions for the mean and the variance as given in (8.2.3), we can verify that the mean is 1.383333 minutes and the variance is 1.903056 minutes^2.

Answer

Note that the PH-renewal process is related to the irreducible Markov chain obtained by instantaneously restarting the absorbing Markov chain (whose generator is as given in (8.2.1)) with the same initial probability vector $\boldsymbol{\beta}$ every time an absorption occurs. Thus, the PH-renewal process is related to the irreducible Markov chain with generator

$$\boldsymbol{Q}^* = \boldsymbol{S} + \boldsymbol{S}^0\boldsymbol{\beta}. \tag{8.2.4}$$

The stationary probability vector, $\boldsymbol{\pi}^*$ of $\boldsymbol{Q}^*$ satisfying $\boldsymbol{\pi}^*\boldsymbol{Q}^* = \mathbf{0}$, $\boldsymbol{\pi}^*\boldsymbol{e} = 1$, is given by

$$\boldsymbol{\pi}^* = \frac{1}{\mu_X}\boldsymbol{\beta}(-\boldsymbol{S})^{-1}, \tag{8.2.5}$$

where μ_X is the mean of X and is given in (8.2.3).

The renewal function, $U(t) = E[N(t)]$, and its density for the PH-renewal process are given by (see Chakravarthy (2010))

$$\boldsymbol{H}(t) = \frac{1}{\mu_x}\left\{t + \frac{\sigma_X^2 + \mu_X^2}{\mu_x} + \boldsymbol{\beta}[\boldsymbol{e}\boldsymbol{\pi}^* - e^{\boldsymbol{Q}^*t}]\boldsymbol{S}^{-1}e\right\}, h(t) = H'(t) = \boldsymbol{\beta}e^{\boldsymbol{Q}^*t}\boldsymbol{S}^0. \tag{8.2.6}$$

8.3 Markovian Arrival Process

In practice, we come across many situations where the arrival processes are not necessarily renewal processes. For example, consider a queueing network consisting of 2 nodes. The output of node 1 will form the input to node 2. If we consider non-Poisson arrivals, say, to the first node, the output process will not necessarily be a renewal process. Also, in production line problems where arrivals from different sources form input to an assembly system, the arrival process may not necessarily be a renewal process. So, how do we model these point processes? As mentioned earlier, the answer is to use MAP.

Consider an irreducible CTMC with m transient states. At the end of a sojourn in state i that is exponentially distributed with parameter λ_i, there are two possibilities. The first possibility corresponds to an "event" (or an arrival) in which the CTMC visits any of the m transient states including the state from which this event occurred with probability p_{ij}. The second possibility corresponds to no arrival and the CTMC visits any of the $m-1$ transient states (all m except i) with probability q_{ij}. Thus, the CTMC can go from state i to state i only through an arrival.

Define matrices $\mathbf{D}_0 = (d_{i,j}^{(0)})$ and $\mathbf{D}_1 = (d_{i,j}^{(1)})$ such that $d_{i,i}^{(0)} = -\lambda_i$, $1 \le i \le m$, $d_{i,j}^{(0)} = \lambda_i q_{i,j}$, $j \ne i$, $1 \le i,j \le m$, and $d_{i,j}^{(1)} = \lambda_i p_{i,j}^{(1)}$, $1 \le i,j \le m$, with $\sum_j p_{i,j}^{(1)} + \sum_{j\ne i} q_{i,j} = 1$, for $1 \le i \le m$. By assuming that $\mathbf{D}_0$ to be a nonsingular matrix, the successive times between "event" (or arrivals) will be finite with probability one and that the process will not terminate. A pictorial description of this process is given in Figure 8.1 below.

Thus, a MAP is described by the parameter matrices $(\mathbf{D}_0, \mathbf{D}_1)$ of order m such that the transitions corresponding to no arrivals are governed by D_0 and the transitions corresponding to arrivals of batch size k are governed by D_1. The underlying CTMC has the generator given by $\boldsymbol{Q} = \boldsymbol{D}_0 + \boldsymbol{D}_1$.

The point process described by the MAP is a special class of semi-Markov process with transition probability matrix given by

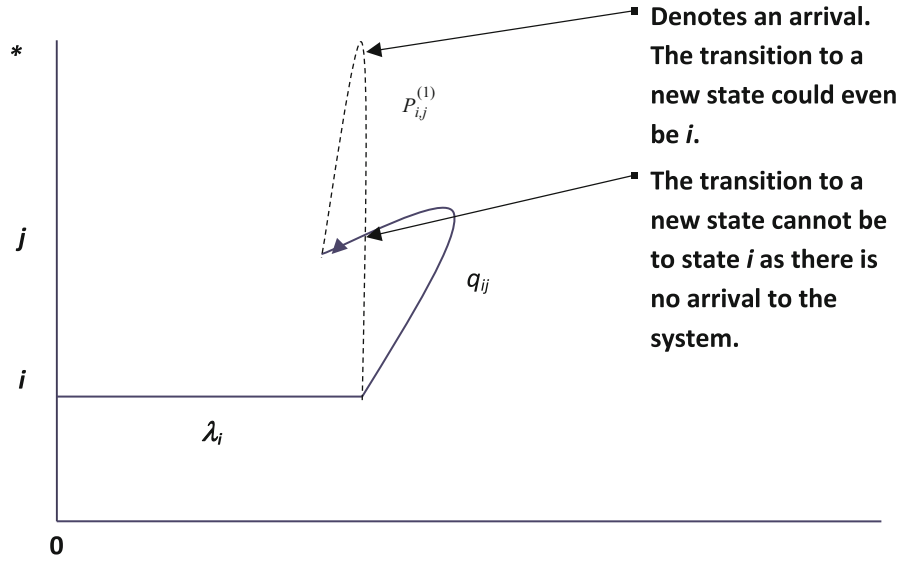


Figure 8.1 Description of MAP

$$\int_0^t e^{\boldsymbol{D}_0 x}\boldsymbol{D}_1 dx = [I - e^{\boldsymbol{D}_0 t}](-\boldsymbol{D}_0)^{-1}\boldsymbol{D}_1. \tag{8.3.1}$$

Let $\boldsymbol{\pi}$ be the steady state probability vector of the generator $\boldsymbol{Q}$ governing the underlying CTMC satisfying the equation

$$\boldsymbol{\pi Q} = \boldsymbol{0}, \boldsymbol{\pi e} = 1. \tag{8.3.2}$$

Let $\boldsymbol{\alpha}$ be the initial probability vector of the underlying CTMC with generator $\boldsymbol{Q}$. We can choose $\boldsymbol{\alpha}$ in a variety of ways to model different scenarios. For example, if we want the time origin to coincide with an arbitrary arrival point we can take $\boldsymbol{\alpha} = c\boldsymbol{\pi}\boldsymbol{D}_1$, where c is the normalizing constant to make the vector $\boldsymbol{\alpha}$ a probability vector. Other scenarios include the time origin to be the end of an interval during which there are at least k arrivals, the point at which the system is in specific state such as the end of a busy period or the beginning of a busy period begins. A useful case for model comparisons is the one where we get the stationary version of the MAP by using $\boldsymbol{\alpha} = \boldsymbol{\pi}$. The fundamental rate (the rate of arrivals per unit of time), λ, is given by

$$\lambda = \boldsymbol{\pi}\boldsymbol{D}_1\boldsymbol{e}. \tag{8.3.3}$$

Often in model comparisons it is convenient to select the time scale of the MAP so that λ has a certain value. This is accomplished by multiplying the parameter matrices $\mathbf{D}_0$ and $\mathbf{D}_1$ by the appropriate common constant.

MAP is a special case of batch MAP (BMAP), where we have a sequence, $\{\boldsymbol{D}_k\}$, of parameter matrices of order m such that the transitions corresponding

to no arrivals are governed by $\mathbf{D}_0$ and the transitions corresponding to arrivals of batch size k are governed by $\boldsymbol{D}_k$. While $\mathbf{D}_0$ has the same form as that for MAP, $\boldsymbol{D}_k = (d_{i,j}^{(k)})$ such that $d_{i,j}^{(k)} = \lambda_i p_{i,j}^{(k)}$, $1 \leq i, j \leq m$. The underlying CTMC has the generator given by $\boldsymbol{Q} = \sum_k \boldsymbol{D}_k$. By assuming that $\mathbf{D}_0$ to be a nonsingular matrix, the successive times between "events" (or arrivals) will be finite with probability one and that the process will not terminate. A pictorial description of a BMAP process is given in Figure 8.2 below.

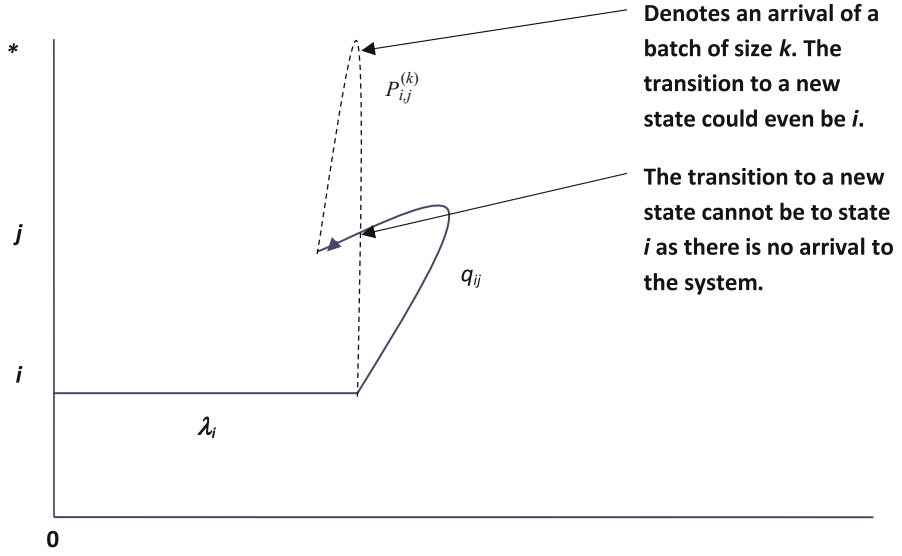


Figure 8.2 Description of BMAP

For details on MAP and BMAP, and their usefulness in stochastic modeling, we refer the reader to Lucantoni (1991), Neuts (1974, 1989), and for a review and recent work on MAP and BMAP we refer the reader to Chakravarthy (2001, 2010) and Artalejo et al. (2010).

8.4 Analysis of Queueing Models Using MAM

In order to illustrate the use of matrix-analytic methods in the analysis of queueing systems we consider some basic models such as $M/PH/1$, $PH/M/1$, and $MAP/PH/1$. Through these models, the reader will be exposed to the key steps involved in constructing, analyzing, and interpreting the results. First we present the model $MAP/PH/1$, and then consider $M/PH/1$, $PH/M/1$, and $M/M/1$ as special cases and exploit the special structure of these models. A good background in matrix theory (Steeb and Hardy 2011) will help the reader in following these analyses.

Assume that arrivals occur according to a Markovian arrival process with representation $(\mathbf{D}_0, \mathbf{D}_1)$ of order m. Let $\boldsymbol{\pi}$ be the steady state probability vector of the generator $\boldsymbol{Q} = \boldsymbol{D}_0 + \boldsymbol{D}_1$ governing the underlying CTMC and let λ denote the arrival rate (see equations (8.3.2) and (8.3.3)). Assume the service times to be of phase type with irreducible representation given by $(\boldsymbol{\beta}, \boldsymbol{S})$ of order n. Let the service rate be denoted by μ and recall that $\mu = [\boldsymbol{\beta}(-\boldsymbol{S})^{-1}\boldsymbol{e}]^{-1}$.

Let $X_1(t), X_2(t)$, and $X_3(t)$ denote, respectively, the number of customers in the system, the phase of the service (if the server is busy), and the phase of the arrival process at time t. The three-dimensional process $\{(X_1(t), X_2(t), X_3(t)) : t \geq 0\}$ is CTMC on the state space Ω given by

$$\Omega = \{(0,k) : 1 \leq k \leq m\} \cup \{(i,j,k) : i \geq 1, 1 \leq j \leq n, 1 \leq k \leq m\}. \quad (8.4.1)$$

Defining the set, $\mathbf{0}$, to be the set of states $\{(0,k) : 1 \leq k \leq m\}$ and $\boldsymbol{i}$ to be the set of states $\{(i,j,k) : i \geq 1, 1 \leq j \leq n, 1 \leq k \leq m\}$, for $i \geq 1$, the generator of the CTMC with state space Ω is given by

$$\mathbf{Q} = \begin{array}{c} 0 \\ \mathbf{1} \\ \mathbf{2} \\ \mathbf{3} \\ \vdots \end{array} \begin{bmatrix} \boldsymbol{D}_0 & \boldsymbol{\beta} \otimes \boldsymbol{D} & 0 & 0 & 0 & \dots \\ \boldsymbol{S}^0 \otimes \boldsymbol{I} & \boldsymbol{S} \oplus \boldsymbol{D}_0 & \boldsymbol{I} \otimes \boldsymbol{D}_1 & 0 & 0 & \dots \\ 0 & \boldsymbol{S}^0\boldsymbol{\beta} \otimes \boldsymbol{I} & S \oplus \boldsymbol{D}_0 & \boldsymbol{I} \otimes \boldsymbol{D}_1 & 0 & \dots \\ 0 & 0 & \boldsymbol{S}^0\boldsymbol{\beta} \otimes \boldsymbol{I} & \boldsymbol{S} \oplus \boldsymbol{D}_0 & \boldsymbol{I} \otimes \boldsymbol{D}_1 & \dots \\ \vdots & \vdots & \vdots & \ddots & \ddots & \ddots \end{bmatrix}, \quad (8.4.2)$$

where $\boldsymbol{S}^0$ is a column vector of dimension m such that $\boldsymbol{S}\boldsymbol{e} + \boldsymbol{S}^0 = \mathbf{0}$. That is, $\boldsymbol{S}^0$ is obtained as the negative of the row sums of the matrix $\boldsymbol{S}$.

The entries in (8.4.2) are derived as follows. Starting from an empty system with the arrival process in phase l, (i.e., starting in state $(0,l)$) the CTMC under study can go to state $(1,j,k)$, $1 \leq j \leq n$, $1 \leq k \leq m$, through an arrival which triggers the service to start in phase j and this event occurs at a rate $d_{lk}^{(1)}\beta_j$. Note that when the system is in level $\mathbf{0}$, the arrival process can possibly change its phase without producing an arrival and the transitions are governed by the entries of $\mathbf{D}_0$. When the system is in state $(1,j,l)$, $1 \leq j \leq n$, $1 \leq l \leq m$, the system can either go to state $(2,j,k)$, $1 \leq j \leq n$, $1 \leq k \leq m$, through an arrival (note that the service phase cannot change as we are talking about transitions in an infinitesimal time here) and this event occurs at a rate $d_{lk}^{(1)}$; or the system can either go to state $(0,l)$, $1 \leq l \leq m$, through a service completion (note that the arrival phase cannot change as we are talking about transitions in an infinitesimal time here) which occurs at a rate $\boldsymbol{S}_j^0$, $1 \leq j \leq n$; or the system can remain in level $\mathbf{1}$ through no arrival and no service completion but possible change in the arrival or service (but not both) phase and the rates are governed by the entries of the matrix $\mathbf{S} \oplus \mathbf{D}_0$. For states away from the boundary, a service completion will result in initiating the customer at the head of the queue service immediately. Hence, we see that when the system is in state (i,r,l), it can go to $(i-1,j,1)$, $i \geq 2, 1 \leq r, j \leq n, 1 \leq l \leq m$, with rate $\boldsymbol{S}_r^0\beta_j$; or it can go to state $(i+1,r,k)$, $i \geq 2, 1 \leq r \leq n, 1 \leq l,k \leq m$, through an

arrival and this event occurs with rate $d_{lk}^{(1)}$; or it can remain in level $\boldsymbol{i}$ through no arrival and no service completion but possible change in the arrival or service phase (but not both) and the rates are governed by the entries of the matrix $\mathbf{S} \oplus \mathbf{D}_0$.

Suppose that $\boldsymbol{x}$ partitioned as $\boldsymbol{x} = (\boldsymbol{x}_0, \boldsymbol{x}_1, \boldsymbol{x}_2 ...)$ denotes the steady-state probability vector of the generator $\boldsymbol{Q}$. The vector x satisfies:

$$\boldsymbol{x}\boldsymbol{Q} = \mathbf{0}, \quad \boldsymbol{x}\boldsymbol{e} = 1. \tag{8.4.3}$$

First note that the kth component of the vector $\boldsymbol{x}_0$ that is of dimension m gives the steady-state probability that the system is empty at an arbitrary time with the arrival process in phase k; the vectors $\boldsymbol{x}_i, i \geq 1$, are of dimension m n such that the jth block vector of dimension m gives the steady-state probability that at an arbitrary time there are i customers in the system and the current service phase is in j and the arrival process is in one of m phases.

Under the stability condition, $\lambda < \mu$, the steady-state equations in (8.4.3) are solved as follows. Expanding the equations in (8.4.3), we get

$$\begin{aligned} \boldsymbol{x}_0\boldsymbol{D}_0 + \boldsymbol{x}_1(\boldsymbol{S}^0 \otimes \boldsymbol{I}) &= \mathbf{0}, && (8.4.4)\\ \boldsymbol{x}_0(\boldsymbol{\beta} \otimes \boldsymbol{D}_1) + \boldsymbol{x}_1(\boldsymbol{S} \oplus \boldsymbol{D}_0) + \boldsymbol{x}_2(\boldsymbol{S}^0\boldsymbol{\beta} \otimes \boldsymbol{I}) &= \mathbf{0} && (8.4.5)\\ \boldsymbol{x}_{i-1}(\boldsymbol{I} \otimes \boldsymbol{D}_1) + \boldsymbol{x}_i(\boldsymbol{S} \oplus \boldsymbol{D}_0) + \boldsymbol{x}_{i+1}(\boldsymbol{S}^0\boldsymbol{\beta} \otimes \boldsymbol{I}) &= \mathbf{0}, \quad i \geq 2, && (8.4.6) \end{aligned}$$

with the normalizing equation

$$\sum_{i=0}^{\infty} \boldsymbol{x}_i \boldsymbol{e} = 1. \tag{8.4.7}$$

Since the generator $\boldsymbol{Q}$ in (8.4.2) has quasi-birth and death (QBD) structure, it follows from Neuts (1981) that the solution to the equations in (8.4.4)–(8.4.6) can be obtained as

$$\boldsymbol{x}_0\boldsymbol{D}_0 + \boldsymbol{x}_1(\boldsymbol{S}^0 \otimes \boldsymbol{I}) = \mathbf{0}, \tag{8.4.8}$$

$$\boldsymbol{x}_0(\boldsymbol{\beta} \otimes \boldsymbol{D}_1) + \boldsymbol{x}_1(\boldsymbol{S} \oplus \boldsymbol{D}_0) + \boldsymbol{x}_2(\boldsymbol{S}^0\boldsymbol{\beta} \otimes \boldsymbol{I}) = \mathbf{0}, \tag{8.4.9}$$

$$\boldsymbol{x}_i = \boldsymbol{x}_1\boldsymbol{R}^{i-1}, \quad i \geq 2, \tag{8.4.10}$$

where $\boldsymbol{R}$ satisfies the matrix-quadratic equation

$$\boldsymbol{R}^2(\boldsymbol{S}^0\boldsymbol{\beta} \otimes \boldsymbol{I}) + \boldsymbol{R}(\boldsymbol{S} \oplus \boldsymbol{D}_0) + (\boldsymbol{I} \otimes \boldsymbol{D}_1) = 0. \tag{8.4.11}$$

The special structure of the coefficient matrices appearing in (8.4.8), (8.4.9), and (8.4.11) can be exploited, especially when m and n are large, and employ (block) Gauss–Seidel iterative procedure (see Stewart (1994)) in solving these equations. In the case when either m or n is of reasonable size, one can use the more efficient logarithmic reduction algorithm (Latouche and Ramaswami 1999) or the cyclic reduction algorithm (see Bini and Meini (1995)) for computing $\mathbf{R}$ directly. Very briefly, we outline the logarithmic reduction algorithm which is used here to discuss some numerical examples.

Logarithmic Reduction Algorithm Latouche and Ramaswami (1999) The key steps involved in computing the $\mathbf{R}$ matrix which is the solution to the matrix-quadratic equation of the form $\boldsymbol{R}^2\boldsymbol{A}_2 + \boldsymbol{R}\boldsymbol{A}_1 + \boldsymbol{A}_0 = 0$ are as follows:

Step 0: Let $\boldsymbol{H} \leftarrow (-\boldsymbol{A}_1)^{-1}\boldsymbol{A}_0$; $\boldsymbol{L} \leftarrow (-\boldsymbol{A}_1)^{-1}\boldsymbol{A}_2$; $\boldsymbol{G} = \boldsymbol{L}$ and $\boldsymbol{T} = \boldsymbol{H}$.

Step 1: Let $\boldsymbol{U} = \boldsymbol{H}\boldsymbol{L} + \boldsymbol{L}\boldsymbol{H}$ and $\boldsymbol{M} = \boldsymbol{H}^2$; and

$$\begin{aligned} \boldsymbol{H} &\leftarrow (\boldsymbol{I} - \boldsymbol{U})^{-1}\boldsymbol{M} \\ \boldsymbol{M} &\leftarrow \boldsymbol{L}^2 \\ \boldsymbol{L} &\leftarrow (\boldsymbol{I} - \boldsymbol{U})^{-1}\boldsymbol{M} \\ \boldsymbol{G} &\leftarrow \boldsymbol{G} + \boldsymbol{T}\boldsymbol{L} \\ \boldsymbol{T} &\leftarrow \boldsymbol{T}\boldsymbol{H} \end{aligned}$$

Continue Step 1 until $||\boldsymbol{e} - \boldsymbol{G}e||_\infty < \epsilon$, where ϵ is a very small prespecified value. Usually, ϵ is taken to be 10^{-7} or so.

Step 2: The matrix $\mathbf{R}$ is obtained as $\boldsymbol{R} = -\boldsymbol{A}_0(\boldsymbol{A}_1 + \boldsymbol{A}_0\boldsymbol{G})^{-1}$.

Note: (1) In the process of obtaining the matrix $\mathbf{R}$ one also gets the matrix $\mathbf{G}$, the matrix of the conditional probabilities of the first passage time. Recall that $\mathbf{G}$ is obtained as the minimal nonnegative solution to a nonlinear matrix equation: $\boldsymbol{A}_0\boldsymbol{G}^2 + \boldsymbol{A}_1\boldsymbol{G} + \boldsymbol{A}_2 = 0$.

(2) It can easily be verified that $\boldsymbol{G} = -(\boldsymbol{A}_1 + \boldsymbol{R}\boldsymbol{A}_2)^{-1}\boldsymbol{A}_2$.

Once $\mathbf{R}$ matrix is obtained (either by using logarithmic reduction or by any other method), the equations (8.4.8) and (8.4.9) along with the normalizing equation (8.4.7) can be obtained by solving the following equations. This is due to the fact that $\boldsymbol{x}_0 = \boldsymbol{x}_1(\boldsymbol{S}^0\boldsymbol{\beta} \otimes (-\boldsymbol{D}_0^{-1}))$.

$$\boldsymbol{x}_1[(\boldsymbol{S}^0\boldsymbol{\beta} \otimes (-\boldsymbol{D}_0^{-1}\boldsymbol{D}_1)) + (\boldsymbol{S} \oplus \boldsymbol{D}_0) + \boldsymbol{R}(\boldsymbol{S}^0\boldsymbol{\beta} \otimes \boldsymbol{I})] = \boldsymbol{0}, \tag{8.4.12}$$

subject to

$$\boldsymbol{x}_1[(\boldsymbol{S}^0 \otimes (-\boldsymbol{D}_0^{-1}\boldsymbol{e})) + (\boldsymbol{I} - \boldsymbol{R})^{-1}\boldsymbol{e}] = 1. \tag{8.4.13}$$

As mentioned earlier, the equations where the coefficient matrices have any special structure should be exploited.

In any computational scheme one should have a set of internal accuracy checks so as to make sure the computations are properly and accurately computed. In the current model, one can use the following intuitively obvious results.

$$\boldsymbol{x}_0 + \sum_{i=1}^{\infty} \boldsymbol{x}_i(\boldsymbol{e} \otimes \boldsymbol{I}) = \boldsymbol{\pi}, \tag{8.4.14}$$

$$\sum_{i=1}^{\infty} \boldsymbol{x}_i(\boldsymbol{I} \otimes \boldsymbol{e}) = \lambda\boldsymbol{\beta}(-S)^{-1}, \tag{8.4.15}$$

where $\boldsymbol{\pi}$ is the steady-state probability vector of the arrival process as defined in (8.3.2). The results are intuitive. For example, the left-hand side of (8.4.14) gives the steady-state probability vector of the arrival process and it should be

equal to $\boldsymbol{\pi}$. The right-hand side quantities of (8.4.14) and (8.4.15) are in terms of the data and hence these two intuitively obvious results will serve as accuracy check in the computation of the steady-state probability vector, $\boldsymbol{x}$.

Even though the results in (8.4.14) and (8.4.15) are intuitive, one can check these algebraically. To see this, post-multiplying equations (8.4.5) and (8.4.6) by $(\boldsymbol{e} \otimes \boldsymbol{I})$ and adding the resulting equations along with (8.4.4), we get

$$\left[\boldsymbol{x}_0 + \sum_{i=1}^{\infty} \boldsymbol{x}_i(\boldsymbol{e} \otimes \boldsymbol{I})\right](\boldsymbol{D}_0 + \boldsymbol{D}_1) = \boldsymbol{0},$$

from which the equation (8.4.14) follows by the uniqueness of $\boldsymbol{\pi}$.

Similarly, post-multiplying equation (8.4.4) by $\boldsymbol{e}$ and the equations (8.4.5) and (8.4.6) by $(\boldsymbol{I} \otimes \boldsymbol{e})$ and adding the resulting equations, we get

$$\sum_{i=1}^{\infty} \boldsymbol{x}_i(\boldsymbol{I} \otimes \boldsymbol{e})(\boldsymbol{S} + \boldsymbol{S}^0\boldsymbol{\beta}) = \boldsymbol{0},$$

from which using the uniqueness of the steady-state probability vector of $\boldsymbol{S} + \boldsymbol{S}^0\boldsymbol{\beta}$, we get

$$\sum_{i=1}^{\infty} \boldsymbol{x}_i(\boldsymbol{I} \otimes \boldsymbol{e}) = c\boldsymbol{\beta}(-\boldsymbol{S})^{-1},$$

where c is the normalizing constant and is given by λ.

To show that the normalizing constant c is given by λ, we post-multiply equation (8.4.4) by $\boldsymbol{e}$ and the equations (8.4.5) and (8.4.6) by $(\boldsymbol{e} \otimes \boldsymbol{e})$. Now, adding the resulting equations we get

$$\boldsymbol{x}_0\boldsymbol{D}_1\boldsymbol{e} + \sum_{i=1}^{\infty} \boldsymbol{x}_i(\boldsymbol{e} \otimes \boldsymbol{D}_1\boldsymbol{e}) = \sum_{i=1}^{\infty} \boldsymbol{x}_i(\boldsymbol{S}^0 \otimes \boldsymbol{e}),$$

from which the stated result follows. Again, intuitively it is obvious that $\sum_{i=1}^{\infty} \boldsymbol{x}_i(\boldsymbol{S}^0 \otimes \boldsymbol{e}) = \boldsymbol{\lambda}$ since in steady-state the rate at which customers leave should be equal to the rate at which they enter.

Observe that equation (8.4.14) implies that the probability that the system is idle is, as expected, given by $1 - \lambda/\mu$.

Once the steady-state probability vector is obtained, a number of system performance measures can be computed for comparing various scenarios and models useful in practical applications.

Writing L and Lq as the mean number of customers in the system and the number in the queue, respectively, we may derive them as follows:

$$\begin{aligned} L &= \sum_{i=1}^{\infty} i\boldsymbol{x}_i\boldsymbol{e} \\ &= \boldsymbol{x}_1 \sum_{i=1}^{\infty} i\boldsymbol{R}^{i-1}\boldsymbol{e} \\ &= \boldsymbol{x}_1(\boldsymbol{I} - \boldsymbol{R})^{-2}\boldsymbol{e}. \end{aligned} \tag{8.4.16}$$

$$\begin{aligned} L_q &= \sum_{i=1}^{\infty}(i-1)\boldsymbol{x}_i\boldsymbol{e} \\ &= \boldsymbol{x}_1\sum_{i=1}^{\infty}(i-1)\boldsymbol{R}^{i-1}\boldsymbol{e} \\ &= \boldsymbol{x}_1 R(\boldsymbol{I}-\boldsymbol{R})^{-2}\boldsymbol{e}. \end{aligned} \tag{8.4.17}$$

Noting that $\boldsymbol{x}_1\boldsymbol{R}(\boldsymbol{I}-\boldsymbol{R})^{-2}\boldsymbol{e} = \boldsymbol{x}_1(\boldsymbol{I}-\boldsymbol{R})^{-2}\boldsymbol{e} - \boldsymbol{x}_1(\boldsymbol{I}-\boldsymbol{R})^{-1}\boldsymbol{e}$ and the fact that $\boldsymbol{x}_1(\boldsymbol{I}-\boldsymbol{R})^{-1}\boldsymbol{e} = \lambda/\mu$, we see that $L = L_q + \frac{\lambda}{\mu}$, as expected.

Steady-State Probability at Arrival Epoch Now, we look at the steady-state probability vector at arrival epochs. Since we have non-Poisson arrivals, we know that PASTA property does not hold good. We need this probability to discuss stationary waiting time distribution of an arriving customer.

Suppose that $\boldsymbol{y}$ partitioned as $\boldsymbol{y} = (\boldsymbol{y}_0, \boldsymbol{y}_1, \boldsymbol{y}_2, \cdots)$ denotes the steady-state probability vector at arrival epoch. That is, the vector of $\boldsymbol{y}_0$ dimension m gives the steady-state probability that an arriving customer will find the server idle; the vector of $\boldsymbol{y}_i$ dimension mn gives the steady-state probability that an arriving customer will find i customers in the system with the server busy serving a customer. It is easy to verify that

$$\begin{aligned} \boldsymbol{y}_0 &= \frac{1}{\lambda}\boldsymbol{x}_0\boldsymbol{D}_1, \\ \boldsymbol{y}_i &= \frac{1}{\lambda}\boldsymbol{x}_i(I\otimes\boldsymbol{D}_1), \quad i \geq 1. \end{aligned} \tag{8.4.18}$$

Stationary Waiting Time Distribution of an Arriving Customer Suppose that Y_q denotes the random variable that an arriving customer has to wait in the queue before entering into service. Let $w_q(s)$ denote the Laplace–Stieltjes transform (LST) of Y_q and let $f(s)$ denote the LST of the service time distribution. First, note that $f(s) = \boldsymbol{\beta}(s\boldsymbol{I}-\boldsymbol{S})^{-1}\boldsymbol{S}^0$. Using the law of total probability, it is easy to verify that

$$\begin{aligned} w_q(s) &= \frac{1}{\lambda}\boldsymbol{x}_0\boldsymbol{D}_1\boldsymbol{e} + \frac{1}{\lambda}\sum_{i=1}^{\infty}\boldsymbol{x}_i(\boldsymbol{1}\otimes\boldsymbol{D}_1\boldsymbol{e})(s\boldsymbol{I}-\boldsymbol{S})^{-1}\boldsymbol{S}^0[f(s)]^{i-1} \\ &= \frac{1}{\lambda}\boldsymbol{x}_0\boldsymbol{D}_1\boldsymbol{e} + \frac{1}{\lambda}\boldsymbol{x}_1[\boldsymbol{I}-f(s)\boldsymbol{R}]^{-1}[(sI-\boldsymbol{S})^{-1}\boldsymbol{S}^0\otimes\boldsymbol{D}_1\boldsymbol{e}]. \end{aligned} \tag{8.4.19}$$

The mean, W_q, waiting time in the queue is obtained as

$$\begin{aligned} W_q &= -w_q'(s)]_{s=0} \\ &= \frac{1}{\lambda\mu}\boldsymbol{x}_1(\boldsymbol{I}-\boldsymbol{R})^{-2}\boldsymbol{R}(\boldsymbol{e}\otimes\boldsymbol{D}_1\boldsymbol{e}) + \frac{1}{\lambda}\boldsymbol{x}_1(I-\boldsymbol{R})^{-1}(-\boldsymbol{S}^{-1}\boldsymbol{e}\otimes\boldsymbol{D}_1\boldsymbol{e}). \end{aligned} \tag{8.4.20}$$

Using the facts that

$$R(\boldsymbol{S}^0\otimes\boldsymbol{e}) = \boldsymbol{e}\otimes\boldsymbol{D}_1\boldsymbol{e}, \tag{8.4.21}$$

and

$$\frac{1}{\mu}R^2(\boldsymbol{S}^0 \otimes \boldsymbol{e}) = R[\boldsymbol{e} \otimes \boldsymbol{e} + (-\boldsymbol{S}^{-1})\boldsymbol{e} \otimes \boldsymbol{D}_1\boldsymbol{e}] - [(-\boldsymbol{S}^{-1})\boldsymbol{e} \otimes \boldsymbol{D}_1\boldsymbol{e}], \quad (8.4.22)$$

Little's Law, $L_q = \lambda W_q$, can be easily verified.

Given below are some special cases of $MAP/PH/1$ queue.

M/PH/1 Queue Assume that the arrivals occur according to a Poisson process of rate λ and the service times are assumed to be of phase type with irreducible representation given by $(\boldsymbol{\beta}, \boldsymbol{S})$ of order m. Let the service rate be denoted by μ and recall that $\boldsymbol{\mu} = [\boldsymbol{\beta}(-\boldsymbol{S})^{-1}\boldsymbol{e}]^{-1}$. Note that by setting the parameters of MAP representation as:

$$m = 1, \quad \boldsymbol{D}_0 = -\lambda, \quad \boldsymbol{D}_1 = \lambda,$$

the arrival process reduces to a Poisson process with rate λ.

Let $X_1(t)$ and $X_2(t)$ denote, respectively, the number of customers in the system and the phase of the service (if the server is busy) at time t. Verify that the two-dimensional process $\{X_1(t), X_2(t)) : t \geq 0\}$ is CTMC on the state space Ω given by

$$\Omega = \{0\} \cup \{(i, j) : i \geq 1, 1 \leq j \leq m\}. \quad (8.4.23)$$

Defining the set, $\boldsymbol{i}$, to be the set of states $\{(i, j) : i \geq 1, 1 \leq j \leq m\}$, for $i \geq 1$, the generator of the CTMC with state space Ω is given by

$$\begin{array}{c} \\ 0 \\ \mathbf{1} \\ \mathbf{Q} = \mathbf{2} \\ \mathbf{3} \\ \vdots \end{array} \begin{array}{c} \begin{array}{ccccccc} & & & & & & \end{array} \\ \begin{bmatrix} -\lambda & \lambda\boldsymbol{\beta} & 0 & 0 & 0 & \dots \\ \boldsymbol{S}^0 & \boldsymbol{S} - \lambda\boldsymbol{I} & \lambda\boldsymbol{I} & 0 & 0 & \dots \\ 0 & \boldsymbol{S}^0\boldsymbol{\beta} & \boldsymbol{S} - \lambda\boldsymbol{I} & \lambda\boldsymbol{I} & 0 & \dots \\ 0 & 0 & \boldsymbol{S}^0\boldsymbol{\beta} & \boldsymbol{S} - \lambda\boldsymbol{I} & \lambda\boldsymbol{I} & \dots \\ \vdots & \vdots & \dots & \ddots & \ddots & \ddots \end{bmatrix} \end{array}. \quad (8.4.24)$$

The steady-state vector $\boldsymbol{x} = (x_0, \boldsymbol{x}_1, \boldsymbol{x}_2,)$ for the generator given in (8.4.24) can be explicitly solved for the current model as follows:

$$x_0 = 1 - \rho, \quad (8.4.25)$$

$$\boldsymbol{x}_i = (1 - \rho)\boldsymbol{\beta}\boldsymbol{R}^j, \quad \text{for } i \geq 1, \quad (8.4.26)$$

and the matrix $\mathbf{R}$ is explicitly obtained as

$$\boldsymbol{R} = \lambda(\lambda\boldsymbol{I} - \lambda\boldsymbol{e}\boldsymbol{\beta} - \boldsymbol{S})^{-1}. \quad (8.4.27)$$

Note that the inverse appearing in the R matrix given in (8.4.27) exists since $(\boldsymbol{S} - \lambda\boldsymbol{I} + \lambda\boldsymbol{e}\boldsymbol{\beta})$ is a stable matrix.

The waiting distribution of the time spent in the queue of an arriving customer is of phase type with representation given by $(\lambda\boldsymbol{\beta}(-\boldsymbol{S})^{-1}, \boldsymbol{S} + \lambda\boldsymbol{S}^0\boldsymbol{\beta}(-\boldsymbol{S})^{-1})$.

The matrix $\mathbf{G}$ is of the form, $\boldsymbol{G} = \boldsymbol{e}\boldsymbol{\beta}$.

PH/M/1 Queue Assume that the inter-arrival times follow a PH-distribution with representation given by $(\boldsymbol{\alpha}, \boldsymbol{T})$ of order m and the service times are exponential with rate μ. Note that by setting the parameters of MAP representation as:

$$\boldsymbol{D}_0 = \boldsymbol{T}, \quad \boldsymbol{D}_1 = \boldsymbol{T}^0\boldsymbol{\alpha},$$

the arrival process reduces to a phase type process with representation $(\boldsymbol{\alpha}, T)$ of order m.

Let $X_1(t)$ and $X_2(t)$ denote, respectively, the number of customers in the system and the phase of the arrival process at time t. Verify that the two-dimensional process $\{(X_1(t), X_2(t)) : t \geq 0$ is CTMC on the state space Ω given by

$$\Omega = \{0\} \cup \{(i,j) : i \geq 1, 1 \leq j \leq m\}. \tag{8.4.28}$$

Defining the set, $\boldsymbol{i}$, to be the set of states $\{(i,j) : i \geq 1, 1 \leq j \leq m\}$, for $i \geq 1$, the generator of the CTMC with state space Ω is given by

$$\begin{array}{c} \\ 0 \\ \mathbf{1} \\ \mathbf{Q} = \mathbf{2} \\ \mathbf{3} \\ \vdots \end{array} \begin{array}{c} \begin{array}{ccccccc} & & & & & & \end{array} \\ \begin{bmatrix} T & \boldsymbol{T}^0\boldsymbol{\alpha} & 0 & 0 & 0 & \dots \\ \mu I & \boldsymbol{T} - \mu\boldsymbol{I} & \boldsymbol{T}^0\boldsymbol{\alpha} & 0 & 0 & \dots \\ 0 & \mu I & \boldsymbol{T} - \mu\boldsymbol{I} & \boldsymbol{T}^0\boldsymbol{\alpha} & 0 & \dots \\ 0 & 0 & \mu I & \boldsymbol{T} - \mu\boldsymbol{I} & \boldsymbol{T}^0\boldsymbol{\alpha} & \dots \\ \vdots & \vdots & \dots & \ddots & \ddots & \ddots \end{bmatrix} \end{array}. \tag{8.4.29}$$

The steady-state vector $\boldsymbol{x} = (\boldsymbol{x}_0, \boldsymbol{x}_1, \boldsymbol{x}_2,)$ for the generator given in (8.4.29) can be solved for the current model as follows:

$$\boldsymbol{x}_i = \boldsymbol{x}_0\boldsymbol{R}^i, \text{ for } i \geq 1, \tag{8.4.30}$$

where $\boldsymbol{x}_0$ is solved using the following equations

$$\boldsymbol{x}_0(\boldsymbol{T} + \mu\boldsymbol{R}) = 0 \text{ and } \boldsymbol{x}_0(\boldsymbol{I} - \boldsymbol{R})^{-1}\boldsymbol{e} = 1, \tag{8.4.31}$$

and the matrix R satisfies the matrix quadratic equation

$$\mu\boldsymbol{R}^2 + \boldsymbol{R}(\boldsymbol{T} - \mu\boldsymbol{I}) + \boldsymbol{T}^0\boldsymbol{\alpha} = 0. \tag{8.4.32}$$

The matrix G is of explicitly obtained as:

$$G = \mu[\mu\boldsymbol{I} - \boldsymbol{T} - \mu\boldsymbol{T}^0\boldsymbol{\alpha}(-\boldsymbol{T})^{-1}]^{-1} + \frac{1 - \rho^{-1}}{\boldsymbol{\alpha}(\mu\boldsymbol{I} - \boldsymbol{T})^{-1}\boldsymbol{T}^0}(\mu\boldsymbol{I} - \boldsymbol{T})^{-1}\boldsymbol{T}^0\boldsymbol{g}, \tag{8.4.33}$$

where $\boldsymbol{g}$ satisfies $\boldsymbol{g}\boldsymbol{G} = \boldsymbol{g}$ and $\boldsymbol{g}\boldsymbol{e} = 1$. Further, $\boldsymbol{g}$ is the left eigenvector of $\mu[\mu\boldsymbol{I} - \boldsymbol{T}^0\boldsymbol{\alpha}(-\boldsymbol{T})^{-1}]^{-1}$corresponding to the positive eigenvalue, say, η, given by

$$\eta = 1 + \frac{\rho^{-1} - 1}{\boldsymbol{\alpha}(\mu\boldsymbol{I} - \boldsymbol{T})^{-1}\boldsymbol{T}^0}\boldsymbol{g}(\mu\boldsymbol{I} - \boldsymbol{T})^{-1}\boldsymbol{T}^0.$$

M/M/1 Queue Assume that the inter-arrival times are exponential with parameter λ and service times are exponential with rate μ. Taking $m = 1$, $\boldsymbol{D}_0 = -\lambda$, $\boldsymbol{D}_1 = -\lambda$, $n = 1$, $\beta = 1$, $\boldsymbol{S} = -\mu$ in the $MAP/PH/1$ queue or $n = 1$, $\beta = 1$, $\boldsymbol{S} = -\mu$ in the $M/PH/1$ queue, it can be verified that $R = \rho$ and hence

$$x_0 = 1 - \rho,$$

$$x_i = (1 - \rho)\rho^i, \quad i \geq 1,$$

and the waiting time in the queue is of phase type with representation $(\rho, -(\mu - \lambda))$ of dimension 1. This agrees, as it should, with the results from Section 4.2 dealing with $M/M/1$ queue. Note that here $\boldsymbol{G}$ is a scalar and its value is 1.

8.5 Numerical Examples

In this section, we look at different single server queueing models and discuss the qualitative aspects of these models numerically. Toward the end, we consider a number of special cases.

