BIRKHÄUSER

Statistics for Industry and Technology

U. Narayan Bhat

An Introduction to Queueing Theory

Modeling and Analysis in Applications

Birkhäuser
Boston • Basel • Berlin

U. Narayan Bhat
Professor *Emeritus*
Statistical Science & Operations Research
Southern Methodist University
Dallas, TX 75275-0332
USA

ISBN: 978-0-8176-4724-7 e-ISBN: 978-0-8176-4725-4
DOI: 10.1007/978-0-8176-4725-4

Library of Congress Control Number: 2007941114

Mathematics Subject Classification (2000): 60J27, 60K25, 60K30, 68M20, 90B22, 90B36

Cover design: Dutton and Sherman, Hamden, CT.

Printed on acid-free paper.

9 8 7 6 5 4 3 2 1

www.birkhauser.com

In memory of my parents,
Vaidya P. Ishwar and Parvati Bhat

Contents

Preface

There are several books on queueing theory available for students as well as researchers. At the low end of mathematical sophistication, some provide usable formulas in a recipe fashion. At the high end there are research monographs on specific topics and books with an emphasis on theoretical analysis. In between there are a few textbooks with one common feature: all of them require an adequate background knowledge of probability and Markov processes that can be acquired normally with a semester-length graduate course. Consequently, most people who deal with the modeling and analysis of queueing systems either do not take a course on the subject because it would require 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 using 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 an 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 references for advanced topics. It gives a comprehensive discussion of statistical inference techniques usable 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 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. This book is the product of the author's experience in teaching queueing theory for 40 years at various levels to students with or without the necessary background in stochastic processes.

The mathematical background assumed here 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. An appendix on mathematical results provides some of the essential theorems for reference. Instructors may request a guide to the solutions of exercises via the Birkhäuser website at www.birkhauser.com/978-0-8176-4724-7.

The book does not advocate any specific software for the numerical analysis of queueing problems. The one chapter on modeling and analysis using computational tools employs 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 the 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 the numerical analysis of queueing systems (in which the author's expertise is limited). Special acknowledgement of indebtedness is also made to the reviewers' comments, which have helped to improve 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 Birkhäuser. 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 this 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 perserverance.

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, TX
January 2008

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 that incorporate characteristics of a queue.

We identify the unit demanding service, whether it is human or otherwise, as the *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.

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 proceed strictly according to schedule, a queue can be avoided. But in practice this does not happen.

U.N. Bhat, *An Introduction to Queueing Theory*, DOI: 10.1007/978-0-8176-4725-4_1,

In most cases the arrivals are a product of external factors. Therefore, the best one can do is to describe the input process in terms of random variables that can 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 being 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 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, 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 (RR)" and "shortest processing time" may need some elaboration, which is provided in later chapters. In many situations customers in some classes have priority in service over others. There are many other queue disciplines that 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 made up of symbols representing three elements: input/service/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 the following:

$M/G/1$: 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 briefly define a few terms that will be 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 in many other branches of science, the models are studied analytically with the hope that the information obtained from such a 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 amount of time a new arrival has to wait until its service begins, which we call the *waiting time*, and the length of time during which the server is continuously busy, which we call the *busy period*, are major characteristics of interest. Note 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 and 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 the 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 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. 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), 33). In this paper, he lays the foundation for the place of Poisson (and hence exponential) distribution in queueing theory. His papers written during the next 20 years contain some of the most important concepts and techniques; the notion of statistical equilibrium and the method of writing 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.

In Erlang's work, as well as the work done by others in the 1920s and 1930s, 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 towards 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

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

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 explicitly 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:

1. using existing mathematical techniques or developing new ones for the analysis of the underlying processes; and
2. 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 on the direction of research in queueing theory.

The queue $M/M/1$ (Poisson arrival, exponential service, single server) was 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 Ledermann 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 analysis 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 $GI/M/s$. (The widely used symbolic notation to identify queueing systems was used by Kendall (1953).) 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 usable results in applications, various types of approximations have appeared in the literature. For an extensive bibliography, see Bhat et al. (1979). 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 on 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 (1970). See also 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 the 1960s most of the basic queueing systems that could be considered as reasonable models of real-world phenomena had been analyzed; 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: methodological 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, 1989), 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 usable results can be derived.

(ii) Transform inversion. The traditional method of analysis of queueing systems depends on inverting generating functions and/or Laplace transforms to derive usable results. The complexities of transform inversion has spurred more research, 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 in 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 literature, 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-form solutions are valid for networks more general than those with simple $M/M/s$-type nodes, such as state-dependent service; processor-sharing discipline; heterogeneous service times; Coxian service time distributions; and the preemptive resume LCFS discipline.

Since the publication of Baskett et al., a large body of literature has grown in the performance modeling of queueing networks. The works of Courtois (1977), Kelly (1979), Sauer and Chandy (1981), Lavenberg (1983), Disney and Kiessler (1987), Molloy (1989), Perros (1994), and Gelenbe and Pujolle (1998) 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 the references cited earlier) and books such as those by 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 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.