Example 8.5.1 Customers arrive at processing center according to a MAP with $(\boldsymbol{D}_0, \boldsymbol{D}_1)$ of order m. We consider different arrival processes by varying the values of the entries of these representation matrices. The service times are independent and identically distributed with a common PH-distribution with representation $(\boldsymbol{\beta}, \boldsymbol{S})$ of order n. We look at different service times by varying the values of the representation parameters. Use $MAP/PH/1$ queueing model to discuss how the performance measure, the mean queue length, behaves under different special cases. For arrival process consider the following five MAP representations:

ERL (Erlang of order 2):

$$\mathbf{D}_0 = \begin{bmatrix} -2 & 2 \\ 0 & -2 \end{bmatrix}, \mathbf{D}_1 = \begin{bmatrix} 0 & 0 \\ 2 & 0 \end{bmatrix}.$$

EXP (Exponential):

$$\mathbf{D}_0 = [-1], \quad \mathbf{D}_1 = [1].$$

HEX (Hyperexponential):

$$\mathbf{D}_0 = \begin{bmatrix} -1.90 & 0 \\ 0 & -0.19 \end{bmatrix}, \mathbf{D}_1 = \begin{bmatrix} 1.710 & 0.190 \\ 1.171 & 0.019 \end{bmatrix}$$

MNC (MAP with negative correlation):

$$\mathbf{D}_0 = \begin{bmatrix} -1.00222 & 1.00222 & 0 \\ 0 & -1.00222 & 0 \\ 0 & 0 & -225.75 \end{bmatrix}, \mathbf{D}_1 = \begin{bmatrix} 0 & 0 & 0 \\ 0.01002 & 0 & 0.99220 \\ 223.4925 & 0 & 2.2575 \end{bmatrix}.$$

MPC (MAP with negative correlation):

$$\mathbf{D}_0=\begin{bmatrix} -1.00222 & 1.00222 & 0 \\ 0 & -1.00222 & 0 \\ 0 & 0 & -225.75 \end{bmatrix}, \mathbf{D}_1=\begin{bmatrix} 0 & 0 & 0 \\ 0.99220 & 0 & 0.01002 \\ 2,2575 & 0 & 223.4925 \end{bmatrix}.$$

Note that all these five MAP processes have the arrival rate of 1. However, these are qualitatively different in that they have different variance and correlation structure. The first three arrival processes, namely ERL, EXP, and HEX, correspond to renewal processes and so the correlation is 0. The arrival processes labeled MNC has correlated arrivals with correlation between two successive inter-arrival times given by -0.4889, and the arrivals corresponding to the process labeled MAP has positive correlation with a value 0.4889. The ratio of the standard deviations of the inter-arrival times of these five arrival processes with respect to ERL are, respectively, 1, 1.4142, 3.1745, 1.9934, and 1.9934.

For service times, we consider the following three PH-distributions:
ERS (Erlang of order 2):

$$\boldsymbol{\beta} = (1, 0), \quad \mathbf{S} = \begin{bmatrix} -2 & 2 \\ 0 & -2 \end{bmatrix}.$$

EXS (Exponential):

$$\boldsymbol{\beta} = (1), \quad \mathbf{S} = [-1].$$

HES (Hyperexponential):

$$\boldsymbol{\beta} = (0.8, 0.2), \quad \mathbf{S} = \begin{bmatrix} -2.8 & 0 \\ 0 & -0.28 \end{bmatrix}.$$

Note that these three PH-distributions will be normalized so as to have a specific service rate. For example, these all have a rate 1. If we want to have the service rate to be 1.2, the matrix S will be normalized for these three distributions, respectively, as follows:

$$\mathbf{S} = \begin{bmatrix} -2.4 & 2.4 \\ 0 & -2.4 \end{bmatrix}, \quad \mathbf{S} = [-1.2], \quad \mathbf{S} = \begin{bmatrix} -3.36 & 0 \\ 0 & -0.336 \end{bmatrix}.$$

Further, these are qualitatively different in that they have different variance structure. The ratio of the standard deviations of the service times of these three PH-distributions with respect to ERS are, respectively, 1, 1.4142, and 2.9347.

In the following table we list the performance measure, the mean queue length, for various combinations of arrival and service times. To have a specific value for the traffic intensity we will adjust the service rate.

Table 8.1 Mean queue length of $MAP/PH/1$ model

ρ		ERL	EXP	HEX	MNC	MPC
0.10	ERS	0.002	0.008	0.015	0.050	5.264
	EXS	0.003	0.011	0.021	0.052	5.266
	HES	0.016	0.030	0.052	0.063	5.281
0.50	ERS	0.195	0.375	1.075	0.532	49.305
	EXS	0.309	0.500	1.294	0.641	49.419
	HES	1.117	1.327	2.429	1.427	50.222
0.80	ERS	1.492	2.400	8.642	2.618	198.544
	EXS	2.275	3.200	9.575	3.393	199.326
	HES	7.542	8.490	15.288	8.629	204.583
0.90	ERS	3.923	6.075	22.331	6.322	447.530
	EXS	5.930	8.100	24.487	8.319	449.535
	HES	19.295	21.490	38.294	21.652	462.889
0.95	ERS	8.888	13.538	49.956	13.819	945.599
	EXS	13.381	18.050	54.596	18.302	950.091
	HES	43.195	47.888	84.845	48.082	979.892
0.99	ERS	48.861	73.508	271.455	74.018	4930.391
	EXS	73.343	98.010	296.083	98.491	4954.871
	HES	235.335	260.027	458.505	260.447	5116.850

A quick look at Table 8.1 reveals the following observations.

- As the traffic intensity increases, the mean queue length increases for all combinations of arrival and service times.
- As the variation in service times increases, the mean queue length appears to increase. The rate of increase depends not only on the traffic intensity but also on the type of arrival process. For example, the rate of increase (as a function of the variability in the service times) is much higher for all arrival processes except MPC one, as the traffic intensity is increased. However, for the MPC arrival process the rate (as a function of the variability in the service times) is not that significant.
- Looking at the arrival processes (ERL, EXP, and HEX) whose inter-arrival times form a renewal process, the mean queue length appears to increase as the variation in the inter-arrival times increases. This appears to be the case for all cases. However, with regards to MNC and MPC arrivals, which are such that the successive inter-arrival times are, respectively, negatively and positively correlated, we see the mean queue length appears to behave very differently. It should be pointed out that these two arrival processes have same variance also.
- In summary, not only the variance of inter-arrival as well as service times play a key role but also the correlation of the inter-arrival times.

Table 8.2 R and **G** matrices for *ERL/ERS*/1 model

Model	**R**	**G**
$\rho = 0.10$	$\begin{bmatrix} 0 & 0 & 0 & 0 \\ 0.093 & 0.009 & 0.084 & 0.016 \\ 0 & 0 & 0 & 0 \\ 0.001 & 0 & 0.092 & 0.008 \end{bmatrix}$	$\begin{bmatrix} 0.844 & 0.156 & 0 & 0 \\ 0.131 & 0.869 & 0 & 0 \\ 0.916 & 0.084 & 0 & 0 \\ 0.071 & 0.929 & 0 & 0 \end{bmatrix}$
$\rho = 0.50$	$\begin{bmatrix} 0 & 0 & 0 & 0 \\ 0.421 & 0.188 & 0.281 & 0.219 \\ 0 & 0 & 0 & 0 \\ 0.039 & 0.031 & 0.360 & 0.140 \end{bmatrix}$	$\begin{bmatrix} 0.562 & 0.438 & 0 & 0 \\ 0.246 & 0.754 & 0 & 0 \\ 0.719 & 0.281 & 0 & 0 \\ 0.158 & 0.842 & 0 & 0 \end{bmatrix}$
$\rho = 0.95$	$\begin{bmatrix} 0 & 0 & 0 & 0 \\ 0.787 & 0.682 & 0.404 & 0.546 \\ 0 & 0 & 0 & 0 \\ 0.155 & 0.210 & 0.567 & 0.383 \end{bmatrix}$	$\begin{bmatrix} 0.425 & 0.575 & 0 & 0 \\ 0.244 & 0.756 & 0 & 0 \\ 0.596 & 0.404 & 0 & 0 \\ 0.171 & 0.829 & 0 & 0 \end{bmatrix}$
$\rho = 0.99$	$\begin{bmatrix} 0 & 0 & 0 & 0 \\ 0.820 & 0.742 & 0.412 & 0.578 \\ 0 & 0 & 0 & 0 \\ 0.168 & 0.236 & 0.582 & 0.408 \end{bmatrix}$	$\begin{bmatrix} 0.416 & 0.584 & 0 & 0 \\ 0.243 & 0.757 & 0 & 0 \\ 0.588 & 0.412 & 0 & 0 \\ 0.172 & 0.828 & 0 & 0 \end{bmatrix}$

Table 8.2 displays **R** and **G** matrices for one particular case (others will be left for the reader as exercises). Before we discuss the entries of Table 8.2, recall that the (j,k)th entry of **R** matrix gives the mean time spent in state $(i+1,k)$ before the first return to level $\boldsymbol{i}$, given that the process starts in state (i,j) expressed in the time unit measured in terms of the mean time spent in state (i,j). The (j,k)th entry of **G** matrix gives the conditional probability of returning to level i for the first time by visiting the state (i,k) given that the process started in state $(i+1,j)$.

Now looking at the entries of Table 8.2, we observe the following:

- The only way the system can move to level $\boldsymbol{i}+1$ starting from level $\boldsymbol{i}$ is for the arrival process to be in phase 2 (due to arrival process being an Erlang of order 2) and the next transition to occur is not a service completion. Thus, starting from level $\boldsymbol{i}$ with the arrival process being in phase 1, the event "first return to level $\boldsymbol{i}$" by moving to level $\boldsymbol{i}+1$ can never occur and that leads to seeing the first and third rows of R to have zero entries. Note, however, that the process can return to level $\boldsymbol{i}$ for the first time by not visiting level $\boldsymbol{i}+1$ and hence the mean time spent in state $(i+1,k)$ is zero.

- As the traffic intensity increases, the nonzero entries of **R** appear to increase as is to be expected. The intuitive reasoning is as follows. An

increase in the arrival rate results in the process needing to take more time to return to level $\boldsymbol{i}$ for the first time by visiting the states to the right of the level $\boldsymbol{i}$ more often than in cases where the arrival rate is small.

- Since a return to level $\boldsymbol{i}$ for the first time after spending time in state $(i+1,k)$ can occur only through a service completion and since the next customers service always has to start in service phase 1 (due to Erlang services), it is clear that the third and fourth columns of $\boldsymbol{G}$ should have zero entries.

Comparing $\mathbf{R}$ matrices for two selected cases the qualitative nature of the arrival process can be observed. In one case the successive inter-arrival times are negatively correlated and in the other they are positively correlated. The matrices for these two cases under different traffic conditions are displayed in Table 8.3. Similarly, we display the $\mathbf{G}$ matrix under similar circumstances in Table 8.4.

Looking at the entries of Table 8.3 and Table 8.4, we observe the following:

- As the traffic intensity increases, the nonzero entries of $\mathbf{R}$ appear to increase in all cases (subject to round-off errors due to three decimal places) as is to be expected.

- Observe first that for the MNC arrival process an arrival occurring from phase 3 will result in the CTMC (governing the arrival process) starting in phase 1 with a higher probability than starting in phase 3; for the MPC arrival process it is the other way; that is, an arrival from phase 3 will result in starting the CTMC in phase 3 with a higher probability.

Table 8.3 Comparing $\mathbf{R}$ matrix for $MNC/M/1$ and $MPC/M/1$ models

Model	$MNC/M/1$	$MPC/M/1$
$\rho = 0.10$	$\begin{bmatrix} 0 & 0 & 0 \\ 0.081 & 0.015 & 0.004 \\ 20.641 & 1.924 & 0.010 \end{bmatrix}$	$\begin{bmatrix} 0 & 0 & 0 \\ 0.091 & 0.009 & 0 \\ 10.879 & 10.707 & 0.990 \end{bmatrix}$
$\rho = 0.50$	$\begin{bmatrix} 0 & 0 & 0 \\ 0.279 & 0.217 & 0.0004 \\ 80.976 & 31.888 & 0.011 \end{bmatrix}$	$\begin{bmatrix} 0 & 0 & 0 \\ 0.365 & 0.136 & 0 \\ 47.897 & 63.988 & 0.990 \end{bmatrix}$
$\rho = 0.95$	$\begin{bmatrix} 0 & 0 & 0 \\ 0.404 & 0.544 & 0.004 \\ 127.516 & 86.934 & 0.012 \end{bmatrix}$	$\begin{bmatrix} 0 & 0 & 0 \\ 0.594 & 0.358 & 0 \\ 82.842 & 130.631 & 0.990 \end{bmatrix}$
$\rho = 0.99$	$\begin{bmatrix} 0 & 0 & 0 \\ 0.413 & 0.575 & 0.0004 \\ 130.989 & 92.491 & 0.012 \end{bmatrix}$	$\begin{bmatrix} 0 & 0 & 0 \\ 0.612 & 0.380 & 0 \\ 85.724 & 136.779 & 0.990 \end{bmatrix}$

Table 8.4 Comparing **G** matrix for $MNC/M/1$ and $MPC/M/1$ models

Model	$MNC/M/1$	$MPC/M/1$
$\rho = 0.10$	$\begin{bmatrix} 0.915 & 0.085 & 0 \\ 0.072 & 0.928 & 0 \\ 0.808 & 0.150 & 0.042 \end{bmatrix}$	$\begin{bmatrix} 0.916 & 0.084 & 0 \\ 0.077 & 0.923 & 0 \\ 0.478 & 0.478 & 0.044 \end{bmatrix}$
$\rho = 0.50$	$\begin{bmatrix} 0.719 & 0.281 & 0 \\ 0.158 & 0.842 & 0 \\ 0.556 & 0.435 & 0.009 \end{bmatrix}$	$\begin{bmatrix} 0.731 & 0.269 & 0 \\ 0.196 & 0.804 & 0 \\ 0.421 & 0.570 & 0.009 \end{bmatrix}$
$\rho = 0.95$	$\begin{bmatrix} 0.596 & 0.404 & 0 \\ 0.172 & 0.828 & 0 \\ 0.422 & 0.573 & 0.005 \end{bmatrix}$	$\begin{bmatrix} 0.625 & 0.375 & 0 \\ 0.233 & 0.767 & 0 \\ 0.384 & 0.611 & 0.005 \end{bmatrix}$
$\rho = 0.99$	$\begin{bmatrix} 0.588 & 0.412 & 0 \\ 0.172 & 0.828 & 0 \\ 0.414 & 0.581 & 0.004 \end{bmatrix}$	$\begin{bmatrix} 0.619 & 0.381 & 0 \\ 0.235 & 0.765 & 0 \\ 0.381 & 0.614 & 0.004 \end{bmatrix}$

Thus, in the case of MNC arrivals, the system is very likely to spend more time in phase 1 of the arrival process and less in phase 3 of the arrival process. A similar intuitive interpretation can be given for the MPC case but observing that when the arrival process is in phase 3 it gets out of that phase at a faster rate also through an arrival and so has more excursions to the right of the level $\boldsymbol{i}+1$ before returning to level $\boldsymbol{i}$.

- Interpretations for the entries of G matrix are left to the reader as an exercise.

8.6 Simulation of MAP

Simulating a MAP with representation $(\mathbf{D}_0, \mathbf{D}_1)$ of order m is carried out as follows. For the sake of simplicity in describing the simulation process, we will write these parameter matrices as follows:

$$\mathbf{D}_0 = \begin{bmatrix} -\lambda_1 & \lambda_1 q_{12} & \dots & \lambda_1 q_{1m} \\ \lambda_2 q_{21} & -\lambda_2 & \dots & \lambda_2 q_{2m} \\ \vdots & \vdots & \dots & \vdots \\ \lambda_m q_{m1} & \lambda_m q_{m2} & \dots & -\lambda_m \end{bmatrix}, \mathbf{D}_1 = \begin{bmatrix} \lambda_1 p_{11} & \lambda_1 p_{12} & \dots & \lambda_1 p_{1m} \\ \lambda_2 p_{21} & \lambda_2 p_{22} & \dots & \lambda_2 p_{2m} \\ \vdots & \vdots & \dots & \vdots \\ \lambda_m p_{m1} & \lambda_m p_{m2} & \dots & \lambda_m p_{mm} \end{bmatrix} \tag{8.6.1}$$

Step 0: Choose an initial probability vector, $\boldsymbol{a} = (a_1, ..., a_m)$ for the underlying CTMC with generator $\boldsymbol{Q} = \boldsymbol{D}_0 + \boldsymbol{D}_1$. If one is interested to simulate a stationary version, then solve for the steady state probability of $\boldsymbol{Q}$.

Step 1: Choose a random sample from the set $\{1, 2, ..., m\}$ with probabilities given by the components of $\boldsymbol{a}$, generate the starting state. Let this state be denoted by i. The MAP will be in this state for an exponential amount of time with parameter λ_i. Choose a random sample from this exponential distribution. This sample determines the sojourn time of the MAP in this state during this visit.

Step 2: Choose a random sample from the discrete probability function on the set $\{1^*, 2^*, ..., (i-1)^*, (i+1)^*, ..., m^*, 1, 2, ..., m\}$ with probabilities $\{q_{i1}, q_{i2}, ..., q_{i,i-1}, q_{i,i+1}, ..., q_{im}, p_{i1}, p_{i2}, ..., p_{im}\}$. Let the state chosen be j.

Step 3: If j belongs to the set $\{1^*, 2^*, ..., (i-1)^*, (i+1)^*, ..., m^*\}$, then the MAP made a transition without an arrival and the MAP spends in state $j (j \neq i)$ for an exponential amount of time with parameter λ_j. Otherwise, there is an arrival to the system and the MAP will be in phase $j, 1 \leq j \leq m$, for an exponential amount of time with parameter λ_j.

Step 4: If the desired number of samples from the MAP is taken or the simulation run time is expired, stop. Otherwise go to step 2 by replacing state i with state visited in step 3.

8.7 Exercises

1. Prove that the mean and variance of a random variable that follows a phase type distribution with representation $(\boldsymbol{\beta}, \boldsymbol{S})$ of order m are as given in (8.2.3).

2. Using the expressions for the mean and variance of a phase type distribution as given in (8.2.3) and the representation for Erlang of order m as given in Section 8.2, verify the simple expressions for the mean and variance for Erlang. [Hint: See Appendix A.4 for the simple expressions.]

3. Suppose that the random variable, X, denotes the inter-arrival time (i.e., times between two successive arrivals) in MAP process with representation $(\mathbf{D}_0, \mathbf{D}_1)$ of order 3 is as given below.

$$\mathbf{D}_0 = \begin{bmatrix} -2 & 2 & 0 \\ 0 & -2 & 0 \\ 0 & 0 & -450.5 \end{bmatrix}, \quad \mathbf{D}_1 = \begin{bmatrix} 0 & 0 & 0 \\ 1.98 & 0 & 0.02 \\ 4.505 & 0 & 445.995 \end{bmatrix}.$$

   Show that the correlation coefficient between two successive inter-arrival times is 0.4889.

4. For the $MAP/PH/1$ queueing model use the expressions for L_q and W_q as given in (8.4.17) and (8.4.20), respectively, to prove the Littles law for this model.

5. Show that for the $M/PH/1$ queueing model with an arrival rate λ and the service times with irreducible representation $(\lambda, \boldsymbol{S})$ of order m, the rate matrix R is given by $\lambda(\lambda I - \lambda \boldsymbol{e}\boldsymbol{\beta} - \boldsymbol{S})^{-1}$.

6. Suppose that the customers arrive according to a Poisson process with a rate of 10 per hour to a single server system. If the service times are Erlang of order 3 such that the rate of service is 12 per hour, show that R matrix for this model is given by

$$\mathbf{R} = \begin{bmatrix} 0.454 & 0.355 & 0.278 \\ 0.176 & 0.355 & 0.278 \\ 0.099 & 0.077 & 0.278 \end{bmatrix}.$$

7. For $MAP/PH/1$ model prove the expression given in (8.4.20) using (8.4.19).

8. Prove the expression in equation (8.2.5) for the steady-state vector of the generator of the PH-renewal process.

9. Suppose that the inter-arrival times of customers arriving at a single server station follow an Erlang distribution of order 3. The service times are exponential with a mean of 2 minutes. If the traffic intensity is 0.95, show that **R** and **G** matrices are given by

$$\mathbf{R}= \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0.975 & 0.950 & 0.926 \end{bmatrix} \text{ and } \mathbf{G}= \begin{bmatrix} 0.342 & 0.333 & 0.325 \\ 0.111 & 0.450 & 0.439 \\ 0.150 & 0.257 & 0.593 \end{bmatrix}.$$

10. Suppose that the inter-arrival times of customers arriving at a single server station follow an Erlang distribution of order 2. The service times are also Erlang of order 2. If the arrival rate is 4 per minute and the mean service time is 0.2 minute. For this queueing model, show that R and G matrices are given by

$$\mathbf{R} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0.664 & 0.481 & 0.369 & 0.431 \\ 0 & 0 & 0 & 0 \\ 0.109 & 0.127 & 0.505 & 0.295 \end{bmatrix} \text{ and}$$

$$\mathbf{G} = \begin{bmatrix} 0.461 & 0.539 & 0 & 0 \\ 0.248 & 0.752 & 0 & 0 \\ 0.631 & 0.3569 & 0 & 0 \\ 0.170 & 0.830 & 0 & 0 \end{bmatrix}.$$

Also, explain the zeros appearing in the rows of **R** and in the columns of **G**.

11. Messages arrive from different sources and the pooled input is modeled using MAP with representation $(\mathbf{D}_0, \mathbf{D}_1)$ of order 3 is as given below.

$$\mathbf{D}_0 = \begin{bmatrix} -1.00222 & 1.00222 & 0 \\ 0 & -1.00222 & 0 \\ 0 & 0 & -225.75 \end{bmatrix},$$

$$\mathbf{D}_1 = \begin{bmatrix} 0 & 0 & 0 \\ 0.99220 & 0 & 0.01002 \\ 2.2575 & 0 & 223.4925 \end{bmatrix}.$$

The service times are of phase type with representation $(\boldsymbol{\beta}, \boldsymbol{S})$ of order 2 given by

$$\boldsymbol{\beta} = (0.7, 0.3), \quad \mathbf{S} = \begin{bmatrix} -3.648 & 1.824 \\ 1.824 & -2.432 \end{bmatrix}.$$

For this queueing model, $MAP/PH/1$, answer the following questions.

(a) Show that the traffic intensity is 5/6. (b) The correlation coefficient of any two successive inter-arrival times is 0.4889. (c) The standard deviation of the service times is 0.8788 unit. (c) The mean queue length is 249.5861 (d) Compare this queueing model to $M/M/1$ and write a short note on your findings.

12. Referring to Table 8.4 give interpretations to the entries of G matrix for the various scenarios.

Chapter 9

The General Queue $G/G/1$ and Approximations

The use of Markov models in queueing theory is very common because they are appropriate for basic systems and lend themselves for easy applications. But often the real-world systems are so complex and so general that simple Markov and renewal process models do not represent them well. The presentation of matrix-analytic models of Chapter 8 is an introductory attempt to go beyond the basic models discussed earlier. The computer and communication systems which have had a major role in advancing technology in the past three decades require queueing models that go well beyond those we have seen so far in the last eight chapters. Their full discussion is beyond the scope of this text. Here we provide an introduction to the analysis of the waiting time process in the general queue and a few approximation techniques that have proved useful in handling emerging complex applications. Readers who are not prepared for the complexities of the derivations may use this chapter for the fundamental concepts and the results it presents.

9.1 Bounds for Mean Waiting Time

Consider the general queue $G/G/1$ (also known as $GI/G/1$ in the literature) with the following description. Customers arrive at time points t_n $(n = 0, 1, 2, \ldots)$ and let the inter-arrival times $T_n = t_{n+1} - t_n$ be independent and identically distributed (i.i.d.) random variables with distribution function $A(\cdot)$. Let the service time of the nth customer be S_n and let $\{S_n,\ n = 1, 2, \ldots\}$ be i.i.d. random variables with distribution function $B(\cdot)$. We represent the means and variances of these random variables as follows:

$$E(T_n) = \frac{1}{\lambda} \qquad E(S_n) = \frac{1}{\mu}$$
$$V(T_n) = \sigma_A^2 \qquad V(S_n) = \sigma_B^2 \tag{9.1.1}$$

© Springer Science+Business Media New York 2015

U. N. Bhat, *An Introduction to Queueing Theory*, Statistics for Industry and Technology, DOI 10.1007/978-0-8176-8421-1_9

Note that for the means of inter-arrival times and service times, we have used the same notations of $\frac{1}{\lambda}$ and $\frac{1}{\mu}$ as in $M/M/1$ queues but with broader interpretation so as to make comparisons simple. We also define traffic intensity $\rho = \frac{\lambda}{\mu}$ as before.

Let W_n $(n = 1, 2, \ldots)$ be the waiting time of the nth customer and $W(t)$, the waiting time of the customer if it were to arrive at time t. Since there may or may not be a customer arrival at t, the process $W(t)$ is known as the *virtual waiting time process*. However, $\{W_n\}$ $(n = 1, 2, \ldots)$ are the actual waiting times of the arrivals at $(t_1, t_2, \ldots)$, and the process W_n is a subset of the $W(t)$ process. These are graphically illustrated in Figure 9.1.

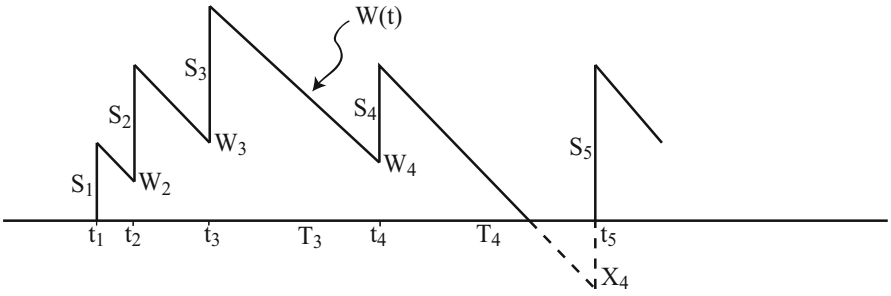


Figure 9.1 Waiting time processes

As shown in the figure, we may write the following relations:

$$\begin{aligned} W_1 &= 0 \\ W_2 &= W_1 + S_1 - T_1 \\ W_3 &= W_2 + S_2 - T_2 \\ W_4 &= W_3 + S_3 - T_3 \\ W_5 &= 0 = W_4 + S_4 - T_4 + X_4 \end{aligned} \tag{9.1.2}$$

In writing these relations, we have used the fact that in between arrivals the $W(t)$ process decreases at a unit rate because of the service provided to the customer. This will be clear if we interpret W_n as the service load in the system just before the arrival at t_n, and by providing service, the load $W_n + S_n$ gets depleted at a unit rate until the arrival at t_{n+1}, when its value is equal to $W_n + S_n - T_n$. When this amount becomes negative (by an amount X_n), to show $W_{n+1} = 0$, we write $W_n + S_n - T_n + X_n$. Hence, generalizing (9.1.2) we have

$$W_{n+1} = \begin{cases} W_n + S_n - T_n & \text{if } W_n + S_n - T_n > 0 \\ 0 & \text{if } W_n + S_n - T_n \leq 0 \end{cases} \tag{9.1.3}$$

or

$$W_{n+1} = W_n + S_n - T_n + X_n \tag{9.1.4}$$

where X_n can be defined as

$$X_n = \text{-}min(0, W_n + S_n - T_n). \tag{9.1.5}$$

We observe that X_n is the length of the idle time after the departure of the nth arrival. Note that X_n is nonzero only when W_{n+1} is zero and vice versa.

Using the random variable relation (9.1.3), we may write

$$\begin{aligned} P(W_{n+1} \le t) &= P(W_{n+1} = 0) + P(0 < W_{n+1} \le t) \\ &= P(W_n + S_n - T_n \le 0) + P(0 < W_n + S_n - T_n \le t) \\ &= P(W_n + S_n - T_n \le t). \end{aligned} \tag{9.1.6}$$

Define $F_n(t) = P(W_n \le t)$, $S_n - T_n = U_n$, and $U_n(t) = P(U_n \le t)$. With these notations (9.1.6) can be written as

$$F_{n+1}(t) = \int_{-\infty}^{t} F_n(t-x)dU_n(x) \quad 0 \le t < \infty. \tag{9.1.7}$$

For the existence of the steady state, we need the traffic intensity $\rho < 1$. This is the same as $E(U_n) = E(S_n) - E(T_n) < 0$. Under this condition, dropping the subscripts notationally in (9.1.7), we get

$$F(t) = \int_{-\infty}^{t} F(t-x)dU(x) \tag{9.1.8}$$

where

$$U(x) = \int_{x}^{\infty} B(y)dA(y-x). \tag{9.1.9}$$

Equation (9.1.8) was first established by Lindley (1952). It is one of the fundamental equations in queueing theory. Unfortunately, its solution requires the use of the Wiener–Hopf method which has been well illustrated in Kleinrock (1975). Also see Gross et al. (2008) for a summary of the solution technique and an illustration.

Instead of the distribution of W_n we now look at its mean. As $n \to \infty$, we may write $E(W_{n+1}) = E(W_n)$. Dropping subscripts, taking expectations of both sides of (9.1.4), we get

$$E(S) - E(T) = E(U) = -E(X). \tag{9.1.10}$$

Since X is the length of the idle period and the idle period, say I, ends with an arrival which finds the system empty, we may write

$$E(X) = E(I)P(\text{an arrival finds the system empty}). \tag{9.1.11}$$

Let us denote the probability on the right-hand side of (9.1.11) as a_0. Then we have

$$\begin{aligned} E(I) &= \frac{E(X)}{a_0} = \frac{-E(U)}{a_0} \\ &= \frac{1-\rho}{\lambda a_0}. \end{aligned} \tag{9.1.12}$$

Going back to (9.1.4) and rewriting, we have

$$W_{n+1} - X_n = W_n + U_n. \tag{9.1.13}$$

Squaring both sides and taking expectations

$$\begin{aligned} E(W_{n+1}^2) &+ E(X_n^2) - 2E(X_n W_{n+1}) \\ &= E(W_n^2) + E(U_n^2) + 2E(W_n U_n). \end{aligned}$$

Observe that $E[W_{n+1}^2] = E[W_n^2]$ as $n \to \infty$, W_n and U_n are independent of each other, and $X_n W_{n+1} = 0$. Thus as $n \to \infty$, we have

$$\begin{aligned} E(X^2) &= E[U^2] + 2E[W]E[U] \\ E[W] &= \frac{E[X^2] - E[U^2]}{2E(U)}. \end{aligned} \tag{9.1.14}$$

Defining $E(X^2)$ in a manner similar to (9.1.11) we may write $E(X^2) = a_0 E(I^2)$. From (9.1.11) we also get

$$\begin{aligned} E(U) &= \frac{1}{\mu} - \frac{1}{\lambda} \\ [E(U)]^2 &= \frac{1}{\lambda^2}(1-\rho)^2 \\ V(U) &= \sigma_A^2 + \sigma_B^2 \\ E(U^2) &= V(U) + [E(U)]^2 \\ &= \sigma_A^2 + \sigma_B^2 + \frac{1}{\lambda^2}(1-\rho)^2 \end{aligned} \tag{9.1.15}$$

Rewriting (9.1.14) as

$$E(W) = \frac{E(X^2)}{2E(U)} - \frac{E(U^2)}{2E(U)}$$

and using (9.1.12) and (9.1.15), we get

$$\begin{aligned} E(W) &= \frac{a_0 E(I^2)}{2[-a_0 E(I)]} - \frac{\sigma_A^2 + \sigma_B^2 + \frac{1}{\lambda^2}(1-\rho)^2}{2(\frac{1}{\mu} - \frac{1}{\lambda})} \\ &= \frac{\lambda^2(\sigma_A^2 + \sigma_B^2) + (1-\rho)^2}{2\lambda(1-\rho)} - \frac{E(I^2)}{2E(I)}. \end{aligned} \tag{9.1.16}$$

This result leads us to the important upper bound for $E(W)$ in the general queue $G/G/1$.

The expression (9.1.16) for $E(W)$, includes $E(I^2)$ and $E(I)$ which cannot be determined without a complete analysis of the system. Nevertheless, to obtain a lower bound for $E(I^2)/2E(I)$ (in order to get an upper bound for $E(W)$) we proceed as follows.

Setting $a_0 = 1$ in $E(I) = \frac{-E(S-T)}{a_0}$ of (9.1.12), we get

$$E(I) > \frac{1}{\lambda} - \frac{1}{\mu}. \tag{9.1.17}$$

Also

$$E(I^2) = V(I) + [E(I)]^2.$$

Since $V(I)$ is a positive quantity

$$E(I^2) \geq [E(I)]^2. \tag{9.1.18}$$

Using these two results in (9.1.16), we get

$$\begin{aligned} E(W) &\leq \frac{\lambda^2(\sigma_A^2 + \sigma_B^2)}{2\lambda(1-\rho)} + \frac{1}{2\lambda}(1-\rho) - \frac{[E(I)]^2}{2E(I)} \\ &= \frac{\lambda^2(\sigma_A^2 + \sigma_B^2)}{2\lambda(1-\rho)} + \frac{1-\rho}{2\lambda} - \frac{E(I)}{2} \end{aligned}$$

giving

$$E(W) \leq \frac{\lambda(\sigma_A^2 + \sigma_B^2)}{2(1-\rho)}. \tag{9.1.19}$$

These results are due to Kingman (1962a, b) and Marshall (1968). They have also provided lower bounds. Unfortunately, the lower bounds given by these authors are not easy to obtain. A simpler lower bound has been given by Marchal (1978) as

$$E(W) \geq \frac{\rho^2 + \lambda^2\sigma_B^2 - 2\rho}{2\lambda(1-\rho)}. \tag{9.1.20}$$

In the case of multiserver queues $G/G/s$ getting the bounds for $E(W)$ gets more complicated. The only result we mention here is by Kingman (1962a, b) which has the form

$$E(W) \leq \frac{\lambda(\sigma_A^2 + \sigma_B^2) + (s-1)\frac{\rho}{\mu}}{2s(1-\rho)}, \tag{9.1.21}$$

where $\rho = \frac{\lambda}{s\mu}$ is the traffic intensity. Also see Suzuki and Yoshida (1970).

The relationship (9.1.3) between W_n and W_{n+1} establishes the Markov property of the process $\{W_n, n = 0, 1, 2, \ldots\}$. It is a discrete time, continuous state Markov process and all techniques applicable to Markov processes can be used for its analysis. See Prabhu (1998) for results providing the time dependent as well as limiting distributions of the process.

9.2 Little's Law $L = \lambda W$

One of the most important and useful relationships in queueing theory is what is commonly known as *Little's Law*, named after J. D. C. Little (1961) who gave its first formal proof. It relates the long-term mean number of customers in the system to the mean amount of time customers spend in the system provided the number of customers entering the system is equivalent to the number of customers departing from it. Using common notations we write it as

$$L = \lambda W. \tag{9.2.1}$$

If we are looking at the number of customers waiting, we may write it as

$$L_q = \lambda W_q. \tag{9.2.2}$$

A formal proof of this result is beyond the scope of this text. Nevertheless, we may understand the plausibility of the results using intuitive arguments. Since we are considering steady state, i.e., when the traffic intensity is < 1, the number of customers waiting at the end of a service are those who arrive during the time the customer leaving after service has entered the system. That duration includes the waiting time and the service time. Using averages, notationally, the number in the system is L and the waiting time plus service time is W. The arrival rate is λ. Thus, ignoring all assumptions regarding the random variables and their distributions, we have the result $L = \lambda W$. A similar statement can be made to justify the result $Lq = \lambda W_q$. The reasoning in this argument is illustrated in the following example.

Example 9.2.1 Consider a queueing system in which customers arrive with rate λ. In Chapter 7, we have described how a queueing process can be considered a renewal process with busy cycle as the renewal period. The start of a busy cycle is a renewal epoch and renewal periods are probabilistic replicas of each other. Consequently, the properties that can be established in one such period should hold throughout the process.

Now suppose a busy cycle is of 10σ units of time with the description given in Table 9.1. The customers in the busy cycle are identified as $(C_1, C_2, C_3, C_4, C_5)$ and the number in the system is counted just before an arrival or departure. The 10σ point is the start of the next busy cycle.

Let $L(BC)$, $W(BC)$, and $\lambda(BC)$ be, respectively, the average number in the system, average time in system for a customer, and the average rate of arrival in the busy cycle considered here. We get

$$L(BC) = \frac{16}{10}.$$

The amounts of time the four customers have spent in the system are (in σ units of time)

$$C_1 : 4; \quad C_2 : 4; \quad C_3 : 5; \text{ and } C_4 : 3$$

Table 9.1 Arrival and departures in the busy cycle

Time(σ)	Arrival	Departure	Number in system
0	C_1		0
1	C_2		1
2			2
3	C_3		2
4		C_1	3
5		C_2	2
6	C_4		1
7			2
8		C_3	2
9		C_4	1
10	C_5		0

for a total of 16σ units of time. Thus, we get

$$W(BC) = \frac{16\sigma}{4} = 4\sigma \text{ units.}$$

The arrival rate is obtained as

$$\lambda(BC) = \frac{4}{10} \text{ per } \sigma \text{ unit}$$

verifying the relationship

$$L(BC) = \lambda(BC)\dot{W}(BC).$$

A similar relationship can be verified for $L_q(BC)$, $W_q(BC)$, and $\lambda(BC)$ as well. Thus, in general, we have the relationships

$$L = \lambda W;\ L_q = \lambda W_q$$

(Also see Jewell (1967)).

9.3 Approximations

Architectural models are exact; mathematical models can be exact; but probability models of random phenomena are always approximations. Since we use probability models for queueing systems, their usefulness can be gauged only by noting how closely the model approximates the real random phenomenon.

Three different stages may be identified in the modeling and analysis of a queueing system. At the first stage, a suitable mathematical model for the

system is developed. The second stage concerns the identification of and investigation into the basic process underlying the model. At the third stage, results useful for understanding the system are obtained in forms convenient for numerical and computational evaluations. Corresponding to these three stages we may identify three types of approximations: approximating the systems, the process, and the result.

Approximating the system is mainly a simplification of the system under study without undermining the basic structure, while making the analysis manageable. The four main elements of a queueing system are the arrival process, the service process, the queue discipline, and the system structure. These elements are described by their properties and attributes. Also, due to the complexity of some systems, such as networks of queues, we need to add a set of relations among these elements. Hence, a system approximation may be characterized either by simplifying the elements or relaxing the relational assumptions or both.

Often simplification of system elements is essential in order to be able to apply the results obtained from theory. It may not be possible to derive results for a model with the closest approximation to the element model (such as the distribution of the inter-arrival time or service time). Then the best useable model is employed to derive the best approximate result. The predominant use of the exponential distribution and the Markov model in practice is due to this approximating process. Other examples of approximating through simpler distribution models are the use of Erlangian and hyper exponential distributions, and the emergence of phase-type distributions and the matrix-analytic method. Structural simplifications include approximating dependent subsystems with independent subsystems, replacing weakly dependent subsystems in queueing networks with single nodes, dividing a nonstationary process into segments which are fairly stationary, and using bounding systems whose properties are easy to derive.

System approximations are generally heuristic in nature. Their quality depends very much on the practical insight of the analyst and a thorough understanding of the system behavior. Therefore, validation of the model is always necessary in probability modeling. It is also essential to confirm the applicability of the technique and the reliability of the results. An analyst has to evaluate constantly the trade-off between the ease of application of a particular technique and the accuracy of the ensuing result. Thus, the validation procedure must, in some way, involve a comparison between the approximate and the expected results. Generally, validation of approximation can be achieved through statistical tests, error analysis, experimentation, and simulation.

We have already seen one example of the process of approximating the result in Section 9.1. There, being unable to get a closed form expression for the mean waiting time of an arriving customer in a $G/G/1$ queue, we obtained an upper bound that can be used in its place in applications. There are other examples of approximating the result, either analytically or numerically in the queueing theory literature. We consider them beyond the scope of this text. For some early references, the readers may go to Bhat et al. (1979), which provides a comprehensive discussion of approximations in queueing theory.

In approximating the underlying process we use a process which is simpler for analysis while retaining as much of the original properties as possible. An example that has found wide application is the heavy traffic approximation in the general queue $G/G/1$. The relationship (9.1.3) between W_{n+1} and W_n can be stated as

$$\begin{aligned} W_{n+1} &= \max(0, W_n + S_n - T_n) \\ &= \max(0, W_n + U_n) \end{aligned} \tag{9.3.1}$$

where $U_n = S_n - T_n$, $n = 0, 1, 2, \ldots$.

For $n = 0, 1, 2, \ldots$, we have

$$\begin{aligned} W_1 &= \max(0, W_0 + U_0) \\ W_2 &= \max(0, W_1 + U_1) \\ W_3 &= \max(0, W_2 + U_2) \end{aligned}$$

Thus, we have $W_2 > 0$ only if $U_1 > 0$ (note that we have assumed that the first customer enters an empty queue; otherwise, it would be $W_1 + U_1 > 0$); $W_3 > 0$ only if $W_2 + U_2 > 0$ and so on. When the traffic is heavy we may assume that the arrival rate and the service rate are nearly equal to each other. Let $\frac{1}{\lambda} - \frac{1}{\mu} = \alpha$ and $\sigma_A^2 + \sigma_B^2 = \sigma^2$ giving $E(U) = E(U_n) = -\alpha$ and $V(U) = V(U_n) = \sigma^2$.

Under heavy traffic we may write (using $\cong$ to indicate approximate equivalence)

$$\begin{aligned} W_2 &\cong U_1 \\ W_3 &\cong U_1 + U_2 \\ &\vdots \\ W_{n+1} &\cong \sum_{r=1}^{n} U_r = U^{(r)}, \quad \text{say.} \end{aligned} \tag{9.3.2}$$

$U^{(n)}$, $n = 1, 2, \ldots$ are known as partial sums of $\{U_n\}$, and we have

$$\begin{aligned} E[U^{(n)}] &= -n\alpha \\ V(U^n) &= n\sigma^2. \end{aligned} \tag{9.3.3}$$

Since $\{U_n,\ n = 1, 2, \ldots\}$ are i.i.d. random variables, for n large, using central limit theorem we may write

$$\frac{U^{(n)} + n\alpha}{\sqrt{n}\sigma} \sim N(0, 1) \tag{9.3.4}$$

indicating that the left-hand side of (9.3.4) has a normal distribution with zero mean and unit variance.

When α/σ is small, Kingman (1962a, b, 1965) has shown that as $n \to \infty$ the waiting time W_n has approximately an exponential distribution with mean $\sigma^2/2\alpha$. (The details are beyond the scope of this text.) Now

$$\begin{aligned} \frac{\sigma^2}{2\alpha} &= \frac{1}{2}\left(\frac{V(U)}{-E(U)}\right) \\ &= \frac{1}{2}\left[\frac{V(T) + V(S)}{E(T) - E(S)}\right]. \end{aligned} \tag{9.3.5}$$

Referring back to (9.1.19), we may note that this is exactly the upper bound for $E(W)$ derived in Section 9.1. We should emphasize that (9.3.5) is a heavy traffic approximation for the mean of the limiting waiting time in the queue $G/G/1$ and it is useful only for larger values of traffic intensity.

In the case of the multiserver queue $G/G/s$, Kingman suggests two possible approximations for its mean waiting time. The first approximate result is obtained by extending the approximation to the mean waiting time under heavy traffic in the queue $G/M/s$:

$$E(W) \cong \frac{V(T) + V(S/s)}{2[E(T) - E(S/s)]}. \tag{9.3.6}$$

This result can also be obtained by considering the performance of $G/G/s$ queue in heavy traffic as being approximately the same as that of a $G/G/1$ queue whose service rate is s times the former (See Gross et al. (2008)).

The second approximation suggested by Kingman (1962a, b) is obtained by considering the performance of the $G/G/s$ queue in heavy traffic as being similar to that of a set of s parallel $G/G/1$ queues that are fed by an arrival process with mean inter-arrival time $sE(T) = \frac{s}{\lambda}$.

$$\begin{aligned} E(W) &\cong \frac{sV(T) + V(S)}{2[sE(T) - E(S)]} \\ &= \frac{V(T) + sV(\frac{S}{s})}{2[E(T) - \frac{1}{s}E(S)]} \end{aligned} \tag{9.3.7}$$

Clearly (9.3.7) is larger than (9.3.6) by $(s-1)V(\frac{S}{s})$ and therefore these results must be used with caution. Since these are not upper bounds but approximate values obtained by considering an underlying process for which results are available, both results may be considered as legitimate candidates for use.

9.4 Diffusion Approximation

Using a diffusion process to represent the underlying process in a queueing system is another example of the process approximation introduced in the last section. A *diffusion process* is a continuous state and parameter Markov process with the following properties:

(a) The process changes its state continually, but only small changes occur in small intervals of time.

(b) The mean and variance of the displacement during a small interval of time are finite.

These two properties can be formally stated using the transition distribution function $F(x,t;y,s) = P(X(s) \le y | X(t) = x)$. The property (a) can be stated as:

$$\lim_{\Delta t \to 0} \frac{1}{\Delta t} \int_{|y-x|>\delta} d_y F(x,t;y,t+\Delta t) = 0. \tag{9.4.1}$$

The following two equations mathematically describe property (b).

$$\lim_{\Delta t \to 0} \frac{1}{\Delta t} \int_{|y-x| \le \delta} (y-x) d_y F(x,t;y,t+\Delta t) = a(x,t) \tag{9.4.2}$$

$$\lim_{\Delta t \to 0} \frac{1}{\Delta t} \int_{|y-x| \le \delta} (y-x)^2 d_y F(x,t;y,t+\Delta t) = b(x,t) > 0. \tag{9.4.3}$$

Applying these properties in the derivation of forward Kolmogorov equation for the Markov process, we can get the diffusion equation, called the *Fokker–Planck equation*, as

$$\frac{\partial f(x,t)}{\partial t} = a(x,t) \frac{\partial f(x,t)}{\partial x} + \frac{b(x,t)}{2} \frac{\partial^2 f(x,t)}{\partial x^2}. \tag{9.4.4}$$

(See Prabhu (1965) or other books on stochastic processes.) The transition density function $f(x,t)$ is determined by solving the differential equation (9.4.4) with appropriate boundary conditions.

Suppose in a queueing process $X(t)$, the mean and variance are defined as follows:

$$\begin{aligned} \alpha(t)\Delta t &\cong E[X(t+\Delta t) - X(t) | X(t)] \\ \sigma^2(t)\Delta t &\cong V[X(t+\Delta t) - X(t) | X(t)]. \end{aligned} \tag{9.4.5}$$

These values are inserted in (9.4.4) to particularize the diffusion equation.

Gaver (1968) uses this approximation to determine the time-dependent distribution and the mean waiting time in the queue $M/G/1$. Let $W(t)$ be the waiting time process as shown in Figure 9.1. For a small interval of time Δt, the changes occurring in $W(t)$ are as follows:

$$\begin{aligned} W(t+\Delta t) - W(t) &= -\Delta t \qquad \text{with probability } 1 - \lambda \Delta t + o(\Delta t) \\ W(t+\Delta t) - W(t) &= S - \Delta t \qquad \text{with probability } \lambda \Delta t + o(\Delta t). \end{aligned} \tag{9.4.6}$$

In the statements of (9.4.6) we have used the following assumptions:

(a) Since the arrivals are Poisson with rate λ, the Δt interval includes an arrival point with probability $\lambda \Delta t + o(\Delta t)$ and does not include the arrival point with probability $1 - \lambda \Delta t + o(\Delta t)$. (We may recall here the definition of $o(\Delta t)$ given in Section 4.1: $\frac{o(\Delta t)}{\Delta t} \to 0$ as $\Delta t \to 0$ and $o(\Delta t)$ can be positive or negative.)

(b) The value of $W(t)$ decreases at a unit rate with time.

(c) When an arrival occurs, $W(t)$ increases by an amount equivalent to the service time of the customer. We have used a generic symbol S to denote the service time.

Using (9.4.6) the mean and variance of $W(t)$ can be obtained as

$$\begin{aligned} \alpha(t)\Delta t = \alpha\Delta t &\cong E[W(t+\Delta t) - W(t)|W(t)] \\ &= E[(S-\Delta t)(\lambda\Delta t + o(\Delta t) \\ &\quad +(-\Delta t)(1-\lambda\Delta t + o(\Delta t)] \\ &= \lambda E(S)\Delta t - \Delta t + o(\Delta t) \end{aligned}$$

giving, when $\Delta t \to 0$,

$$\alpha = \lambda E(S) - 1. \tag{9.4.7}$$

$$\begin{aligned} \sigma^2(t)\Delta t \cong\ & (\Delta t)^2[1-\lambda\Delta t + o(\Delta t)] + E[(S-\Delta t)^2][\lambda\Delta o(\Delta t))] \\ & - [\lambda E(S)\Delta t - \Delta t + o(\Delta t)]^2 \\ & = \lambda E(S^2)\Delta t - [\lambda E(S) - 1]^2(\Delta t)^2 + o(\Delta t) \end{aligned}$$

giving, when $\Delta t \to 0$,

$$\sigma^2 = \lambda E(S^2). \tag{9.4.8}$$

Substituting these values in the diffusion equation (9.4.4), we get

$$\frac{\partial f(x,t)}{\partial t} = -\alpha\frac{\partial f(x,t)}{\partial x} + \frac{\sigma^2}{2}\frac{\partial^2 f(x,t)}{\partial x^2} \tag{9.4.9}$$

with conditions

$$\begin{aligned} f(x,t|x_0) &\geq 0 \\ \int_0^\infty f(x,t|x_0)dx &= \frac{\lambda}{\mu} \\ \lim_{t\to 0} f(x,t|x_0) &= 0 \qquad x \neq x_0 \\ f(x,t|x_0) &= 0 \quad \text{for } x \leq 0 \text{ and } t \geq 0. \end{aligned} \tag{9.4.10}$$

Solving the diffusion equation, in addition to the Laplace transform of the waiting time distribution, Gaver finds an explicit expression for the mean waiting time, as

$$E[W(t)|W(0) = x_0] = \alpha t + x_0 + \frac{\sigma^2}{2\alpha}e^{-(\frac{2\alpha}{\sigma^2})x_0} \tag{9.4.11}$$

for large t.

The discontinuities and jumps in queueing processes make a diffusion approximation less than ideal for direct applications, as illustrated above. However, diffusion approximation has played a major role in obtaining weak convergence of functional limits in dealing with complex or unstable systems.

A special case of the general diffusion process defined by (9.4.4) is the *Brownian motion process* (also known as the *Wiener process*) $\{X(t), t \geq 0\}$ which has a normal distribution for specific values of $t > 0$ and has stationary independent increments, and for which $E[X(t)] = 0$ for $t > 0$, $V[X(t) - X(S)] = \sigma^2|t-s|$.

There has been a large amount of literature on functional limit theorems on various queueing processes when they are hard to analyze because of their complexity or lack of stability. The Brownian motion process plays a significant role in such limits. For instance, one of the earliest investigations is by Iglehart and Whitt (1970) who obtained weak convergence results of functionals of queue length, waiting time, and other related processes in a $G/G/$s queue (the first paper) and sequences of $G/G/$s queues (the second paper) when the traffic intensity is larger than 1. For a survey of investigations into such topics, including extensions to queueing networks readers are referrred to Glynn (1990).

9.5 Fluid Approximation

We introduce fluid approximation with an example from road traffic by ignoring the randomness in the arrival and service processes. Starting with an engineering approach, the approximation procedure developed by Newell (1971, 1992) has the advantage of being able to handle time-dependent queueing processes, especially when they are oversaturated (i.e., when the arrival rate exceeds the service rate). Lately, a combination of the fluid approximation along with the use of diffusion processes has proved useful in investigations into communication traffic.

Let $A(t)$ and $D(t)$ represent the number of arrivals and number of departures respectively in $(0, t)$. These are assumed to be continuous variables, not random; let us assume the arrival and service rates to be $\lambda(t)$ and $\mu(t)$, defined as

$$\frac{dA(t)}{dt} = \lambda(t); \qquad \frac{dD(t)}{dt} = \mu(t). \tag{9.5.1}$$

The rate $\lambda(t)$ is likely to be time dependent and the rate $\mu(t)$ is likely to be a constant or piecewise constant.

Consider $\lambda(t) = \lambda$ and $\mu(t) = \mu$, both constants. Define $Q(t) = A(t) - D(t)$. When $\lambda < \mu$ and $Q(0) >> 1$, $Q(t)$ will gradually decrease until it hits zero. On the other hand if $\lambda > \mu$, $Q(t)$ will grow progressively larger and larger and will go to infinity as $t \to \infty$.

Modeling the rush hour traffic when $\lambda(t)$ is likely to be time dependent, Newell assumes the form

$$\lambda(t) = \lambda(t_1) - \beta(t - t_1)^2 \tag{9.5.2}$$

where t_1 is the point at which it achieves the maximum. Also let t_0, t_2, and t_3 be such that the $\lambda(t_0) = \mu$; $\lambda(t_2) = \mu$ and $Q(t_3) = 0$. This means during this rush hour $Q(t) = 0$ at t_0 and t_3. With these assumptions,

Newell obtains

$$\text{Total delay } = \frac{9[\lambda(t_1) - \mu]^2}{4\beta}. \tag{9.5.3}$$

For details the readers are referred to Newell (1971).

In the monograph (1971, 1982) G. F. Newell extends the fluid approximation technique for use in applications such as transportation problems. The results are based on purely deterministic assumptions on system elements. Also, the results are very much dependent on the state of the system at specific time points. One way of addressing these problems is to use stochastic differential equations for representing the transitions in the underlying process. Then both the arrival and service processes can be made random and we can consider the limiting properties of the process as well. Examples of such models are provided in the review paper by Kulkarni (1997) for a buffer content process in communication traffic.

9.6 Remarks

An in-depth discussion of the topics covered in this chapter is beyond the scope of a book at an introductory level. In fact, a large amount of cutting edge ongoing research solving increasingly complex problems related to computer and communication traffic covers the area of approximations. Readers interested in gaining better knowledge of topics pertaining to general queues, approximations, limit theorems, etc., are recommended to look up more recent issues of research journals and books on such topics.

Chapter 10

Statistical Inference for Queueing Models

10.1 Introduction

Statistical analysis of data is essential to initiate probability modeling. Statistical inference completes the process by linking the model with the random phenomenon. Thus, for using the queueing models developed in earlier chapters, we need to estimate model parameters and make sure that we have the right model. In the next few sections, we discuss methods of parameter estimation appropriate for various data collection procedures.

We have not discussed any data collection and analysis procedures in this text simply because there are several books on them in the literature on statistics. Statistical inference procedures are also well established and on the face of it, a chapter on statistical inference of queueing systems may seem superfluous. However, in queueing systems standard data collection procedures may not be possible and those are the cases we plan to consider in this chapter.

When estimating parameters of a probability model, which define the input process or the service time distribution, there are two issues to be settled first: the sampling plan and the method of estimation. The sampling plan specifies the data collection procedure: how long to observe the system (for a specific length of time or until a specified number of events occur); what type of observations are to be made (the length inter-arrival times, the number of arrivals, length of service times, number of departures, etc.); and how these data elements are to be collected. The job of estimating a parameter can follow standard statistical procedures if we can collect all the necessary information from the system. For instance, if data are available on the arrival times of customers such that information on a specified number of inter-arrival times can be obtained, then the parameters of the distribution of the inter-arrival times can be estimated using standard statistical procedures. On the other hand, if the information

© Springer Science+Business Media New York 2015

U. N. Bhat, *An Introduction to Queueing Theory*, Statistics for Industry and Technology, DOI 10.1007/978-0-8176-8421-1_10

available provides only the number of arriving customers and the number of departing customers during a period of specified length, standard statistical procedures do not work. This sampling plan can be used only if an appropriate statistical procedure is available.

Random samples of observations are used in estimating parameters of distributions. In a similar fashion, for estimating parameters of a stochastic process we use sample paths, which are samples of realizations of the stochastic process.

A stochastic process is *ergodic* when its time average converges to its ensemble average as time $\to \infty$. A Markov process in which state space is irreducible, positive recurrent, and aperiodic belongs to the class of ergodic processes. When a stochastic process is ergodic, estimates obtained using one long sample path have been found to be equally accurate as estimates obtained from a large number of shorter sample paths.

There are two estimation procedures widely used in queueing applications: the method of moments and that of maximum likelihood (m.l.). The method of moments, as the name indicates, provides estimates by equating sample moments with the moments of the distribution. The number of equations to be used depends on the number of parameters to be estimated. Inspite of its simplicity the major drawback of this procedure is the lack of desirable properties of the estimators that make them reliable. Also, the estimators are not unique. For instance, one could use either the raw moments or the central moments. To guard against the unreliability of the estimates, it would, therefore, be necessary to obtain the properties of the estimators themselves (such as asymptotic normality, minimum variance, etc.).

To avoid the problems associated with the method of moments, the preferred procedure of estimation is the method of m.l. In this method, a likelihood function is constructed using observations from a random sample. When they are from a discrete distribution, the likelihood function is the probability of obtaining that particular sample and is constructed as the product of probability mass at the sample points. When the observations are from a continuous distribution, likewise, the likelihood function is the product of probability densities evaluated at the sample points. The parameter estimates are now those values that maximize the likelihood function. For details of the procedure the readers are referred to introductory textbooks on statistical theory. The properties that make m.l. estimation preferable are:

(1) Consistency (the variance of the estimator $\to 0$ as sample size $n \to \infty$)

(2) Asymptotic normality (the estimator has a normal distribution when the sample size is large)

(3) Invariance (the m.l. estimator of a function of the parameter is the corresponding function of the m.l. estimator)

However, the m.l. estimation is not perfect either. The estimate obtained by this procedure can be biased.

As indicated earlier, if random samples of inter-arrival times and service times are available, the parameters of their parent distributions can be estimated separately using the m.l. method. However, obtaining such random samples from the sample path of a queueing process presents problems. For instance, if the sample path is observed for a specific length of time, the sample sizes of both inter-arrival and service times are random and the stopping time is unlikely to be an arrival or a departure time. These factors need to be taken into account in the estimation procedure. For such reasons special sampling plans have been developed for inference on stochastic processes. These are discussed in the next two sections for queueing systems which allow birth and death process models and imbedded Markov chain models.

10.2 Birth and Death Process Models

The estimation of parameters using the m.l. method is similar for all queueing systems which can be modeled as birth and death processes. Therefore, for the sake of simplicity we use m.l. estimation in the simple queue $M/M/1$ as given by Clarke (1957) for illustration.

Let λ and μ be the arrival and service rates, respectively. Suppose the system is observed for a length of time T after it has achieved steady state. Let n_0 be the number of customers in the system at the start of observations. The four components of the sampling plan are: the initial number of customers in the system (n_0), the number of arrivals (n), the number of departures (m), and the length of time during $(0, T]$ the system has been busy (T_b). With these elements in the final result, the m.l. estimation procedure is developed as follows.

If we observe the sample path of the number of customers in the system we see the following features:

1. Changes of state occur due to arrivals or departures. Using results from Section 4.2, during a busy period, the amount of time the process resides in a specific state (sojourn time in a state) has an exponential distribution with mean $1/(\lambda + \mu)$.

2. When a change of state occurs during a busy period using property (d) leading to the result (A.1.2) of Appendix A of the exponential distribution, we conclude that

$$\begin{aligned} P(\text{an upward jump; i.e., an arrival}) &= \frac{\lambda}{\lambda + \mu} \\ P(\text{a downward jump; i.e., a departure}) &= \frac{\mu}{\lambda + \mu} \end{aligned}$$

   Thus, the jump event has a Bernoulli distribution with probabilities given above.

3. If at any time the system is empty the amount of time until the next arrival has an exponential distribution with mean $1/\lambda$. Then the probability that the process takes an upward jump = 1.

4. If the stopping time T for observations during a busy period is of length x_ℓ from the last change of state, then the probability element to be associated with that event is $e^{-(\lambda+\mu)x_\ell}$.

5. If the stopping time T is during the idle period, and if x_ℓ is the corresponding time from the last change of state, then the probability element to be associated with that event is $e^{-\lambda x_\ell}$.

6. Due to the Markovian nature of the process, the intervals of time representing the inter-event times as identified in items 1 and 3–5 above are independent of each other and also of the events identified in 2, and the nature of jumps are independent of all other events. Thus, the sample path is made up of independent realizations of various random variables, which can be used for the purposes of constructing a likelihood function for the m.l. estimation.

For the purpose of deriving the m.l. estimator, we define

$$\begin{aligned}
n_{ae} &= \text{number of arrivals to an empty system} \\
n_{ab} &= \text{number of arrivals to a busy system} \\
m &= \text{number of departures from the system} \\
x_i &= \text{intervals of time spent in state } i \text{ when the system is busy} \\
&\qquad i = 0, 1, 2, \ldots, (n_{ab} + m) \\
x_j &= \text{intervals of time the system has been empty} \\
&\qquad j = 0, 1, 2, \ldots, n_{ae} \\
x_\ell &= \text{the very last interval terminating in } T \\
n &= n_{ab} + n_{ae} \\
T_b &= \sum x_i + x_\ell \\
T - T_b &= \sum x_j + x_\ell
\end{aligned}$$

The likelihood function can now be constructed with the following components:

(a) The probability distribution of the initial queue size n_0

(b) The probability distribution of n_{ab} arrivals and m departures out of a total of $n_{ab} + m$ Bernoulli events

(c) Likelihood elements corresponding to x_i, $(i = 0, 1, 2, \ldots, (n_{ab} + m))$, x_j $j = 0, 1, 2, \ldots, n_{ae})$, and x_ℓ

(d) A combinatorial term reflecting the restrictions on the sequence of arrivals and departures, so that departures can occur only when there are customers in the system; since this term does not involve the parameters λ and μ, we denote it as a constant C

Then we have the likelihood function as

$$\begin{aligned} f(\lambda,\mu) &= C\left(1-\frac{\lambda}{\mu}\right)\left(\frac{\lambda}{\mu}\right)^{n_0}\binom{n_{ab}+m}{n_{ab}}\left(\frac{\lambda}{\lambda+\mu}\right)^{n_{ab}}\left(\frac{\mu}{\lambda+\mu}\right)^{m} \\ &\times \Pi_{i=1}^{n_{ab}+m}(\lambda+\mu)e^{-(\lambda+\mu)^{x_i}} \\ &\times \Pi_{j=1}^{n_{ae}}\lambda e^{-\lambda x_j}e^{-(\lambda+\mu)x_\ell}, \end{aligned} \tag{10.2.1}$$

if the last interval is part of a busy period. Otherwise, the last term $e^{-(\lambda+\mu)x_\ell}$ will be replaced by $e^{-\lambda x_\ell}$. Simplifying the terms in (10.2.1), we get

$$f(\lambda,\mu) = C'\left(1-\frac{\lambda}{\mu}\right)\left(\frac{\lambda}{\mu}\right)^{n_0}\lambda^n\mu^m e^{-\lambda T}e^{-\mu T_b} \tag{10.2.2}$$

where C' includes C and the combinatorial term of (10.2.1). If the initial number in the system is ignored we have the likelihood function as

$$f(\lambda,\mu) = C'\lambda^n\mu^m e^{-\lambda T}e^{-\mu T_b}. \tag{10.2.3}$$

Taking logarithms, differentiating with respect to λ and μ, equating to zero, and solving the resulting equations, we get (estimators of λ and μ are denoted as $\hat{\lambda}$ and $\hat{\mu}$, respectively)

$$\hat{\lambda}_{\text{crude}} = \frac{n}{T}; \quad \hat{\mu}_{\text{crude}} = \frac{m}{T_b}. \tag{10.2.4}$$

If the information provided by the initial queue length is included in the likelihood function we have to use (10.2.2) in the maximization process. (The more information we use in estimating a parameter, the better will be the accuracy of the estimate.) Taking logarithms, differentiating with respect to λ and μ, equating the resulting expressions to zero, and simplifying, we find that the estimated $\hat{\lambda}$ and $\hat{\mu}$ of λ and μ must satisfy the following equations:

$$\begin{aligned} \hat{\lambda} &= (\hat{\mu}-\hat{\lambda})(n+n_0-\hat{\lambda}T) \\ \hat{\lambda} &= (\hat{\lambda}-\hat{\mu})(m-n_0-\hat{\mu}T_b). \end{aligned} \tag{10.2.5}$$

Nonlinearity of these equations compels us to use indirect methods of solution. Writing $\hat{\lambda} = \hat{\mu}\hat{\rho}$ in the two equations of (10.2.5), we get

$$\begin{aligned} \hat{\rho} &= (1-\hat{\rho})(n+n_0-\hat{\mu}\hat{\rho}T) \\ \hat{\rho} &= (\hat{\rho}-1)(m-n_0-\hat{\mu}T_b). \end{aligned} \tag{10.2.6}$$

These equations give

$$\begin{aligned} \hat{\mu} &= \frac{n+m}{\hat{\rho}T+T_b} \\ \hat{\lambda} &= \frac{(n+m)\hat{\rho}}{\hat{\rho}T+T_b}. \end{aligned} \tag{10.2.7}$$

The problem is solved if we can get $\hat{\rho}$ from (10.2.6). Eliminating $\hat{\mu}$ from these two equations (rearranging and dividing one equation by the other), we get

$$\frac{\hat{\rho} - (n+n_0)(1-\hat{\rho})}{\hat{\rho} - (m-n_0)(\hat{\rho}-1)} = -\frac{\hat{\rho}T}{T_b} \tag{10.2.8}$$

which gives a quadratic equation in $\hat{\rho}$,

$$f(\hat{\rho}) = T(m-n_0-1)\hat{\rho}^2 - [(m-n_0)T + (n+n_0+1)T_b]\hat{\rho} + (n+n_0)T_b = 0. \tag{10.2.9}$$

This has exactly one admissible root $\hat{\rho}_1$ (say) since $f(0) = (n+n_0)T_b > 0$ and $f(1) = -T - T_b < 0$. Clearly $\hat{\rho}_1$ is therefore the required estimate. Now $\hat{\lambda}$ and $\hat{\mu}$ are obtained by substituting this value back in (10.2.7).

A simple approximation to $\hat{\rho}_1$, can be obtained by replacing $m - n_0 - 1$ by $m - n_0$ and $n + n_0 + 1$ by $n + n_0$ in (10.2.9). The corresponding quadratic equation,

$$f^*(\hat{\rho}) = T(m-n_0)\hat{\rho}^2 - [(m-n_0)T + (m+n_0)T_b]\hat{\rho} + (n+n_0)T_b = 0 \tag{10.2.10}$$

yields the two roots

$$\hat{\rho}_1^*,\ \hat{\rho}_2^* = \frac{(n+n_0)T_b}{(m-n_0)T},\ 1. \tag{10.2.11}$$

Substituting $\hat{\rho}_1^*$ from (10.2.11) in (10.2.7), we get

$$\hat{\lambda}_{\text{approx}} \cong \frac{n+n_0}{T}, \quad \hat{\mu}_{\text{approx}} \cong \frac{m-n_0}{T_b}. \tag{10.2.12}$$

By comparing $\hat{\rho}_1$ obtained from (10.2.9) with $\hat{\rho}_1^*$ as obtained above, we can also show that

$$\hat{\rho}_1 < \hat{\rho}_1^*$$

and

$$0 < \hat{\rho}_1^* - \hat{\rho}_1 < \frac{2\hat{\rho}_1^*}{(1-\hat{\rho}_1^*)(m-n_0)}. \tag{10.2.13}$$

We may therefore conclude that $\hat{\rho}_1^*$ is a good approximation of $\hat{\rho}_1$ when it is bounded away from 1 (i.e., $<<1$) and $m - n_0$ is large.

Example 10.2.1 Observations of a theater ticket counter for 30 minutes (T) yielded the following results:

- Number of customers at the start of observation $(n_0) = 2$
- Number of arrivals during $(0, T)$ $(n) = 75$
- Number of departures during $(0, T)$ $(m) = 70$
- Amount of time the system was busy $(T_b) = 25$ mins

Assume that at the time of observation the system was in steady state.

Without using the initial value, from (10.2.4), we get ($\hat{\lambda}$ and $\hat{\mu}$ are the estimates for the arrival and service rates, respectively)

$$\begin{aligned} \hat{\lambda}_{\text{crude}} &= \frac{75}{30} = 2.5, \\ \hat{\mu}_{\text{crude}} &= \frac{70}{25} = 2.8. \end{aligned}$$

Evaluating the admissible root in (0,1) of (10.2.9), we get

$$\hat{\rho}_{\text{exact}} = 0.827$$

from which substituting back in (10.2.7), we get

$$\hat{\lambda}_{\text{exact}} = 2.407, \quad \hat{\mu}_{\text{exact}} = 2.911.$$

When the initial value $n_0 = 2$ and the approximation are used in the estimation process, from (10.2.12) the approximate estimates are obtained as

$$\hat{\lambda}_{\text{approx}} = 2.567, \quad \mu_{\text{approx}} = 2.720.$$

Table 10.1 summarizes these results.

Table 10.1 Summary of results

	$\hat{\rho}$	$\hat{\lambda}$	$\hat{\mu}$
Exact	0.827	2.407	2.911
Approximate	0.944	2.567	2.720
Crude	0.893	2.500	2.800

Answer.

The m.l. method used in the foregoing discussion can be easily expanded for use in other birth and death process models of queueing systems. For instance, in the generalized model with arrival parameters λ_n $(n = 0, 1, 2, \ldots)$ and service parameter μ_n $(n = 1, 2, 3, \ldots)$, ignoring the information on the initial state ($\hat{\lambda}_n$ and $\hat{\mu}_n$ are the corresponding estimates), we get

$$\begin{aligned} \hat{\lambda}_n &= \frac{\text{No. of arrivals when the process is in state } n}{\text{Total time the process is in state } n}, \\ \hat{\mu}_n &= \frac{\text{No. of departures when the process is in state } n}{\text{Total time the process is in state } n}. \end{aligned} \tag{10.2.14}$$

(See Wolff (1965).)

10.3 Imbedded Markov Chain Models for $M/G/1$ and $G/M/1$

In the birth and death process models, because of the Markovian structure of the queue length process (number of customers in the system), we are able to construct a likelihood function using information on macroelements such as number of arrivals, number of departures, etc. The queue length process in $M/G/1$ and $G/M/1$ is Markovian only at certain time epochs (departure points in $M/G/1$ and arrival points in $G/M/1$). Consequently, we have to use the information provided by a realization of the resulting Markov chain. Such a realization is known as its sample path.

Let $\boldsymbol{\theta}$ represent the parameters of the inter-arrival and service time distributions in an $M/G/1$ queue. Recalling definitions and notations from Section 5.2, for the one step transition probability P_{ij} of the imbedded Markov chain, we have

$$P_{ij} = \begin{cases} k_{j-i+1} & \text{if } i > 0 \\ k_j & \text{if } i = 0 \end{cases}, \tag{10.3.1}$$

where

$$k_j = \int_0^\infty e^{-\lambda t} \frac{(\lambda t^j)}{j!} dB(t) \quad j = 0, 1, 2, \ldots. \tag{10.3.2}$$

Note that $\boldsymbol{\theta}$ includes arrival rate λ and the parameters of the distribution $B(\cdot)$.

Suppose the sampling plan is to observe the queueing system until N departures have occurred and note down the number of customers in the system at the start of observations, which we assume to be a departure point, and at the subsequent departure points, soon after departure. These are the values of Q_n in the sample path and let n_{ij} be the number of transitions of observed values of Q_n from state i to state j $(i, j = 0, 1, 2, \ldots)$. Let $(r_0, r_1, \ldots, r_N)$ be the observed values of Q_n $(n = 0, 1, 2, \ldots, N)$. Using the observed values of the sample path, the likelihood function may be written down as (ignoring the distribution of r_0)

$$f(\boldsymbol{\theta}) = \Pi_{n=1}^N P\left(Q_n = r_n | Q_{n-1} = r_{n-1}\right).$$

Taking logarithms

$$\ln f(\boldsymbol{\theta}) = \sum_{n=1}^N \ln P\left(Q_n = r_n | Q_{n-1} = r_{n-1}\right). \tag{10.3.3}$$

Expressing the transition probabilities in terms of k_j's defined in (10.3.2) and using the transition counts n_{ij}, (10.3.3) simplifies to

$$\begin{aligned} \ln f(\boldsymbol{\theta}) &= \sum_{j=0}^\infty (n_{0j} + n_{1j}) \ln k_j \\ &\quad + \sum_{i=2}^\infty \sum_{j=i-1}^\infty n_{ij} k_{j-i+1}. \end{aligned} \tag{10.3.4}$$

The m.l. estimates of $\boldsymbol{\theta}$ (λ and parameters of $B(\cdot)$) can be determined by specializing (10.3.4) in particular cases. In most cases, the maximization of $\ln f(\boldsymbol{\theta})$ will have to be carried out using numerical methods. For illustrations of this approach see Goyal and Harris (1972).

A similar approach can be used in the case of the $G/M/1$ queue. But the expressions are more complicated because the transition probability $P_{i0} = \sum_{r=i+1}^{\infty} b_r$, as shown in (5.3.3), involves a sum of integrals (See Bhat (2003)).

Harishchandra and Rao (1988) have suggested another way of using the queue length information from the sample path in the queue $M/G/1$. Equation (5.2.2) of Section 5.2 can be rearranged as

$$X_{n+1} = \begin{cases} Q_{n+1} - Q_n + 1 & \text{if } Q_n > 0 \\ Q_{n+1} & \text{if } Q_n = 0 \end{cases}, \tag{10.3.5}$$

where $\{X_n, n = 1, 2, \ldots\}$ are independent and identically distributed (i.i.d.) random variables representing the number of arrivals during service times with the distribution given by (10.3.2). Thus from successive observations of Q_n, $n = 0, 1, \ldots, N$, we get a corresponding sample of $\{X_n\}$ that are i.i.d's, suitable for use as a random sample. Now the product of corresponding density elements gives the likelihood function. However, as discussed in Bhat (2003), this likelihood function may not have enough information to estimate all parameters. For instance, in the queue $M/E_k/1$, only the traffic intensity ρ can be estimated by this method. To estimate the arrival and service rates separately, we need additional information, such as the amount of time the server has been busy, say τ, during the period N customers have been served. Then the service rate μ can be independently estimated as $\hat{\mu} = \frac{N}{\tau}$, from which the estimate of λ is obtained from the relationship $\frac{\lambda}{\mu} = \rho$.

A similar approach to the queue $G/M/1$ does not work because, equation (5.3.2), when rearranged, yields the relation

$$\begin{aligned} X_{n+1} &= Q_n + 1 - Q_{n+1} && \text{when } Q_{n+1} > 0 \\ &\geq Q_n + 1 && \text{when } Q_{n+1} = 0, \end{aligned} \tag{10.3.6}$$

which does not provide a complete sample on $\{X_n\}$ since when $Q_{n+1} = 0$, X_{n+1} is available only as larger than Q_n.

10.4 The Queue $G/G/1$

As mentioned earlier, a sampling plan that collects data for a specified length of time or until a specified number of events have occurred (these are known as *stopping rules*) presents problems because of the randomness of the sample size. Nevertheless, it is possible to get at least approximate estimates of parameters of the distributions using the m.l. method with most of the asymptotic properties of m.l. estimates intact (Basawa and Prabhu (1981)). The idea of using only

sequences of inter-arrival and service times in estimation is originally due to Cox (1965).

Let $a(\boldsymbol{u};\boldsymbol{\theta})$ and $b(\boldsymbol{v};\boldsymbol{\phi})$ be the inter-arrival time and service time densities respectively, with $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ representing the respective parameters. Let the corresponding distribution functions be denoted as $A(\cdot)$ and $B(\cdot)$, respectively. Let the system be observed until n departures have occurred, and let N_A be the number of arrivals during that period. Note that N_A is a random variable. We assume that the initial customer arrival is at $t = 0$. Let $\boldsymbol{u} = (u_1, u_2, \ldots, u_{NA})$ and $\boldsymbol{v} = (v_1, v_2, \ldots, v_n)$ be the sample data. Also, let x_n be the time difference between the stopping time (nth departure point) and the last arrival epoch. The likelihood function $f(\boldsymbol{\theta}, \boldsymbol{\phi})$ can be written as

$$f(\boldsymbol{\theta},\boldsymbol{\phi}) = \left[\Pi_{i=1}^{N_A} a(u_i;\boldsymbol{\theta})\right]\left[\Pi_{j=1}^{n} b(v_j;\boldsymbol{\phi})\right][1 - A(x_n;\boldsymbol{\theta})]. \tag{10.4.1}$$

Since the factor $[1 - A(x_n;\boldsymbol{\theta})]$ causes difficulty in obtaining simple estimates, consider the alternative approximate likelihood function, sometimes called *conditional likelihood function*, obtained by dropping the last term in (10.4.1)

$$f_c(\boldsymbol{\theta},\boldsymbol{\phi}) = \left[\Pi_{i=1}^{N_A} a(u_i;\boldsymbol{\theta})\right]\left[\Pi_{j=1}^{n} b(v_j;\boldsymbol{\phi})\right]. \tag{10.4.2}$$

The m.l. estimators are obtained from (10.4.2) by solving the following two equations

$$\begin{aligned} \sum_{i=1}^{N_A} \frac{\partial}{\partial\boldsymbol{\theta}} \ln a(u_i;\boldsymbol{\theta}) &= 0, \\ \sum_{j=1}^{n} \frac{\partial}{\partial\boldsymbol{\phi}} \ln b(v_j;\boldsymbol{\phi}) &= 0. \end{aligned} \tag{10.4.3}$$

For large samples, estimators of $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ can be obtained from (10.4.3), at least numerically, if not in closed form.

Basawa and Prabhu (1981) show that the estimators determined using the conditional likelihood function (10.4.2) have the requisite properties of the true m.l. estimators. Also, if $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\phi}}$ are estimators based on the full likelihood (10.4.1) and $\hat{\boldsymbol{\theta}}_C$ and $\hat{\boldsymbol{\phi}}_C$ are estimators based on (10.4.2), they have also shown that $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\theta}}_C$ have the same limiting distribution whenever

$$\frac{1}{\sqrt{n}} \frac{\partial}{\partial\boldsymbol{\theta}} \ln[1 - A(x_n;\boldsymbol{\theta})] \to 0 \quad \text{in probability.} \tag{10.4.4}$$

Referring back to the second term in (10.4.1) and (10.4.2), and noting that the corresponding equation to solve for the estimators is the same as (10.4.3) in both cases, we can conclude $\hat{\boldsymbol{\phi}} = \hat{\boldsymbol{\phi}}_c$.

In a subsequent paper, Basawa and Prabhu (1988) extend the results using four different stopping rules: (1) Observe until a fixed time T, (2) observe until

n arrivals have occurred, (3) observe until n departures have occurred, and (4) observe until n transitions have occurred. Conditions have been established for the approximate m.l. estimators to be asymptotically equivalent to the m.l. estimators one gets by using the likelihood functions corresponding to the four stopping rules in the sampling plan.

10.5 Other Methods of Estimation

In other methods of estimation, the method of moments plays a major role. One such method using data on inter-departure intervals to estimate parameters of the service time distribution in the queue $M/G/1$ is given by Cox (1965). Let λ and μ be the arrival and service rates in such a system, with $B(\cdot)$ as the distribution function of the service time. Let $C(\cdot)$ be the distribution function of the inter-departure interval. Note that in steady state $1-\frac{\lambda}{\mu}$ is the probability that the system is empty and $\frac{\lambda}{\mu}$ $(=\rho)$ is the probability it is busy. Also, when it is busy the inter-departure interval is the service time itself.

With this information, it is not difficult to write

$$C(t) = \frac{\lambda}{\mu}B(t) + (1-\frac{\lambda}{\mu})\int_0^t B(t-x)\lambda e^{-\lambda x}dx. \tag{10.5.1}$$

When the service time is exponential, we do not get any additional information on μ from (10.5.1) since, the departure process in $M/M/1$ has the same distribution as the inter-arrival time as $t \to \infty$ (see Section 4.2.1). On the other hand, if the service time is a constant $= \frac{1}{\mu}$, we have

$$\begin{aligned} C(t) &= 0 && t < \frac{1}{\mu} \\ &= \frac{\lambda}{\mu} && t = \frac{1}{\mu} \\ &= \frac{\lambda}{\mu} + \left(1-\frac{\lambda}{\mu}\right)\left(1-e^{-\lambda(t-1/\mu)}\right) && t > \frac{1}{\mu}, \end{aligned} \tag{10.5.2}$$

Due to the nonzero probability associated with $t = 1/\mu$, we may use the minimum observed inter-departure time as the estimate of $1/\mu$. (Even without the help of (10.5.2), this is the best conclusion because the length of service time is the minimum of inter-departure times).

When the service time distribution is different from the exponential or the deterministic, the parameters of the distribution can be estimated by equating the appropriate cumulants of the inter-departure time distribution with those of the cumulants from observed data. (See Cox (1965) for details and a discussion on the problems arising out of dependent observations.)

When data are available on the time in the system for customers, a similar approach can be used by noting down their arrival and departure epochs. The time in system (waiting + service time) has the Laplace–Stieljes transform given

by (5.2.36). Its moments can be determined by differentiation and setting $\theta = 0$ in the resulting expressions. Now the parameters of the service time distribution $B(\cdot)$ are determined by solving equations resulting from equating these moments with those from the data (see Cox (1965)).

While estimating parameters in $M/G/1$ queueing systems, we need to assume a parametric form for the service time distribution to specify parameters. What if we are not certain about the prametric form itself? In Chapter 2 and Appendix A, we saw that the Erlangian family of distributions for different values of k have coefficient of variation (CV = standard deviation/mean) less than 1 and the distributions belonging to the hyperexponential family have CV greater than 1. If one looks at these two families of distributions as belonging to a large family with CV varying in the range $[0, \infty)$, we can say that the exponential distribution with CV equal to 1, divides them in two groups. Also because of their relationship with the exponential distribution they are easy for analysis as models for inter-arrival or service time distributions. Furthermore, the Erlangian family with different values of k and the hyperexponential family with different mixing parameters, together cover a wide variety of distribution forms that can be used in modeling in most of the applications. Thus, estimating the value of CV from the data can lead into the selection of the right distribution model for service time in an $M/G/1$ queue.

To estimate the coefficient of variation of the service time in an $M/G/1$ queue, we start with equation (10.3.5) where $\{X_n, n = 1, 2, \ldots\}$ are i.i.d. random variables representing the number of arrivals occurring during service periods. The random variable X_n has the distribution $\{k_j\}$ given by (10.3.2). Let μ_1 and μ_2 be the first and second moments of this distribution. The PGF of k_j can be obtained as (see derivations leading to (5.2.9))

$$K(z) = \psi(\lambda - \lambda z), \tag{10.5.3}$$

where $\psi(\theta)$ is the Laplace–Stieltjes transform of the service time distribution $B(\cdot)$. Clearly we have

$$\begin{aligned} K'(1) &= \mu_1 = -\lambda\psi'(0), \\ K''(1) &= \mu_2 - \mu_1 = \lambda^2\psi''(0). \end{aligned} \tag{10.5.4}$$

Let μ_1^s and μ_2^s be the first two moments of the service time distribution, with σ^2 as its variance. The CV of the service time distribution is now given by $C = \sigma/\mu_1^s$.

From (10.5.3) we get $\mu_1^s = -\psi'(0)$ and $\mu_2^s = \psi''(0)$. Thus

$$K''(1) = \lambda^2 \left[\sigma^2 + (\mu_1^s)^2\right] = \lambda^2[\sigma^2 + (K'(1))^2] \tag{10.5.5}$$

which leads to

$$\sigma^2 = \lambda^{-2} \left[K''(1) - (K'(1))^2\right].$$

But $\mu_1^s = \lambda^{-1}K'(1)$. Hence, we get

$$C^2 = \frac{K''(1) - [K'(1)]^2}{[K'(1)]^2}. \tag{10.5.6}$$

Substituting from (10.5.4), we have

$$C^2 = \frac{\mu_2}{\mu_1^2} - \frac{1}{\mu_1} - 1. \tag{10.5.7}$$

Let m_1 and m_2 be the first two sample moments of X_n as observed from the system. For the estimator of C, $\hat{C}$, we get

$$\hat{C} = \sqrt{\frac{m_2}{m_1^2} - \frac{1}{m_1} - 1}. \tag{10.5.8}$$

Asymptotic properties of this estimator (consistency and normality) have been established by Miller and Bhat (2002). A simulation study used to determine the working rules for distribution selection provide the following guidelines: When $\hat{C} << 1$ use Erlang; $\hat{C} >> 1$ use hyperexponential; and when $\hat{C} \cong 1$ use exponential. The last conclusion is based on the fact that when k is close to 1, using the exponential distribuion in the model is likely to be more cost effective in further analysis than either Erlang (if $\hat{C}$ is slightly less than 1) or hyperexponential (if $\hat{C}$ is slightly greater than 1) distribution.

If the decision is to adopt an Erlangian distribution, its parameters, μ and k can be determined using the m.l. method. In the case of the scale parameter k, however, the integer m.l. method should be used. For details see Miller and Bhat (1997) and Miller (1999). If the hyperexponential distribution H_2 is chosen, the m.l. method becomes unwieldy. For such circumstances, Miller (1996) has developed an estimation procedure for the mixing parameter p, using moments of the distribution.

In queueing theory very often estimates of performance measures are the major objectives; e.g., system utilization and probability of blocking in a communication system. Since the theory has grown along with applications, over the years researchers in industrial laboratories have developed various methods of estimating such measures. Also, there are other investigations that provide additional methods of estimation of parameters. For a comprehensive survey of these procedures and results, the readers are referred to Bhat et al. (1997).