(v) Manufacturing 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 jobshop is a network of machines. (See also Jackson (1963).) Simulation studies reported in Conway et al. (1967) provide excellent examples of the incorporation of queueing models with jobshop scheduling. Since the 1970s, with the advent of new processes in manufacturing incorporating computers in their various stages, not only the application of queueing theory results, but also the development of new techniques have occurred at a phenomenal rate. Three articles in Buzacott and Shanthikumar (1992) and the book by Buzacott and Shanthikumar (1993) bring together most of the important developments in the application of queueing theory in manufacturing systems.

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 an infinite number of servers. Other applications include job-flow lines as tandem queues, and jobshops and flexible manufacturing systems as queueing networks. For recent articles on the applications of queueing theory in manufacturing system modeling, readers may refer to various journals such as *Management Science*, *European Journal of Operational Research*, *IIE Transactions*, and *Computers and Industrial Engineering*.

(vi) Specialized models. Some of the special 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 in 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 bibliographies 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 models is the ability to decompose the results into those corresponding to systems without vacations and those 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, the concept of customers being denied entry to the system and 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 appropriate 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, it may not be possible to implement sampling plans appropriate for data collection to estimate parameters of the constituent elements. 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 in this area. See also 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 the following: 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 (1996) extended the maximum likelihood estimation to include Jackson networks; and Pitts (1994) considered the queue as a functional that maps the service and interarrival 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 interarrival 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 became interested in queueing problems, operational problems were being handled using primarily behavioral results. Note 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 (1954), which 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. See also 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 may be, 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. See also 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 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 an FCFS 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.
6. In store-and-forward communication networks messages for transmission are stored in buffers of fixed size. Each message may need one or more buffers. The message is transmitted through several identical channels. Knowing the characteristics of the arrival process, transmission rate, and the message, we are interested in the storage requirements of a network node.

   Describe how the long-run storage requirements can be estimated for this type of a system.

7. In a warehouse items are stacked in such a way that the most recently stacked item is 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 input–output (I/O) with a certain probability, say p, and the job is complete with 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 the 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 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.

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, the number of servers, system capacity, and discipline are normally deterministic (unless the number of available servers becomes a random variable—which is possible in some cases). But there are uncertainties related to arrivals and service which result in the underlying process being stochastic.

The similarity of the arrival and service processes can be brought out by identifying similar components, such as interarrival times and service times, or 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 system (exceptions are possible). Therefore, first we discuss the possibilities of using certain probability distributions to represent the process of interarrival times and service times. In the case of the 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 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. (The algebra of discrete distributions could be cumbersome.) Nevertheless, continuous and discrete models are mutual analogues and most of the properties carry through in both cases.

Using a common notation we take $Z_1, Z_2, \ldots$ as nonnegative random variables representing either interarrival times or service times of consecutive customers. Further, let

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

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

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

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

U.N. Bhat, *An Introduction to Queueing Theory*, DOI: 10.1007/978-0-8176-4725-4_2,

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

Clearly, we get

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

Note that when b is the mean interoccurrence time, $1/b$ is the rate of occurrence of the event.

In considering the suitability of a probability model for a random phenomenon, the moment properties of the model distribution become useful. Often 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., using s.d. to represent standard deviation, coefficient of variation = 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 specified time epochs or interarrival times are of constant length); exponential (as distribution models for interarrival times or service times); Poisson (as the distribution of the number of arrivals during a specified length of time); Erlang (as distribution models for interarrival 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.

2.1.1 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 and late or early 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 that the occurrence of the kth scheduled event can be later than the occurrence of the $(k+1)$th scheduled event.

2.1.2 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

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

and

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

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

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

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

Let

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

where $F_n(t) = P(Z_1 + Z_2 + \cdots + Z_n \leq t)$ 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

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

$$- \left[1 - \sum_{r=0}^{n} e^{-\lambda t} \frac{(\lambda t)^r}{r!} \right]$$
$$= e^{-\lambda t} \frac{(\lambda t)^n}{n!}, \tag{2.1.4}$$

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

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

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

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 interoccurrence 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 $P(Z_n \le x) = 1 - e^{-\lambda x}$ $(\lambda > 0)$

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

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 (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 interoccurrence times, then

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

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. (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, the Poisson process is considered an appropriate model for events occurring "at random." The reasons for such a characterization rest 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 i.i.d. nature of the interoccurrence times, and its relationship with the uniform distribution as expressed in (A.1.4) in 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 interarrival time between effective arrivals to the queueing system is the sum of k exponential random variables with mean $1/\lambda$, and hence it has the distribution given by (2.1.6) with n replaced by k. 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 $1/\lambda$ and scale parameter k. This can be accomplished by starting with an exponential distribution with parameter $k\lambda$. Then we get

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

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

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

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 λ, 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 interarrival 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 rates) 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 $1 - \rho$. See also 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 has 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 Cox (1962), p. 86, 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 general and 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 may 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 (1955). 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 Xs and Ys, 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 $M/G/\infty$ queue, where G is assumed to be known.