10.6 Tests of Hypotheses

Hypothesis testing is an integral part of inference in statistical theory. It involves analytical procedures to determine whether hypotheses made regarding the characteristics of the random phenomena are true. In queueing theory, since the objective is to set up a suitable probability model as an aid to decision-making, the use of hypothesis testing is limited. Therefore, we restrict ourselves to providing only references and the type of problems considered in them.

Most of the circumstances where hypothesis testing can be used in queueing theory are when there is some prior information on parameter values of the process or when the goodness of fit of a distribution form for the inter-arrival

time or the service time has to be ascertained. In all these cases, if we can get enough information from the sample path of the process, standard techniques from statistical theory can be used. But there are circumstances where complete information is not available. For instance, Clarke's (1957) estimation of parameters in Section 10.2 for the queue $M/M/1$ used only the number of arrivals and departures, amount of time the system was busy, and the total time. Using a similar sampling plan Wolff (1965) develops likelihood ratio tests for parameter values. Thiagarajan and Harris (1979) have developed a procedure to test whether the service time distribution is exponential in an $M/G/1$ queue based only on information on waiting times. Using information derived from (10.3.5) for the number of customers arriving during a service period in an $M/E_k/1$ queue, Harischandra and Rao (1988) have developed a likelihood ratio test for the traffic intensity ρ.

Another form of test that can be used in queueing theory is the sequential probability ratio test which is described in the next section.

10.7 Control of Traffic Intensity in $M/G/1$ and $G/M/1$

Confidence intervals are useful in determining whether a parameter can be assumed to lie within some specified limits. As pointed out by Cox (1965), using the notation from Section 10.2, confidence intervals for λ, μ, and ρ in an $M/M/1$ queue can be obtained by observing that $2\hat{\lambda}T$ can be treated as a chi-square variate with $2n$ degrees of freedom and $2\hat{\mu}T_b$ as a chi-square variate with $2m$ degrees of freedom. It is well known that the ratio of two chi-square variates has an F distribution. The confidence intervals now follow using the known values of this distribution (Also see Lilliefors (1966)).

When operating a queueing system, monitoring and controlling the parameter values are essential to ensure that the system performance is consistent with design standards, and in order to respond to exigencies of the environment. The parameter control problem, in effect, involves the problem of testing the hypotheses $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$, where $\boldsymbol{\theta}$ is the vector of parameters, with $\boldsymbol{\theta}^0$ as the set of desired values, against a suitable alternative say, $H_1 : \boldsymbol{\theta} = \boldsymbol{\theta}_1$. If the hypothesis is not rejected at a chosen level of significance, we conclude that the system parameters have not changed, while the rejection of the hypothesis is indicative of change in parameter values. Once the change is detected, an appropraite control action can be taken.

When the difference between parameter values under the null and alternative hypothesis is large, a sequential test has the advantage of using a considerably smaller sample size. With this objective, Rao et al. (1984) have developed a procedure for testing the hypothesis $H_0 : \rho = \rho_0$ versus $H_1 : \rho = \rho_1$ using Wald's sequential probability ratio test (SPRT) for the systems $M/G/1$ and $G/M/$s in which the queue length process $\{Q_n,\ n = 0, 1, 2, \ldots\}$, respresenting the number of customers in the system at departure epochs (in $M/G/1$) or

arrival epochs (in $G/M/$s), has an imbedded Markov chain. Let the transition probabilities of the chain be $P_{ij}(\rho)$ when ρ is the traffic intensity, and let n_{ij} be the number of transitions $i \to j$ of $\{Q_n\}$ up to and including the nth transition. Then, the likelihood ratio for the SPRT is ($n = \sum \sum n_{ij}$)

$$L_n = \Pi_{i,j} P_{ij}^{n_{ij}}(\rho_1) / \Pi_{i,j} P_{ij}^{n_{ij}}(\rho_0). \tag{10.7.1}$$

Let $A = (1-\beta)/\alpha$ and $B = \beta(1-\alpha)$, where α and β are the probabilities of Type I and Type II errors, respectively. The SPRT procedure is: Accept H_1 if $L_n \geq A$, accept H_0 if $L_n \leq B$, and observe the next queue length Q_{n+1} and compute L_{n+1} and repeat the procedure if $B < L_n < A$. The mechanics of applying the test are easier if logarithms are used. In the case of the systems $M/M/1$, $M/E_k/1$, $E_k/M/$s, $M/M/$s/s, and the machine interference problem, the logarithm of (10.7.1) takes the form $\ln L_n = an + \sum_{i,j} n_{ij} c_{ij}$, where a and c_{ij} are constants depending upon ρ_0, ρ_1, and the transition probabilities of the imbedded Markov chain.

For details of the procedure see Rao et al. (1984). The paper also provides the operating characteristic function and the average sample number for the SPRT. Even though the procedure uses a finite Markov chain, its validity for denumerable infinite chains has been established in Rao and Bhat (1991).

An alternative procedure in parameter control in $M/G/1$ and $G/M/1$ queues is to use the limiting distribution of the number of customers in the system as outlined in Bhat (1987). Let $t_0, t_1, \ldots$ be the departure epochs in an $M/G/1$ (or arrival epochs in a $G/M/1$) queue, and let Q_n be the number of customers at these points (appropriately defined). The control technique has two phases. The first phase (a warning phase) indicates the time at which the sample function gets out of the region covered by the upper and lower control limits c_u and c_ℓ; the second phase (the testing phase) is intended to see whether the process returns to the control region within a specific amount of time, and involves two limits say, d_u and d_ℓ.

The procedure here is similar to the control chart technique of industrial quality control, but with the addition of a second set of limits. The second phase has been introduced in order to avoid errors in decision-making, which may result because of fluctuations in the sample path of the process.

The first set of limits is determined using the limiting distribution of $\{Q_n, n = 0, 1, 2, \ldots\}$. Let $Q^* = \lim_{n\to\infty} Q_n$ and let α_u and α_ℓ be two specified probabilities. Then c_u and c_ℓ are integers such that

$$\begin{aligned} c_u &= \min\{k | P(Q^* \geq k) \leq \alpha_u\}, \\ c_\ell &= \max\{k | P(Q^* \leq k) \leq \alpha_\ell\}. \end{aligned} \tag{10.7.2}$$

A simple procedure suggested in Bhat (1987) for the determination of the second set of limits d_u and d_ℓ, makes use of those service periods in which no customer arrivals occur in the queue $M/G/1$ and inter-arrival periods in which no service completion occurs in the queue $G/M/1$. Clearly these are Bernoulli events with

probability of success k_0 in the queue $M/G/1$ and b_0 in the queue $G/M/1$. The second phase limits d_u and d_ℓ are then defined with associated probabilities β_u and β_ℓ as follows.

In the queue $M/G/1$ when $\{Q_n\}$ hits or goes beyond the upper limit c_u, we do not conclude that the traffic intensity $> \rho_0$ unless the process stays at or beyond c_u for a minimum number of d_u transitions. Hence, given a probability β_u, d_u is the smallest number n such that the probability of the number of arrivals being at least 1 in n consecutive transitions $\leq \beta_u$. This can be stated as

$$d_u = \min\{n|(1-k_0)^n \leq \beta_u\}. \tag{10.7.3}$$

When $\{Q_n\}$ reaches c_ℓ, we do not conclude that the traffic intensity $< \rho_0$ unless it stays at or below c_ℓ, for a minimum number of d_ℓ transitions. Hence, given a probability β_ℓ, d_ℓ is the smallest number n such that the probability that the number of arrivals is zero for n consecutive transitions $\leq \beta_\ell$. This can be stated as

$$d_\ell = \min\{n|k_0^n \leq \beta_\ell\}. \tag{10.7.4}$$

In the case of the $G/M/1$ queue, similar expressions can be obtained by noting that b_0 is the probability of no service completion during an inter-arrival period. This will be accomplished by replacing $1-k_0$ with b_0 in (10.7.3) and (10.7.4).

Since $1-k_0$ is the probability of one or more arrivals in $M/G/1$, the second phase limits derived as described above are very conservative and provide enough protection from the wrong conclusion that the traffic intensity has changed.

Thus, once the limits (c_u, c_ℓ; d_u, d_ℓ) are determined as given in (10.7.2)–(10.7.4), the procedure to monitor and control traffic intensity in $M/G/1$ and $G/M/1$ can be described as follows:

1. Starting with an initial queue length i and traffic intensity ρ_0, leave the system alone as long as Q_n lies between c_u and c_ℓ, or when it goes out of these limits if it returns within bounds before d_u and d_ℓ transitions, respectively.

2. If the queue length does not return within bounds between d_u or d_ℓ consecutive transitions, as the case may be, conclude that the traffic intensity has changed from ρ_0 and reset the system to bring the traffic intensity back to the level ρ_0.

3. Repeat 1 and 2 using the last state of the system as the initial state.

10.8 Remarks

Statistical inference for queueing models is often ignored in textbooks on queueing theory. One exception in a limited form is Gross et al. (2008) starting from its first edition in 1974. Generally, it seems, queueing models are applied without going beyond the method of moments for estimation of model parameters.

However, the author of this text believes that an adequate use of statistical inference is necessary for a rigorous application of any probability model. For this reason, we have incorporated several inference topics beyond what is given in Gross et al. For a comprehensive discussion of these and other topics in inference on queueing systems, the readers may consult Bhat et al. (1997), which also includes an extensive bibliography.

Chapter 11

Decision Problems in Queueing Theory

11.1 Introduction

In Chapter 1 we identified three types of problems occurring in queueing theory. These related to the behavioral, statistical, and operational decision-making aspects of queueing systems. In Chapters 4–9 we described probability models used in understanding system behavior and in Chapter 10 we discussed how statistical techniques can be employed to choose the right models. In this chapter, we address some of the simpler decision problems that arise in the operation of queueing systems.

If we recall the origins of queueing theory recounted in Chapter 1, A. K. Erlang used the Poisson model for call arrivals with the objective of improving the operation of the system. His 1924 paper, "On the Rational Determination of the Number of Circuits" (see Brockmeyer et al. (1960)) specifically addressed a decision problem.

The use of behavioral results derived from probability models in decision-making has played a major role in queueing theory. Since the 1950s, with the development of optimization techniques for decision-making, operations researchers have introduced design and control procedures into the field. However, the amount of work on these topics makes up only a small fraction of the volume of research on the subject.

In his introduction to a special issue of the journal *Queueing Systems* on design and control, Stidham (1995) provides two reasons for the paucity of research on these topics in queueing theory: the well-developed nature of models and the availability of explicit performance measures in them. We may add a third reason as well: the complexity of models required in representing the advanced systems in areas such as computers and communications.

© Springer Science+Business Media New York 2015

U. N. Bhat, *An Introduction to Queueing Theory*, Statistics for Industry and Technology, DOI 10.1007/978-0-8176-8421-1_11

In the next three sections, we introduce the three modes of decision-making: (1) using performance measures, (2) design problems, and (3) control problems. We use the categorization of decision problems as design and control, as provided by Crabill et al. (1977). According to them, the use of static optimization to determine the best system for optimizing some long-run average criterion, such as cost or profit, characterizes a design problem. In a control problem, the optimization is dynamic and the system operating characteristics are allowed to change over time. In all three cases our discussion will be minimal because using performance measures in decision-making is a natural process in probability modeling, and real-world applications of the queueing theory do not make extensive use of design and control procedures.

11.2 Performance Measures for Decision-Making

The first half of the twentieth century was the formative period for the queueing theory. Model development occurred for improving the operations of queueing systems starting with the work of A. K. Erlang. Since the early applications were in telephone industry, graphs and charts were developed for using information on performance measures such as, probability of blocking and mean waiting times, in decision-making. Examples of such charts can be found in Cooper (1981), Hillier and Lieberman (1986), or in issues of *Bell Systems Technical Journal* of earlier times. With the advent of computers, such preprepared charts and graphs have become unnecessary.

As indicated earlier, with the availability of performance measures from models developed specifically for the systems in question, most of the decisions are based on such measures. System performance is measured against specified objectives and changes are made in the parameters of the system elements in order to achieve them. See Edie (1954) for an example of this procedure in the context of traffic delays in toll booths.

An additional aid to decision-making developed in the past 30 years or so is the use of computer simulations. They can be used to validate models as well as to determine the best characteristics of the system in specific scenarios. Since there is enough published material on this subject we do not go into it in detail in this text. An introduction to the simulation of queueing systems and some examples are provided in Chapter 14. Also see books on the subject by Law and Kelton (1991) or Schriber (1991).

11.3 Design Problems in Decision-Making

In a design problem, cost functions are used to establish optimum values of the parameters or optimum structural configurations to achieve a desired performance in the system. The cost functions could be based on monetary costs or performance measures. These problems are also known as *economic models*.

The optimization is static (i.e., varying values of the parameters are not considered), and it is achieved using established procedures. Unfortunately, when queueing system models become complex, the expressions for performance measures may not be tractable for optimization procedures. In such cases, trial and error or numerical procedures may be needed.

Three investigations published in the 1950s and 1960s illustrate the economic model approach. Brigham (1955) determined the optimum number of clerks to be placed behind tool crib counters in an aircraft factory. After determining that the arrivals follow a Poisson process and the service times are exponentially distributed, Brigham uses Erlang's formula for the probability of blocking to get an expression for the waiting time of arriving customers. The cost function includes the cost per clerk and the cost per customer per unit time. The best value for the number s of clerks is obtained with the help of graphs of the ratio of the two costs for each value of s. To complete the determination of cost savings, Brigham uses what he calls the "obverse" queue, in which the cost of idleness of the clerks is obtained.

Morse (1958) tackles the problem of determining the optimum value of the number admitted to an $M/M/1/N$ queueing system by balancing the service cost with the cost of losing customers. He uses a cost of $E\mu$ dollars per unit time to provide service, when the mean cost per unit serviced is E, rate of service is μ, and a gross profit of G dollars per single service operation. With λ as the Poisson arrival rate, the net profit per unit time is obtained as

$$P = \frac{\lambda G(1-\rho^N)}{1-\rho^{N+1}} - E\mu. \tag{11.3.1}$$

Differentiating this expression with respect to μ and setting the result equal to zero, Morse obtains the following equation for the maximum value of μ:

$$\rho^{N+1}\left[\frac{N-(N+1)\rho+\rho^{N+1}}{(1-\rho^{N+1})^2}\right] = \frac{E}{G} \tag{11.3.2}$$

Plotting this equation for E/G against ρ we get graphs that can be used to determine the number N of customers to be admitted to the system for varying cost structure and service rates.

In the infinite waiting room case $M/M/1$, Morse is able to obtain the optimum service rate μ with the standard approach to optimization. He uses the cost function

$$D\mu + CW = D\mu + \frac{C}{\mu-\lambda}, \tag{11.3.3}$$

where C is the cost of wait per unit time, D is the cost of service per unit time, and W is the mean waiting time. Optimizing this cost function by differentiating with respect to μ and equating to zero, he gets

$$\mu = \lambda + \sqrt{C/D}. \tag{11.3.4}$$

In the multiple server case, however, for the determination of the optimum number of servers, the optimization is carried out with a trial-and-error method. For details the readers are referred to Morse (1958).

Hillier's (1963) study of economic models for waiting lines is much more general than the previous two models. He considers three multiserver models, with models 2 and 3 having two variants each. The arrivals in all models are Poisson and the queue discipline is FCFS. All models assume that the cost of waiting is proportional to the time in system, and the cost of service is a linear function of the number of servers. Let λ and μ be the arrival rate and the service rate per server, respectively, and let s be the number of servers. Three basic models determine optimum values for λ, μ, and s as noted below with various cost structures.

Model 1: Find s; Model 2: Find λ and s; Model 3: Find μ and s. Under model 2, travel time is also considered. Because of the multiserver structure, when the service time is other than exponential, individual queues in front of servers become necessary.

The usual method of solution is trial and error, except in cases where the service times are exponential, when explicit expressions that are mathematically tractable for optimization are available. For details the readers are referred to Hillier (1963). These problems have been discussed in a more general framework in Hillier and Lieberman (1986).

The following example illustrates the use of cost considerations in a static decision model.

Example 11.3.1 Customer arrivals at a department store can be assumed to be Poisson at the rate of λ per unit time. After picking up their merchandise, the customers queue up in front of checkout counters. The time spent in doing so can be assumed to have an exponential distribution. The checkout time for each customer has a distribution with mean b_1 and second moment b_2. Suppose we have to determine the optimum number of checkout counters under the following cost structure: (i) C_1 per unit time due to a waiting customer and (ii) C_2 per unit time for maintaining service at a counter.

Because of the exponential distribution of the time spent in picking up merchandise, the arrival process at the checkout counters can be assumed to be Poisson as well. (See Section 4.2.1). When there are s counters, assuming that the customers choose the counters at random, the arrival rate at each counter can now be assumed to be Poisson with rate λ/s. Using the expression for the waiting time in queue for a customer in an $M/G/1$ system from (5.2.43) we have

$$\begin{aligned} W_q &= \left(\frac{\lambda}{s}\right) b_2/2 \left(1 - \frac{\lambda b_1}{s}\right) \\ &= \frac{\lambda b_2}{2(s - \lambda b_1).} \end{aligned} \tag{11.3.5}$$

Let C be the total cost per unit time. We have

$$E(C) = \frac{\lambda b_2 C_1}{2s - 2\lambda b_1} + sC_2. \tag{11.3.6}$$

Minimizing this cost function with respect to s in the usual manner, we find that

$$s = \lambda b_1 + \sqrt{\frac{\lambda b_2}{2}\left(\frac{C_1}{C_2}\right)} \tag{11.3.7}$$

minimizes $E(C)$ as given by (11.3.6). Since the optimal value s must be an integer, it is determined by evaluating $E(C)$ at $[s]$, the integer part of s, and at $[s]+1$, and choosing the one that gives the smallest $E[C]$.

Answer

For a numerical example, use $\lambda = 2$ per minute, the checkout time as exponential with mean $(b_1) = 3$ minutes. Then $b_2 = 9$.

Further let $C_1 = 0.5$, $C_2 = 2.5$. Substituting in (11.3.7) we get

$$s = 7.34$$

with $E(C)|_{s=7} = 22$ and $E(C)|_{s=8} = 22.25$.

Hence, the optimum value of $s = 7$.

Answer

11.4 Control Problems in Decision-Making

Under control problems we include decision problems that require optimization in a dynamic setting. One of the earliest investigations is by Moder and Phillips (1962) in which the authors consider a multiserver queue with a single waiting line and waiting room of infinite size. The number of servers varies between a minimum of s and a maximum of S. Between s and S a server is added whenever the queue length reaches N and a server is removed from service every time the queue length drops below N. Performance measures of the model provide the effectiveness of such a policy in the operation of the system.

The optimality of increasing the service rate with the increasing number of customers in the system has been formally established by Crabill (1972).

In a queueing system $M/M/1$, let λ be the arrival rate and μ_i $(i = 1, 2, \ldots, K)$ be K possible service rates. Then Crabill uses two cost rates:

$C(i)$ = the customer cost rate incurred when there are i customers in the system

r_i = cost rate incurred when the service rate μ_i is being used.

The general policy stated by Crabill is as follows: If $C(i)$ is nondecreasing and $\to \infty$ as $i \to \infty$; $0 < \mu_1 < \mu_2 < \ldots < \mu_k$; $0 \le r_1 < r_2 < \ldots < r_k$; $\lambda < \mu_k$, and

$$\sum_{i=0}^{\infty} C(i)\left(\frac{\lambda}{\mu_k}\right)^i < \infty$$

then the optimal stationary policy is given by the specification of $K+1$ numbers $0 = d_1 \leq d_2 \leq d_3 \leq \ldots \leq d_k \leq d_{k+1} = \infty$ and the use of service rate μ_j when the number of customers in the system is $\geq d_j$ but $< d_{j+1}$. If $d_j = d_{j+1}$, then service rate μ_j is not used in the optimal policy. Crabill (1972) provides a proof of this policy for $K = 2$.

Another type of control problem that has been investigated in the queueing literature considers whether, given a cost structure, it is optimal to start serving when there is at least one customer in the system. For an $M/G/1$ queue, with the cost structure that includes a server start-up cost, a server shutdown cost, a cost per unit time when the server is turned on, and a holding cost per unit time spent in the system for each customer, Heyman (1968) has obtained a stationary optimal policy of turning the server on when a specified number of customers are present and turning it off when the system is empty. Balachandran (1973) derives a similar policy based on the workload in the system. Because of the esoteric nature of these investigations, we shall not explore them any further.

A substantial number of papers have been written on various optimal design and control problems. Readers interested in them are referred to the survey papers by Sobel (1974), Stidham and Prabhu (1974), Crabill at al. (1977), and the special issue of the journal *Queueing Systems* edited by Stidham (1995). These articles provide extensive bibliographies, though overlapping somewhat at times.

As mentioned in the introduction, because of the nature of queueing theory, design and control policies used in applications are relatively few. As the systems become complex, the representative models are also complex and the resulting performance measures become intractable for deriving useable policies. For these reasons, we have given only a few examples of such investigations. The survey articles cited above can be used to build an appropriate bibliography on topics of readers' interest.

Chapter 12

Queueing Theory Applications in Manufacturing Systems

Contributing Author: Professor Andrew Junfang Yu[1]

12.1 Introduction

The manufacturing process involves various operational steps in converting raw materials (used in a generic sense) into finished products. In order to make the process efficient and cost effective, analytical tools such as queueing theory have been used extensively with the advancement of technology. They play an important role in the performance analysis, design, and planning and control of manufacturing processes (see Govil and Fu (1999)). The objective of this chapter is not to review all important contributions that queueing theory has made to manufacturing. Here we provide a few glimpses of such contributions and give three illustrative examples.

There have been several surveys of applications of queueing theory in manufacturing systems. Buzacott and Yao (1986), Bitran and Dasu (1992), Kouvelis et al. (1992), Rao et al. (1998), and Kumar and Kumar (2001) are some of them. The variety of problems in manufacturing where queueing theory has been applied, identified in the comprehensive survey of Rao et al. (1998), are the following: transfer line and flow line production systems, job shop, advanced manufacturing systems that include flexible manufacturing systems (FMSs) and

[1]Department of Industrial and Systems Engineering, The University of Tennessee, Knoxville, TN 37996, USA.

© Springer Science+Business Media New York 2015

U. N. Bhat, *An Introduction to Queueing Theory*, Statistics for Industry and Technology, DOI 10.1007/978-0-8176-8421-1_12

Just-in-Time (JIT) operations, capacity planning and control, production support functions, and other manufacturing situations such as automated guided-vehicle systems (AGVs) and computer controlled storage facilities. Of these flow line and transflow line systems, job shops, and FMS can be easily identified as obvious application areas for queueing theory. As noted by the authors even other applications are no less significant.

The Jackson networks, including open and closed queueing networks (OQN and CQN, see Chapter 7) were analyzed initially to solve problems related to job-shop scheduling. Those analyses were further expanded into the analyses of FMS. Thus, queueing networks have become fundamental to the modeling and analysis of manufacturing systems. Before going to specific cases, we provide a brief overview of some research available in the literature in general terms.

Production processes on multistage assembly lines in which workpieces from two or more input stations are merged to form a new one for further processing is the subject of study in Manitz (2008). The ultimate objective is the determination of the throughput and other performance measures of such assembly lines while considering finite buffer capacities and generally distributed processing times. One of the illustrative examples (see Section 3) used later gives some details of the procedure used in the article.

Significant amount of research work in queueing theory application can be found in the literature related to production planning and control. Many of those cases are beyond a single manufacturer and are involved with multiple organizations that form, what we call supply chain, which is covered later in this section. Boucherie et al. (2003) introduce a new class of queueing networks called *arrival first networks*. The new model is developed for the production systems operating under a Kanban protocol. Karrer et al. (2012) propose a framework for developing production control strategy. The framework is used to address some important questions in production control such as, how to limit work in process and position order penetration point, and to cope with demand uncertainty. It is formulated as a queueing network model and solved numerically using simulation. Gayon et al. (2009) study a production planning problem for a single item make-to-stock production system with backorders. The problem is to determine optimal stock and capacity allocation policies for the cases where production may be interrupted and restarted. An $M/E_r/1$ queueing model is used and a heuristic policy is developed and assessed based on the results of the model analysis.

FMS was very popular in 1980s and 1990s with the rising of technologies such as robots, CNC machines, material handling systems, sensors, and computers. This is also the time when the bulk of the research on queueing theory application in FMS can be found. Buzacott and Yao (1986) outline state-of-the-art studies in FMS using analytical queueing network models. Their focus is to identify the major features of the models that are related to the operational

characteristics of FMS (see Section 2). Kouvelis et al. (1992) survey layout problems in an FMS environment, which deals with the allocation of workstations to equal number or more candidate locations in order to meet the throughput requirement of the system. One of their emphases is the queueing and dynamic aspects of the layout decisions. A key part of any FMS is the common conveyor subsystem that interconnects all the workstations. Coffman et al. (1988) find a new queueing problem in their study of FMS conveyor system, which has input–output dependencies resulting from the fact that the conveyor transports items both to and from a workstation. Dallery and Stecke (1990) consider single-class, multi-server closed queueing networks for the design and planning problems of flexible manufacturing. Their results are useful for characterizing optimal allocations of servers and workload using the optimal configuration of subnetworks that maximize the overall network or FMS throughput. Lin et al. (1994) also study the closed queueing network for FMS. However, their focus is to model a maintenance float network problem, which integrates with a cost optimization model to determine the best number of standby units and repair stations.

Semiconductor manufacturing systems is another leading industrial sector where queueing theory has been extensively applied. The readers may refer to surveys by Kumar and Kumar (2001) and Shanthikumar et al. (2007) for details.

The globalization of world economy has transformed the traditional manufacturing system into a distributed manufacturing and supply chain system. Most of the companies no longer manufacture their products completely in a single location and totally by themselves. Outsourcing and offshoring have been popular for a couple of decades already as companies have been looking for ways to reduce their cost and innovate their products in order to gain financial and technological edge in their competitive markets. In a distributed manufacturing environment supported by supply chains, lead time and uncertainty in and between operations typically are of concern more than in a stand-alone environment. This trend is evidenced by the high volume of research in the area of supply chain, including the application of queueing theory. The majority of articles that can be found in the literature related to the application of queueing theory in supply chain can be categorized as follows: supply chain design, supply chain planning and control, inventory control in supply chain, supply chain performance, product and process design for supply chain, logistics and transportation, and maintenance and spare parts management.

Kerbache and Smith (2004) develop a methodology based on analytical queueing network coupled with nonlinear optimization to design supply chain topologies and evaluate various performance measures. Their approach has proved useful for analyzing congestion problems and evaluating the performance of supply chains. Most of the applications of queueing theory are found in the area of supply chain planning and control, including inventory control. Bhaskar and Lallement (2010) look at a supply chain as a two-input, three-stage queueing network. Orders to the supply chain are modeled as two stochastic variables, one for the order arrival time and the other for the order quantity. The objec-

tive of the study is to obtain the minimum response time for the delivery of the orders along the three stages of the network. Thus, the optimum capacity of the queueing network can be obtained as the average number of order quantity that can be delivered within this minimum response time. Vericourt et al. (2002) study a capacitated supply chain that produces a single product demanded by several classes of customers with different backorder costs. The supply system is modeled as a multi-customer make-to-stock queue with stock allocation as a key decision problem. The problem is solved for optimal allocation policy using dynamic programming and heuristic algorithm. Bai et al. (2004) consider an inventory-queue, which is an inventory system controlled by a processing station with queueing. The most important issue for inventory-queues is the behavior of their departure processes that are triggered either by a new job arrival when the output buffer is not empty or when a service completion occurs. Their objective is to obtain the probability distribution and squared coefficient of variation of inter-departure times. Liu et al. (2004) develop a multistage inventory-queue model with an approach of a job-queue decomposition that evaluates the performance of serial manufacturing and supply systems with inventory control at each stage. The objective of their research is to find an efficient procedure to minimize the overall inventory in the system while meeting the required service level. Arda and Hennet (2006) analyze an enterprise network for an end-product manufacturer that makes production and supply plan to minimize its total holding and stock-out costs. The manufacturer has more than one supplier for each of its main components. Both the customer order arrival time and supplier delivery time are random. Arda and Hennet model the supply system as a queueing network with the inventory position level and the supply allocation as decision variables.

In addition to design, planning, and inventory control, the applications of queueing theory can also be found in performance evaluation and improvement, logistics and transportation, and other areas of supply chain. Viswanadham and Raghavan (2000) have investigated a dynamic modeling technique for analyzing supply chain network using generalized stochastic petri nets (GSPNs). They have used the framework of integrated GSPN-queueing network modeling, with the GSPN at the higher level and a generalized queueing network at the lower level, to solve the decoupling point location problem in supply chain. Their objective is to minimize the total relevant cost: inventory carrying cost and the delay costs. Wu and Dong (2008) develop a new methodology for performance analysis of multiproduct supply chain by combining multi-class queueing networks and inventory models. They employ a job-queue decomposition strategy to analyze the major performance measures and propose an approach for aggregating input streams and separating output streams to link all the sites or nodes in the supply chain together. Woensel et al. (2008) consider a vehicle routing problem with dynamic travel times due to traffic congestion. Their approach uses queueing theory to capture travel times. This approach is compared with other approaches and its benefits are evaluated and quantified. Lieckens and Vandaele (2012) consider a single product reverse logistics network design prob-

lem with multiple layers and multiple routings. They build a new advanced strategic planning model with integrated queueing relationships. The model takes into account stochastic delays due to various processes like collection, production, and transportation, as well as disturbances due to various sources of variability like uncertain supply, uncertain process times, unknown quality, breakdowns, etc. One of many other areas of application of queueing theory in supply chain is maintenance and spare parts management. Sleptchenko et al. (2005) examine the impact of repair priorities in spare part network. They model repair shops by multi-class, multi-server priority queues. Their objective is to reduce the inventory investment, which is necessary to attain target system availability, and to increase the utilization of repair shop.

Next, we select three articles from the literature as illustrative examples of applications of queueing theory in the analysis of manufacturing systems. In these examples we emphasize the modeling part of the article, rather than complete results with the assumption that interested readers can go to the articles to understand the complete analysis.

The readers should note that the notations used in the illustrative examples are mostly those of the authors of the articles and may not be the same as those used in earlier chapters.

12.2 Modeling a Flexible Manufacturing System Using Jackson Networks
(Buzacott and Yao 1986)

Queueing theory has long been applied in the modeling of FMS which is an integrated system that consists of several workstations. An FMS differs from the traditional manufacturing system in its flexibility. FMS has a set of versatile machines that can perform a variety of different types of operations with negligible setup times. Thus, the system can process a variety of jobs simultaneously. An integrated FMS is often controlled by computers, so jobs can have flexible routings, resulting in different paths of successive machine visits to complete their operations. In order to fully exploit the flexibility of an FMS to enhance its productivity, it is necessary to develop some models that can be used to predict its performance, and provide the guidelines for its design and control. Buzacott and Yao (1986) have discussed some of queueing network models that are applied to FMS.

An FMS is basically a flexible job shop in which machines or workstations are interconnected with automated conveyers controlled by computers. Jobs can enter and leave the system at any workstation, and their arrival and operation times at workstations are usually random. Based on these characteristics, it is natural to model the FMS as a Jackson queueing network. We assume that there are M workstations in the FMS modeled as an open queueing network.

Let J_i denote the total number of jobs at workstation i, which includes both the jobs in the queue and the jobs in service. Let $\mathbf{J} = (J_1, J_2, , J_M)$ and

$|\mathbf{J}| = \sum_{i=1}^{M} J_i$, where $\mathbf{J}$ is a job vector for the jobs in all workstations and $|\mathbf{J}|$ is the total number of jobs in the system. Assume that jobs arrive at the system in a Poisson process and their service times follow exponential distributions. The job arrival rate is a function of total number of jobs in the system, $\lambda(|\mathbf{J}|)$. The service rate at each workstation is a function of its queue length, $\mu_i(J_i)$, $i = 1, \ldots, M$. A new job could arrive at any workstation with certain probability. Let α_{0j} be such a probability for workstation j. Once a job enters the system, it will route through a number of workstations to complete its operations. The routing of jobs follows a Markov chain. Let $[\alpha_{ij}]$,$(i, j = 1, \ldots, M)$ be the routing matrix, where α_{ij} is the routing or transition probability from workstation i to workstation j. Let α_{i0} be the probability that a job leaves the system through workstation i. It is obvious that $\alpha_{i0} + \sum_{j=1}^{M} \alpha_{ij} = 1$.

The main differences between the open Jackson network discussed in Chapter 7 and the FMS model used here are the number of servers at each network node, in this case the workstation, the initial arrival rate at the system, and the service rates at different nodes or workstations. In the general open Jackson network discussion, we assumed multiple servers, constant new customer arrival rate, and constant service rate at each node. However, in the FMS model being discussed, we assume that there is only one server at each workstation; new job arrival rate is the same for all workstations, but not constant; and service rates are different for different workstations and not constant. The actual jobs arriving at a workstation include the new jobs from outside the system and routed jobs from other workstations. So the effective job arrival rate at a workstation is not the same as the new job arrival rate. Let γ_j denote the effective job arrival rate at workstation j. The effective job arrival rates can be obtained by solving the following traffic equations.

$$\gamma_j = \alpha_{0j} + \sum_{i=1}^{M} \gamma_i \alpha_{ij}, \quad j = 1, ..., M. \tag{12.2.1}$$

Let $\mathbf{n} = (n_1, \ldots, n_M)$ and $|\mathbf{n}| = \sum_{i=1}^{M} n_i$, where $\mathbf{n}$ is a vector with non-negative integers as elements. The equilibrium probability distribution of the number of jobs in the FMS, the most important property of the underlying Jackson network, can be expressed as follows:

$$P[\mathbf{J} = \mathbf{n}] = G^{-1} \Pi_{j=0}^{|n|-1} \lambda(j) \Pi_{i=1}^{M} f_i(n_i), \tag{12.2.2}$$

where

$$f_i(n_i) = \gamma_i^{n_i} \Pi_{j=1}^{n_i} \mu_i^{-1}(j), \quad i = 1, ..., M, \tag{12.2.3}$$

and G is a normalizing constant $(< \alpha)$,

$$G = \sum_{k=0}^{\infty} \Pi_{j=0}^{k-1} \lambda(j) \sum_{|n|=k} \Pi_{i=1}^{M} f_i(n_i). \tag{12.2.4}$$

Certain data are required to apply the above FMS model. First is the job routing matrix $[\alpha_{ij}]$, $(i, j = 1, \ldots, M)$, where the probability α_{ij} can be

interpreted as the proportion of jobs leaving workstation i for next operation at workstation j. The second is the mean service times at the workstations and mean inter-arrival times of new jobs to the system. Certain control mechanisms may be exercised to optimize the system performance through the dispatching of arriving jobs to artificially adjust the job arrival rate.

There are several special cases of the FMS model described above. The first is the fixed arrival rate for new jobs. The second is the closed queueing network. The third is the restricted open queueing network. We only discuss the first two cases here and refer the readers to Buzacott and Yao (1986) for the third case.

Constant Arrival Rate

Let $\lambda(\cdot) = \lambda$ be the constant arrival rate of new jobs from outside. The distribution function in (12.2.2) can be simplified as:

$$P[\mathbf{J} = \mathbf{n}] = P[\mathbf{Q} = \mathbf{n}] = \Pi_{i=1}^{M} P[Q_i = n_i], \tag{12.2.5}$$

where $\mathbf{Q} = (Q_1, \ldots, Q_M)$ is a random vector for the equilibrium number of jobs at different workstations in the system.

Closed Queueing Network

For a certain positive integer number N, let $\lambda(j) = 0$ if $j \geq N$ and $\lambda(j) = \infty$ otherwise. So we have, $\Pi_{j=0}^{k} \lambda(j) = 0$ for all $k \geq N$. This special case implies that there are a fixed number of jobs, N, in the FMS. As soon as one job completes its operations and leaves the system, a new job will be dispatched into the system. With the infinite arrival rates appearing in both numerator and denominator in (12.2.2), taking limits, we have

$$P[\mathbf{J} = \mathbf{n}] = G^{-1}(N) \Pi_{i=1}^{M} f_i(n_i), \tag{12.2.6}$$

where

$$G(N) = \sum_{|n|=N} \Pi_{i=1}^{M} f_i(n_i), \tag{12.2.7}$$

and $f_i(n_i)$ remains the same as in (12.2.3).

With $\alpha_{0j} = 0$, $a_{i0} = 0$, and $\sum_{j=1}^{M} \alpha_{ij} = 1$ for all $i, j = 1, \ldots, M$, the traffic equations in (12.2.1) will need another equation to have a unique solution. The equation $\sum_{i=1}^{M} \gamma_i = 1$ will serve the purpose. In this case, γ_i can be interpreted as the job visit frequency to workstation i. The probability distribution in equation (12.2.6) can also be expressed as,

$$P[\mathbf{J} = \mathbf{n}] = \frac{\Pi_{i=1}^{M} P[Q_i = n_i]}{P[|\mathbf{Q}| = |\mathbf{n}|]} = P[\mathbf{Q} = \mathbf{n} || \mathbf{Q}| = |\mathbf{n}|]. \tag{12.2.8}$$

From the above equation or (12.2.6), the marginal job length distribution can be expressed as follows:

$$P[J_i = n_i] = \frac{f_i(n_i) G_i(N - n_i)}{G(N)}, \quad (n_i \leq N) \tag{12.2.9}$$

where $G_i(N - n_i)$ can be considered as the normalizing constant for a closed queueing network with $M - 1$ workstations, excluding station i in the original M station network, with the reduced network having $N - n_i$ jobs.

From (12.2.3), it can be derived that $\mu_i(n_i)f_i(n_i) = \gamma_i f_i(n_i - 1)$. By the definition of $G_i(N - n_i)$ and (12.2.7), the following can also be derived:

$$\sum_{n_i=1}^{N} f_i(n_i - 1)G_i(N - n_i) = G(N - 1). \tag{12.2.10}$$

Let TH_i be the throughput of workstation i. The throughput can be expressed as follows:

$$TH_i = \sum_{n_i=1}^{N} \mu_i(n_i)P[J_i = n_i] = \frac{\gamma_i G(N-1)}{G(N)}, \quad (i = 1, ..., M). \tag{12.2.11}$$

Let TH be the FMS's throughput. With the assumption that $\sum_{i=1}^{M} \gamma_i = 1$, $TH = \sum_{i=1}^{M} TH_i$.

Buzacott and Yao (1986) outline the state of the art in the use of queueing theory in FMS at that time. What is given here is just an introduction and the article includes other topics such as reversible networks in FMS, approximate queueing networks, and performance models. Even though the article is nearly 30-years-old, for beginning readers in the area of queueing theory applications in FMS, it is highly recommended.

12.3 Assembly Lines with Finite Buffers (Manitz 2008)

Queueing phenomenon can be observed on multistage assembly lines where work-in-process (WIP) parts are merged at certain assembly stations from multiple sourcing stations. Material flows are asynchronous, processing times are random, and there are buffers in-between stations. Manitz (2008) develops a queueing model for analyzing such assembly lines by decomposing them into a series of two-station subsystems. The two are then analyzed by using G/G/1/N stopped-arrival queueing models.

Assume that an assembly line has M stations. There is a finite buffer between any two successive stations. Let B_{im} denote the buffer between stations i and m for WIP parts of station i waiting for processing at station m. Let C_{im} be the capacity of buffer B_{im}.

When multiple parts from multiple predecessor stations are merged at an assembly station, the operation at the station can only be started after all the parts are available. This synchronization constraint extends the holding time of a part at the server. It is assumed that parts can be loaded onto the server of the assembly station independently. This implies that the waiting part will not consume the input buffer. An assembly station is considered as blocked if

it cannot transfer a processed part into the following buffer because it is fully occupied. The blocked part is added to the queue in the buffer which effectively increases its capacity by one. An assembly station is defined as starving if it is completely empty.

We assume that the processing times can be represented by random variables. Let T_m^S be the processing time for a part at station m, which is distributed with a general distribution with the following mean:

$$E\{T_m^S\} = \frac{1}{\mu_m} \quad (m = 1, 2, \ldots, M), \tag{12.3.1}$$

where μ_m is the processing rate at station m. Its coefficient of variation is defined as,

$$CV\{T_m^S\} = \zeta_m \quad (m = 1, 2, \ldots, M). \tag{12.3.2}$$

Let X be the effective production or output rate of the finished product, which is defined as the mean number of finished product out of the last assembly station, say M, per unit time. The effective production rate, which is also called the throughput of the assembly line, can be expressed as the follows:

$$X = \mu_m P\{\text{station } M \text{ is busy}\}. \tag{12.3.3}$$

The main question of the problem is how to determine the probability that the last assembly station is busy. The busy period of the station excludes the repair, starving, waiting for synchronization, or other idle times. Manitz presents an algorithm to estimate the blocking and starving probabilities, and then the effective production rate of the assembly line.

The assembly line with M stations can be virtually decomposed into $M-1$ subsystems, each of such subsystems consisting of two adjacent stations connected by a buffer. Let (i, m) denote a subsystem with stations i and m, connected by the buffer B_{im}. The upstream station of the subsystem is denoted by $M_u(i, m)$ and the downstream station by $M_d(i, m)$. Consider a given station m which has a succeeding station s_m. Let $\mathbf{S}_m$ be the set of station m's successor stations and $\mathbf{P}_m$ be the set of station m's predecessor stations. Set $\mathbf{S}_m$ contains at most one station while set $\mathbf{P}_m$ contains at least one station. Thus, the multi-predecessor subsystem can be decomposed into several two-station subsystems. Let p_{mj} denote the jth predecessor station of the assembly line station m, where $j = 1, \ldots, |\mathbf{P}_m|$.

A decomposed two-station subsystem (i, m) can be modeled as a G/G/1/N queueing system. When its buffer B_{im} is full, the processed part at its preceding station i has to wait at the station. Also it stops processing any other part at that station. This stoppage switches off the arrival process to the subsystem. The maximum possible number of parts in the subsystem is the buffer capacity plus the two additional parts, one blocked in the upstream station i and one in the downstream station m, either being processed or waiting for synchronization.

The upstream station $M_u(i, m)$ in the decomposed subsystem (i, m) represents the segment of the assembly line upstream of the buffer B_{im}. The virtual

arrival rate of the arrival process into the buffer B_{im} is denoted as $\mu_u(i,m)$, which is the effective arrival rate when the arrival process is switched on. This effective arrival rate is not the same as the processing rate μ_i of station i as the former includes the factors of starving and synchronization at the station. The coefficient of variation for the inter-arrival time to the subsystem is denoted as $\zeta_u(i,m)$. For an assembly station like station m, it may appear as a downstream station in multiple subsystems in parallel. The downstream station $M_d(i,m)$ represents the segment of the assembly line downstream of the buffer B_{im} including these parallel subsystems. Let $\mu_d(i,m)$ denote the effective service rate of the subsystem, the rate at which the processed parts leave the subsystem (i,m) when it is not starved. Similar to the effective arrival rate, $\mu_d(i,m)$ differs from the processing rate μ_m of station m as the former includes the effects of blocking and waiting for synchronization. The coefficient of variation for the time experienced by parts at the downstream station is denoted as $\zeta_d(i,m)$.