2.2.3 Tests for Independence

While formulating a queueing model, for simplification and convenience, several assumptions of independence are made about its elements. Thus most of the models assume that interarrival times and service times are independent sequences of i.i.d. 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 interarrival and service times, nonparametric tests seem appropriate. Spearman's rho and Kendall's tau (Conover (1971), Randles and Wolfe (1979)) are used to test for 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.

2.2.4 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 the data and the 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 the analysis under such assumptions is considerably similar.

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 the 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, the Erlang with a coefficient of variation ≤ 1 and the hyperexponential with a 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 and 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.

2.3 Review Exercises

1. Determine the mean, variance, and coefficient of variation (CV) for the following distributions introduced in this chapter and Appendix A:
   (a) deterministic, (2.1.1);
   (b) exponential, (2.1.2);
   (c) hyperexponential, (A.3.1);
   (d) Erlang, (2.1.6), (A.4.1);
   (e) mixed Erlang, (A.5.1), (A.5.2);
   (f) geometric;
   (g) binomial;
   (h) negative binomial.
2. Determine the Laplace transform or the PGF, as the case may be, for the distributions listed in Exercise 1.
3. Determine the PGF for
   (a) the Poisson process;
   (b) the compund Poisson process.
4. Redo Exercise 1 using the Laplace transform or PGFs, as the case may be.
5. Determine for a specific value of t, the mean, variance, and coefficient of variation for a
   (a) Poisson process;
   (b) 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 in Exercise 1.

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 time by $t \in T$ when it is continuous and by $n \in T$ when it is discrete. Then the family of random variables $\{X(t),\ t \in T\}$ or the sequence of random variables $\{X_n,\ n \in T\}$ is known as a *stochastic process*. (A sample value of a random variable can be thought of 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 this 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 products 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 embedded 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 normally used in the analysis of queueing models we do not provide any information on them in our discussion.

3.2 Point, Regenerative, and Renewal Processes

Point process. Consider a randomly located discrete set of points in the parameter space T. These points may represent events such as arrivals in a queueing system or

U.N. Bhat, *An Introduction to Queueing Theory*, DOI: 10.1007/978-0-8176-4725-4_3,

accidents on a stretch of road. Let $N(t), t \in T$ 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. Let us consider a stochastic process $\{X(t),\ t \in T\}$ and a discrete set of points $t_1 < t_2 < \cdots < t_n \in T$. 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 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 $(0, t]$. This 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.

In the context of queueing systems, when the interarrival times are i.i.d., the arrivals form a renewal process. But, since a departure cannot take place when there are no customers in the system, the departure process is not renewal even when service times are i.i.d. random variables. They form a renewal process only during the period when customers are continuously busy. (Periods when customers are continuously busy are known as *busy periods*. They are followed by *idle periods* during which the server is idle.) 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 another set of renewal points, with the sequence of busy period-idle period pairs forming the renewal periods. The renewal process framework is useful in analyzing some advanced classes of queueing systems.

3.3 Markov Process

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

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

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 will 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 value of 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 notation depending on the nature of state and parameter spaces:

(i) Discrete state, discrete parameter:

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

(ii) Discrete state, continuous parameter:

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

(iii) Continuous state, discrete parameter:

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

(iv) Continuous state, continuous parameter:

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

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)
$$P_{ij}^{(m,n)} = \sum_{k \in S} P_{ik}^{(m,r)} P_{kj}(r,n), \qquad m < r < n; \tag{3.3.7}$$

(ii)
$$P_{ij}(s,t) = \sum_{k \in S} P_{ik}(s,u) P_{kj}(u,t), \qquad s < u < t; \tag{3.3.8}$$

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

(iv)
$$F(x_s, x; s, t) = \int_{y \in S} d_y F(x_s, y; s, u) \cdot F(y, x; u, t), \qquad s < u < t. \tag{3.3.10}$$

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 $(k \in S)$ in time period (m, r) is $P_{ik}^{(m,r)}$ and from state k to state j in time period (r, n) is $P_{kj}^{(r,n)}$. (3.3.7) now follows by multiplying these two probabilities and summing over all values of $k \in S$. 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 (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 $P_{ij}^{(m+k,n+k)}$ 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 $\mathbf{P} = \mathbf{P}^{(1)}$ 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., $\sum_{j\in S} P_{ij}^{(n)}$) are equal to 1 for all values of n. With these notational simplifications, (3.3.7) can be written as

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

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 a time-homogeneous Markov process in which transition probabilities $P_{ij}(s, t)$ and $P_{ij}(s + u, t + u)$ 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 $i, j \in S$ 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, note the following properties, which are either obvious or assumed:

(a) $P_{ij}(t) \geq 0$;
(b) $\sum_{j \in S} P_{ij}(t) = 1$;
(c) $P_{ij}(s + t) = \sum_{k \in S} P_{ik}(s) P_{kj}(t)$;
(d) $P_{ij}(t)$ is continuous;
(e) $\lim_{t \to 0} P_{ij}(t) = 1$ if $i = j$ and $= 0$ otherwise.

Note that properties (a) and (b) are obvious from the transition structure and (c) is a restatement of the Chapman–Kolmogorov relation. Properties (d) and (e) are necessary (hence assumed) for deriving the differential equations.

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

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

Setting $t = 0$

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

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 $\sum_{j \in S} P_{ij}(t) = 1$. These transition rates are also known as *generators*, displayed in a matrix as

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

In continuous-time Markov processes with discrete states, the generator matrix **A** plays the same role in its analysis as that of the transition probability matrix **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 $\{X(t),\ t \in T\}$ be a Poisson process with parameter λ, such that

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

(See B.2.2 in 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 these relations as

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

Thus the analysis of the behavior of a queueing system that 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.

4

Simple Markovian Queueing Systems

Poisson arrivals and exponential service enable us to use Markovian queueing models that are easy to analyze and that produce usable results. Historically, these have also been the models used in the early stages of queueing theory to help decision making in the 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 increases and decreases 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

Again using the birth (arrival)–death (departure) terminology, when the 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 in the system, λ_n and μ_n indicate that the arrival and service rates depend on the number in the system. Based on the properties of the Poisson process, i.e., when arrivals are in a Poisson process and service times are exponential, we can make the following probability statements for a transition during $(t, t + \Delta t]$:

birth $(n \geq 0)$:

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

death $(n > 0)$:

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

U.N. Bhat, *An Introduction to Queueing Theory*, DOI: 10.1007/978-0-8176-4725-4_4,

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 the $o(\Delta t)$ terms do not specify actual values. In each of the two cases, the $o(\Delta t)$ 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

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

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) leads to the following forward Kolmogorov equations for $P_{in}(t)$; see equations (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.)

$$\begin{aligned} P'_0(t) &= -\lambda_0 P_0(t) + \mu_1 P_1(t), \\ P'_n(t) &= -(\lambda_n + \mu_n) P_n(t) + \lambda_{n-1} P_{n-1}(t) \\ &\quad + \mu_{n+1} P_{n+1}(t), \quad n = 1, 2, \ldots. \end{aligned} \tag{4.1.3}$$

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

$$\begin{aligned} P_0(t + \Delta t) &= P_0(t)[1 - \lambda_0 \Delta t + o(\Delta t)] + P_1(t)[\mu_1 \Delta t + o(\Delta t)], \\ P_n(t + \Delta t) &= P_n(t)[1 - \lambda_n \Delta t - \mu_n \Delta t + o(\Delta t)] \end{aligned}$$

$$
\begin{aligned}
&+ P_{n-1}(t)[\lambda_{n-1}\Delta t + o(\Delta t)] \\
&+ P_{n+1}(t)[\mu_{n+1}\Delta t + o(\Delta t)] \\
&+ o(\Delta t), \quad n = 1, 2, \ldots.
\end{aligned} \tag{4.1.4}
$$

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 condition $P_i(0) = 1$, $P_n(0) = 0$ 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* $\{p_n,\ n \in \mathbf{S}\}$ *are such that they either vanish identically* (*i.e.,* $p_n = 0$ *for all* $n \in \mathbf{S}$) *or are all positive and form a probability distribution* (*i.e.,* $p_n > 0$ *for all* $n \in \mathbf{S}$, $\sum_{n\in\mathbf{S}} p_n = 1$).

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

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

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

and therefore

$$P'_n(t) \to 0 \quad \text{as } t \to \infty.$$

Using these results in (4.1.3), we get

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

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

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

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

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

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

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

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

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

and

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

For the birth-and-death queueing model, the generator matrix **A** is given by (4.1.2).

Another way of looking at (4.1.5) is to consider the equations 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 4.1.1.

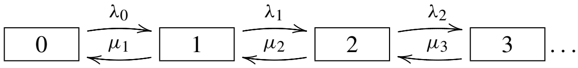

Fig. 4.1.1. Transition diagram.

Noting that the λs and μs represent infinitesimal transition rates in and out of the states, the equalities in (4.1.12) can be interpreted as (long-term probability of being in state n) × (transition rates out of state n) $= \sum_{i=n-1,n+1}$ (long-term probability of being in state i) × (transition rate from state i to state n).

Such state balance equations can be easily written using the transition diagram of Figure 4.1.1.

Thus we have given three ways of determining the state balance equations:

1. taking the appropriate limits as $t \to \infty$ in the forward Kolmogorov equations;
2. using the equation $\mathbf{pA} = 0$; and
3. with the help of the transition diagram.