Consider a subsystem (m, s_m), in which station m is an upstream station that generates the arrival process to the subsystem. If the arrival process is not blocked, the effective time between two consecutive service completions can be expressed as $T_{mj}^{IS} + T_{mj}^{IW} + T_m^S$, where T_{mj}^{IS} is the starving time of station m for parts from the jth preceding station p_{mj}; T_{mj}^{IW} is the waiting time of a part from station p_{mj} at station m for synchronization; T_m^S is the processing time at station m. We may identify these three components as phases, all of which may not be necessary for every part. The expected virtual inter-arrival time in the queueing system (m, s_m) and its squared coefficient of variation can be written as

$$\begin{aligned} \frac{1}{\mu_u(m,s_m)} &= E[T_{mj}^{IS} + T_{mj}^{IW} + T_m^S] \\ &\quad (m \in \{\{1,\ldots,M\} | \boldsymbol{S}_m \neq \mathbf{0}\}; j = 1,\ldots,|\mathbf{P}_m|), \quad (12.3.4) \end{aligned}$$

$$\begin{aligned} \zeta_u^2(m,s_m) &= \mu_u^2(m,s_m) Var[T_{mj}^{IS} + T_{mj}^{IW} + T_m^S] \\ &\quad (m \in \{\{1,\ldots,M\} | \boldsymbol{S}_m \neq \mathbf{0}\}; \quad j = 1,...,|P_m|). \quad (12.3.5) \end{aligned}$$

We assume that the phase lengths for each component of the completion time are independent and identically distributed (i.i.d) for all parts and all three time components are independent of each other for a particular part. We now need to determine the expected values and variances of the phase lengths. For the processing time T_m^S, the expected value and variance can be computed from the given data:

$$E[T_m^S] = \frac{1}{\mu_m} \quad (m \in \{1,...,M\}), \quad (12.3.6)$$

$$Var[T_m^S] = \frac{\zeta_m^2}{\mu_m^2} \quad (m \in 1,...,M). \quad (12.3.7)$$

The starving time T_{mj}^{IS} is the time that station m is starved for parts from station p_{mj} until the next service completion at the station. This time period is basically the remaining inter-arrival time in the subsystem (p_{mj}, m). As an approximation, assuming that the probability distribution of the remaining inter-arrival time is the same as the full inter-arrival time, a memoryless characteristic of exponential distribution, the expected value and variance of T_{mj}^{IS} can be derived as,

$$\begin{aligned} E[T_{mj}^{IS}] &= p_s^*(p_{mj}, m)\frac{1}{\mu_u(p_{mj}, m)} \\ &\quad (m \in \{\{1, ..., M\}|\mathbf{P}_m \neq \mathbf{0}\}; \quad j = 1, ..., |P_m|), \end{aligned} \tag{12.3.8}$$

$$\begin{aligned} Var[T_{mj}^{IS}] &= p_S^*(p_{mj}, m)(\frac{\zeta_u^2(p_{mj}, m) + 1}{\mu_u^2(p_{mj}, m)} - p_s^*(p_{mj}, m)\frac{1}{\mu_u^2(p_{mj}, m)}) \\ &\quad (m \in \{\{1, ..., M\}|\mathbf{P}_m \neq \mathbf{0}\}; \quad j = 1, ..., |\mathbf{P}_m|), \end{aligned} \tag{12.3.9}$$

where $p_s^*(p_{mj}, m)$ is the probability that station m is starved for parts from p_{mj} at its service completion.

The waiting time of a part from station p_{mj} at station m for synchronization, T_{mj}^{IW}, is the longest remaining starving time among the parallel subsystems, $(i, m), i \in \mathbf{P}_m/\{p_{mj}\}$. In addition to the assumption of the memorylessness of the starving time, Manitz also assumes that the moment of a maximum of some random variables can be approximated by the maximum of their moments. (See article for the derivation of $E[T_{mj}^{IW}]$ and $Var[T_{mj}^{IW}]$.)

The virtual service time of the subsystem (p_{mj}, m) is the time that a part from station p_{mj} experiences at station m, which should include the waiting time T_{mj}^{IW} for synchronization, the original processing time T_m^S, and the blocking time T_m^B. Accordingly, the expected value of the virtual service time and its squared coefficient of variation can be defined as,

$$\begin{aligned} \frac{1}{\mu_d(p_{mj}, m)} &= E[T_{mj}^{IW} + T_m^S + T_m^B] \\ (m \in \{\{1, ..., M\}||\mathbf{P}_m| &\geq 2, \mathbf{S}_m \neq \mathbf{0}\}; \quad j = 1, ..., |\mathbf{P}_m|), \end{aligned} \tag{12.3.10}$$

$$\begin{aligned} \zeta_d^2(p_{mj}, m) &= \mu_d^2(p_{mj}, m)Var[T_{mj}^{IW} + T_m^S + T_m^B] \\ (m \in \{\{1, ..., M\}||P_m| \geq 2, \mathbf{S}_m &\neq \mathbf{0}\}; \quad j = 1, ..., |P_m|). \end{aligned} \tag{12.3.11}$$

The only unknown component in the above equations is the blocking time T_m^B at station m, which is the remaining time of a part at station s_m, which, in turn, is the remaining virtual service time in the corresponding queueing system. The expected value and variance of the blocking time at station m can be shown as,

$$\begin{aligned} E[T_m^B] &= p_B^*(m, s_m)\frac{1}{\mu_d(m, s_m)} \\ &\quad (m \in \{\{1, ..., M\}|\mathbf{S}_m \neq \mathbf{0}\}), \end{aligned} \tag{12.3.12}$$

$$Var[T_m^B] = p_B^*(m,s_m)\left(\frac{\zeta_d^2(m,s_m)+1}{\mu_d^2(m,s_m)} - p_B^*(m,s_m)\frac{1}{\mu_d^2(m,s_m)}\right) \quad (m \in \{\{1,...,M\}|\mathbf{S}_m \neq \mathbf{0}\}), \tag{12.3.13}$$

where $p_B^*(m,s_m)$ is the probability that station m is blocked at the time of a service completion. The expected value and squared coefficient of variation of the virtual service time as defined in (12.3.10) and (12.3.11) can be derived from (12.3.12) and (12.3.13), as described in the article.

Some of the performance measures of the assembly line can now be calculated with the parameter values of the virtual queueing systems derived above. One of the key performance measures is the production rate of the assembly line, which is the effective output rate of the last station. This is equivalent to the virtual service rate of the last subsystem in the assembly line when it is not starved. Assume that the downstream station in the last subsystem is station M and the upstream station is p_{M1}. The production rate X can be expressed as,

$$X = \mu_d(p_{M1},M)(1-p_S(p_{M1},M)), \tag{12.3.14}$$

where $p_S(p_{M1},M)$ is the probability that the subsystem (p_{M1},M) is empty. This probability can be expressed as,

$$p_s(p_{M1},M)=P_0(\mu_u(p_{M1},M),\zeta_u(p_{M1},M),\mu_d(p_{M1},M),\zeta_d(p_{M1},M),C_{p_{M1},M}+2), \tag{12.3.15}$$

where $P_0(\lambda,c_A,\mu,c_S,N)$ is the long-term probability that a queueing system, with arrival rate λ, service rate μ, coefficient of variations of inter-arrival times and service times c_A and c_S respectively, and capacity N, is empty. For probability calculations, Manitz uses the results given by Buzacott and Shanthikumar (1993). He also provides expressions for other performance measures such as, effective output rates and blocking probabilities of other subsystems.

With the derived virtual parameters for the arrival and service processes of the $M-1$ subsystems, the author develops an algorithm that iteratively calculates the upstream parameters in a forward pass from station 1 to M, and downstream parameters in a backward pass. The algorithm also includes the procedures to calculate the blocking and starving probabilities for each subsystem. Finally, the production rate of the assembly line and other performance metrics can be calculated using the iteratively computed virtual parameters of the subsystems and the associated blocking and starving probabilities. Manitz has also conducted several numerical tests for his algorithm which can be found in his original article.

12.4 A Supply Chain with Multiple Suppliers
(Toktas-Palut and Ulengin 2011)

Supply chains consist of suppliers, manufacturers, warehouses, and distribution centers. In Toktas-Palut and Ulengin (2011) the authors consider a supply

chain with two stages: multiple independent suppliers and a manufacturer with limited production capacities. The following assumptions are made:

- The number of suppliers is $n(\geq 2)$.
- The suppliers operate on a make-to-stock basis and apply base stock policy to manage their inventories.
- S_i is the base stock level of supplier i.
- No inventory is held by the manufacturer (make-to-order strategy).
- The customer demands arrive in a Poisson process with rate λ in single units.
- The service times of supplier i are i.i.d. with exponential distributions with mean $1/\mu_i$.
- The manufacturer's service time is also exponentially distributed with mean $1/\mu_M$.
- The traffic intensity of supplier i $(i = 1, 2, ..., n)$ and the manufacturer are ρ_i and ρ_M, respectively. It is assumed $\rho_i < 1$ and $\rho_M < 1$.

When the manufacturer receives a demand, it triggers each supplier. If the supplier's stock is not empty, the demand is met immediately. If the stock is empty, the component is released to the manufacturer only when its production is complete. It should be understood that when the supplier's stock is empty, there are outstanding back orders. The manufacturer can start production only after receiving components from all suppliers. The customer demand is met when the manufacturer completes its production. All orders are processed on an first-come, first-served (FCFS) basis at all facilities. It is also assumed that the transfer times between facilities are negligible.

The modeling of this seemingly simple system has to be done in two stages. The first stage is when the production is at the supplier. The second stage is the production at the manufacturer, which can start only when all components have arrived from the suppliers. The key is the determination of the inter-arrival time distribution at the manufacturer stage. When that is known the system at that level can be modeled as a G/M/1 queue.

When there is only one supplier Buzacott et al. (1992) have derived the probability density function (p.d.f.) $f_{(\cdot)}(t)$ of the inter-departure times of the supplier (that are the inter-arrival times of the manfacturer) as

$$f_A(t) = \lambda e^{-\lambda t}(1 - \rho_1^{S_1+1}) + \mu_1 e^{-\mu_1 t}\rho_1^{S_1-1} - (\lambda + \mu_1)e^{-(\lambda+\mu_1)t}(1 - \rho_1^2)\rho_1^{S_1-1}, \tag{12.4.1}$$

where A stands for inter-arrival times. When the number of suppliers gets larger, the authors devise a strategy to derive an approximate distribution in place of (12.4.1).

Since the manufacturer cannot start production until all components have arrived, the supplier with the minimum base stock level is likely to affect the inter-arrival times for the manufacturer the most. With this assumption and using (12.4.1) an approximate p.d.f. in the multiple supplier case can be obtained as

$$f_A(t) \simeq \lambda e^{-\lambda t}(1-\rho_j^{S_j+1}) + \mu_j e^{-\mu_j t}\rho_j^{S_j-1} - (\lambda+\mu_j)e^{-(\lambda+\mu_j)t}(1-\rho_j^2)\rho_j^{S_j-1}, \tag{12.4.2}$$

where supplier j is the one with the minimum base stock level among all suppliers. This distribution gives an approximate squared coefficient of variation as

$$C_A^2 \simeq 1 - 2\rho_j^{S_j+1}\frac{1-\rho_j}{1+\rho_j}. \tag{12.4.3}$$

After simulation studies involving two, three, and four suppliers, customer demand rate at one, traffic intensities of the suppliers and the manufacturer at 0.50, 0.67, and 0.80, and with the base stock levels of the suppliers as 3, 5, and 7, the authors conclude the approximate inter-arrival time distribution of the product for the manufacturer given in (12.4.2) as quite appropriate, giving an error of only 2.47 % (as compared to 71.6 % for an approximate exponential distribution).

We give here three performance measures of the system derived by the authors using the results derived above: $E(B_i)$, expected outstanding backorders for supplier i; $E(I_i)$, expected inventory level for supplier i; and $E(N_M)$, expected number of jobs in the manufacturer subsystem. Since at the supplier level, the subsystem acts like an M/M/1 queue, following Buzacott and Shanthikumar (1993), the first two measures can be given as

$$E[B_i] = \frac{\rho_i^{S_i+1}}{1-\rho_i}, \quad i = 1, ..., n \tag{12.4.4}$$

and

$$E[I_i] = S_i - \frac{\rho_i(1-\rho_i^{S_i})}{1-\rho_i}, \quad i = 1, ..., n. \tag{12.4.5}$$

At the manufacturer's subsystem, one can use a G/M/1 queueing model with (12.4.2) as the inter-arrival time distribution, and an exponential distribution with rate μ_M for service times. Again without closed form expressions for $E(N_M)$ when the inter-arrival times have a general distribution of the type (12.4.2), the authors resort to approximate results based on coefficients of variation of the two distributions available in the literature. After comparing different formulas with the help of simulation models used earlier, the authors use Marchal's (1976) approximation formula for the average number of jobs in the manufacturer's subsystem. This is given as

$$E[N_M] \simeq \rho_M + \left(\frac{\rho_M^2(1+C_S^2)}{1+\rho_M^2 C_S^2}\right)\left(\frac{C_A^2+\rho_M^2 C_S^2}{2(1-\rho_M)}\right) \tag{12.4.6}$$

where C_A^2 and C_S^2 are the squared coefficients of variation of the inter-arrival time distribution and the service time distribution, respectively. Substituting from (12.4.3) for C_A^2 and noting $C_S^2 = 1$, we get

$$E[N_M] \simeq \rho_M + (\frac{2\rho_M^2}{1+\rho_M^2})(\frac{(1+\rho_j)(1+\rho_M^2) - 2\rho_j^{S_j+1}(1-\rho_j)}{2(1+\rho_j)(1-\rho_M)}) \qquad (12.4.7)$$

where supplier j is the one with the minimum base stock level.

With this analysis the authors derive results for two other quantities: Expected value of N_{qM}, the number of jobs in the manufacturer's queue and expected value of B_M, the outstanding backorders at the manufacturer.

The authors also explore two decision problems related to costs in this system: (1) A centralized model in which optimization is based on the overall supply chain and (2) A decentralized model in which, the suppliers individually and the manufacturer separately optimize based only on their own entity. Interested readers may refer to the paper for details.

Chapter 13

Queueing Theory Applications in the Analysis of Computer and Communication Systems

Contributing Author: Professor Krishna M. Kavi[1]

Computer and communication systems are very complex and are rapidly evolving. Understanding the behavior of these systems is essential to provide reasonable answers to the questions of cost and performance of the systems. In many cases the utilization of computers owned by most businesses is low, thus wasting their investments. This is the reason for the popularity of Cloud computing. In simple terms, Cloud computing allows organizations to "rent" computing and storage from a Cloud provider such as Amazon, Google, HP, or Microsoft, and pay only for the resources and services they actually use. Businesses and even individuals can store their data on the Cloud disk storage and access them from anywhere. Organizations can also develop services (or programs) and deploy their IT services on the computers of Cloud providers. These services can be made available to customers of these businesses from anywhere in the world. Examples of such services include travel services such as Expedia or auction sites like eBay. In addition to the benefit of paying only for the amount of computing used, Cloud computing offers another advantage. Businesses can request any model or operating system. This makes it easy to run legacy software that was designed for a specific computer system or operating system when those systems are no longer available.

[1]Department of Computer Science and Engineering, University of North Texas, Denton, TX 76203, USA

© Springer Science+Business Media New York 2015

U. N. Bhat, *An Introduction to Queueing Theory*, Statistics for Industry and Technology, DOI 10.1007/978-0-8176-8421-1_13

Cloud providers such as Amazon must maximize their profits while providing their customers promised levels of performance (known as service level agreements) and plan for future capacity needs. A Cloud provider may be required to pay penalties if an agreed upon level of performance is not met. Businesses that use Cloud computing need to understand their workloads to manage their IT budgets and shop for Cloud service providers that offer the best value. Modeling Cloud computing systems is more complex than can be described in this book.

Consider eBays auction business that relies on Cloud computing. eBay rents IT services from a Cloud provider such as Google or Amazon. Typically such business operations rely on three types of systems. Customers are connected to eBay through the Internet using web browsers like Explorer or Safari. There is a second system that performs computations to process user bids, charging users for purchases of merchandise, answering queries about merchandise, etc. A third system maintains a database of the inventory, current bids for individual items, and user accounts. To model the eBay system as a queueing network, we need to model three interacting queues with complex communication paths among the three systems, including matching of job requests as they flow through them. Each system is itself a queueing network with multiple distributed central processing units (CPUs), storage devices, terminals, and communication networks. The system may be simplified and modeled as a closed network. Yet the number of requests generated by the customers in the closed network in terms of bids or database accesses, can vary significantly. We introduce simple models of computer and communication systems as queueing networks so that they can be analyzed using techniques presented in previous chapters. We also introduce operational analysis and mean value analysis techniques, which are amenable to numerical solutions. Interested readers may explore the available literature to find more details on how Cloud computing systems can be analyzed.

13.1 Modeling Computer Systems

To simplify, we normally model an application by representing its workload in terms of processing, disk, and network requirements. Applications are customers in a queueing network requiring various services. Consider a typical single processor system shown in Figure 13.1.

We assume that jobs enter at the CPU and depart from the CPU. A job may request service from a disk (or other I/O devices), for example, requesting data from a file. This model does not allow us to describe realistic features of applications, including simultaneous requests to multiple services or an application creating new jobs during its execution (using *fork* commands). However, this simple model permits us to derive performance measures including response times (the amount of time a job spends in the system) and utilization rates of each service.

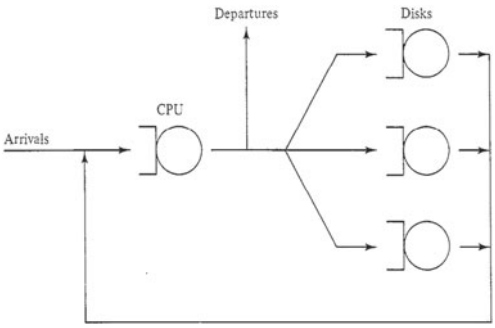


Figure 13.1 A simple computer system as an open queueing network

We can model systems shown in Figure 13.1 using open queueing networks described in Chapter 7. For example, the relative throughput of CPU and I/O devices can be derived as follows. Let λ_i be the arrival rate at node i (we use $i = 0$ for the CPU), and let $\boldsymbol{\lambda} = (\lambda_0, \lambda_1, ..., \lambda_k)$. In a steady state we have $\boldsymbol{\lambda}\mathbf{P} = \boldsymbol{\lambda}$, where $\mathbf{P}$ is the routing matrix describing the probability that a customer at station i will go to station j. The individual throughputs (i.e., λ_i) can be obtained by solving $\boldsymbol{\lambda}\mathbf{P} = \boldsymbol{\lambda}$. We can use the normalizing equation $\lambda_0 + \lambda_1 + ... + \lambda_k = 1$ and compute the relative throughputs of each server. One can also view relative throughputs in terms of the relative number of visits made by a job to each device. We formulate solutions based on visits later in this chapter.

It should be made clear that queues with feedback do not preserve the Poisson arrival process. In Figure 13.1, jobs depart I/O devices and rejoin the CPU queue. The combined arrival process at the CPU (including external arrivals and jobs returning from I/O devices) is not Poisson. Systems with feedback flows, when modeled as open queueing networks, can be solved assuming local balance (see Section 7.3 of Chapter 7).

We can also use closed queueing networks to model a computer center. Consider Figure 13.2, which includes a CPU, memory, and I/O devices. It should be noted that the memory subsystem is nonseparable and cannot be analyzed independently. One can use delay stations to model a memory queue and to include delays incurred in making available an adequate amount of memory to the application. Memory constraints can also be modeled implicitly in closed networks by restricting the number of jobs in the system, assuming all jobs require the same amount of memory. In other words, the number of jobs in the system is determined by the amount of memory needed by the jobs and the total amount of memory available. More accurate representations of memory constraints require the use of simulation techniques. Terminals normally represent "think time" where the job is delayed by a user before reentering the CPU queue.

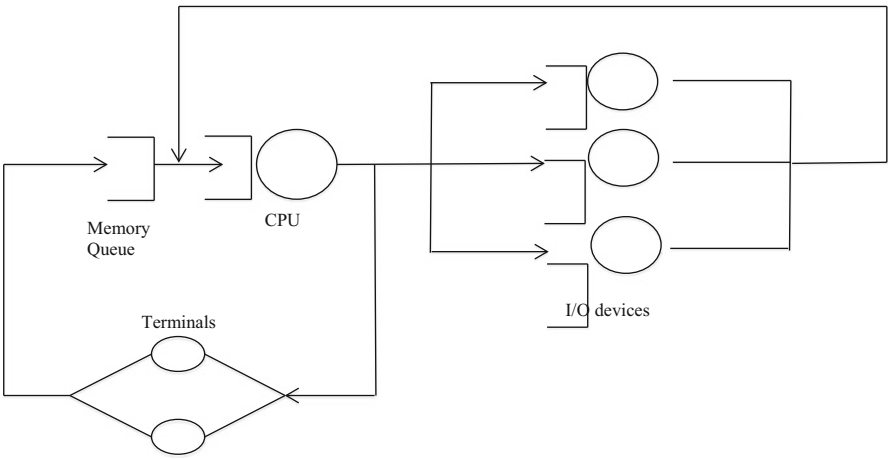


Figure 13.2 A simple computer system as a closed queueing network

We can use operational laws (Section 13.3) to analyze such a system. Suppose we want to compute the overall response time of the system. In a closed network, we assume that the number of jobs in the system is constant: when a job leaves the system, a new job takes its place in the system. Let us consider a system with a single I/O device (or a disk drive). Let us also assume that there are 20 terminals ($N = 20$), the think time is 10 s, and a job typically makes 16 visits to the I/O device. Finally, assume that the utilization of the I/O device is 0.4 and the average service time of the disk is 0.025 s.

We can apply Little's law (Section 9.2 of Chapter 9) to the I/O subsystem to obtain its throughput (or the number of disk requests completed per second):

$$X_{disk} = (U_{disk}/S_{disk}) = 0.4/0.025 = 16$$

Since each job makes 16 visits to the I/O subsystem, the number of jobs in the system is:

$$X_{system} = (N_{disk}/V_{disk}) = 16/16 = 1.0$$

The time (including think time) a job spends in the system is given by $N/X_{system} = 20/1.0 = 20$ seconds. The response time, excluding think time, is $20 - 10 = 10$ seconds.

Implied in the above formulation is the concept of the forced flow law. In a closed system, we assume that the *flows* (or throughputs) in all parts of a system must be proportional to each other.

Multiple Classes of Jobs

In our introduction, we have assumed that jobs exhibit similar behavior and thus can be described with a single set of resource requirements. In reality, computer systems serve different types of jobs with widely varying service requirements. Consider the eBay auction application outlined above. Customers of eBay may either be just browsing the current items available for auction or very intensively involved in the bidding process; and they present entirely different demands on processing, communicating with and accessesing databases. Each job class can be described with its own workload (arrival rates, visit ratios, throughput of each device—or in case of closed networks—think time and number of jobs in the system). We can then analyze the workloads to obtain per class response times or per class utilization of devices. The overall system performance can then be obtained using weighted averages from per class analyses. It should be noted that such performance values are only average values and do not provide distributional information. Simulations can be used to control workloads of different job classes and obtain per class and system wide performance values.

Also see Example 13.3.2 that shows the differences between using single class and multiple class models.

13.2 Modeling Communication Systems

Communication networks can be modeled as a network of servers. A message originates at a source and travels to several intermediate nodes before reaching its destination. To compute the time it takes for a message to reach its destination, we can use an open queueing network with N service centers (representing the source, destination, and $N-2$ intermediate nodes through which the message is routed.)

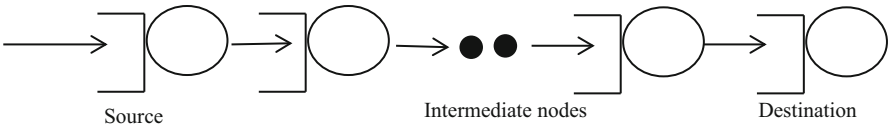


Figure 13.3 An open network for communication systems

Such a model (see Figure 13.3) is too simplistic to be useful. Most systems rely on flow control to regulate the number of messages from source to destination to avoid congestion, as well as buffer allocation and management issues. We can use closed queueing networks to model flow control as shown in Figure 13.4. This representation assumes a total of $2K-1$ messages (or jobs) in the system with N nodes including source, destination, and intermediate nodes as before. The pacing station waits until it receives K messages and then places the K messages in the generation station queue. The pacing station can be viewed as implementing the flow control whereby there are at most K messages being

serviced by the N nodes, limiting the maximum number of messages inflight. The generation station is associated with a service rate to represent the arrival rate of messages into the communication system. The pacing station makes such a system nonseparable. Global balance techniques can be used to obtain response times for our system. The system can also be simplified by replacing the N nodes of the communication system with a single flow equivalent aggregate node. To construct a flow equivalent aggregate, we assume that the

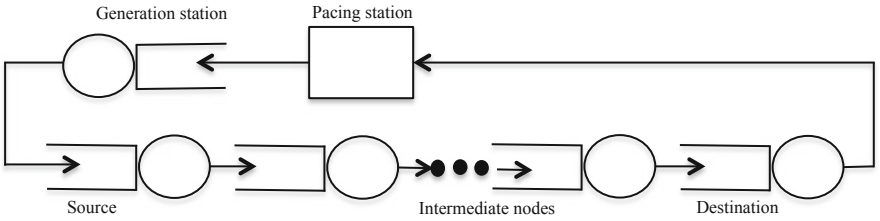


Figure 13.4 A closed network for communication systems

average rate at which customers leave the aggregate system (i.e., the throughput of the aggregate system) depends only on the customer population within the aggregate and not where these customers are waiting. This is based on the assumption that a local balance exists in the aggregate system. The aggregate system will then exhibit load-dependent service rates and we must compute the rates for all possible populations from 1 to $2K - 1$.

Once we replace the N-node communication system with a single aggregate node, the overall system will be simple enough for global balance techniques.

13.3 Modeling and Analysis using Computational Tools

In this section, we present techniques that can be used to numerically analyze queueing systems. While it is possible to use numerical analysis techniques for solving differential and integral functions to compute values associated with the various mathematical equations presented in previous chapters, here we take a different approach to analyze queueing systems. Although the techniques presented here are based on the same underlying principles used earlier, the solutions are based on simplifying assumptions and in some cases yield only mean values. These approaches provide practical solutions to systems that are intractable or when the behaviors cannot be easily modeled using simple probability distributions. We provide algorithms and techniques for analyzing

simple systems. Several computational tools are available to permit modeling and analysis of more complex systems.

Operational Laws for Performance Analysis

Operational laws describe relationships among various performance measures. Our discussion follows an excellent survey article on the topic by Denning and Buzen (1978). For ease of understanding and to avoid confusion, when new notation is introduced, we use the same notation as given in that article.

Consider a computer system modeled abstractly, where we can observe the system in terms of jobs entering and completed jobs leaving. Assume that over an observation period T, A jobs enter the system and D jobs depart from the system. We can then define the arrival rate $\gamma = A/T$ and throughput $X = D/T$. Suppose it is also observed that the computing system has been busy for B time units, then we can define the utilization of the system as $U = B/T$. The average service requirement per job is obtained as $S = 1/\mu = B/D$. From these relationships we can show that $U = X * S$. In a stable system, when the observation period T is very long, all arriving jobs will leave the system and $\gamma = X$.

Little's law (Section 9.2 of Chapter 9) is a more general operational law. Let N be the average number of jobs in the system, which can be obtained by observing the number of jobs in the system at regular intervals, say every second, and calculating the average of these observations. Little's law can then be used to relate the average number of jobs in the system to the average response time (or the average time a job is resident in the system) R, as $N = \gamma * R$. As previously shown, Little's law can be applied to a computer system with a CPU and I/O disks.

Consider a computer system with a CPU and k I/O devices. We assume the number of jobs in the system is fixed (i.e., a closed network where a new job enters a system as soon as a job departs from the system.) The system is observed for a period of time T. Let B_i be the time that device i is busy providing service. Let C_{ij} be the number of times a job requests service at device j immediately after completing service at device i, and $C_i = \sum_{j=0}^{k} C_{ij}$.

Using these quantities, we can estimate the following measures with respect to each device that we have seen in earlier sections.

Utilization $U_i = \rho_i = \frac{B_i}{T}$

Effective output rate $\gamma_i = \frac{C_i}{T}$

Routing probability $\alpha_{ij} = \frac{C_{ij}}{C}$

Note that we are assuming input and output flow balance in these expressions. Since the CPU is where the job is initiated and completed (as shown in Figure 13.1), using subscript 0 to indicate its status, we have the job-flow

balance equations

$$\gamma_0 = \sum_{i=1}^{k} \gamma_i \alpha_{i0}. \tag{13.3.1}$$

We use the term *response time* that is common in computer science literature for the total time a job spends in the system (i.e., waiting+service). The response time R_i at device i can be estimated as (the total amount of time accumulated at a device)/(the number of services completed at the device). If Q_i represents the number at device i waiting for or being serviced in the long run, using Littles law ($L = \gamma R$), for each device we get

$$E(Q_i) = \gamma_i E(R_i). \tag{13.3.2}$$

Since in Markovian networks job flows are balanced, γ_i can be identified as the device *throughput*. These quantities also give us *visit ratios,* which are the mean number of service requests per job for a device relative to the mean number of jobs coming to the systems. The visit ratio V_i for device i can be defined as $V_i = \frac{\gamma_i}{\gamma_0}$, and estimated as $\frac{C_i}{C_0}$, remembering that device 0 is the CPU. The relation

$$\gamma_i = V_i \gamma_0 \tag{13.3.3}$$

is known as the *forced flow law,* which states that the flow in any one part of the system determines the flows everywhere in the system. Substituting from (13.3.3) in (13.3.1), we obtain the visit ratio equations

$$\begin{aligned} V_0 &= 1 \\ V_i &= \gamma_{0j} + \sum_{i=1}^{k} V_i \gamma_{ij}, \qquad j = 1, 2, ..., k. \end{aligned} \tag{13.3.4}$$

The system response time R is obtained by pooling the response times of all devices. From Littles law, writing $E(Q) = \sum_{i=1}^{k} E(Q_i)$, we have

$$E(R) = \frac{E(Q)}{\gamma_0}.$$

Using (13.3.2) and (13.3.3), we get

$$E(R) = \sum_{i=1}^{i} V_i E(R_i) \tag{13.3.5}$$

which is known as the *general response time law.* This result is valid even when the network is not Markovian.

In a closed network (see Figure 13.4) with M jobs and a think time of Z (the time spent at the terminal by a user before submitting a job), the total time includes the response time and think time. When the job flow is balanced, we have

$$M = [E(Z) + E(R)]\gamma_0$$

giving

$$E(R) = \frac{M}{\gamma_0} - E(Z) \tag{13.3.6}$$

This relationship is known as the *interactive response time law.*

Denning and Buzen (1978) include several illustrative examples of these relationships. The reader should also see Jain (1991) for further elaboration of these laws.

Mean Value Analysis

Another important analysis procedure used in applications is the *mean value analysis* (MVA). It applies to closed queueing networks and provides their performance in mean values. MVA can be used only if a queueing network has a product form solution. We limit ourselves to simple service centers with a fixed limit on the queue size and a single class of customers (or jobs).

Ordinarily to determine response times in networks, one has to get the mean queue lengths and the corresponding throughputs (effective arrival rates) and use Little's law. In their article simplifying this procedure for applications, Reiser and Lavenberg (1980) show that in closed queueing networks the mean queue sizes, the mean waiting times, and throughputs can be computed recursively without computing the product terms and normalizing constants. The key result in this computation is the seemingly simple result relating the mean waiting time of a closed system with N customers with the mean waiting time of a system with $N-1$ customers, thus providing a recursion. Let $R_j(N)$ be the response time at station j when there are N customers in the closed network, and let $Q_j(N)$ be the number of customers in that station $t \to \infty$. The recursion established by Reiser and Lavenberg is the relationship

$$E[R_j(N)] = E(S_j)\{1 + E[Q_j(N-1)]\}.^{2}$$

The derivation of this relationship is omitted because of its complexity. Instead, the usefulness of MVA is illustrated with the following numerical example. Readers are cautioned to note that in the numerical illustration notations are simplified by dropping the expected value operator E.

To set the framework for illustrating MVA, consider the following network shown in Figure 13.5, representing a computer system with a single CPU and several I/O devices (or file servers). Each of these devices represents a service station. A task (or a computer program) starts at the CPU, visits a file server, returns to CPU for more service, and repeats this process of visiting a file server and CPU, until the task is completed. Thus a job makes V_j visits to service station S_j. If jobs are not lost, the arrival rate at each service station is the same as the departure rate, and the arrival rate into the computer system is the same as the departure rate from the system. For such systems V_j can be

[2] $E[S_j]$ is the expected service time at j.

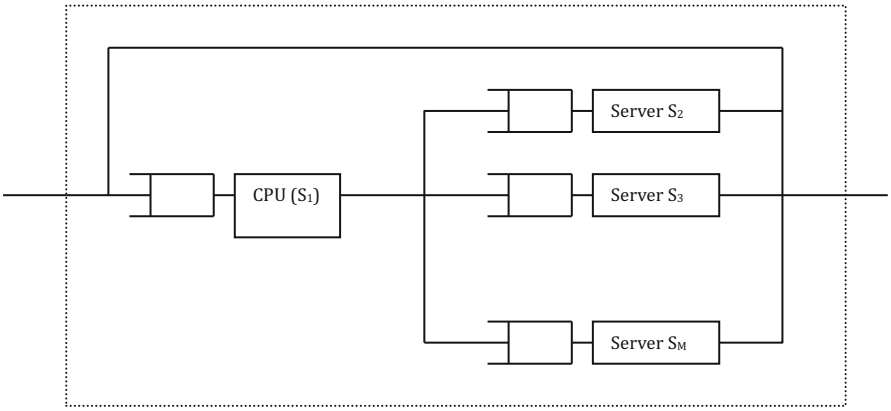


Figure 13.5 A simple computing system

computed as $V_j = \frac{\gamma_j}{\gamma_0}$, where γ_0 is the arrival rate of jobs entering the system (and also leaving the system assuming job flow balance) and γ_j is the arrival rate of jobs at jth service center. The number of visits to CPU is given by $1 + \sum_{j=2}^{j=M} V_j$.

These formulations are based on *operational laws* (Denning and Buzen 1978) introduced in the previous section and that can be verified by direct observations.

In a closed network, the number of jobs in the system is fixed. This can be a model for a system where a new job arrives soon after a completed job leaves the system. Such models are used to represent time-sharing computer systems where the number of terminals connected to the system represents the total number of jobs in it. One can insert a delay before a job reenters the system to represent think time of a user sitting at a terminal.

It has been shown in Reiser and Lavenberg (1980) that the mean response time for service at the jth service station in a closed network with N jobs is given by

$$R_j(N) = (1/\mu_j) * [1 + Q_j(N-1)], \tag{13.3.7}$$

where μ_j is the service rate and $Q_j(N)$ is the mean number of jobs at the jth service station. This relationship is intuitive. The Nth job arriving at the jth service center will see a queue with a mean number of jobs (including the one being serviced) given by $Q_j(N-1)$, and must wait for these jobs to be serviced. It should be noted that this formulation assumes that the service distribution is exponential. The response time shown in the equation above can be solved iteratively, by starting with $Q_j(0) = 0$.

To compute the mean response time $R(N)$ of the system with N jobs and M service centers, we use operational laws that specify that

$$R(N) = \sum_{j=1}^{M} R_j(N) V_j \tag{13.3.8}$$

Here V_j is the number of visits that a job makes to the jth service center.

Using Littles law, we can obtain the mean throughput rate and mean number of jobs at service station j. In case of a delay representing think time, the system throughput is given by $X(N) = \frac{N}{(R+Z)}$ where Z is the mean think time of a user.

The queue lengths of each service station can be calculated as

$$Q_j(N) = X(N) * R_j * V_j \tag{13.3.9}$$

Example 13.3.1 Consider a computer system with a CPU (C) and three file servers (labeled F1, F2, F3) that can perform file reads and writes. Let us assume that each job visits F1 10 times, F2 20 times and F3 30 times. After each visit to a file server, the job comes back to CPU (thus the number of visits each program makes to CPU is 1+10+20+30 = 61). We are also given the following data. The mean service times per visit to the various service stations are given as: CPU = 1; F1= 2; F2= 3; F3 = 4.

Initialization: $N = 0$

$Q_C = Q_{F1} = Q_{F2} = Q_{F3} = 0$

Iteration 1: $N = 1$

$R_C(1) = (1/\mu_C)[1 + Q_C(0)] = 1 * [1 + 0] = 1$
$R_{F1}(1) = (1/\mu_{F1}(0))[1 + Q_{F1}(0)] = 2 * [1 + 0] = 2$
$R_{F2}(1) = (1/\mu_{F2})[1 + Q_{F2}(0)] = 3 * [1 + 0] = 3$
$R_{F3}(1) = (1/\mu_{F3})[1 + Q_{F3}(0)] = 4 * [1 + 0] = 4$

System Response time

$R(1) = R_C(1) * V_C + R_{F1}(1) * V_{F1} + R_{F2}(1) * V_{F2} + R_{F3}(1) * V_{F3}$
$= 1 * 61 + 2 * 10 + 3 * 20 + 4 * 30 = 261$

Queue lengths at each service station are computed as follows

$Q_j(N) = [N/R(N)] * R_j(N) * V_j$
$Q_C(1) = [1/R(1)] * R_C(1) * V_C = (1/261) * 1 * 61 = 0.234$
$Q_{F1}(1) = [1/R(1)] * R_{F1}(1) * V_{F1} = (1.261) * 2 * 10 = 0.077$
$Q_{F2}(1) = [1/R(1)] * R_{F2}(1) * V_{F2} = (1/261) * 3 * 20 = 0.230$
$F3(1) = [1/R(1)] * R_{F3}(1) * V_{F3} = (1/261) * 4 * 30 = 0.460$

Iteration 2. $N = 2$

$R_C(2) = (1/\mu_C)[1 + Q_C(1)] = 1 * [1 + 0.234] = 1.234$
$R_{F1}(2) = (1/\mu_{F1})[1 + Q_{F1}(1)] = 2 * [1 + 0.077] = 2.154$
$R_{F2}(2) = (1/\mu_{F2})[1 + Q_{F2}(1)] = 3 * [1 + 0.230] = 3.69$
$R_{F3}(2) = (1/\mu_{F3})[1 + Q_{F3}(1)] = 4 * [1 + 0.460] = 5.84$

System Response time

$R(2) = RC(2) * VC + R_{F1}(2) * V_{F1} + R_{F2}(2) * V_{F2} + R_{F3}(2) * V_{F3}$
$= 1.234 * 61 + 2.154 * 10 + 3.69 * 20 + 5.84 * 30 = 345.814$

Queue lengths

$Q_C(2) = [2/R(2)] * R_C(2) * V_C = (2/345.814) * 1.234 * 61 = 0.435$
$Q_{F1}(2) = [2/R(2)] * R_{F1}(2) * V_{F1} = (2/345.814) * 2.154 * 10 = 0.125$

$$Q_{F2}(2) = [2/R(2)] * R_{F2}(2) * V_{F2} = (2/345.814) * 3.69 * 20 = 0.427$$
$$Q_{F3}(2) = [2/R(2)] * R_{F3}(2) * V_{F3} = (2/345.814) * 5.84 * 30 = 1.013$$

We can continue the iterative process to find response times and queue lengths for higher numbers of jobs (N) in the system. Table 13.1 below shows some values.

Table 13.1 Mean response times and queue lengths

N	R	Q_C	Q_{F1}	Q_{F2}	Q_{F3}
1	261.00	0.234	0.077	0.230	0.460
2	345.814	0.435	0.125	0.427	1.013
3	437.215	0.601	0.154	0.587	1.657
4	534.875	0.730	0.173	0.712	2.358
5	637.915	0.827	0.184	0.805	3.184
6	745.494	0.897	0.191	0.872	4.041
7	856.711	0.946	0.195	0.918	4.942
8	970.700	0.978	0.197	0.4948	5.877
9	1086.709	0.999	0.198	0.968	6.834
10	1204.129	1.013	0.199	0.981	7.807
20	1322.500	1.857	0.363	1.797	15.983
30	2407.349	2.172	0.340	2.092	25.397
40	3573.418	2.166	0.300	2.076	35.458
50	4778.653	2.021	0.272	1.931	45.776
100	5998.709	3.072	0.424	2.932	93.572

Example 13.3.2 In order to appreciate the difference between single class and multiple class models, consider a system with two classes of jobs (A and B) and two service stations, a CPU and a Disk. We are given the following information based on observations. The average service time required by A class jobs at CPU and Disk are 1.0 and 0.09 seconds, and for B class jobs they are 0.1 and 0.9 seconds. If we model the system as a single job class, we observe that the average service times for jobs are 0.6 and 0.4 seconds (these numbers are based on different number of A and B class jobs, and the number of visit to the Disk server). Using MVA analysis we observe that the response time for the single class model is 1.53 seconds while the response time for A and B classes are 2 and 1.85 seconds. This is due to the fact that class A jobs are CPU intensive while class B jobs are disk intensive.

The following pseudo algorithm using C programming notation can be used for MVA.

```
for (j=1; i<=M; j++) //Initialization
     Q[j] = 0.0;
for (k=1; k<=N; k++) // Main loop
   {
     for (j=1; j<=M; j++) //Compute new response times
       R[j] = (1.0 / mu[j]) * (Q[j] + 1.0);
R = 0.0;
       for (j=1; j<=M; j++) //Compute system response time
           R = R + R[j];
          for (j=1; j<=M; j++) //Update queue lengths
              Q[j] = (k/R)*R[j];
       }
```

When dealing with networks containing delay centers, where a job arriving at a center is serviced immediately without having to wait, the only change that needs to be made to MVA is the response time computation. For delay centers we use $R_j(N) = 1/\mu_j$.

In order to use MVA for multiple classes of customers, we iterate the MVA for each class of customers. In other words, we find the average queue lengths iteratively for each customer class.

Approximation Solutions To MVA

As can be seen from Examples 13.3.1 and 13.3.2, MVA is a recursive algorithm and it can be computationally expensive for systems with very large number of jobs. There have been many extensions and approximate solutions proposed with MVA so that it can be used with other types of queues to obtain upper bounds on response times, or to improve the computational efficiency of the analyses. It is beyond the scope of this book to discuss these extensions. Here we describe one approximate algorithm. We follow the convention used in Jain (1991). The algorithm is due to Schweitzer (1979), which is based on the assumption that the queue length at each station (or device) is proportional to the number of jobs in the system. In other words, as the number of jobs increases, so does the queue length at every device. This implies that

$$\frac{Q_i(N-1)}{(N-1)} = \frac{Q_i(N)}{N} \quad \text{for every device } i.$$

We can write the MVA equations as follows

$$R_i(N) = \begin{cases} (1/\mu_i)[1 + \frac{N-1}{N} Q_i(N)] \\ (1/\mu_i) \end{cases} \tag{13.3.10}$$

(the bottom expression is for delay centers while the top expression is for fixed capacity service centers).

$$\begin{aligned} X(N) &= \frac{N}{Z + \sum V_i R_i(N)} \\ Q_i(N) &= X(N) V_i R_i(N) \end{aligned} \tag{13.3.11}$$

where $X(N)$ is the system throughput and Z is the think time in a closed network

The algorithm starts with some value for $Q_i(N)$ and computes new $Q_i(N)$ values using above expressions, until the values converge.

Convolution

The MVA presented so far provides an easy way to obtain average (mean) response times and queue lengths; but MVA is not useful for obtaining more detailed analysis, such as the distribution of queue lengths or response times. In this section we introduce how some of these analyses can be made using convolution techniques.

Chapter 7 included an analysis of both closed and open networks of queues. These analyses can be used to solve for the distribution of jobs in a system, $p_{n_1,n_2,...,n_M}$[3] where there are n_j jobs at service station j (including the job being serviced) and $[n_1, n_2, ..., n_M]$ denotes the state of the system. Note that the system state represents an element of the set defined here

$$\underset{N}{\rightarrow} = \{[n_1, n_2, ..., n_M] | \sum_{j=1}^{j=M} n_j = N|\}.$$

In this chapter we restrict ourselves to closed networks of queues and provide a technique that can be implemented as a computer program.

For systems where the service time per job is independent of the queue lengths (load-independent service), we can use the following result (Gordon and Newel 1967)

$$p_{n_1,n_2,...,n_M} = \frac{(d_1^{n_1}, d_2^{n_2} ... d_M^{n_M})}{G(N)} \tag{13.3.12}$$

where d_j is the total service demand per job at the jth device and $N = \sum_{j=1}^{j=m} n_j$. The total demand for service by a job at a service station is the combined service requirements for all visits a job makes to the service station. $G(N)$ is a normalizing constant such that the probabilities that the system is in any one of the possible states add to 1. This is very complex since we need to find probabilities for all possible states of the system where the number of states is given by $\binom{N+M-1}{M-1}$ with N as the number of jobs in the system and M as the number of service stations.

Buzen's (1973) iterative solution method for $G(N)$, described in Section 7.6 of Chapter 7, is based on the following observation.

$$\sum_{\underset{N}{\rightarrow}} \Pi_{j=1}^{M} (d_j)^{n_j} = \sum_{\underset{N}{\rightarrow} | n_M = 0} \Pi_{j=1}^{M} (d_j)^{n_j} + \sum_{\underset{N}{\rightarrow} | n_M > 0} \Pi_{j=1}^{M} (d_j)^{n_j} \tag{13.3.13}$$

[3] To be consistent with Buzen (1973), we use subscript 1 for CPU, unlike subscript zero used previously.

where the summation is over the set of all possible states $[n_1, n_2, ...n_M]$, such that $\sum_{j=1}^{M} n_j = N$. The first term on the right-hand side is the case when there are zero customers at service station M, which can be viewed as a system with one less service station. The second term indicates that there is at least one customer at service station M, and places one service demand on that server. Thus the second term can be rewritten as $d_M \sum_{\overrightarrow{N-1}} \Pi_{j=1}^{M} (d_j)^{n_j}$ where the summation is over the set of all possible vectors $[n_1, n_2, ...n_M]$, such that $\sum_{j=1}^{M} n_j = N - 1$. Note that since there is at least one customer at service station M, we factored d_M out. The summation now deals with a system with one less customer. Thus we have

$$\sum_{\overrightarrow{N}} \Pi_{j=1}^{M} (d_j)^{n_j} = \sum_{\overrightarrow{N}|n_M=0} \Pi_{j=1}^{M} (d_j)^{n_j} + d_m \sum_{\overrightarrow{N-1}} \Pi_{j=1}^{M} (d_j)^{n_j}. \tag{13.3.14}$$

If we use $g(n, m)$ for $\sum_{\overrightarrow{N}} \Pi_{j=1}^{M} (d_j)^{n_j}$ then the normalizing constant $G(N)$ is given by $g(N, M)$.

But as we have seen

$$g(n, m) = g(n, m-1) + d_m * g(n-1, m). \tag{13.3.15}$$

The initial conditions are:

$$g(j, 0) = 0 \text{ for } j = 1, 2, ..., n$$

$$g(0, k) = 1 \text{ for } k = 1, 2, ...m$$

(13.3.15) provides the basis of the iterative convolution algorithm for computing the normalizing constant $G(N)$. This can be used to compute the state probabilities as shown in (13.3.12).

Example 13.3.3 Let us use the same example (Example 13.3.1) as the one used for computing MVA. Here we have four service stations (CPU and three file servers). Using the service times and the number of visits at each service station, we can obtain the service demands as shown here.

$$\begin{aligned} d_1 &= d_{cpu} = 1 * 61 = 61 \\ d_2 &= d_{F1} = 2 * 10 = 20 \\ d_3 &= d_{F2} = 3 * 20 = 60 \\ d_4 &= f_{F3} = 4 * 30 = 120 \end{aligned}$$

Table 13.2 shows $g(n, m)$ values for $n = 0, 1..., 10$ and $m = 1, 2, 3, 4$.

Thus $G(N)$ when $N = 10$ is $3.01 * 10^{21}$.

Using this value for the normalizing constant, we can find the probability distribution given by (13.3.12). For example, the probability that all ten customers are waiting at the CPU is given by

$$p_{10,0,0,0} = \frac{(61^{10})}{3.00989 * 10^{21}} = 2.027 * 10^{-10}.$$

Table 13.2 Interative convolution algorithm for $G(N)$

n	g(n,1)	g(n,2)	g(n,3)	g(n,4)
0	1	1	1	1
1	61	81	141	261
2	372	5341	13801	45121
3	226981	333801	1161861	6576381
4	13845841	20521861	90233521	879399241
5	844596301	1255033521	6669044781	1.12197E+11
6	51520374361	76621044781	4.76764E+11	1.39404E+13
7	3.14274E+12	4.67516E+12	3.3281E+13	1.70613E+15
8	1.91707E+14	2.85211E+14	2.28207E+15	2.07018E+17
9	1.16941E+16	1.73984E+16	1.54323E+17	2.49964E+19
10	7.13343E+17	1.0613E+18	1.03207E+19	3.00989E+21

As can be seen from this example, using total service demands for service stations may lead to a $G(N)$ that is too large (or too small in some systems) to provide accurate results in a computer (although one can use higher precision arithmetic such as double precision floating point arithmetic). In such cases, the service demands can be scaled up or down by writing $y_j = (\frac{d_j}{k})$ and using the scaled value y_j in the convolution algorithm.

Computing Other Performance Measures

Once $G(N)$ is known for a closed queueing network, we can obtain other performance measures including queue lengths, utilizations, and response times of individual service stations. It should be noted that the convolution algorithm not only computes $G(N)$ but also computes several intermediate values including $G(N-i)$.

Queue Length The probability that there are k or more jobs at service station j is given by

$$P(n_j \geq k) = \sum_{\vec{N}|n_j \geq j} \frac{d_1^{n_1} d_2^{n_2}, ..., d_M^{n_M}}{G(N)} = d_j^k \frac{G(N-k)}{G(N)}. \tag{13.3.16}$$

Note that if we use a scaled value $y_i = (\frac{d_j}{k})$ while computing $G(N)$, we will use y_j in the above equation as well.

In the previous example,

$$P(n_1 \geq 5) = d_1^5 \frac{G(10-5)}{G(10)} = 0.00023.$$

This gives the probability that there are five or more jobs at the CPU.

Using this method, we can find the entire distribution for the number of jobs at each service station. Consider first computing $P(n_j \geq 0)$, then computing $P(n_j \geq 1)$. We can find the probability of $P(n_j = 0) = P(n_j \geq 1) - P(n_j \geq 0)$. Likewise, we can compute $P(n_j = k)$ for all k. From these probabilities we can compute the expected values for queue lengths.

Utilization The utilization of service station j is the probability that there is at least one customer at that service station. In other words

$$U_j = P(n_j \geq 1) = d_j \frac{G(N-1)}{G(N)}.$$

In the previous example, the utilizations of the various service stations are

$$U_{CPU} = 0.508, \qquad U_{F1} = 0.167, \qquad U_{F2} = 0.50, \qquad U_{F3} = 1.00$$

As can be seen, file server F3 is a bottleneck since it reached 100 % utilization. Note that for the purpose of simplifying the examples, we picked service times and visits that are whole numbers. These numbers should not be viewed as representative of a real computing system.

Throughput The throughput of service station j is given by $X_j = \frac{U_j}{(1/\mu_j)}$. Since closed queueing networks are based on forced flows, the system throughput is given by $\frac{X_j}{V_j} = \frac{U_j}{d_j}$. For the given example the system throughput is 0.0083 jobs per unit time.

In this chapter we have considered only simple queueing systems. MVA and convolution algorithms for more complex queueing networks are available in the literature. Interested readers should consult more advanced sources for such techniques.

13.4 Remarks

Most of the material described in this chapter is based on traditional modeling and analytical techniques that were developed more than two decades ago. Several of the classical books about computer and communication systems performance modeling and evaluation are no longer available in print. However the books by Jain (1991) and Kant (1992) are by no means old. Jain (1991) provides comprehensive coverage of performance modeling, benchmarking, simulation, and capacity planning of computer systems. A new book by Le Boudec (2011) is more theoretical in its coverage and includes topics such as data collection, modeling, fitting, and testing. The book Obaidat and Boudriga (2011) provides the basic queueing models and simulation techniques that can be used for both computer systems and communication networks. There are other books that cover related material on computer and communication systems performance modeling and evaluation and they include Dattatreya (2008) and Bolch (2006).

Modern computer and communication systems are very complex and cannot easily be modeled as networks of queues, either because the arrival and service distributions are non-Markovian, where the service distributions may involve multiple classes and priority queues, or the flow of jobs in the network of queues may require matching of multiple service paths. Thus most of the current research utilizes simulation techniques for analyzing the performance of computer and communication systems. The next chapter introduces simulation techniques and informs how to use some of the publicly available software packages for simulations as well as for modeling systems as queueing networks.

Modeling workloads for Cloud computing environments and modeling applications in cloud-based computer systems is an active research area. Interested readers should consult recent publications from various IEEE and ACM conferences (that cover Cloud computing and big data analyses) and related journals.

13.5 Exercises

1. Model the following computing system with a CPU and two disk drives as an open queueing network. The arrival rate is 1 transaction per second, and a job makes 20 visits to disk A and 5 visits to disk B (thus it makes a total 26 visits to the CPU). The service times are 1 second for CPU, and 30 and 25 milliseconds for disks A and B, respectively. Determine the average number of transactions in the system and the average response time.

2. Model the system in Exercise 1 as a closed network and using MVA compute response times and throughput for N=1...10.

3. Using Schweitzer approximation approach, compute the response times and throughput for N=100. Verify the accuracy of the result using exact MVA analysis.

4. Using Buzen's convolution algorithm, find the distribution for queue lengths at CPU for system described in Exercise 2.

Chapter 14

Simulating Queueing Systems

Contributing Author: Professor Krishna M. Kavi[1]

When systems modeled as stochastic processes or queueing systems become complex and dynamic, analytical or numerical solutions outlined in the previous chapter may become intractable. In such cases, a computer program that mimics the behavior of the system (or at least the behaviors of interest) may be used. The computer program (or simulation) is run with several random values and the modeled behaviors are recorded for analysis.

Key to good simulations is the quality of the random number generators used in them. Computer generated random numbers are actually pseudorandom numbers since they all start with a seed that is not random. With the same initial seed, the generators produce the same sequence of random numbers. The numbers in the sequence represent outcomes of a uniform random variable. Repeating the same sequence of random numbers is sometimes useful in reproducing results of a simulation. However, simulations may have to be repeated with different seeds to produce a sample of the population of outcomes. To produce accurate analyses of the system, statistical analysis of these results is required. A good random number generator should have a long period before the random numbers recycle. The correlation between successive numbers in a sequence should also be small. The linear congruential (LC) method is a widely used technique for generating random numbers. In this method, the next random number r_n is generated using the current random number r_{n-1} using the following equation:

$$r_n = (a * r_{n-1} + c) \text{modulo m},$$

where a and c are nonnegative constants. To produce m different numbers the following conditions must hold:

[1]Department of Computer Science and Engineering, University of North Texas, Denton, TX 76203, USA

© Springer Science+Business Media New York 2015

U. N. Bhat, *An Introduction to Queueing Theory*, Statistics for Industry and Technology, DOI 10.1007/978-0-8176-8421-1_14

- The constants m and c are relatively prime.
- All prime factors of m divide $a - 1$.

To increase the range of numbers generated and to reduce the correlation among successive numbers, several variations to the LC method have been proposed. These include multiplicative LC (where $c = 0$) and adaptive LC (where $r_n = (r_{n-1} + r_{n-k})$ modulo m). Due to the growing interest in computer security using cryptography, that requires the generation of random keys, there have been several new techniques for generating long sequences of random numbers.

For most simulations, we recommend using a random number generator that has been tested for its quality (for example, those provided by MATLAB).

Using a random number generator that represents a uniform probability distribution with a range [0,1], other probability distributions can be generated. For example, the following function generates outcomes of a Poisson distribution with an arrival rate of lambda, and a fixed time interval of T.

```
int poisson (float lambda, T)
{
float r, temp;
int n;
n =0;
temp= -1/ (lambda * ln(random_number(seed)));

while ( temp < T)
{  n = n+1;
  temp = temp -1/ (lambda * ln (random_number(seed)));
}
return n;
}
```

The accuracy of a simulation also depends on a clear understanding of the modeled system including interactions among the various subsystems as well as the quality of the developed software. Since complex behaviors lead to complex models and complex programs, they are difficult to validate for correct behaviors. A good simulation should permit variance in data (or simulation parameters) to study the modeled systems under different conditions.

Since simulations of stochastic systems use random numbers, they are known as Monte Carlo simulations. Typically, computer simulators only simulate specific events at discrete times and hence they are also known as discrete event simulators. An event can be viewed as a point in time when the modeled system changes its state. Examples of events include: the arrival of a new customer (or job), the start of a service, or the end of a service. Program defined state variables are used to track the state of the system. Examples of state variables include the number of jobs waiting at each server (or in each queue when multiple job classes are modeled). Other variables are used to define system

parameters including arrival rates, service rates, and maximum sizes of queues. The program will simulate the events by changing the values of system variables and changing the time (or simulated clock) to the time of the event.

The following outlines a generic structure of typical simulators.

```
    Initialize;  //Initialize termination conditions
//Initialize system state variables, clocks
//Schedule an initial event

while (termination is false)
{ set_clock; //move clock to next event time
   simulate_next_event; //execute procedures to simulate the event
    //remove the simulated event
   update_statistics;
}
Analyse_results; //produce statistical reports
```

To develop a simulator for a queueing system (e.g., $M/M/1$), we can select one of the two possible variations. We can create all job arrival events at the very beginning of the simulation. We use a random number generator to generate the time of arrival for each job (by adding inter-arrival time to the time when the last job arrived). Alternatively, we can generate one job at a time. In this case, we randomly generate a new event which can either be an arrival or service. We recommend the first choice because it will be easier to control the simulation, and this approach also permits reproduction of the population such that different queue disciplines (such as priority scheduling, earliest-deadline-first (EDF), shortest-job-first (SJF)) can be applied to the same population.

It is also necessary to decide on a termination test based either on a total number of jobs processed by the simulation or a maximum time period over which the system is simulated. In the first case, all jobs entering the system will be processed, while in the second case, not all entering jobs may be processed by the time the simulation is terminated.

It is necessary to decide on the information to be associated with each job. In a simple $M/M/1$ system using first-in, first-out (FIFO) discipline (also identified as first-come, first-served (FCFS) in earlier chapters), it is only necessary to keep the time of arrival, the time when a service is initiated, and service time with each job. From this information it is possible to calculate waiting times and response times for each job, as well as average waiting times and response times for the system. For real-time systems, it is necessary to maintain deadlines by which a job must be completed. Deadline can be based on service times or created randomly.

Changes in processing the lists of waiting jobs can simulate variations to FIFO queue disciplines. To implement EDF scheduling, it is necessary to sort the waiting list of jobs by their deadlines. To implement SJF scheduling, the list is sorted by the service times of waiting jobs. Priority queues can be simulated by maintaining separate lists for each priority.

To simulate $M/M/1$, the simulation time is set to the arrival time of the next job in the waiting list. If the server is idle, the job is scheduled by setting the service initiation time. If the server is busy, the simulation time is set to the service completion time of the currently serviced job (which is equal to service initiation time plus service time). At this time, the next waiting job is scheduled for service. This process is repeated until the termination condition is met.

$M/M/$n (identified as $M/M/$s in earlier chapters) queues can be simulated as follows. The simulation clock is set to the earliest time when any server completes an assigned job (and becomes idle). A new job (unless the waiting queue is not empty) is assigned to the server.

Programming languages and software libraries are available to simplify the design of simulation programs. They provide ready-made random number generators, functions to generate various probability distributions, data structures to queue events, manage time, record outcomes, and produce common statistical analyses. One of the earliest languages is SIMULA, dating back to the 1960s. Newer versions of SIMULA based on C++ and JAVA have been developed at various universities, often as freeware. Another example is service provisioning markup language (SMPL), developed by MacDougall at MIT, which contains a set of C language functions that can be used to simulate queueing systems. Other commercial languages and tools are available for purchase. In this chapter, we will focus on developing simulation systems using MATLAB.

Even when using available software libraries, it is still necessary to develop programs representing a modeled systems behavior. The behaviors of each modeled component, the connections among the components (how a job moves from one component to another), and how a waiting queue of jobs are processed must be coded into your simulation. In the next section, we provide a basic introduction to MATLAB and how it can be used to model queueing systems.

14.1 Using MATLAB

MATLAB[2] is a high-level technical computing language and interactive environment for algorithm development, data visualization, data analysis, and numeric computation. Using MATLAB, we can solve technical computing problems faster than with traditional programming languages, such as C, C++, and Fortran. MATLAB is available for Windows, Linux, Solaris, and Mac. There is also a student edition that is educationally priced that runs on Windows, Mac, and Linux.

MATLAB's functionality can be extended by adding different toolboxes for optimization, statistics, data analysis, control system design, signal processing, image processing, data acquisition, financial modeling, application deployment, and computational biology.

The statistics toolbox, for instance, provides tools for data organization, statistical plotting and data visualization, analysis of variance, linear and nonlinear

[2] http://www.mathworks.com/products/matlab/

modeling, hypothesis testing, and probability distributions which may be very useful when simulating queues.

The MATLAB program below[3] will perform a discrete event simulation of an $M/M/1$ queue with arrival rate $\lambda = 0.5$ and service rate $\mu = 1$. The variable **nextarrival** gives the time when the next customer will arrive. Similarly, **nextdeparture** gives the time when the customer currently being served will depart (this is set to infinity if the queue is currently empty). The key statement is **if nextarrival < nextdeparture**, which determines whether the next event to occur will be an arrival or a departure. For an arrival, we move the **now** variable forward to the time of the arrival, increase the length of the queue **currentlength** by 1, announce the arrival with a **disp** statement, and schedule the next arrival (after this one) by resetting **nextarrival**. Recall that **(–1/lambda)*log(rand)** generates an exponential (λ) inter-arrival time. If the newly arrived customer is the only one present (i.e., **if currentlength == 1**), the customer can go straight into service, so we also decide how long the service will take by generating a random service time **(−1/mu)*log(rand)** with the exponential (μ) distribution, and setting **nextdeparture** accordingly.

To handle a departure, we decrease the current queue length by 1 and announce the departure with another **disp** statement. This either leaves the queue empty, in which case **nextdeparture** must be set to infinity, or brings another customer into service, in which case **nextdeparture** must be set by generating a service time for that customer.

The complete processing is enclosed in a **while** loop which keeps the simulation going until **targettime**, which is the time when the simulation must end.

$M/M/1$ queue simulation

```
lambda= 0.5;
mu= 1.0;

targettime= 50;

nextarrival= (-1/lambda)*log(rand);
now= 0;
nextdeparture= inf;  % infinity
currentlength= 0;

while now < targettime,

if nextarrival < nextdeparture,
  now= nextarrival;
  currentlength= currentlength + 1;
  disp(sprintf('Arrival at : %f (current length %d)', now,
  currentlength));
  nextarrival= now + (-1/lambda)*log(rand);
```

[3] http://www.stat.uackland.ac.nz/s̃tat320/

```
  if currentlength == 1,
     nextdeparture= now + (-1/mu)*log(rand);
  end
else
  now= nextdeparture;
  currentlength= currentlength - 1;
  disp(sprintf('Departure at : %f (current length %d)', now,
  currentlength));
  if currentlength > 0,
     nextdeparture= now + (-1/mu)*log(rand);
  else
     nextdeparture= inf;
  end
end

end
```

When the program is run, the output is something like:
Arrival at : 0.102314 (current length 1)
Departure at : 0.601800 (current length 0)
Arrival at : 3.031791 (current length 1)
Departure at : 3.146866 (current length 0)
Arrival at : 4.474956 (current length 1)
Arrival at : 5.018319 (current length 2)
Departure at : 5.259194 (current length 1)

Each time the program goes through the main loop, the program generates one line of output, corresponding to an arrival or departure.

The following is another example of a simple $M/M/1$ queue simulation that graphs the average number of clients in the system, the average delay, and the utilization.

Implementation of a simple $M/M/1$

```
queue_lim=200000; % system limit
arrival_mean_time(1:65)=0.01;
service_mean_time=0.01;
sim_packets=750;              %number of clients to be simulated
util(1:65) = 0;
avg_num_in_queue(1:65) = 0;
avg_delay(1:65) = 0;
P(1:65) = 1;

for j=1:64 %loop for increasing the mean arrrival time

    arrival_mean_time(j+1)=arrival_mean_time(j) + 0.001;
```

```
num_events=2;

% initialization
sim_time = 0.0;

server_status=0;
queue_size=0;
time_last_event=0.0;

num_pack_insys=0;
total_delays=0.0;
time_in_queue=0.0;

time_in_server=0.0;
delay = 0.0;

time_next_event(1) = sim_time + exprnd(arrival_mean_time(j+1));

time_next_event(2) = exp(30);

disp(['Launching Simulation...',num2str(j)]);

while(num_pack\insys < sim_packets)

min_time_next_event = exp(29);
  type_of_event=0;
    for i=1:num_events


      if(time_next_event(i)<min_time_next_event)
         min_time_next_event = time_next_event(i);
         type_of_event = i;
      end;

  end

  if(type_of_event == 0)
     disp(['no event in time ',num2str(sim_time)]);
  end

  sim_time = min_time_next_event;

time_since_last_event = sim_time - time_last_event;
time_last_event = sim_time;
```

```
 time_in_queue = time_in_queue + queue_size
                 * time_since_last_event ;

 time_in_server = time_in_server + server_status
                  * time_since_last_event;


if (type_of_event==1)

%disp(['packet arrived']);
% -------------------------arrival-------------------------
time_next_event(1) = sim_time + exprnd(arrival_mean_time(j+1));
% epomenos xronos afiksis

if(server_status == 1)

   num_pack_insys = num_pack_insys + 1;
   queue_size = queue_size + 1 ;

   if(queue_size > queue_lim)
      disp(['queue size = ', num2str(queue_size)]);
      disp(['System Crash at ',num2str(sim_time)]);
      pause
   end

   arr_time(queue_size) = sim_time;

else


   server_status = 1;
   time_next_event(2) = sim_time + exprnd(service_mean_time);


end


elseif (type_of_event==2)

    % ---------------service and departure---------------

    if(queue_size == 0)
      server_status = 0;
      time_next_event(2) = exp(30);
  else
```

```
        queue_size = queue_size - 1;

        delay = sim_time - arr_time(1);
        total_delays = total_delays + delay;

        time_next_event(2) = sim_time + exprnd(service_mean_time);


        for i = 1:queue_size
           arr_time(i)=arr_time(i+1);
        end

    end

 end

end

%results output
util(j+1) = time_in_server/sim_time;
avg_num_in_queue(j+1) = time_in_queue/sim_time;
avg_delay(j+1) = total_delays/num_pack_insys;
P(j+1) = service_mean_time./arrival_mean_time(j+1);


end

%----------------------graphs---------------------------------
figure('name','mean number of clients in system
        diagram(simulated)');
plot(P,avg_num_in_queue,'r');
Xlabel('P');
Ylabel('mean number of clients');
axis([0 0.92 0 15]);

figure('name','mean delay in system diagram (simulated)');
plot(P,avg_delay,'m');
Xlabel('P');
Ylabel('mean delay (hrs)');
axis([0 0.92 0 0.15]);

figure('name', 'UTILIZATION DIAGRAM');
plot(P,util,'b');
Xlabel('P');
Ylabel('Utilization');
```

```
axis([0 0.92 0 1]);
```

Routines[4] to simulate $M/G/1$ and $M/G/\infty$.

```
function [jumptimes, systsize, systtime] = simmg1(tmax, lambda)
% SIMMG1 simulate a M/G/1 queueing system. Poisson arrivals
% of intensity lambda, uniform service times.
%
% [jumptimes, systsize, systtime] = simmd1(tmax, lambda)
%
% Inputs: tmax - simulation interval
%         lambda - arrival intensity
%
% Outputs: jumptimes - time points of arrivals or departures
%          systsize - system size in M/G/1 queue
%          systtime - system times

% set default parameter values if ommited
if (nargin==0)
  tmax=1500;                  % simulation interval
  lambda=0.99;                 % arrival intensity
end

arrtime=-log(rand)/lambda;  % Poisson arrivals
i=1;
  while (min(arrtime(i,:))<=tmax)
    arrtime = [arrtime; arrtime(i, :)-log(rand)/lambda];
    i=i+1;
  end
n=length(arrtime);           % arrival times t_1,...,t_n

servtime=2.*rand(1,n);          % service times s_1,...,s_k
cumservtime=cumsum(servtime);

arrsubtr=arrtime-[0 cumservtime(:,1:n-1)]';           % t_k-(k-1)
arrmatrix=arrsubtr*ones(1,n);
deptime=cumservtime+max(triu(arrmatrix));  % departure times
                                        % u_k=k+max(t_1,..,t_k-k+1)

% Output is system size process N and system waiting
% times W.
B=[ones(n,1) arrtime ; -ones(n,1) deptime'];
Bsort=sortrows(B,2);                % sort jumps in order
```

[4] R. Gaigalas and I. Kaj; http://www.mathworks.com/mathlabcentral/fileexchange/loadFile.do?objectId=2494

```
jumps=Bsort(:,1);
jumptimes=[0;Bsort(:,2)];
systsize=[0;cumsum(jumps)];          % size of M/G/1 queue
systtime=deptime-arrtime';            % system times

figure(1)
stairs(jumptimes,systsize);
xmax=max(systsize)+5;
axis([0 tmax 0 xmax]);
grid

figure(2)
hist(systtime,20);


function [jumptimes, systsize] = simmginfty(tmax, lambda)
% SIMMGINFTY simulate a M/G/infinity queueing system. Arrivals are
% a homogeneous Poisson process of intensity lambda. Service times
% Pareto distributed (can be modified).
%
% [jumptimes, systsize] = simmginfty(tmax, lambda)
%
% Inputs: tmax - simulation interval
%         lambda - arrival intensity
%
% Outputs: jumptimes - times of state changes in the system
%      systsize - number of customers in system
%

  % set default parameter values if ommited
  if (nargin==0)
    tmax=1500;
    lambda=1;
  end

  % generate Poisson arrivals
  % the number of points is Poisson-distributed
  npoints = poissrnd(lambda*tmax);

  % conditioned that number of points is N,
  % the points are uniformly distributed
  if (npoints>0)
    arrt = sort(rand(npoints, 1)*tmax);
  else
    arrt = [];
```

```
  end

% uncomment if not available POISSONRND
% generate Poisson arrivals
% arrt=-log(rand)/lambda;
% i=1;
%   while (min(arrt(i,:))<=tmax)
%     arrt = [arrt; arrt(i, :)-log(rand)/lambda];
%     i=i+1;
%   end
% npoints=length(arrt);        % arrival times t_1,...,t_n

%  servt=50.*rand(n,1);       % uniform service times s_1,...,s_k

  alpha = 1.5;                   % Pareto service times
  servt = rand^(-1/(alpha-1))-1; % stationary renewal process
  servt = [servt; rand(npoints-1,1).^(-1/alpha)-1];
  servt = 10.*servt;             % arbitrary choice of mean

  dept = arrt+servt;            % departure times

  % Output is system size process N.
  B = [ones(npoints, 1) arrt; -ones(npoints, 1) dept];
  Bsort = sortrows(B, 2);                 % sort jumps in order
  jumps = Bsort(:, 1);
  jumptimes = [0; Bsort(:, 2)];
  systsize = [0; cumsum(jumps)];       % M/G/infinity system size
                                       % process

  stairs(jumptimes, systsize);
  xmax = max(systsize)+5;
  axis([0 tmax 0 xmax]);
  grid
%
```

14.2 Other Tools for Simulating and Analyzing Queueing Systems

There are many tools that permit modeling of systems such as queueing networks. They come with different user interfaces from simple web calculators and application programming interfaces (APIs) to sophisticated software packages with graphical user interfaces (GUIs). A semiofficial list of available software tools is maintained at http://web2.uwindsor.ca/math/hlynka/qsoft.html. While the list is long, many of the tools only provide very simple analysis capa-

bilities. This section will introduce a few of the more user-friendly tools and show simple examples to illustrate how to use them. MATLAB is widely used for a variety of numerical analyses including analysis of queueing networks. However, MATLAB is a commercial product and must be purchased. On the other hand, open source software tools allow them to be easily adapted to model computer and communication systems. Some examples of open source tools include Java modeling tools (JMT) and qnetworks (The Octave Queueing Toolbox). We provide examples to show how these tools can be used to analyze systems modeled as simple queueing networks.

The JMT suite comes with a simple user interface and permits a choice of six different tools:

(1) JMVA : exact and approximate solution

(2) JSIMgraph : graphical simulation

(3) JSIMwiz : textual simulation

(4) JMCH : Markov chain

(5) JABA : asymptotic bound analysis

(6) JWAT : workload analyzer tool

In each of the tool suites, modelers can provide necessary data in designated columns to create the queueing model, and proceed to solve it with what-if analysis options. The performance indices are displayed graphically. The output can also be saved in Extensible Markup Language (XML) format for analysis by other tools.

Gnu Octave is an open source version of MATLAB with equivalent capabilities. The qnetworks package provides the queueing network capabilities when using the computation environment. Classes of queueing problems are categorized with designated prefix names and handled with API functions. With the script capabilities in Octave, what-if analysis can be flexibly conducted during modeling. The modeling script and the solution results can be saved with plain text and can be easily manipulated to accommodate modeling processes.

Both JMT and qnetworks can be used to model either open or closed networks, and they allow mean value analysis (MVA) and convolution techniques as well as several approximate algorithms discussed in previous chapters. These tools can be used to obtain many of the performance indices discussed in this book, including queue lengths, response times, and throughput.

To illustrate how to use these tools, consider a simple computer system with a single central processing unit (CPU) and three file I/O servers or disks (FS1, FS2, FS3). Each job visits the three file servers 10, 20, and 30 times respectively and the service times of CPU and the file servers are 1, 2, 3, and 4 seconds. We model this system as a closed network.

In JMT, using JSIMgraph, the example can be described as a queueing network shown in Figure 14.1. We show the results of analysis by varying the number of customers between 1 and 10.

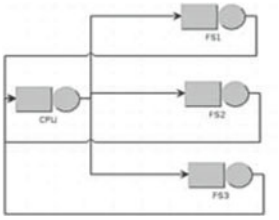


Figure 14.1 Closed network model

The following figures show the results for response times, utilizations, and throughputs of the four devices (Figures 14.2, 14.3, 14.4, 14.5).

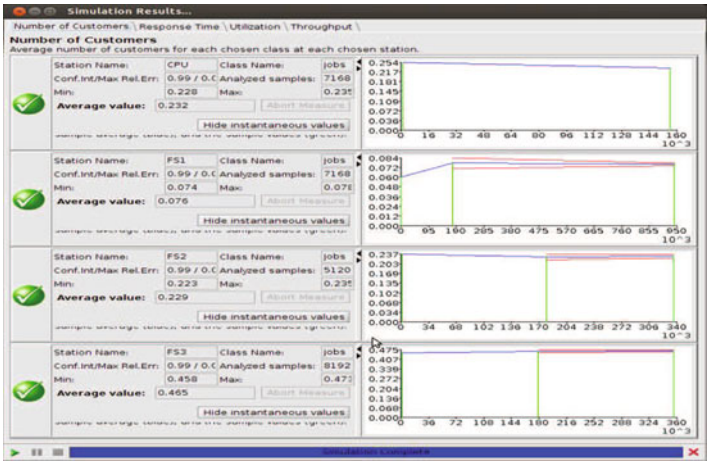


Figure 14.2 Results of analysis

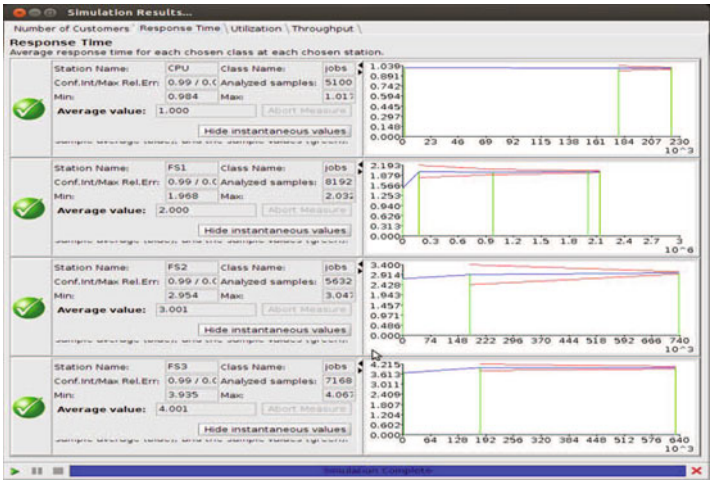


Figure 14.3 Results of analysis

Figure 14.4 Results of analysis

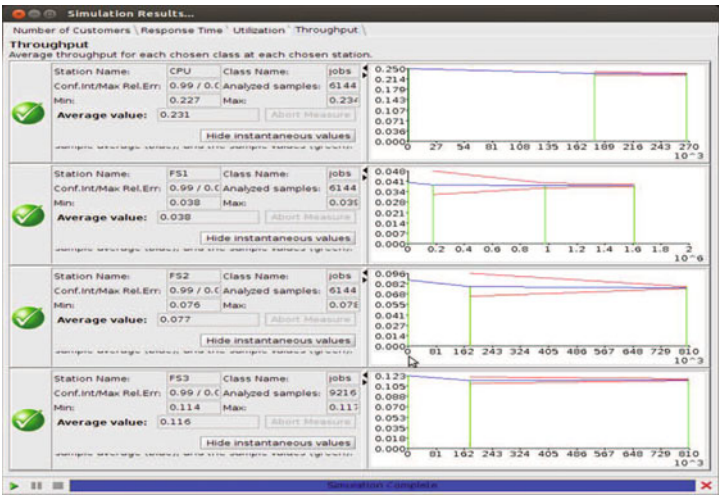


Figure 14.5 Results of analysis

The same example can be modeled in qnetwork in Octave with the script file:

```
pkg load queueing;
S=[1 2 3 4];
V=[1 10 20 30];
for N=1:10
     N=8;
     [U R Q X]=qncsmva(N,S,V);
     X_s=X(1)/V(1);
     R_s=dot(R,V);
     printf("N=%d\n",N)
     printf("\t\tUtil       Qlen       RespT      Tput     \n");
 printf("\t\t--------  --------  --------  --------\n");
for k=1:length(S)
    printf("Station%d\t%8.4f  %8.4f  %8.4f  %8.4f\n",
    k,U(k),R(k),Q(k),X(k));
endfor
printf("\nSystem rt: %8.4f  th: %8.4f\n", R_s, X_s);
endfor
```

The result can be output as plain text: (Station 1 – 4 represents: CPU, FS1, FS2, and FS3 respectively): N=1

	Util	Qlen	RespT	Tput
Station1	0.0050	1.0000	0.0050	0.0050
Station2	0.0995	2.0000	0.0995	0.0498
Station3	0.2985	3.0000	0.2985	0.0995
Station4	0.5970	4.0000	0.5970	0.1493

System rt: 201.0000 th: 0.0050

		Util	Qlen	RespT	Tput
	Station1	0.0068	1.0050	0.0069	0.0068
N=2	Station2	0.1367	2.1990	0.1503	0.0684
	Station3	0.4102	3.8955	0.5326	0.1367
	Station4	0.8204	6.3881	1.3102	0.2051

System rt: 292.5473 th: 0.0068

		Util	Qlen	RespT	Tput
	Station1	0.0076	1.0069	0.0077	0.0076
N=3	Station2	0.1526	2.3007	0.1755	0.0763
	Station3	0.4578	4.5979	0.7016	0.1526
	Station4	0.9156	9.2406	2.1151	0.2289

System rt: 393.1908 th: 0.0076

		Util	Qlen	RespT	Tput
	Station1	0.0080	1.0077	0.0081	0.0080
N=4	Station2	0.1599	2.3511	0.1879	0.0799
	Station3	0.4796	5.1049	0.8161	0.1599
	Station4	0.9592	12.4606	2.9879	0.2398

System rt: 500.4342 th: 0.0080

		Util	Qlen	RespT	Tput
	Station1	0.0082	1.0081	0.0082	0.0082
N=5	Station2	0.1633	2.3758	0.1940	0.0817
	Station3	0.4900	5.4482	0.8898	0.1633
	Station4	0.9799	15.9518	3.9079	0.2450

System rt: 612.2848 th: 0.0082

		Util	Qlen	RespT	Tput
	Station1	0.0083	1.0082	0.0083	0.0083
N=6	Station2	0.1650	2.3880	0.1970	0.0825
	Station3	0.4950	5.6695	0.9355	0.1650
	Station4	0.9901	19.6317	4.8591	0.2475

System rt: 727.2298 th: 0.0083
N=7

	Util	Qlen	RespT	Tput
Station1	0.0083	1.0083	0.0084	0.0083
Station2	0.1658	2.3940	0.1985	0.0829
Station3	0.4975	5.8065	0.9630	0.1658
Station4	0.9951	23.4366	5.8302	0.2488

System rt: 844.1768 th: 0.0083
N=8

	Util	Qlen	RespT	Tput
Station1	0.0083	1.0084	0.0084	0.0083
Station2	0.1663	2.3970	0.1993	0.0831
Station3	0.4988	5.8889	0.9791	0.1663
Station4	0.9975	27.3206	6.8133	0.2494

System rt: 962.3752 th: 0.0083
N=9

	Util	Qlen	RespT	Tput
Station1	0.0083	1.0084	0.0084	0.0083
Station2	0.1665	2.3985	0.1996	0.0832
Station3	0.4994	5.9372	0.9883	0.1665
Station4	0.9988	31.2532	7.8037	0.2497

System rt: 1081.3328 th: 0.0083
N=10

	Util	Qlen	RespT	Tput
Station1	0.0083	1.0084	0.0084	0.0083
Station2	0.1666	2.3993	0.1998	0.0833
Station3	0.4997	5.9649	0.9935	0.1666
Station4	0.9994	35.2147	8.7982	0.2498

System rt: 1200.7395 th: 0.0083

14.3 Remarks

Interested readers should consult MATLAB manuals for details on how to model and simulate computer and communication systems. There are many software tools that can be used to analyze systems modeled as queueing networks.

Most of the textbooks listed in the previous chapter also contain information on how discrete event simulations can be created.

14.4 Exercises

1. Write a simulation program to simulate a traffic intersection with a north-south street crossing an east-west street. You should also permit left turn at the intersections. All cars waiting for a green light will proceed immediately when the light turns green. For safety reason, once a light turns green it will remain green for t_1 seconds. Unless cars are waiting on cross street, once a signal turns green it will stay green. Assume that cars arrive at the intersection as a Poisson process with the mean arrival rate of λ. If there are no cars in the left-turn lane, no turn signal appears.

   Generate a random number indicating how many cars arrive at the intersection from each direction and if a car is requesting left-turn signal or not.

   Simulate the intersection using different values for t_1 and λ. Calculate the average waiting time at the intersection, that is, the interval from when a car arrives at the intersection and until when the light turns green allowing the car to exit the intersection. Note that this time can be zero.

   Using the statistical data, can you derive an empirical relationship between t_1, λ and the average waiting times?

2. Repeat the simulation in Exercise 1 using different arrival rates for each direction of travel.

3. Consider a multithreaded computer system that uses the following model to execute programs. Each thread consists of three phases: preload, execute, and post-store. A memory processor (MP) executes preload and post-store phases, providing access to memory resident data. An execute processor (EP) provides service during execute phase. New threads are enabled as some threads complete their execution and supply data to waiting threads (modeled as Synch service). Consider the following queueing network as a model of the system.

   The service time of MP is based on the average number of load and store instructions, while the service time of EP is based on the average number of non-memory instructions. The service time at Synch depends on the average number of inputs needed by a thread (and provided by other threads).

   Explore the response time of such a system for different number of threads (jobs in the system), by varying the service times at each server. You should examine any available benchmarks to estimate the various service times. Assume that each thread visits EP and Synch once, and visits MP twice. For example, you can try with these service times: MP = 3; EP = 10, Sync = 6 (the time unit is one instruction cycle, and using a 1 GHz processor, the time unit is a nanosecond). Using N = 10, 20, 50 threads, perform MVA for this exercise.

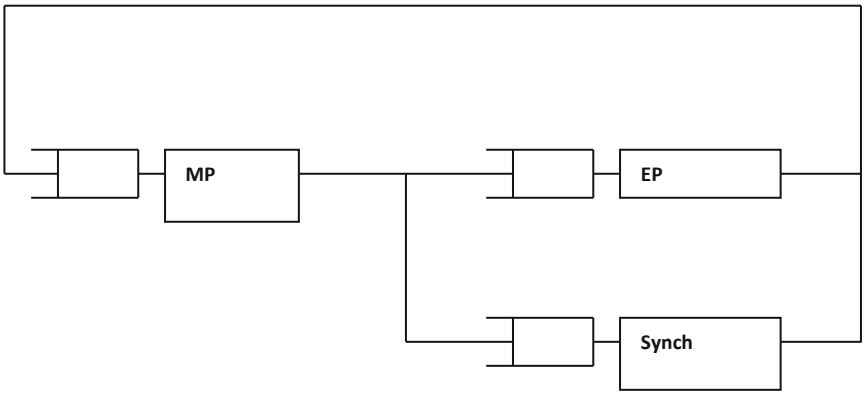


Figure 14.6 Queueing network model

4. Using convolution algorithm, find the expected queue lengths at each processing element of the computer system described in Exercise 3.