When using the last method care should be taken to ensure that all transitions have been accounted for. Also, in applications the readers may use any method with which they are comfortable. 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 that 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

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

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 that 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

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

It also means that the interarrival 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

$$E[\text{interarrival time}] = \frac{1}{\lambda} = \frac{1}{\text{arrival rate}},$$
$$E[\text{service time}] = \frac{1}{\mu} = \frac{1}{\text{service rate}}.$$

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 $\mu_n = \mu$. The generator matrix is given by (state space: 0, 1, 2, ...)

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

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

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

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 and Harris (1998), p. 129, 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))

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

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

$$p_n = (1-\rho)\rho^n, \quad n = 0, 1, 2, \ldots, \quad (4.2.4)$$

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 the 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

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

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 $F_q(t) = P(T_q \le t)$, 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 $dF_q(t) = P(t < T_q \le t + dt)$, 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. Numerous 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 the queue $G/G/1$ in Chapter 9 (Section 9.2).

Busy period. A *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 interarrival time, after the last customer in the busy period leaves after service. With an exponential interarrival 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, the busy period is the duration of time that 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 (4.2.1) 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 $P_n(t)$ $(n = 0, 1, 2, \dots)$ are

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

with the initial condition $P_1(0) = 1$, $P_n(0) = 0$ for $n \neq 1$. Solving these difference-differential equations requires the use of PGFs and Laplace transforms. (See Gross and Harris (1998).)

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

$$\pi_0(\theta) = \int_0^\infty e^{-\theta t} P_0'(t)dt, \quad \text{Re}(\theta) > 0.$$

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 Section C.2 in Appendix C.)

$$\phi(\theta) = \int_0^{\infty} e^{-\theta x} f(x)dx, \quad \text{Re}(\theta) > 0.$$

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 properties of the arrival process we can show that the transition of the underlying Markov process from i to 0 can be considered to be 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

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

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 $\rho \le 1$ 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 the departures as well. Nevertheless, in equilibrium we can derive 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 note here, as we shall see in Chapter 5, that 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 of the value of x, we have

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

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 interdeparture 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 that we have to consider the possibility of only arrivals during $(x, x + \Delta x]$. We have

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

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

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

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 that

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

Hence

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

Now using the initial condition (4.2.34) to determine C, we get

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

The general solution to (4.2.33) can be obtained by induction. Let

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

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 (4.2.34). We get

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

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 interarrival times. Since $\{p_n\}$ is also the distribution of the number of customers in the system at departure points, equation (4.2.36) 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 multiserver $M/M/s$ queue.

The important result from 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:

$$\begin{aligned} \text{arrival rate} &= 15/\text{hour } (\lambda), \\ \text{service rate} &= \frac{60}{3}/\text{hour} = 20/\text{hour } (\mu), \\ \text{utilization} &= \rho = \frac{\lambda}{\mu} = \frac{3}{4}. \end{aligned} \qquad \textsc{Answer}$$

(b) Expected number of airplanes waiting to land:

$$L_q = \frac{\rho^2}{1-\rho} = \frac{(0.75)^2}{0.25} = 2.25. \qquad \textsc{Answer}$$

(c) Expected waiting time:

$$E(W_q) = \frac{\lambda}{\mu(\mu-\lambda)} = \frac{15}{20(20-15)} = \frac{3}{20}\text{ hour} = 9\text{ minutes.} \qquad \textsc{Answer}$$

(d) Probability that the waiting will be more than 5 minutes? 10 minutes? No waiting?

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

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

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

$$= \frac{3}{4} e^{-\frac{25}{60}} = 0.4944, \qquad \textsc{Answer}$$

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

(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 multiserver 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, and service times are assumed to have an exponential distribution. We will let the number of servers be s, providing service independently of each other. We also assume that the arriving customers form a single queue and the one at the head of the waiting line enters 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 interarrival 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 one of the n busy servers will complete service during $(t, t + \Delta t]$ can be determined using the binomial probability distribution as

$$= \binom{n}{1} [\mu\Delta t + o(\Delta t)] [1 - \mu\Delta t + o(\Delta t)]^{n-1}$$

$$= n\mu\Delta t + o(\Delta t). \tag{4.3.1}$$

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

$$= \binom{n}{r} [\mu\Delta t + o(\Delta t)]^r \, [1 - \mu\Delta t + o(\Delta t)]^{n-r}$$

$$= o(\Delta t).$$

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

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

The generator matrix **A** for the process can be given as

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

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 as specializations of (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 $\mathbf{pA} = 0$)

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

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

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

(See also (4.1.7).) Therefore,

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

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

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

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

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

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

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

we may say that when the number of customers in the system is $\ge s$, 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

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

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

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

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.

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 (4.3.6), we get (writing $s\rho = \alpha$ when convenient)

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

Note that the terms inside [ ] above $= p_0^{-1}$ (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 is 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 example, 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 interdeparture 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 multiserver 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 multiserver queue requires further elaboration. If the busy period is the time during which arriving customers have to wait for service, in a multiserver 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 in (4.2.18), but because of the varying service rates, the equations are 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 notation as before, we get

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

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; see 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:

$$\begin{aligned} \text{arrival rate} &= 15/\text{hour } (\lambda), \\ \text{service rate} &= 20/\text{hour } (\mu), \\ \text{number of servers} &= 2\ (s), \\ \text{utilization of each runway} &= \rho = \frac{\lambda}{s\mu} = \frac{3}{8}. \end{aligned}$$

Answer

(b) Expected number of airplanes waiting to land:

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

(note that $\alpha = s\rho = \frac{3}{4}$),

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

Answer

(c) Expected waiting time:

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

Answer

(d) Probability that the waiting will be more than 5 minutes? 10 minutes? No waiting?

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

(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 hour, 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.

(a) 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. The results are listed in Table 4.3.1.

Table 4.3.1. Results from Example 4.3.2(a).

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

(b) What is the effect of operating the two queues as a two-server queue with arrival rate 18/hour and service rate 18/hour? What conclusion can you draw from this operation? See Table 4.3.2.

Table 4.3.2. Results from Example 4.3.2(b).

	Two-server queue	
λ	18/hour	
μ	18/hour	
Number of servers (s)	2	
$\rho = \dfrac{\lambda}{s\mu}$	0.5	
$\alpha = \dfrac{\lambda}{\mu}$	1	
$p_0 = \left[\sum_0^1 \dfrac{\alpha^r}{r!} + \dfrac{\alpha^2}{2(1-\rho)} \right]^{-1}$	0.4	
$L_q = \dfrac{\rho\alpha^2 p_0}{2(1-\rho)^2}$	0.4	ANSWER
$W_q = \dfrac{\alpha^2 p_0}{(2)2\mu(1-\rho)^2}$	1.33 minutes	ANSWER

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

Incidentally, the efficiency of multiserver queues over single-server systems is the reason that multiserver service systems, whenever possible, use single waiting lines feeding multiple counters for service. Airline check-in counters and 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 a 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:

$$\lambda_n = \lambda, \qquad n = 0, 1, 2, \ldots, K-1,$$

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

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:

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

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

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

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

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

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 probability that an arriving customer will join the system is given by $\frac{p_n}{1-p_K}$. Thus, with the notation used earlier for the distribution for the waiting time, we have

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

where

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

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

$$W_q = \int_0^\infty t dF_q(t) = \sum_{n=s}^{K-1} \frac{p_n}{1-p_K} \int_0^\infty e^{-s\mu t} \frac{(s\mu t)^{n-s}}{(n-s)!} s\mu t dt$$

$$= \frac{1}{s\mu(1-p_K)} \sum_{n=s}^{K-1} (n-s+1) p_n. \tag{4.4.6}$$

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) $M/M/1/K$ and
(ii) $M/M/s/s$.

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,

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

$$W_q = \frac{1}{\mu}\left[\frac{\rho}{1-\rho} - \frac{K\rho^K}{1-\rho^K}\right], \qquad \rho \neq 1,$$
$$= \frac{1}{2\mu}(K-1), \qquad \rho = 1, \tag{4.4.13}$$
$$W = \frac{1}{\mu}\left[\frac{1}{1-\rho} - \frac{K\rho^K}{1-\rho^K}\right], \qquad \rho \neq 1,$$
$$= \frac{1}{2\mu}(K+1), \qquad \rho = 1, \tag{4.4.14}$$
$$L_q = \frac{\rho}{1-\rho} - \frac{\rho(1+K\rho^K)}{1-\rho^{K+1}}, \qquad \rho \neq 1,$$
$$= \frac{K(K-1)}{2(K+1)}, \qquad \rho = 1, \tag{4.4.15}$$
$$L = \frac{\rho(1-\rho^K)}{(1-\rho)(1-\rho^{K+1})} - \frac{K\rho^{K+1}}{1-\rho^{K+1}}, \qquad \rho \neq 1,$$
$$= \frac{K}{2}, \qquad \rho = 1. \tag{4.4.16}$$

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 one 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 minutes ($1/\mu$), we have $\rho = 2.5$. Also, $K = 3$. Using the first result from (4.4.15), we get

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

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

$$W_q = 1.4778 \text{ minutes.}$$
Answer

(b) What is the probability that the call has to wait for more than 1.5 minutes before being 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

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

(c) Because of the excessive waiting time, 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 $\alpha = 2.5$, $\rho = 1.25$, and $s = 2$ and $K = 3$. We get

$$p_0 = 0.0950, \qquad p_1 = 0.2374,$$
$$p_2 = 0.2969, \qquad p_3 = 0.3711.$$

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

$$W_q = 0.5902 \text{ minute;}$$ **Answer**
$$L_q = \lambda(1 - p_3)W_q = 0.3712;$$ **Answer**

$P(\text{wait in queue} > 1.5 \text{ minutes})$:

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

$$p_0 = 0.0649, \qquad p_1 = 0.1622,$$
$$p_2 = 0.2028, \qquad p_3 = 0.2535,$$
$$p_4 = 0.3169,$$
$$W_q = 1.2989 \text{ minutes;}$$ **Answer**
$$L_q = 0.8873;$$ **Answer**

$P(\text{wait in queue} > 1.5 \text{ minutes})$:

$$1 - F_q(1.5) = 0.3353.$$ **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 the truncated version of (4.3.3):

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

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

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

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

$$\begin{aligned} p_0 &= \left[1 + \alpha + \frac{\alpha^2}{2!} + \cdots + \frac{\alpha^s}{s!}\right]^{-1}, \\ p_n &= \frac{\alpha^n}{n!} p_0, \quad n = 0, 1, \ldots, s. \end{aligned} \tag{4.4.19}$$

This gives

$$p_s = \frac{\frac{\alpha^s}{s!}}{1 + \alpha + \frac{\alpha^2}{2!} + \cdots + \frac{\alpha^s}{s!}}, \tag{4.4.20}$$

which is the probability that a customer is blocked from entering the system. (Telephones calls are lost.) Also, λp_s gives the expected number of customers who will be blocked from entering the system in unit time. (4.4.20) is commonly known as *Erlang's loss formula* or *Erlang's first formula* and denoted as $E_{1,s}(\alpha)$ 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. Note 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 is lost to the system.

The right-hand side expression in 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 an *infinite-server queue* (an infinite number of servers and consequently no waiting line) 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 then get

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

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

$$\boldsymbol{A} = \begin{matrix} 0 \\ 1 \\ 2 \\ \vdots \end{matrix} \begin{bmatrix} -\lambda & \lambda & & & \\ \mu & -(\lambda+\mu) & \lambda & & \\ & 2\mu & -(\lambda+2\mu) & \lambda & \ldots \\ & & \vdots & & \vdots \end{bmatrix}. \tag{4.5.2}$$

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

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

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

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

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 in a linearly dependent service rate $n\mu$ when there are $n\mu$ customers in the system. Note 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 $\lambda\Delta t + o(\Delta t)$. 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 $k\lambda\Delta t + o(\Delta t)$ (see the discussion preceeding (4.3.1)). When the service times are exponential, we are now 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 ($s \le M$) 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 $M/M/s/s$ 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 the state balance equations as follows:

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

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

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

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

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

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

$$\text{machine availability} = 1 - \frac{L}{M},$$
$$\text{operative utilization} = \sum_{n=0}^{s-1} \frac{np_n}{s} + \sum_{n=s}^{M} p_n.$$

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 get closed-form expressions for the mean number of inoperative machines in the system. 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

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

We get

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

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

Illustrative values of machine availability and operative utilization for different values of α and M are tabulated in Bhat (1984), p. 394. An obvious conclusion we can draw from them is that it is better to use repairmen in a pool, rather than assigning a certain number of machines to each of them.

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 the 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 as

$$M\lambda p_0 = \mu p_1,$$
$$[(M-n)\lambda + n\mu]p_n = (M-n+1)\lambda p_{n-1} + (n+1)\mu p_{n+1}, \quad 1 \le n < s,$$

$$s\mu p_s = (M - s + 1)\lambda p_{s-1}. \tag{4.6.7}$$

Solving these equations recursively, we get

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

and therefore

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

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. 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 the $M/M/s/s$ 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 service times are not exponential.

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

Although 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 a two-state Markov process useful in a large number of applications.

Let customers arrive in a Poisson process with parameter λ and be 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, $\{Q(t),\ t \in T\}$ is a Markov process with the generator matrix

$$\boldsymbol{A} = \begin{array}{c} \\ 0 \\ 1 \end{array} \begin{array}{c} \begin{array}{cc} 0 & 1 \end{array} \\ \begin{bmatrix} -\lambda & \lambda \\ \mu & -\mu \end{bmatrix} \end{array}. \tag{4.7.1}$$

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

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

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

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

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

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

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 that there are N such multiple processes undergoing transitions between alternate states independently of each other with the transition structure as described above. The probability distribution of the number of processes in state 0 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 several 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)

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

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)

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

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

$$\begin{aligned} p_0 &= \left[1 + \frac{\lambda}{\mu} + \frac{1}{2}\left(\frac{\lambda}{\mu}\right)^2 + \frac{1}{3!}\left(\frac{\lambda}{\mu}\right)^3 + \cdots\right]^{-1} \\ &= e^{-\rho}, \\ p_n &= \frac{1}{n!}\left(\frac{\lambda}{\mu}\right)^n \rho_0 \\ &= e^{-\rho}\frac{\rho^n}{n!}, \quad n = 0, 1, 2, \ldots, \end{aligned} \tag{4.7.13}$$

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

(iii)

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

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

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

(iv)

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

The arrival rate here incorporates a fraction that reflects an estimate of the 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

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

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

4.7.3 Markovian Queues with Reneging

After joining the queue, if a customer abandons its desire to be 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 events. 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

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

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

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

with $\rho < 1$.