5. In real-time systems, it is necessary to assure that jobs (or tasks) complete before a specified deadline, otherwise the task is considered to have failed. A well-known algorithm used for such systems is called the EDF. As the name implies, tasks are scheduled based on their deadlines. In this exercise, you are asked to simulate an EDF-based system. You need to generate tasks using an arrival process, task service times, and deadlines. Note that the deadlines should be greater than task arrival time plus its service time. Once a job is created, the waiting list of jobs will be sorted based on the task deadlines. A task is scheduled only if it can meet its deadline. A performance measure of EDF is the percentage of jobs that meet their deadlines, known as the success ratio of EDF. It has been shown that when the system is heavily loaded (the sum of execution times of all waiting jobs exceeds the total system or simulation time) the success ratio of EDF drops dramatically. Explore the relationship between system utilization and the tightness of deadlines on the success ratios. It has been shown that if the utilization is low, EDF achieves high success ratios; but when the utilization is high (or the system is heavily loaded), the success ratio drops. Verify this claim using your simulations.

6. Another variation of EDF used in real-time systems is known as the least-laxity first (LLF) algorithm. Defining laxity as the deadline of a task minus its execution time, the job with smallest laxity is scheduled first. Repeat the experiment of Exercise 5 with LLF and compare its success ratio with that of EDF.

7. A commonly scheduling method to increase throughput of a system is known as the SJF. As the name implies, tasks are scheduled based on

their execution times; schedule jobs with smaller execution times. As in the previous exercise, create jobs using an arrival time and service time. Compute average response times using SJF and FIFO. SJF unfairly delays large jobs (jobs with large service times). Check how this unfairness affects the average response time of SJF when compared to a FIFO scheduling method.

APPENDIX A
Poisson Process Properties and Related Distributions

The Poisson process and exponential distribution occupy an important place in the modeling and analysis of queueing systems. This Appendix provides properties additional to these given in Section 2.1 of Chapter 2 and related distributions that are often used in applications. Distributions other than those mentioned here but that are some times used in applications can be found in standard texts in statistics.

A.1 Properties of the Poisson Process

(**a**) In reliability theory it is common to identify the failure rate of a component as an instantaneous failure rate, called the *hazard rate*, say $h(t)$. With $f(t)$ as the probability density of the life distribution of a component, $h(t)$ is defined as

$$h(t) = \frac{f(t)}{1 - F(t)}. \tag{A.1.1}$$

It should be noted that probabilistically $f(t)dt$ is approximately the probability that the component fails during $(t, t + dt]$ and $1 - F(t)$ is the probability that it is at least of age t. Thus $h(t)dt$ represents the approximate probability that the component fails during $(t, t + dt]$, given that it is of age t. Hence it is also called the instantaneous failure rate, or simply the *failure rate*.

When $f(x)$ is exponential with parameter λ and given as $f(x) = \lambda e^{-\lambda x}$ $(x > 0)$, then $h(t) = \lambda$, a constant.

(**b**) Let $Z_1, Z_2, \ldots$ be the random variables representing the interoccurrence times of a Poisson process. Apart from the fact that $\{Z_n,\ n = 1, 2, \ldots\}$ have exponential distributions, it can also be shown that they are independent and identically distributed.

© Springer Science+Business Media New York 2015

U. N. Bhat, *An Introduction to Queueing Theory*, Statistics for Industry and Technology, DOI 10.1007/978-0-8176-8421-1

(**c**) The memoryless property of the exponential distribution and properties (a) and (b) above imply that the Poisson process has the following characteristics.

- Events occurring in nonoverlapping intervals of time are independent of each other.
- There is a constant λ such that the probabilities of occurrence of events in a small interval of length Δt can be given as follows:

$$P\{\text{Number of events of occurring in } (t, t+\Delta t] = 0\} = 1 - \lambda\Delta t + o(\Delta t)$$

$$P\{\text{Number of events occurring in } (t, t+\Delta t] = 1\} = \lambda\Delta t + o(\Delta t)$$

$$P\{\text{Number of events occurring in } (t, t+\Delta t] > 1\} = o(\Delta t)$$

where $o(\Delta t)$ is such that $o(\Delta t)/\Delta t \to 0$ as $\Delta t \to 0$.

With these properties, λ is the mean number of events occurring per unit time.

(**d**) Consider two independent exponential random variables X_1 and X_2 with parameters λ_1 and λ_2, respectively. Then, we have

$$P(X_1 < X_2) = \int_{x=0}^{\infty} P(X_1 < X_2 | X_2 = x) f_2(x) dx.$$

We get

$$\begin{aligned} P(X_1 < X_2) &= \int_{x=0}^{\infty} P(X_1 < x) f_2(x) dx \\ &= \int_{x=0}^{\infty} (1 - e^{-\lambda_1 x}) \lambda_2 e^{-\lambda_2 x} dx \\ &= \frac{\lambda_1}{\lambda_1 + \lambda_2} \end{aligned} \tag{A.1.2}$$

(Note that $f_1(x)$ and $f_2(x)$ are the density functions of the random variables X_1 and X_2, respectively.)

Thus, if two types of Poisson events occur independently of each other, the probability that the first type of event occurs before the second is given by (A.1.2).

(**e**) The additive property of the Poisson distribution carries through to the Poisson process as well. Let $X_1(t)$ and $X_2(t)$ be two Poisson processes with parameters λ_1 and λ_2, respectively. Let $X(t) = X_1(t) + X_2(t)$. For $t \geq 0$, let

$$P[X_1(t) = n_1] = e^{-\lambda_1 t} \frac{(\lambda_1 t)^{n_1}}{n_1!}$$

$$P[X_2(t) = n_2] = e^{-\lambda_2 t} \frac{(\lambda_2 t)^{n_2}}{n_2!}.$$

Using these results and writing $n_1 + n_2 = n$ we can show that $X(t)$ is also Poisson for $t \geq 0$:

$$\begin{aligned} P\left[X(t) = n\right] &= \sum_{n_2=0}^{n} P\left[X_1(t) = n - n_2\right] P\left[X_2(t) = n_2\right] \\ &= e^{-(\lambda_1+\lambda_2)t} \frac{[(\lambda_1 + \lambda_2)t]^n}{n!}. \end{aligned} \tag{A.1.3}$$

Clearly, this property can be extended to any number of Poisson processes.

(**f**) Another useful property of the Poisson process is its relationship to the uniform distribution. Let n Poisson events occur at epochs $t_1 < t_2 < t_3 < \ldots < t_n$ in the interval $[0, T]$. Then the random variables $t_1, t_2, \ldots, t_n$ have the same distribution as the n order statistics corresponding to the independent random variables $U_1, U_2, \ldots, U_n$, uniformly distributed in the interval $[0, T]$. If $f_{t_1,t_2,\ldots,t_n}(x_1, \ldots, x_n)$ is the joint probability density function of $t_1, t_2, \ldots, t_n$, this property shows that

$$f_{t_1,t_2,\ldots,t_n}(x_1, x_2, \ldots, x_n) = \frac{n!}{T^n} \quad 0 \leq x_1 \leq x_2 \leq \ldots \leq x_n \leq T. \tag{A.1.4}$$

A.2 Variants of the Poisson Process

The Poisson process assumes that the events occur one at a time. But in real systems the occurrence of arrivals and service in groups is not uncommon. To accommodate such situations we may assume that each Poisson event spawns a group of subevents. If arrival is the event, customers in the group are the subevents. Using this terminology for convenience, let arrivals occur in a Poisson process with rate of occurrence λ, and assume that the nth arrival epoch brings in G_n customers, where $\{G_n, n = 1, 2, ...\}$ are independent and identically distributed (i.i.d.) random variables. Let

$$\Pr(G_n = r) = g_r.$$

Then the probability distribution of $X(t)$, the number of customers arriving during $(0, t]$, is given by

$$P_n(t) = \sum_{r=0}^{n} e^{-\lambda t} \frac{(\lambda t)^r}{r!} g_n^{(r)} \tag{A.2.1}$$

where $g_n^{(r)}$ is the r-fold convolution of g_n with itself, with $g_n^{(0)} = 1$ if $n = 0$ and 0 otherwise. Let $\gamma(z)$ be the PGF of $\{G_n\}_{n=1}^{\infty}$; then we get

$$\Pi(z,t) = \sum_{n=0}^{\infty} z^n P_n(t) = e^{-\lambda(1-\gamma(z))t}$$

and

$$E[X(t)] = \lambda\gamma'(1)t$$

where $\gamma'(1)$ is the mean size of the arriving group.

When the arriving groups consist of continuous units that can be represented by continuous random variables with distribution function $H(x)$, let $X(t)$ be the number of such arrivals and $Y(t)$ be the resultant input (e.g., number of claims $X(t)$ and the total amount of claims $Y(t)$ in insurance risk theory). Then, we get

$$Pr[X(t) = n, Y(t) \le x] = e^{-\lambda t}\frac{(\lambda t)^n}{n!}H_n(x) \tag{A.2.2}$$

and

$$\sum_{n=0}^{\infty} z^n \int_0^{\infty} e^{-\theta t} d_x Pr[X(t) = n, Y(t) \le x] = e^{-\lambda(1-z\eta(\theta))t}$$

where we have used $\eta(\theta)$ to represent the Laplace–Stieltjes transform of $H(x)$, and $H_n(x)$ for the n-fold convolution of $H(x)$ with itself. Clearly we get

$$E[Y(t)] = [-\eta'(0)]\lambda t$$

where $-\eta'(0)$ is the mean input per arrival.

The processes given in (A.2.1) and (A.2.2) are known as *compound Poisson processes* (also known as *stuttering Poisson processes*). These turn out to be good approximating models for a wide variety of arrival processes (See Haight (1967) and Johnson and Kotz (1969)).

Another class of Poisson related processes can be generated by assuming that the Poisson parameter λ itself is a random variable (Λ). Let $L(x)$ be its distribution function. Then $X(t)$, the number of arrivals occuring in $(0,t]$ can be given as

$$P_n(t) = P[X(t) = n] = \int_0^{\infty} e^{-\lambda t}\frac{(\lambda t)^n}{n!}dL(\lambda). \tag{A.2.3}$$

When the range of Λ is other than $(0,\infty)$, suitable range is to be used for integration. From (A.2.3), we get

$$E[X(t)] = tE[\Lambda].$$

For instance, when Λ has the Erlang distribution (see (A.4.1))

$$dL(\lambda) = e^{-\mu\lambda}\frac{\mu^k\lambda^{k-1}}{(k-1)!}d\lambda \quad (0 \le \lambda < \infty).$$

We get

$$P_n(t) = \binom{n+k-1}{k-1}\left(\frac{\mu}{t+\mu}\right)^k\left(\frac{t}{t+\mu}\right)^n, \quad n = 0, 1, 2\ldots \tag{A.2.4}$$

which is in a negative binomial form. The underlying process is called a Polya process.

The Polya process belongs to the class of mixed Poisson distributions which can be used to represent variations in the arrival or service intensity. Other useful mixing patterns would be to assume $L(x)$ as normal in the positive range or a discrete distribution of the type,

$$P(\Lambda = \lambda) = p_\lambda \quad \lambda = \lambda_1, \lambda_2, \ldots$$

A.3 Hyperexponential Distribution (HE)

Let random variables $\{Z_1, Z_2, \ldots\}$ be distributed as

$$\begin{aligned} F(x) &= 1 - \sum_{i=1}^{K} p_i e^{-\lambda_i x} \qquad 0 \le x < \infty \\ \lambda_i &> 0 \qquad\qquad \text{for all } i \text{ for which } \ p_i > 0; \\ 1 \ge p_i &\ge 0, \qquad\qquad \sum_{1}^{K} p_i = 1. \end{aligned} \tag{A.3.1}$$

We get

$$E(Z_n) = \sum_{1}^{K}(p_i/\lambda_i)$$

and its Laplace transform

$$\psi(\theta) = \sum_{1}^{K} p_i\left(\frac{\lambda_i}{\theta+\lambda_i}\right). \tag{A.3.2}$$

Also

$$E[Z_n^2] = \sum_{i=1}^{K} 2p_i/\lambda_i^2 \quad \text{and} \quad CV(Z_n) = \left[\frac{2\sum_{i=1}^{K} p_i/\lambda_i^2}{(\sum p_i/\lambda_i)^2} - 1\right]^{1/2}.$$

This distribution is generated if events fall into identifiable classes $(1, 2, ..., k)$ and an event belonging to class i arrives with probability p_i with an interoccurrence time that is exponential with mean $1/\lambda_i$. Depending on the values of p_i, λ_i and the possible values of i, a wide variety of distributions can be generated.

In order to retain the same mean $1/\lambda$ the following form of the HE distribution can be used.

$$F(x) = 1 - \sum_{i=1}^{K} p_i e^{-Kp_i\lambda x} \quad (x \geq 0) \tag{A.3.3}$$

$$\lambda > 0,\ 1 \geq p_i \geq 0, \quad \sum_{1}^{K} p_i = 1.$$

Then

$$\begin{aligned} E[Z_n] &= 1/\lambda \\ E[Z_n^2] &= \frac{2}{(K\lambda)^2} \sum_{i=1}^{K} (1/p_i) \\ CV[Z_n] &= \left[\left(\sum_{i=1}^{K} 1/p_i \right) \frac{2}{K^2} - 1 \right]^{1/2}. \end{aligned}$$

The value of K commonly used in applications is 2.

A.4 Erlang Distribution (E_k)

This distribution has been introduced in Chapter 2 (see (2.1.6).

Let random variables $\{Z_1, Z_2, \ldots\}$ be distributed as

$$\begin{aligned} F(x) &= \int_0^x e^{-\lambda y} \frac{\lambda^k y^{k-1}}{(k-1)!} dy \quad 0 \leq x \leq \infty, \quad \lambda > 0 \\ &= 1 - \sum_{r=0}^{k-1} e^{-\lambda x} \frac{(\lambda x)^r}{r!}. \end{aligned} \tag{A.4.1}$$

We get

$$E[Z_n] = k/\lambda$$

$$\psi(\theta) = \left(\frac{\lambda}{\theta + \lambda} \right)^k$$

and

$$E[Z_n^2] = (1 + 1/k)\frac{k^2}{\lambda^2} \text{ and } CV[Z_n] = \frac{1}{\sqrt{k}}.$$

The distribution $F(x)$ is a two parameter distribution, and is commonly known as the Erlang distribution (A. K. Erlang demonstrated its use in the analysis of telephone system congestion), or the gamma distribution, or the Pearson type III distribution with integral values for the parameter k. (It is also a particular case of the χ^2 distribution.)

A.5 Mixed Erlang Distributions

The HE distribution of Section A.3 above is obtained by using a finite mixture of exponential distributions. In a similar manner, in order to provide versatility, we can get mixed Erlang distributions.

(a) **Constant λ; varying k** $(k_1, k_2, ..., k_N)$. Let

$$F(x) = \int_0^x \sum_{i=1}^{N} p_i e^{-k_i \lambda y} \frac{(k_i\lambda)^{k_i} y^{k_i - 1}}{(k_i - 1)!} dy \quad 0 \le x < \infty,\ \lambda > 0. \tag{A.5.1}$$

We get

$$\begin{aligned} E[Z_n] &= \frac{1}{\lambda} \\ \psi(\theta) &= \sum_{i=1}^{N} p_i \left(\frac{k_i \lambda}{\theta + k_i \lambda} \right)^{k_i} \end{aligned} \tag{A.5.2}$$

and

$$\begin{aligned} E[Z_n^2] &= \frac{1}{\lambda^2} \sum_{i=1}^{N} p_i (1 + 1/k_i) \\ CV(Z_n) &= \left[\sum_{i=1}^{N} (p_i / k_i) \right]^{1/2}. \end{aligned}$$

This adds another dimension of generality to the Erlang distribution. It has been shown by several authors that this distribution approximates very well nearly all distributions of practical interest. A finite limit for the values of N has also been found satisfactory. (See Luchak (1956).)

(b) **Both λ and k varying**.

$$F(x) = \int_0^x \sum_{i=1}^{N} p_i e^{-k_i \lambda_i y} \frac{(k_i \lambda_i)^{k_i} y^{k_i - 1}}{(k_i - 1)!} dy. \tag{A.5.3}$$

This general form admits both the hyperexponential and Erlang distribution as special cases.

$$\begin{aligned} \text{HE}: \ k_i &= 1 \quad \text{for } i = 1, 2, \ldots, N \\ \text{Mixed Erlang}: \lambda_i &= \lambda \quad \text{for } i = 1, 2, \ldots, N. \end{aligned}$$

Assuming the coefficient of variation as a measure providing an adequate representation of the model, Erlang (with CV ≤ 1) and HE (with CV ≥ 1) distributions offer a wide spectrum of choice for purposes of model selection. In the Erlang model, the CV is decreased by increasing the value of the parameter k and in the HE model with $N = 2$, CV is increased by moving p_1 and p_2 away from 1/2.

A.6 Coxian Distributions; Phase Type Distribution (PH)

Generalizing the Laplace transform of the Erlang distribution (A.4.1), Cox (1955) has proposed a class of distributions whose Laplace transforms are rational functions. A member of this class is the generalized Erlang which has the Laplace transform

$$\psi(\theta) = \Pi_{i=1}^{N}\left(\frac{\lambda_i}{\theta+\lambda_i}\right). \tag{A.6.1}$$

The corresponding distribution can be thought of as the distribution of time a process takes to pass through N phases $(X_1, X_2, \ldots, X_N)$ with X_i having an exponential distribution with mean $1/\lambda_i$. It is obtained as the convolution of N exponentials with parameters $\lambda_1, \lambda_2, \ldots, \lambda_N$. Another large subset of Coxian distributions is the phase-type distribution introduced by Neuts (1975, 1981), which can be considered to be a natural probabilistic generalization of the Erlang. The underlying process generating the distribution undergoes transitions on a Markov chain with an absorbing state. Further discussion of this distribution is given in the Chapter 8.

Using Coxian distributions in their generality in queueing models leads to highly complicated analytical expressions and requires the use of complex variables in their analysis. For instance, see Cohen (1969). In striking a balance, between generality and practical use, Neuts' phase-type distributions have found wide use because of their versatility in modeling leading to algorithmic solutions. (See Chapter 8.)

A.7 A General Distribution

Let $F(x)$ be a continuous distribution function, with probability density function $f(x)$. We have

$$f(x) = -\frac{d}{dx}[1-F(x)]. \tag{A.7.1}$$

Using the hazard function concept introduced in (A.1.1), we have

$$\begin{aligned} h(x) &= \frac{1}{1-F(x)}\left\{-\frac{d}{dx}[1-F(x)]\right\} \\ &= -\frac{d}{dx}\ln[1-F(x)]. \end{aligned} \tag{A.7.2}$$

Integrating we find

$$\begin{aligned} \int_0^x h(y)dy &= -\ln[1-F(x)] \\ F(x) &= 1-e^{-\int_0^x h(y)dy} \end{aligned}$$

and

$$f(x) = h(x)e^{-\int_0^x h(y)dy} \tag{A.7.3}$$

which is in a generalized exponential form and is very convenient for use in the study of queueing systems with arbitrary inter-arrival time and/or service time distributions.

A.8 Some Discrete Distributions

Let $0, \sigma, 2\sigma, 3\sigma, \ldots$ be discrete equidistant points along the time axis. We assume that events occur only at these time points. (Even when events occur at other points, we may think of a counter that registers the events only at these time points.) If the value of σ is small enough, the discrete time axis is a convenient base to represent most systems of practical interest. Furthermore in systems such as computer systems, time is discrete, and a discrete queueing system is the most natural outcome.

As before, let $Z_1, Z_2, \ldots$ be nonnegative (integer valued) random variables, representing the inter-occurrence times of events. Define

$$p_k = P(Z_n = k) \quad n = 1, 2, \ldots \qquad k = 0, 1, 2, \ldots$$

and

$$\phi(z) = \sum_k p_k z^k \quad |z| \leq 1$$

as the PGF of $\{p_k\}$.

Three discrete distributions which are analogs of exponential, Poisson, and Erlang distributions are given below.

(i) *The geometric distribution.* Let events occur one at a time independent of each other, and the probability that an event occurs at a time point be α and does not occur be $(1-\alpha)$. Let p_k be the probability that the event occurs at time point k for the first time. Then

$$p_k = \alpha(1-\alpha)^{k-1} \quad k = 1, 2, \ldots \tag{A.8.1}$$

We get,

$$E[Z] = \frac{1}{\alpha} \quad \text{and} \quad V[Z] = \frac{1-\alpha}{\alpha^2}$$

and

$$\phi(z) = \frac{\alpha z}{1-(1-\alpha)z}. \tag{A.8.2}$$

The distribution in (A.8.1) is called the *geometric distribution* and it gives the discrete version of the exponential.

(ii) *The binomial distribution.* Consider the distribution of $X(n\sigma)$, the number of events occurring in the interval $(0, n\sigma)$. Let $p_k(n) = P[X(n\sigma) = k]$. Then, using the properties of the binomial distribution,

$$p_k(n) = \binom{n}{k}\alpha^k(1-\alpha)^{n-k} \quad k = 0, 1, 2, \ldots, n. \tag{A.8.3}$$

We get

$$E[X(n\sigma)] = n\alpha \quad \text{and} \quad V[X(n\sigma)] = n\alpha(1-\alpha)$$

and

$$\phi_n(z) = \sum_{k=0}^{n} p_k(n)z^k = (1-\alpha+\alpha z)^n. \tag{A.8.4}$$

Clearly, (A.8.3) is the discrete analog of the Poisson distribution. (Recall the method of derivation of the Poisson distribution as a limit of the binomial in statistics texts.)

(iii) *The negative binomial distribution.* Let $p_k(n)$ be the probability that the event occurs for the kth time at time point n. This means that the event occurs $k-1$ times during $[0, (n-1)\sigma]$; this event has a binomial probability given in (A.8.3). Since the event has to occur at the nth time point, we have

$$\begin{aligned} p_k^{(n)} &= \binom{n-1}{k-1}\alpha^{k-1}(1-\alpha)^{n-k}\alpha \\ &= \binom{n-1}{k-1}\alpha^k(1-\alpha)^{n-k} \quad n = k, k+1, \ldots \end{aligned} \tag{A.8.5}$$

We get

$$E[Z] = \frac{k}{\alpha} \quad \text{and} \quad V[Z] = \frac{k(1-\alpha)}{\alpha^2}$$

and

$$\phi(z) = \left[\frac{\alpha z}{1-(1-\alpha)z}\right]^k. \tag{A.8.6}$$

As in the Erlang, this distribution is generated by counting every kth event as an effective event for the queueing system.

APPENDIX B
Markov Process

This appendix builds on the basic concepts introduced in Section 3.3 of the main text and provides additional background material that may be needed for further work on modeling and analysis of queueing systems. The notations used in the material given below are consistent with those used in Chapter 3.

B.1 Kolmogorov Equations

Let $\{X(t), t\epsilon T\}$ be a time homogeneous Markov process and (see (3.3.14))

$$P_{ij}(t) = P[X(t) = j | X(0) = i]. \tag{B.1.1}$$

There are two types of differential equations for the determination of $P_{ij}(t)$ in Markov processes. These are *forward Kolmogorov equations* and *backward Kolmogorow equations*. Forward Kolmogorov equations are the ones commonly used in applications because of their convenient structure, even though backward equations are considered to be more fundamental due to the nature of limiting properties used in their derivation. In order to derive these equations we proceed as follows. In a time-homogeneous Markov process, (3.3.8) of Chapter 3 representing the Chapman–Kolmogorov relation, can be written down as

$$P_{ij}(t+s) = \sum_{k\epsilon S} P_{ik}(t) P_{kj}(s).$$

Set $s = \Delta t$; then

$$P_{ij}(t+\Delta t) = \sum_{k\epsilon S} P_{ik}(t) P_{kj}(\Delta t).$$

Subtracting $P_{ij}(t)$ from both sides of the equation and dividing by Δt,

$$\begin{aligned} \frac{P_{ij}(t+\Delta t) - P_{ij}(t)}{\Delta t} &= \sum_{k \neq j} \frac{P_{ik}(t) P_{kj}(\Delta t)}{\Delta t} \\ &+ P_{ij}(t) \frac{P_{jj}(\Delta t) - 1}{\Delta t}. \end{aligned}$$

© Springer Science+Business Media New York 2015

U. N. Bhat, *An Introduction to Queueing Theory*, Statistics for Industry and Technology, DOI 10.1007/978-0-8176-8421-1

Let $\Delta t \to 0$; we get

$$P'_{ij}(t) = -\lambda_{jj} P_{ij}(t) + \sum_{k \neq j} \lambda_{kj} P_{ik}(t). \tag{B.1.2}$$

In deriving (B.1.2) we have used the definition of λ_{ij} given in (3.3.15) and (3.3.16). Equation (B.1.2) for $i, j \epsilon S$ is known as *forward Kolmogorov equation.*

In matrix notation we can write them as

$$\boldsymbol{P'(t) = P(t)A} \tag{B.1.3}$$

where $\boldsymbol{A}$ is given by (3.3.18).

The transition probability $P_{ij}(t)$ can be determined by solving these differential equations along with the boundary condition $\boldsymbol{P}(0) = I$.

Backward Kolmorov equations can be obtained in a similar manner, by starting with the relation

$$P_{ij}(\Delta t + t) = \sum_{k \epsilon S} P_{ik}(\Delta t) P_{kj}(t).$$

The corresponding matrix equation can be given as

$$\boldsymbol{P'}(t) = \boldsymbol{AP}(t).$$

Formally, the solution for both sets of equations can be given as

$$\boldsymbol{P}(t) = e^{\boldsymbol{A}t} = \boldsymbol{I} + \sum_{n=1}^{\infty} \boldsymbol{A}^n \frac{t^n}{n!}. \tag{B.1.4}$$

B.2 The Poisson Process

Here we show how forward Kolmogorov equations can be used to determine the transition probability distribution of the Poisson process. In Chapter 2, we have introduced events whose inter-occurrence times are exponential. In Section A.1 of Appendix A, we have also listed the following properties:

1. Events occurring in nonoverlapping intervals of time are independent of each other.

2. There is a constant λ such that the probabilities of occurrence of events in a small interval of length Δt are given as follows:

   (a) $P\{$ Number of events occurring in $(t, t + \Delta t] = 0\} = 1 - \lambda \Delta t + o(\Delta t)$

   (b) $P\{$Number of events occurring in $(t, t + \Delta t] = 1\} = \lambda \Delta t + o(\Delta t)$

   (c) $P\{$Number of events occurring in $(t, t + \Delta t] > 1\} = o(\Delta t)$

where $o(\Delta t)$ is such that $o(\Delta t)/\Delta t \to 0$ as $\Delta t \to 0$.

Using the notations and equations developed for Markov processes, in this context we have

$$P'_{ij}(0) = \lambda; \quad P'_{ii}(0) = -\lambda$$

resulting in the generator matrix

$$\mathbf{A} = \begin{bmatrix} -\lambda & \lambda & 0 & 0 & 0 & \dots \\ 0 & -\lambda & \lambda & 0 & 0 & \dots \\ 0 & 0 & -\lambda & \lambda & 0 & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \end{bmatrix}. \tag{B.2.1}$$

The Poisson process is a counting process whose initial value is 0, i.e., $X(0) = 0$. Writing $P_{0n}(t) = P_n(t)$ for convenience and noting that $\boldsymbol{P}(t) = (P_0(t), P_1(t), \dots)$ and $\boldsymbol{P}'(t) = (P'_0(t), P'_1(t), \dots)$ the individual equations in (B.1.2) can be written out as

$$\begin{aligned} P'_0(t) &= -\lambda P_0(t) \\ P'_n(t) &= -\lambda P_n(t) + \lambda P_{n-1}(t) \quad n > 0 \end{aligned}$$

with $P_0(0) = 1$ and $P_n(0) = 0$ for $n > 0$. Solving these differential equations we get

$$P_n(t) = e^{-\lambda t} \frac{(\lambda t)^n}{n!} \quad n = 0, 1, 2, \dots \tag{B.2.2}$$

which is the result we stated in (3.3.19). As we have seen in Chapters 4 and 6, solutions to equations in (B.1.2) are not always easily determined. In the case of the Poisson process, because of the bi-diagonal structure of $\boldsymbol{A}$ and the constant element λ, the differential equations could be solved using standard methods. When such simplifications are not available, in simpler cases we may use Laplace transforms and PGFs in their solutions. When $\boldsymbol{A}$ is finite and diagnoalizable the eigenvalue method can be used to determine the solution in the form (B.1.4). Also, there are computational methods to obtain solutions from the differential equations directly. (See Bailey (1964), Stewart (1994) and Bhat and Miller (2002).)

In the modeling of queueing systems, it helps to understand what the elements of matrix $\boldsymbol{A}$ of (3.3.18) stand for. As indicated earlier, by definition (see (3.3.15) and (3.3.16)), λ_{ij}, $j \neq i$, is the instantaneous rate for the transition $i \to j$. From (3.3.17) we also know that $\sum_{j \neq i} \lambda_{ij} = \lambda_{ii}$. That means λ_{ii} is also the sum of all the instantaneous transition rates out of state i.

This allows us to interpret $1/\lambda_{ii}$ as the mean length of time the process stays in state i during a visit. The length of time the process stays in a state during a visit is known as the *sojourn time* in that state. This sojourn time of the Markov process in state i has been shown to have an exponential distribution with mean $1/\lambda_{ii}$.

For a proof of this result and for a discussion of special forms of Markov processes used in stochastic modeling, the readers are referred to Bhat and Miller (2002) and advanced books cited in them.

B.3 Classification of States

In order to describe a stochastic process we need to specify the state space and the parameter space. The parameter space is easily categorized as being discrete or continuous. The state space, however, in addition to being discrete or continuous, may include states or groups of states with special properties.

The states of a discrete state stochastic process fall into groups depending on how they interact with each other. The basic property defining this interaction is communication. If state i can be reached from state j, i is said to be *accessible* from j. If i and j are accessible to each other, they are said to *communicate*. It is not hard to visualize all communicating states forming a single group, known as an *equivalence class*. If a Markov process has all its states belonging to a single equivalence class, it is said to be *irreducible*.

For instance, consider the number of customers, Q_n, in a queueing system, at discrete time points t_n, $n = 0, 1, 2, \ldots$. Assume that t_n are such that $\{Q_n,\ n = 0, 1, 2 \ldots\}$ can be modeled as a Markov chain. When no restrictions are imposed on the transitions of $\{Q_n\}$ it is easy to note that all states of the Markov chain communicate with each other and hence form a single equivalence class. Alternatively, we may think of a finite queueing system which ceases to operate when Q_n hits the value, say M. As an example, consider M machines that are in operation in a service facility. The facility ceases its operation when all machines become inoperative. Let the number of failed machines be the state of the process. Now the state M of the Markov chain is accessible from all other states $[0, 1, 2, \ldots, M-1]$; but other states are not accessible from M. Then we have two equivalence classes: $[M]$, and $[0, 1, 2, \ldots, M-1]$. Since the process stops in M, it is known as an *absorbing state*.

Suppose now, the system is modified such that the facility is not shut down when all M machines are inoperative. One or more of them are repaired to bring the facility back into operation. Now all states $[0, 1, 2, \ldots, M]$ belong to the same class. Comparing the states in the two systems, the first with an absorbing state and the second with all communicating states, we can make the following observation. The Markov chain starting from any one of the states in the class $[0, 1, \ldots, M-1]$ in the first system will not remain in any of these states when $n \to \infty$, because at some stage it is bound to get absorbed in M. On the other hand, the Markov chain of the second system will remain in the class $[0, 1, ..., M]$ even when $n \to \infty$. This behavior of the Markov chain allows us to classify the states, and the equivalence classes themselves, into being *recurrent* or *transient*.

(1) Starting from state i, if the Markov chain is certain to return to i, the state is said to be *recurrent*. Since all states in the equivalence class communicate with each other, the class itself is recurrent. A further classification is made based on the length of *recurrence time*, which is the mean time the process takes to return to the same state for the first time. If the recurrence time is finite, the state (and the class to which it belongs) is

known as *positive recurrent*. If it is infinite, the state and the class are known as *null recurrent*. Note that an absorbing state is recurrent.

(2) Starting from state i, if the Markov chain's return to that state is not certain, it is said to be *transient*. Since all states in the equivalence class communicate with each other, then the class itself is transient.

The classification of states of a stochastic process such as queue length (number of customers in the system) plays a major role in understanding its behavior. We give below some of the properties that can be deduced from the nature of the states of the process.

1. If there are transient states in the state space of the process, in the long run ($n \to \infty$), the process will not be found in those states. Thus, if there are transient as well as recurrent states in the state space, the process will always end up in the recurrent states.

2. A process starting out in a recurrent state i will always remain in the recurrent equivalence class to which state i belongs.

3. Because of Properties 1 and 2 above, only processes with irreducible Markov process models need to be considered to understand the long run behavior of the system. As we have seen in earlier chapters, we can establish conditions under which limiting distributions exist for such processes.

4. When the state space includes both transient and recurrent states, one of the characteristics of interest is the transition from the transient states to a state in the recurrent class. For instance the distribution properties of the busy period in a queueing system can be determined by considering 0 as an absorbing state for the queue length process, while all other states are transient.

For an elaboration on the classification of states and their usefulness in stochastic modeling, readers are referred to Bhat and Miller (2002).

APPENDIX C
Results from Mathematics

In this appendix[5], we present useful results from several areas of mathematics that have been used in the book. For further reading on these topics, references are given at the end of each section.

C.1 Reimann–Stieltjes Integral

Let $f(x)$ and $\phi(x)$ be real-valued functions on $[a, b]$, and suppose that $f(x)$ is bounded on $[a, b]$ and $\phi(x)$ is monotonically increasing there. By a partition $\boldsymbol{P}$ of $[a, b]$, we mean a finite sequence of points $x_0, x_1, \ldots, x_n$ such that

$$a = x_0 \leq x_1 \leq x_2 \leq \ldots \leq x_{n-1} \leq x_n = b.$$

For any partition $\boldsymbol{P}$ of the closed interval $[a, b]$, we define

$$\begin{aligned} N_1 &= \text{least upper bound } f(x) \quad \text{where } x \,\epsilon\, [x_{i-1}, x_i] \\ n_i &= \text{greatest lower bound } f(x) \quad \text{where } x \,\epsilon\, [x_{i-1}, x_i] \\ \Delta\phi_i &= \phi(x_i) - \phi(x_{i-1}) \\ U(\boldsymbol{P}, f, \phi) &= \sum_{i=1}^{n} N_i \Delta\phi_i \\ L(\boldsymbol{P}, f, \phi) &= \sum_{i=1}^{n} n_i \Delta\phi_i. \end{aligned}$$

[5]Reprinted with permission from John Wiley & Sons, publishers, from Bhat and Miller (2002).

© Springer Science+Business Media New York 2015

U. N. Bhat, *An Introduction to Queueing Theory*, Statistics for Industry and Technology, DOI 10.1007/978-0-8176-8421-1

Then

$$\int_a^{b^-} f(x)d\phi(x) = \text{greatest lower bound } U(\boldsymbol{P}, f, \phi)$$
$$\int_{a^-}^{b} f(x)d\phi(x) = \text{least upper bound } L(\boldsymbol{P}, f, \phi)$$

where greatest lower bound and least upper bound are taken over all partitions of $[a, b]$.

We say that $f(x)$ is Reimann–Stieltjes integrable with respect to $\phi(x)$ over $[a, b]$ if and only if

$$\int_{a^-}^{b} f(x)d\phi(x) = \int_a^{b^-} f(x)d\phi(x).$$

When $f(x)$ is Riemann–Stieltjes integrable with respect to $\phi(x)$ over $[a, b]$, we write its integral as

$$\int_a^b f(x)d\phi(x).$$

It should be pointed out that one may define the Riemann–Stieltjes integral with respect to a function $\phi(x)$, where $\phi(x)$ is of bounded variation on $[a, b]$. A function $\phi(x)$ is of bounded variation on $[a, b]$ iff

$$V(\phi; a, b) = \text{least upper bound} \sum_{i=1}^{n} |\Delta\phi_i| < +\infty$$

where the least upper bound is taken over all partitions of $[a, b]$ (Rudin 1964).

C.2 Laplace Transforms

The proofs of the properties have been omitted, and all operations are assumed to be well defined.

Definition 1. *Let $f(t)$ be a real-valued function in $[0, \infty)$. We define the Laplace transform of $f(t)$ as*

$$L\{f(t)\} = \phi(s) = \int_0^\infty e^{-st} f(t)dt, \quad \text{Re}(s) > 0.$$

If $f(t)$ is piecewise continuous on every interval $[0, N]$ and of exponential order α (i.e., there exist constants M_1, M_2, and α such that for all $t > M_2$, we have $|f(t)| < M_1 e^{\alpha t}$). Then it can be shown that $L\{f(t)\} = \phi(x)$ exists. In Section 1, we defined what is meant by the Riemann–Stieltjes integral; in turn, we may also define the Laplace–Stieltjes transform of $F(t)$.

Definition 2. *Let $F(t)$ be a real-valued function; then we define the Laplace–Stieltjes transform of $F(t)$ as*

$$\int_0^\infty e^{-st} dF(t), \quad \text{Re}(s) > 0.$$

We note that if $F(t)$ is absolutely continuous and its Laplace–Stieltjes transform exists, then $F(t)$ is a differentiable monotonically increasing function and

$$dF(t) = F'(t)dt.$$

The Laplace–Stieltjes transform of $F(t)$ then equals the Laplace transform for this case.

The following properties apply only to Laplace transforms, although analogous properties hold for Laplace–Stieltjes transforms. Let $L\{f(t)\} = \phi(s)$, and assume that all operations are well defined.

Property C.2.1.

(1) *If* $L\{f_i(t)\} = \phi_i(s)$ *and* $f(t) = \sum_{i=1}^{\infty} \xi_i f_i(t)$, *where* ξ_i *is a constant* ($i = 1, 2, \ldots$), *then* $\phi(s) = \sum_{i=1}^{\infty} \xi_i \phi_i(s)$. (2) *If* $g(t) = e^{\xi t} f(t)$, *then* $L\{g(t)\} = \phi(s - \xi)$. (3) *If*

$$g(t) = \begin{cases} f(t-\xi) & \text{for } t > \xi \\ 0 & \text{for } t \leq \xi \end{cases}$$

then $L\{g(t)\} = e^{-\xi s}\phi(s)$. (4) *If* $\xi \neq 0$ *and* $g(t) = f(\xi t)$, *then*

$$L\{f(\xi t)\} = \frac{1}{\xi}\phi\left(\frac{s}{\xi}\right).$$

(5) *If* $g(t) = d^n[f(t)]/dt^n = f^{(n)}(t)$, *then*

$$L\{g(t)\} = s^n\phi(s) - s^{n-1}f(0) - s^{n-2}f^{(1)}(0) - \ldots - sf^{n-2}(0) - f^{(n-1)}(0).$$

Here the continuity at 0 of $f^{(n)}t$ *is assumed for each n.* (6) *If* $g(t) = t^n f(t)$, *then* $L\{g(t)\} = (-1)^n \phi^{(n)}(s)$. (7) *When the indicated limit exists, we have*

$$\begin{aligned} \lim_{s\to\infty} \phi(s) &= 0 \\ \lim_{t\to 0} f(t) &= \lim_{s\to\infty} s\phi(s) \\ \lim_{t\to\infty} f(t) &= \lim_{s\to 0} s\phi(s). \end{aligned}$$

(8) *Let* $f(t)$ *be the probability density function of a continuous random variable* T; *then* $E(T) = -\phi^{(1)}(0)$. (9) *Let*

$$f(t) = f_1(t) * f_2(t) = \int_{\tau=0}^{t} f_1(\tau) f_2(t-\tau) d\tau$$

and $L\{f_i(t)\} = \phi_i(s) (i = 1, 2)$; *then* $\phi(s) = \phi_1(s) \cdot \phi_2(s)$. (10) *If* $f(t)$ *is such that*

$$\int_0^x f(t)dt = 0$$

for all $x > 0$, *then* $f(t)$ *is called* a null function *and* $\phi(s) = 0$.

Perhaps a word about the uniqueness of the Laplace transform of $f(t)$ is in order. Suppose $f_2(t)$ is a null function and $f(t) = f_1(t) + f_2(t)$; then by properties (1) and (10), we have

$$\phi(s) = \phi_1(s) + \phi_2(s) = \phi_1(s) = L\{f_1(t)\}.$$

One can see that several different functions may have the same Laplace transforms, but if we do not consider null functions, the Laplace transform of a function is unique.

Finally, we give two theorems that are useful in limiting operations dealing with transforms.

Theorem 1 (An Abelian Theorem) *If for some nonnegative number* $\xi(\geq 0)$ *we have*

$$\lim_{t\to\infty} \frac{F(t)}{t^{\xi}} = \frac{C}{\Gamma(\xi+1)}$$

and

$$\psi(s) = \int_0^{\infty} e^{-st} dF(t)$$

exists for $\mathrm{Re}(s) > 0$, *then*

$$\lim_{s\to 0^+} s^{\xi}\psi(s) = C.$$

Theorem 2 (A Tauberian Theorem) *If* $F(t)$ *is nondecreasing and*

$$\psi(s) = \int_0^{\infty} e^{-st} dF(t)$$

exists for $\mathrm{Re}(s) > 0$, *and if for some constant* $\xi(\geq 0)$

$$\lim_{s\to 0} s^{\xi}\psi(s) = C$$

then

$$\lim_{t\to\infty} \frac{F(t)}{t^{\xi}} = \frac{C}{\Gamma(\xi+1)}$$

(Widder 1946).

C.3 Generating Functions

Analogous to the transform of a function is the transform of a sequence of real numbers $\{a_n\}_{n=0}^{\infty}$. This is commonly called a Z-transform or generating function of $\{a_n\}_{n=0}^{\infty}$.

Definition 1. *Let* $\{a_n\}_{n=0}^{\infty}$ *be a sequence of real numbers. If*

$$A(z) = \sum_{n=0}^{\infty} a_n z^n$$

exists, then $A(z)$ *is called the* generating function *of the sequence* $\{a_n\}_{n=0}^{\infty}$.

Since the series $A(z)$ converges to a unique number, the generating function of a sequence of real numbers is unique. The similarity between generating functions and Laplace transforms is obvious and is further exemplified by the properties of generating functions. We again assume that all operations are well defined. Let the generating functions of $\{a_n\}_{n=0}^{\infty}$ and $\{b_n\}_{n=0}^{\infty}$ be $A(z)$ and $B(z)$, respectively.

Property C.3.1.

(1) *If* $c_n = \xi_1 a_n + \xi_2 b_n$ *for each* n, *where* ξ_1 *and* ξ_2 *are constants, then* $C(z) = \sum_{n=0}^{\infty} c_n z^n = \xi_1 A(z) + \xi_2 B(z)$. (2) *If* $b_n = a_{n+k}$, *then* $B(z) = z^{-k}A(z) - \sum_{r=0}^{k-1} b_r z^{r-k}$. (3) *If* $a_n = n^k$ *and* $b_n = n^{k-1}$ *for* $k \geq 1$, *then* $A(z) = zB^{(1)}(z) = zdB(z)/dz$. (4) *If* $c_n = \sum_{r=0}^{n} a_r b_{n-r}$, *then* $C(z) = \sum_{n=0}^{\infty} c_n z^n = A(z) \cdot B(z)$. (5) *Let* X *be a nonnegative, discrete random variable, and let* $P(X = n) = p_n$ *and* $P(X > n) = q_n$; *if* $P(z) = \sum_{n=0}^{\infty} p_n z^n$ *and* $Q(z) = \sum_{n=0}^{\infty} q_n z^n$, *then (a)* $Q(z) = [1 - P(z)]/(1 - z)$, *(b)* $E(X) = P^{(1)}(1) = Q(1)$, *(c)* $V(X) = P^{(2)}(1) + P^{(1)} - [P^{(1)}(1)]^2 = 2Q^{(1)}(1) + Q(1) - [Q(1)]^2$. *Note that when* $p_n = P(X = n)$, *we may write* $P(z) = E[z^X]$.

Finally, we give three theorems that are useful in analyzing stochastic systems.

Theorem 1 (Abel's Theorem) *If* $\lim_{n\to\infty} a_n = a$, *then*

$$\lim_{z\to 1^-} \left[(1 - z) \sum_{n=0}^{\infty} a_n z^n \right] = a.$$

Theorem 2 (Tauber's Theorem): *If* $\lim_{z\to 1^-} (1 - z) \sum_{n=0}^{\infty} a_n z^n = a$ *and* $lim_{n\to\infty} n(a_n - a_{(n-1)}) = 0$, *then*

$$\lim_{n\to\infty} a_n = a.$$

Theorem 3 *Let* $\{a_n\}_{n=0}^{\infty}$ *be a nonnegative sequence of real numbers whose generating function is*

$$A(z) = \sum_{n=0}^{\infty} a_n z_n, \quad |z| < 1.$$

The following hold (for a and c real constants):

$$\sum_{n=0}^{\infty} a_n = a \quad \text{iff} \quad \lim_{z \to 1^-} A(z) = a$$

$$\lim_{m \to \infty} \left(\frac{1}{m} \sum_{n=0}^{m} a_n \right) = c \quad \text{iff} \quad \lim_{z \to 1^-} [(1 - z)A(z)] = c$$

(Beightler et al. 1961; Feller 1968; Hardy 1949; Whittaker and Watson 1992).

References

[1] J. Abate and H. Dubner (1968), A new method of generating power series expansions of functions, *SIAM J. Numer. Anal.*, **5**, 102-112.

[2] J. Abate and W. Whitt (1992), The Fourier-series method for inverting transforms of probability distributions, *Queueing Systems*, **10**, 5-88.

[3] J. Abate, H. Dubner and S.B. Weinberg (1968), Queueing analysis for the BIM 2314 disk storage facility, *J. ACM.*, **15**, 577-589.

[4] S. K. Acharya (1999), On normal approximation for maximum likelihood estimation from single server queues, *Queueing Systems*, **31**, 207-216.

[5] I.J.B.F. Adan, O.J. Boxma and J.A.C. Resing (2001), Queueing models with multiple waiting lines, *Queueing Systems*, **37**, 65-98.

[6] A.S. Alfa (2003), Vacation models in discrete time, *Queueing Systems*, **44**, 5-30.

[7] A.O. Allen (1990), *Probability, Statistics, and Queueing Theory with Computer Science Applications, 2nd Ed.*, Academic Press, Inc. Boston.

[8] Y. Arda and J.C. Hennet (2006), Inventory control in a multi-supplier system, *International Journal of Production Economics*, **104**, 249-259.

[9] C. Armero (1994), Bayesian inference in Markovian queues, *Queueing Systems*, **15**, 419-426.

[10] C. Armero and D. Conesa (2000), Prediction in Markovian bulk arrival queues, *Queueing System* **34**, 327-350.

[11] J. Artalejo, A. Gomez-Corral, and Q. He (2010), Markovian arrivals in stochastic modelling: a survey and some new results. SORT, 34, 101-144.

[12] S. Asmussen and G. Kobe (1993), Marked point processes as limits of Markovian arrivals streams, *J. Appl. Prob.,* **30**, 365-372.

[13] L. Bai, B. Fralix, L. Liu and W. Shang (2004), Inter-departure times in base-stock inventory-queues, *Queueing Systems*, **47**, 345-361.

© Springer Science+Business Media New York 2015

U. N. Bhat, *An Introduction to Queueing Theory*, Statistics for Industry and Technology, DOI 10.1007/978-0-8176-8421-1

[14] N.T.J. Bailey (1952), Study of queues and appointment systems in outpatient departments with special reference to waiting times, *J. Roy. Stat. Soc.* B, **14**, 185-199.

[15] N.T.J. Bailey (1954), A continuous time treatment of a simple queue using generating functions 1, *J. Roy. Stat. Soc.* B, **16**, 288-291.

[16] N.T.J. Bailey (1964), *The Elements of Stochastic Processes with Applications to the Natural Sciences*, John Wiley & Sons, New York.

[17] K.R. Balachandran. (1973), Control policies for a single server system. *Management Sci.*, **19**, 1013-1018.

[18] I.V. Basawa and N.U. Prabhu (1981), Estimation in single server queues. *Nav. Res. Logist. Quart.*, **28**, 475-487.

[19] I.V. Basawa and N.U. Prabhu (1988), Large sample inference from single server queues. *Queueing Systems*, **3**, 289-304.

[20] I.V. Basawa, U.N. Bhat, and J. Zhou (2008), Parameter estimation using partial information with applications to queueing and related models, *Stat. and Prob. Letters*, 78, 1375-1383.

[21] F. Baskett, K.M. Chandy, R.R. Muntz, and F.G. Palacios (1975), Open, closed, and mixed networks of queues with different classes of customers, *J. ACM*, **22**, 248-260.

[22] M. Bäuerle (2002), Optimal control of queueing networks: An approach via fluid models, *Adv. Appl. Prob.*, **34**, 313-328.

[23] C.S. Beightler, L.G. Mitten, and G.L. Nemhauser (1961), A short table of Z-transforms and generating functions, *Operations Res.*, **9**, 547-578.

[24] F. Benson and D.R. Cox (1951, 1952), The productivity of machines requiring attention at random intervals, *J. Roy. Stat. Soc.* B, **13**, 65-82; **14**, 200-210.

[25] V. Bhaskar, I. Kattankulathur and P. Lallement (2010), Modeling a supply chain using a network of queues, *Applied Mathematical Modelling*, **34**, 2074-2088.

[26] U.N. Bhat (1968), *A Study of the Queueing Systems M/G/1 and GI/M/1.* Lecture notes in Operations Research and Mathematical Economics, No. 2, Springer-Verlag, New York.

[27] U.N. Bhat (1969), Sixty years of queueing theory, *Management Sci.*, **15**, B280-B294.

[28] U.N. Bhat (1984), *Elements of Applied Stochastic Processes*, 2nd ed. John Wiley & Sons, New York.

[29] U.N. Bhat (1987), A sequential technique for the control of traffic intensity in Markovian queues. *Annals of Operations Res.*, **8**, 151-164.

[30] U.N. Bhat. (2003), Parameter estimation in M/G/1 and GI/M/1 queues using queue length data, *Stochastic Point Processes*. (Eds. S. K. Srinivasan and A. Vijayakumar) Narosa Publ. New Delhi, 96-107.

[31] U.N. Bhat and I.V. Basawa (2002), Maximum likelihood estimation in queueing systems, *Advances on Methodological and Applied Aspects of Probability and Statistics*, (Ed. N. Balakrishnan), Taylor & Francis, N.Y., 13-29.

[32] U.N. Bhat and K.M. Kavi (1987), Reliability Models for computer systems: An overview including dataflow graphs. *Sadhana*, **11**, Parts 1 & 2, 167-186. *Reliability and Fault-tolerant Issues in Realtime Systems*. (Ed. N. Viswanadham), Indian Academy of Sciences, Bangalore, India.

[33] U.N. Bhat and G.K. Miller (2002), *Elements of Applied Stochastic Processes*, 3rd ed. John Wiley & Sons, New York.

[34] U.N. Bhat, G.K. Miller and S.S. Rao (1997), Statistical analysis of queueing systems, *Frontiers in Queueing* (Ed. J. H. Dshalalow), Ch. 13, CRC Press, New York, 351-393.

[35] U. N. Bhat, R. E. Nance, and R. R. Korfhage (1975), Information networks: A probabilistic model for hierarchical message transfer, *Inform. Sci.,* **9**, 169-184.

[36] U.N. Bhat, M. Shalaby, and M.J. Fischer (1979), Approximation techniques in the solution of queueing problems, *Nav. Res. Logist. Quart.*, **26**, 311-326.

[37] P. Billingsley (1961), *Statistical Inference for Markov Processes*. University of Chicago Press.

[38] D. Bini and B. Meini (1995), On cyclic reduction applied to a class of Toeplitz matrices arising in queueing problems, *Computations with Markov Chains*, (Ed., W. J. Stewart), Kluwer Academic Publisher, 21-38.

[39] A. Birnbaum (1954), Statistical methods for Poisson processes and exponential populations, *J. Amer. Stat. Assoc.*, **49**, 254-266.

[40] G.R. Bitran and S. Dasu (1992), A review of open queueing network models of manufacturing systems, *Queueing Systems*, **12**, 95-133.

[41] G.R. Bitran and R. Morabito (1996), Open queueing networks: optimization and performance evaluation models for discrete manufacturing systems, *Production and Operations Management* **5**, 163-193.

[42] G.R. Bitran and R. Morabito (1999), An overview of tradeoff curves in manufacturing systems design, *Production and Operations Management*, **8**, 56-75.

[43] G. Bolch, S. Greiner, H. de Meer, and K. Tviedi (2006), *Queueing network and Markov chains: Modeling and performance evaluation with computer science applications*, 2nd ed. Wiley Interscience, New Jersey.

[44] G. Boole (1970), *Calculus of Finite Differences*, 5th Ed. (Ed. J. F. Moulton), Chelsea Pub. Co., New York.

[45] R. J. Boucherie, X. Chao, and M. Miyazawa (2003), Arrival first queueing network with applications in kanban production systems, *Performance Evaluation*, **51**, 83-102.

[46] O.J. Boxma and H. Takagi (1992), Editorial introduction, *Queueing Systems*, **11** (Special issue on Polling Models), 1-5.

[47] G. Brigham(1955), On a congestion problem in an aircraft factory. *Operations Res.*, **3**, 412-428.

[48] E. Brockmeyer, H.L. Halstrom, and A. Jensen (1960), *The Life and Works of A. K. Erlang*, Acta Polytechnica Scandinavica, Applied Math. and Comp. Machinery Series No. 6, Copenhagen.

[49] P.J. Burke (1956), The output of a queueing system, *Operations Res.*, **4**, 699-714.

[50] P.J. Burke (1976), Proof of a conjecture on the inter-arrival time distribution in an M/M/1 queue with feedback, *IEEE Trans. Comm.*, **COM-24**, 575-576.

[51] J.A. Buzacott and J.G. Shanthikumar (1992), Editorial introduction, *Queueing Systems*, **12** (Special issue on Manufacturing Systems), 1-2.

[52] J.A. Buzacott and J.G. Shanthikumar (1993), *Stochastic Models of Manufacturing Systems*, Prentice Hall, Upper Saddle River, NJ.

[53] J.A. Buzacott and D.D. Yao (1986), On queueing network models of flexible manufacturing systems, *Queueing Systems*, **1**, 5-27.

[54] J.P. Buzen (1973), Computational algorithms for closed queueing networks with exponential servers, *Comm. ACM*, **16**, 527-531.

[55] S.R. Chakravarthy (2001), The Batch Markovian Arrival Process: A Review and Future Work *Advances in Probability Theory and Stochastic Processes*, (Eds. A. Krishnamoorthy, N. Raju, and V. Ramaswami), Notable Publications, Inc., NJ, 21-49.

[56] S.R. Chakravarthy (2010), Markovian Arrival Processes, *Wiley Encyclopedia of Operations Research and Management Science*. Published Online: 15 JUN 2010.

[57] K.M. Chandy. (1972), The analysis and solutions for general queueing networks, *Proc. of the 6th Annual Princeton Conf. on Inf. Sci. & Syst.*, Princeton Univ., Princeton, NJ.

[58] M.L. Chaudhry (1992), Editorial introduction, *Queueing Systems*, **10**, Special issue on Numerical Computations in Queues, 1-3.

[59] M.L. Chaudhry and J.G.C. Templeton (1983), *A First Course in Bulk Queues*, John Wiley & Sons, New York.

[60] M.L. Chaudhry, M. Agarwal, and J.G.C. Templeton (1992), Exact and approximate numerical solutions of steady-state distributions arising in the queue GI/G/1, *Queueing Systems*, **10**, 105-152.

[61] A.B. Clarke (1957), Maximum likelihood estimates in a simple queue, *Ann. Math. Statist.*, **28**, 1036-1040.

[62] A. Cobham (1954), Priority assignment in waiting line problems, *Operations Res.* **2**, 70-76; Correction, **3**, 547.

[63] E.G. Coffman Jr. and P.J. Denning (1973), *Operating Systems Theory*, Prentice Hall, Englewood Cliffs, NJ.

[64] E.G. Coffman Jr., E. Gelenbe, and E.N. Gilbert (1988), Analysis of a conveyor queue in a flexible manufacturing system, *European Journal of Operational Research* **35**, 382-392.

[65] E.G. Coffman Jr. and M. Hofri (1986), Queueing models of secondary storage devices, *Queueing Systems*, **1**, 129-168.

[66] E.G. Coffman Jr. and L. Kleinrock (1968), Feedback queueing models for time shared systems, *J. ACM*, **15**, 549-576.

[67] J.W. Cohen (1969), *The Single Server Queue*, North Holland, London.

[68] R.W. Conway, W.L. Maxwell, and L.W. Miller (1967), *Theory of Scheduling*, Addison-Wesley, Reading, Massachusetts.

[69] R.B. Cooper (1981), *Introduction to Queueing Theory*, North Holland, New York.

[70] P.J. Courtois (1977), *Decomposability: Queueing and Computer Science Applications*, Academic Press, New York.

[71] D.R. Cox (1955), The analysis of non-Markovian stochastic processes by the inclusion of supplementary variables, *Proc. Comb. Phil. Soc.*, **51**, 433-441.

[72] D.R. Cox (1962), *Renewal Theory*, Methuen, London.

[73] D.R. Cox (1965), Some problems of statistical analysis connected with congestion, *Proc. Symp. on Congestion Theory* (Eds. W. L. Smith and W. B. Wilkinson), Univ. of North Carolina, Chapel Hill, NC, 289-316.

[74] D.R. Cox and P.A.W. Lewis (1966), *The Statistical Analysis of Series of Events*, Methuen, London.

[75] T.B. Crabill (1972), Optimal Control of a service facility with variable exponential service times and constant arrival rate. *Management Sci*, **18**, 560-566.

[76] T.B. Crabill, D. Gross, and R.J. Magazine (1977), A classified bibliography of research on optimal design and control of queues. *Operations Res.*, **25**, 219-232.

[77] Y. Dallery and K.E. Stecke (1990), On the optimal allocation of servers and workloads in closed queueing network, *Operations Res.* **38**, 694-703.

[78] G.R. Dattatreya (2008), *Performance analysis of queueing and computer networks*, Chapman & Hall, CRC Computer & Information Science Series, London.

[79] F. de Vericourt, F. Karaesmen and Y. Dallery (2002), Optimal stock allocation for a capacitated supply system, *Management Sci.,* **48**, 1486-1501.

[80] P.J. Denning and J.P. Buzen. (1978), The operational analysis of queueing network models. *Computing Surv.*, **10(3)**, 225-261.

[81] R.L. Disney and P.C. Kiessler (1987), *Traffic Processes in Queueing Networks*, The Johns Hopkins Univ. Press, Baltimore.

[82] B.T. Doshi (1986), Queueing Systems with vacations – A survey, *Queueing Systems*, **1**, 29-66.

[83] B.T. Doshi and D. Yao (1995), Editorial introduction, *Queueing Systems*, **20**, (Special issue on Telecommunication Systems), 1-5.

[84] J.H. Dshalalow (1997), Queueing systems with state dependent parameters, *Frontiers in Queueing* (Ed. J. H. Dshalalow), Ch. 4, CRC Press, New York, 61-116.

[85] H. Dubner and J. Abate (1968), Numerical inversion of Laplace transforms by relating them to the finite Fourier cosine transform, *J. ACM*, **15**, 115-123.

[86] L.C. Edie (1956), Traffic delays at toll booths, *J. Oper. Res. Soc. Amer.*, **2**, 107-138.

[87] A.K. Erlang (1917), Solution of some problems in the theory of probabilities of significance in automatic telephone exchanges , *Elektroteknikeren*, **13**, 5.

[88] G. Falin (1990), A survey of retrial queues, *Queueing Systems*, **7**, 127-168.

[89] W. Feller (1968), *An Introduction to Probability Theory and its Applications*, Vol. 1, 3rd ed., John Wiley & Sons, New York.

[90] T.C. Fry (1928), *Probability and its Engineering Uses*, Van Nostrand, Princeton, NJ.

[91] S.H. Fuller (1980), Performance evaluation, *Introduction to Computer Architecture*, 2nd ed. (Ed. H. S. Stone), Science Research Associates, Chicago, 527-590.

[92] D.P. Gaver Jr. (1968), Diffusion approximations and models for certain congestion problems , *J. Appl. Prob.*, **5**, 607-623.

[93] J.P. Gayon, F. de Vericourt, and F. Karaesmen (2009), Stock rationing in an $M/E_r/1$ multi-class make-to-stock queue with backorders, *IIE Transactions*, **41**, 1096-1109.

[94] E. Gelenbe and G. Pujolle (1998), *Introduction to Queueing Networks*, John Wiley & Sons, New York.

[95] G. Giambene (2005), *Queueing theory and telecommunications networks and applications*, Springer, New York.

[96] P.W. Glynn (1990), Diffusion approximations, *Stochastic Models, Vol. 2.* (Eds. D. P. Heyman and M. J. Sobol), Ch. 4, Elsevier Science Publishers, Amsterdam, The Netherlands, 145-198.

[97] G. H. Golub and C. Van Loan (1996), *Matrix computations*, 3rd ed., The Johns Hopkins University Press, Baltimore, MD.

[98] W. J. Gordon and G.F. Newell (1967), Closed queueing systems with exponential servers, *Operations Res.* **15**, 254-265.

[99] M.K. Govil and M.C. Fu (1999), Queueing theory in manufacturing: a survey, *Journal of Manufacturing Systems* **18**, 214-240.

[100] T. L. Goyal and C.M. Harris (1972), Maximum likelihood estimation for queues with state dependent service, *Sankhya*, **A34**, 65-80.

[101] A. Graham (1981), *Kronecker Products and Matrix Calculus with Applications*, Ellis Horwood Limited, Chichester, UK.

[102] D. Gross and C.M. Harris (1998), *Fundamentals of Queueing Theory*, 3rd ed., John Wiley & Sons, New York.

[103] D. Gross, J.F. Shortle, J.F. Thompson and C.M. Harris (2008), *Fundamentals of Queueing Theory*, 4th ed. John Wiley & Sons, Hoboken, NJ.

[104] F.A. Haight (1957), Queueing with balking, *Biometrika*, **44**, 360-369.

[105] F. A. Haight (1967), *Handbook of the Poisson Distribution*, John Wiley & Sons, New York.

[106] G. Hardy (1949), *Divergent Series*, Oxford Univ. Press, Oxford, UK.

[107] K. Harishchandra and S.S. Rao (1988), A note on the statistical inference on the traffic intensity parameter in M/E_k/1 queue, *Sankhya*, **A50**, 144-148.

[108] K.J. Hastings (2006), *Introduction to the Mathematics of Operations Research with Mathematica*, 2nd Ed. Chapman & Hall, New York.

[109] D.P. Heyman (1968), Optimal operating policies for M/G/1 queueing systems, *Operations Res.*, **16**, 362-382.

[110] F.B. Hildebrand (1968), *Finite Difference Equations and Simulations*, Prentice Hall, Englewood Cliffs, N. J.

[111] F.S. Hillier (1963), Economic models for industrial waiting line problems, *Management Sci.*, **10**, 119-130.

[112] F.S. Hillier and G.J. Lieberman (1986), *Introduction to Operations Research*, 4th ed., Holden-Day, Oakland, CA.

[113] T. Hirayama, S.J. Hong, and M.M. Krunz (2004), A new approach to analysis of polling systems, *Queueing Systems*, **48**, 135-158.

[114] M. Hollander, D.A. Wolfe and E. Chicken (2013), *Nonparametric Statistical Methods*, 3rd ed., John Wiley & Sons, Hoboken, NJ.

[115] D.L. Iglehart and W. Whitt (1970), Multiple channel queues in heavy traffic I & II, *Adv. Appl. Prob.*, **2**, 150-177, 355-369.

[116] R.R.P. Jackson (1954), Queueing processes with phase type service, *Operations Res. Quart.*, **5**, 109-120.

[117] R.R.P. Jackson (1956), Queueing processes with phase type service, *J. Roy. Stat. Soc.* B, **18**, 129-132.

[118] J.R. Jackson (1957), Networks of waiting lines, *Operations Res.*, **5**, 518-521.

[119] J.R. Jackson (1963), Jobshop-like queueing systems, *Management Sci.*, **10**, 131-142.

[120] A.A. Jagers and E.A. van Doorn (1986), On the continued Erlang loss function, *Operations Res. Letters*, **5**, 43-46.

[121] R. Jain (1991), *The Art of Computer Systems Performance Analysis*, John Wiley & Sons, New York.

[122] N.K. Jaiswal (1968), *Priority Queues*, Academic Press, New York.

[123] W.S. Jewell (1967), A simple proof of $L = \lambda W$, *Operations Res.* **15**, 1109-1116.

[124] N.L. Johnson and S. Kotz (1969), *Distributions in Statistics: Discrete Distributions*, John Wiley & Sons, New York.

[125] L. Joseph, D.B. Wolfson, and C. Wolfson (1990), Is multiple sclerosis an infectious disease? Inference on an input process based on the output, *Biometrics*, **46**, 337-349.

[126] K. Kant (1992), *Introduction to computer system performance evaluation*, McGraw-Hill, New York.

[127] S. Karlin and H.M. Taylor (1975), *A First Course in Stochastic Processes*, 2nd ed. Academic Press, New York.

[128] C. Karrer, K. Alicke, and H.O. Günther (2012), A framework to engineer production control strategies and its application in electronic manufacturing, *International Journal of Production Research*, **50**, 6595-6611.

[129] F.P. Kelly (1979), *Reversibility and Stochastic Networks*, John Wiley & Sons, New York.

[130] D.G. Kendall (1951), Some problems in the theory of queues, *J. Roy. Stat. Soc.* B, **13**, 151-185.

[131] D.G. Kendall (1953), Stochastic processes occurring in the theory of queues and their analysis by the method of imbedded Markov chains, *Ann. Math. Statist.*, **24**, 338-354.

[132] L. Kerbache and J.M. Smith (2004), Queueing networks and the topological design of supply chain system, *International Journal of Production Economics*, **91**, 251-272.

[133] J.F.C. Kingman (1962a), Some inequalities for the GI/G/1 queue, *Biometrika*, **49**, 315-324.

[134] J.F.C. Kingman (1962b), On queues in heavy traffic, *J. Roy. Stat. Soc.* B, **24**, 383-392.

[135] J.F.C. Kingman (1965), The heavy traffic approximation in the theory of queues, *Proceedings of the Symposium on Congestion Theory*. (Eds. W. L. Smith and W. E. Wilkinson), The University of North Carolina Press, Chapel Hill, NC.

[136] J.F.C. Kingman (1969), Markov population processes, *J. Appl. Prob.*, **6**, 1-18.

[137] L. Kleinrock (1975), *Queueing Systems Vol. I: Theory*, John Wiley & Sons, New York.

[138] L. Kleinrock, L. (1976), *Queueing Systems Vol. II: Computer Applications*, John Wiley & Sons, New York.

[139] E. Koenigsberg (1958), Cyclic queues, *Operations Res. Quart.*, **9**, 22-35.

[140] T. Konstantopoulos (1998), Editorial introduction, *Queueing Systems*, **28**. (Special issue on Mathematical and Probabilistic Methods in Communication Networks), 1-5.

[141] P. Kouvelis, W.C. Chiang, and A.S. Kiran (1992), A survey of layout issues in flexible manufacturing systems, *Omega*, **20**, 375-390.

[142] J.A. Koziol and A.F. Nemec (1979), On a Cramer-von Mises type statistic for testing bivariate independence, *Canad. J. Stat.*, **7**, 43-52.

[143] S. Krakowski (1988), *Principles of Operating Systems* (Trans. D. Beeson), MIT Press, Cambridge, MA.

[144] V.G. Kulkarni (1997), Fluid models for single buffer systems, *Frontiers in Queueing* (Ed. J. H. Dshalalow), Ch. 11, CRC Press, New York, 321-338.

[145] V.G. Kulkarni and H.M. Liang (1997), Retrial queues revisited, *Frontiers in Queueing* (Ed. J. H. Dshalalow), Ch. 2, CRC Press, New York, 19-34.

[146] S. Kumar and P.R. Kumar (2001), Queueing network models in the design and analysis of semiconductor wafer fabs, *IEEE Transactions on Robotics and Automation*, **17**, 548-561.

[147] G. Latouche and V. Ramaswami (1999), Introduction to matrix analytic methods in stochastic modeling, ASA SIAM Series on Statistics and Applied Probability, Philadelphia, PA.

[148] S.S. Lavenberg (ed.) (1983), *Computer Performance Modeling Handbook*, Academic Press, New York.

[149] A.M. Law and W.D. Kelton (1991), *Simulation Modeling and Analysis*, 2nd ed., McGraw-Hill, New York.

[150] E. Lazowska (1984), *Quantitative system performance, computer system analysis using queueing network models*, Prentice Hall, NJ.

[151] J. Le Boudec (2011), Performance evaluation of computer and communication systems, EPFL Press, Computer and Communication Systems Series, Lausanne, Switzerland.

[152] W. Ledermann and G.E. Reuter (1956), Spectral theory for the differential equations of simple birth and death processes, *Phil. Trans. Roy. Soc., London*, A, **246**, 321-369.

[153] P.A.W. Lewis (1972), *Stochastic Point Processes* (Ed. P. A. W Lewis), John Wiley & Sons, New York, 1-54.

[154] K. Lieckens and N. Vandaele (2012), Multi-level reverse logistics network design under uncertainty, *International Journal of Production Research* **50**, 365-380.

[155] H.W. Lilliefors (1966), Some confidence intervals for queues, *Operations Res.*, **14**, 723-727.

[156] C. Lin, C. N. Madu, and C. H. Kuei (1994), Closed queueing maintenance network for flexible manufacturing systems, *Microelectronics Reliability*, **34**, 1733-1744.

[157] D.V. Lindley (1952), The theory of queues with a single server, *Proc. Comb. Phil. Soc.*, **48**, 277-289.

[158] J.D.C. Little (1961), A proof for the queueing formula: $L = \lambda W$, *Operations Res.*, **9**, 383-387.

[159] L. Liu, X. Liu, and D.D. Yao, Analysis and optimization of a multistage inventory-queue system, *Management Sci.*, **50**, 365-380.

[160] D. Lucantoni, K.S. Meier-Hellstern, and M.F. Neuts (1990), A single-server queue with server vacations and a class of nonrenewal arrival processes, *Adv. in Appl. Prob.*, **22**, 676-705.

[161] D. Lucantoni (1991), New results on the single server queue with a batch Markovian arrival process, *Stochastic Models*, **7**, 1-46.

[162] G. Luchak (1956), The solution of the single channel queueing equations characterized by a time-dependent Poisson distributed arrival rate and a general class of holding times, *Operations Res.*, **4**, 711-732.

[163] M. Manitz (2008), Queueing-model based analysis of assembly lines with finite buffers and general service times, *Computers and Operations Research*, **35**, 2520-2536.

[164] W.G. Marchal (1978), Some simpler bounds on the mean queueing time, *Operations Res.*, **26**, 1083-1088.

[165] M. Marcus and H. Minc (1964), *A Survey of Matrix Theory and Matrix Inequalities*, Allyn and Bacon, Boston, MA.

[166] K.T. Marshall (1968), Some inequalities in queueing, *Operations Res.*, **16**, 651-665.

[167] J. McKinney (1969), Survey of analytical time sharing models, *Computing Surveys*, **1**, 105-116.

[168] D.R. Miller (1981), Computation of steady state probabilities for M/M/1 priority queues, *Operations Res.*, **29**, 945-958.

[169] G.K. Miller (1996), *Estimation for Renewal Processes with Unobservable Interarrival Times*, Doctoral Dissertation, Southern Methodist University, Dallas, TX, USA (Available through Pro Quest/University Microfilms.)

[170] G.K. Miller (1999), Maximum likelihood estimation for Erlang integer parameter, *Statistics & Prob. Letters*, **43**, 335-341.

[171] G.K. Miller and U.N. Bhat(1997), Estimation for renewal processes with unobservable gamma or Erlang interarrival times, *J. Stat. Planning and Inference*, **61**, 355-372.

[172] G.K. Miller and U.N. Bhat (2002), Estimation of the coefficient of variation for unobservable service times in the M/G/1 queue, *J. of Math. Sciences*, **1**, 1-11.

[173] D. Mitra and I. Mitrani (1991), Editorial introduction, *Queueing Systems*, **9**, (Special issue on Communication Systems), 1-4.

[174] D. Mitra, I. Mitrani, K.G. Ramakrishnan, J.B. Seery, and A. Weiss (1991), A unified set of proposals for control and design of high speed data networks, *Queueing Systems*, **9**, 215-234.

[175] J.J. Moder and C.R. Phillips Jr. (1962), Queueing with fixed and variable channels, *Operations Res.*, **10**, 218-231.

[176] E.C. Molina (1927), Application of the theory of probability to telephone trunking problems, *Bell Systems Tech. J.*, **6**, 461-494.

[177] M.K. Molloy (1989), *Fundamentals of Performance Modeling*, Macmillan, New York.

[178] S.C. Moore (1975), Approximating the behavior of non-stationary single server queues, *Operations Res.*, **23**, 1011-1032.

[179] P.M. Morse (1958), *Queues, Inventories and Maintenance*, John Wiley & Sons, New York.

[180] R.R. Muntz (1973), Poisson departure processes and queueing networks, *Proc. of the 7th Annual Princeton Conf. on Inf. Sci. & Syst.*, Princeton Univ. Press, Princeton, NJ.

[181] M.F. Neuts (1966), The single server queue with Poisson input and semi-Markov service times, *J. Appl. Prob.*, **3**, 202-230.

[182] M.F. Neuts (1967), A general class of bulk queues with Poisson input, *Ann. of Math. Statist.*, **68**, 759-770.

[183] M. F. Neuts (1974), A versatile Markovian point process. *J. Appl. Prob.*, **16**, 764-779.

[184] M.F. Neuts (1975), Probability distributions of phase type, *Liber Amicorum Prof. Emeritus H. Florin*, Department of mathematics, University of Louvain, Belgium, 173-206.

[185] M.F. Neuts (1978), Markov chains with applications in queueing theory, which have a matrix geometric invariant probability vector, *Adv. Appl. Prob.*, **10**, 185-212.

[186] M. F. Neuts (1981), *Matrix-Geometric Solutions in Stochastic Models -An Algorithmic Approach.* Dover Publications, 1995 (originally published by The Johns Hopkins University Press, 1981),

[187] M.F. Neuts (1989), *Structured Stochastic Matrices of M/G/1 Type and Their Applications*, Marcel Dekker, New York.

[188] G.F. Newel (1968), Queues with time dependent arrival rates I - II, *J. Appl. Prob.*, **5**, 436-451, 579-606.

[189] G.F. Newel (1971, 1982), *Applications of Queueing Theory*, Chapman and Hall, London.

[190] M.S. Obaidat and N.A. Boudriga (2010), *Fundamentals of performance evaluation of computer and telecommunications systems*, John Wiley & Sons, Hoboken, New Jersey.

[191] C. Palm (1947), The distribution of repairmen in servicing automatic machines, *Industritidningen Norden*, **75**, 75-80, 90-94, 119-123 (in Swedish).

[192] R.D. Pedersen and J.C. Shah (1972), Multiserver queue storage requirements with unpacked messages, *IEEE Trans. on Comm.*, June, 462-465.

[193] H. Perros (1994), *Queueing Networks with Blocking*, Oxford Univ. Press, New York.

[194] S.M. Pitts (1994), Nonparametric estimation of the stationary waiting time distribution for the GI/G/1 queue, *Ann. Statist.*, **22**, 1428-1446.

[195] F. Pollaczek (1934), Uber das Warteproblem , *Math. Zeitschrift*, **38**, 492-537.

[196] F. Pollaczek (1965), Concerning an analytical method for the treatment of queueing problems, *Proc. Symp. Congestion Theory* (Eds. W. L. Smith and W. E. Wilkinson), Univ. of North Carolina Press, Chapel Hill, 1-42.

[197] N.U. Prabhu (1960), Some results for the queue with Poisson arrivals, *J. Roy. Stat. Soc.* B, **22**, 104-107.

[198] N.U. Prabhu (1965a), *Queues and Inventories*, John Wiley & Sons, New York.

[199] N.U. Prabhu (1965b), *Stochastic Processes*, Macmillian, New York.

[200] N.U. Prabhu (1987), A bibliography of books and survey papers on queueing systems: Theory and applications, *Queueing Systems*, **2**, 393-398.

[201] N.U. Prabhu (1997), *Foundations of Queueing Theory*, Kluwer Academic Publishers, Boston, MA.

[202] N.U. Prabhu (1998), *Stochastic Storage Processes*, Springer, New York.

[203] N.U. Prabhu and U.N. Bhat (1963a), Some first passage problems and their application to queues, *Sankhya, Series* A, **25**, 281-292.

[204] N.U. Prabhu and U.N. Bhat (1963b), Further results for the queue with Poisson arrivals, *Operations Res.*, **11**, 380-386.

[205] J. Putter (1955), The treatment of ties in some nonparametric tests, *Ann. Math. Stat.*, **26**, 368-386.

[206] V. Ramaswami (1980), The N/G/1 queue and its detailed analysis, *Adv. Appl. Prob.*, **12**, 222-261.

[207] V. Ramaswami (1990), From the matrix-geometric to the matrix-exponential, *Queueing Systems*, **6**, 229-260.

[208] V. Ramaswami (2001), The surprising reach of the matrix analytic approach, *Advances in Probability and Stochastic Processes*. (Eds. A. Krishnamoorthy, N. Raju, and V. Ramaswami), Notable Publications, Neshanic Station, NJ, 167-177.

[209] R.H. Randles and D.A. Wolfe (1979), *Introduction to the Theory of Nonparametric Statistics*, John Wiley & Sons, New York.

[210] S.S. Rao and U.N. Bhat (1991), A sequential test for a denumerable Markov chain and an application to queues, *J. Math. Phy. Sci.*, **25**, 521-527.

[211] S.S. Rao, U.N. Bhat, and K. Harishchandra (1984), Control of traffic intensity in a queue – A method based on SPRT, *Opsearch*, **21**, 63-80.

[212] S.S. Rao, A. Gunasekaran, S.K. Goyal, and R. Martikainen (1998), Waiting line model applications in manufacturing, *International Journal of Production Economics*, **54**, 1-28.

[213] E. Reich (1965), Departure processes, *Proc. of the Symp. on Congestion Theory*, (Eds. W. L. Smith and W. E. Wilkinson), The Univ. of North Carolina Press, Chapel Hill, 439-457.

[214] M. Reiser and S.S. Lavenberg (1980) Mean-value analysis of closed Multichain queueing networks, *J. ACM*, **27(2)**, 313-322.

[215] W. Rudin (1964), *Principles of Mathematical Analysis*. McGraw-Hill, New York.

[216] T.L. Saaty (1961), *Elements of Queueing Theory and Applications*, McGraw Hill, New York.

[217] T.L. Saaty (1966), Seven more years of queueing theory: A lament and a bibliography, *Nav. Res. Logist. Quart.*, **13**, 447-476.

[218] C.H. Sauer and K.M. Chandy (1981), *Computer Systems Performance Modeling*, Prentice Hall, Englewood Cliffs, NJ.

[219] T.J. Schriber (1991), *Introduction to Simulation*. John Wiley & Sons, New York.

[220] P. Schweitzer (1979), Approximate analysis of multiclass closed networks of queues, *Proc. International Conference on Stochastic Control and Optimizations*, Amsterdam, The Netherlands.

[221] B. Sengupta (1989), Markov processes whose steady state distribution is matrix-exponential with an application to GI/PH/1, *Adv. Appl. Prob.*, **22**, 159-180.

[222] J.G. Shanthikumar, S. Ding, and M.T. Zhang (2007), Queueing theory for semiconductor manufacturing systems: a survey and open problems, *IEEE Transactions on Automation Science and Engineering*, **4**, 513-522.

[223] V.P. Singh (1970), Two-server Markovian queues with balking: heterogeneous vs. homogeneous servers, *Operations Res.*, **18**, 145-159.

[224] V.P. Singh (1971), Markvoian queues with three heterogeneous servers, *AIIE Trans* III(1), March, 45-48.

[225] A. Sleptchenko, M.C. Heijden, and A. Harten (2005), Using repair priorities to reduce stock investment in spare part networks, *European Journal of Operational Research*, **163**, 733-750.

[226] D.R. Smith and W. Whitt (1981), Resource sharing for efficiency in traffic systems, *Bell Systems Tech. J.*, **60**, 39-65.

[227] M.J. Sobel (1974), Optimal operation of queues. *Mathematical Methods in Queueing Theory*, (Ed. A. B. Clarke), Proceedings of a conference at Western Michigan University, Springer-Verlang, New York.

[228] W-H. Steeb (2006), *Problems and Solutions in Introductory and Advanced Matrix Calculus*, World Scientific Publishing, Singapore.

[229] W-H. Steeb and Y. Hardy (2011), Matrix Calculus and Kronecker Product. World Scientific Publishing, Singapore.

[230] W.J. Stewart (1994), *Introduction to the Numerical Solution of Markov Chains*, Princeton Univ. Press, Princeton, N. J.

[231] S. Stidham Jr. (1995), Editorial introduction, *Queueing Systems*, **21**, (Special issue on Optimal Design and Control of Queueing Systems), 239-243.

[232] S. Stidham. and N.U. Prabhu (1974), Optimal control of queueing systems, *Mathematical Methods in Queueing Theory* (Ed. A. B. Clarke), Proceedings of a conference in Western Michigan University, Springer-Verlag, New York, 263-294.

[233] T. Suzuki and Y. Yoshida (1970), Inequalities for many server queue and other queues, *J. Oper. Res. Soc. of Japan*, **13**, 59-77.

[234] R. Syski (1960), *Introduction to Congestion Theory in Telephone Systems*, Oliver and Boyd, London.

[235] L. Takács (1962), *Introduction to the Theory of Queues*, Oxford University Press, New York.

[236] L. Takács (1967), *Combinatorial Methods in the Theory of Stochastic Processes*, John Wiley & Sons, New York.

[237] H. Takagi (1997), Queueing analysis of polling models: Progress in 1990-1994, *Frontiers in Queueing* (Ed. J. H. Dshalalow), CRC Press, New York, Chapter 5, 119-146.

[238] T.R. Thiagarajan and C.M. Harris (1979), Statistical tests for exponential service from M/G/1 waiting time data, *Nav. Res. Logist. Quart.*, **26**, 511-520.

[239] D. Thiruvaiyaru and I.V. Basawa (1996), Maximum likelihood estimation for queueing networks, *Stochastic Processes and Statistical Inference* (Eds: B. L. S. Prakasa Rao and B. R. Bhat) New Age International Publications, New Delhi, 132-149.

[240] D. Thiruvaiyaru, I.V. Basawa, and U.N. Bhat (1991), Estimation for a class of simple queueing networks, *Queueing Systems*, **9**, 301-312.

[241] P. Toktas-Palut and F. Ulengin (2011), Coordination in a two-stage capacitated supply chain with multiple suppliers, *European Journal of Operational Research*, **212**, 43-53.

[242] K.S. Trivedi (2002), *Probability and Statistics with Reliability, Queueing, and Computer Science Applications*, 2nd ed., John Wiley & Sons, New York.

[243] T. Van Woensel, L. Kerbache, H. Peremans, and N. Vandaele (2008), Vehicle routing with dynamic travel times: a queueing approach, *European Journal of Operational Research*, **186**, 990-1007.

[244] W.N. Venables and B.D. Ripley (2002), *Modern Applied Statistics with S*, 4th ed., Springer, New York.

[245] N. Viswanadham and N.R.S. Raghavan (2000), Performance analysis and design of supply chains: a Petri net approach, *Journal of the Operational Research Society*, **51**, 1158-1169.

[246] W. Whitt (2000), An overview of Brownian and non-Brownian FCLT's for the single server queue, *Queueing Systems*, **36**, 39-70.

[247] P.T. Whittaker and G.N. Watson (1962), *A Course of Modern Analysis*, 4th ed., Cambridge Univ. Press, Cambridge, U. K.

[248] P. Whittle (1967), Nonlinear migration processes, *Bull. Int. Stat. Inst.*, **42**, 642-646.

[249] P. Whittle (1968), Equilibrium distributions for an open migration process, *J. Appl. Prob.*, **5**, 567-571.

[250] D.V. Widder (1946), *The Laplace Transform*, Princeton Univ. Press, Princeton, NJ.

[251] R.W. Wolff (1965), Problems of statistical inference for birth and death queueing models, *Operations Res.*, **13**, 343-357.

[252] R.W. Wolff (1982), Poisson arrivals see time averages, *Operations Res.*, **30**, 223-231.

[253] Y. Wu and M. Dong (2008), Combining multi-class queueing networks and inventory models for performance analysis of multi-product manufacturing logistics chains, *International Journal of Advanced Manufacturing Technology*, **37**, 564-575.

[254] T. Yang and J.G.C. Templeton (1987), A survey of retrial queues, *Queueing Systems*, **2**, 201-233.

[255] S.F. Yashkov (1987), Processor showing queues: some progress in analysis, *Queueing Systems*, **2**, 1-17.

Index

© Springer Science+Business Media New York 2015

U. N. Bhat, *An Introduction to Queueing Theory*, Statistics for Industry and Technology, DOI 10.1007/978-0-8176-8421-1