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 that 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:

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

Let $p_n = (p_0, p_1, p_2, \dots, p_k)$ 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

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

We note that the overall repair time has a generalized Erlang distribution of (A.6.1) 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 the 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 usable 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 representations of real systems, and starting with a Markovian model provides a good starting point for an 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

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

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 minutes. The coach-class passengers arrive at the rate of 18 per hour and the 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:

(a) the distribution of the number of customers picking up merchandise and its mean;
(b) the mean length of time the customers wait at the counter for service;
(c) 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 average.
   (a) Determine the limiting distribution of the number of customers waiting for taxis.
   (b) What is the probability that there are taxis waiting for customers?
   (c) Determine the mean waiting time for a customer.

6. An automobile service station has one station for oil and filter changes. On 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 on 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 changes.

7. Customer arrivals to a service counter are 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 of 0.10 or less.

8. In a manufacturing process production machines break down at the rate of 3 per hour. We may assume that the process of breakdowns is Poisson. The repair times of the 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/hour and 5 per hour at a cost of \$30/hour. Considering the loss of productivity of the machines while they are either waiting for service or being in service, what is the minimum rate of productivity gain that 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 average each customer spends 15 minutes in the store, and we assume that the time the customer spends in the store to have an exponential distribution. Currently, the store provides parking spaces for 15 cars. Overflow cars from the parking lot park elsewhere in the neighborhood. 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 that the arrival and the service rates in Exercise 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 minutes. What is the probability that the calls will not have to wait for more than 2 minutes before getting to the receptionist?

    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 minutes with a 90% probability. Determine K.

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,

$$\rho = \frac{\text{arrival rate}}{\text{service rate}},$$

    $P_B =$ probability that an arriving customer is blocked from entering the system.

13. In a drugstore, customers arrive at the counter (with one server per counter) in a Poisson process at the rate of 48/hour. The service time can be assumed to be exponential with an average of 1 minute. The service is provided by one or more servers 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 customers	4 counters.

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

    Determine the following:
    (a) What is the 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 atmospheric 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 is diffused in an average time of length β. Also assume that the diffusion times are exponential random variables that are i.i.d. Obtain the mean and variance of $A(t)$ as $t \to \infty$.

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

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

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

(b) Let λ^* be the effective arrival rate of machines for repair. Noting that λ^* can also be expressed 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 an FCFS basis and wait in a queue when the line is busy. It has been observed that the data entry job takes on average 100 seconds, and once the terminal is free, it is ready for the next job in 5 seconds on 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 $s = 3$:

(a) Three separate systems.

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

(c) The arrivals are pooled as in (b). 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 ($i = a, b, c$). Show that

$$W_a > W_b > W_c.$$

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 means $1/\mu_1$ and

$1/\mu_2$, 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, which 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 $1/(R\mu_L)$.

    The data field size per buffer is b. Let N be the random variable representing the number of buffers in 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 (Pedersen 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 file drum, described in Exercise 10 of Chapter 1, be connected to a CPU 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 Exercise 10 of Chapter 1, and let 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 multiprogamming) (Fuller (1980)).

21. A simplified model of the drum storage unit described in Exercise 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 $t + \Delta t$ with 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 $p_n, n = 0, 1, 2, \dots, M$.

(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 Exercise 23 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)).

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. Then if the arrival rate is λ and the service rate is μ, at any time point t, the time to the next arrival has 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 nonexponential, 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 are served by a single server. We assume that service times of customers are 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 one is 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., arrival and service times, are completely specified. The earliest investigation to analyze this vector process by itself was performed by Cox (1955), who used information on $R(t)$ as a supplementary variable in constructing the forward Kolmogorov equations given in Chapter 3. Since this method employs analysis techniques beyond the scope of this text, we shall not cover it here.

U.N. Bhat, *An Introduction to Queueing Theory*, DOI: 10.1007/978-0-8176-4725-4_5,

In two papers in the 1950s, 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, and 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. Because of 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 the $M/G/1$ and $G/M/1$ 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.

5.2 The Queue $M/G/1$

Let customers arrive in a Poisson process with parameter λ and are 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. Q_{n+1} 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

$$P_{ij}^{(n)} = P(Q_n = j|Q_0 = i), \quad i, j \in S, \tag{5.2.3}$$

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 **P** for the Markov chain is

$$\boldsymbol{P} = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ 3 \\ \ddots \end{array} \begin{array}{c} \begin{array}{cccc} 0 & 1 & 2 & \ldots \end{array} \\ \left[\begin{array}{cccc} k_0 & k_1 & k_2 & \ldots \\ k_0 & k_1 & k_2 & \ldots \\ & k_0 & k_1 & \ldots \\ & & k_0 & \ldots \\ & & & \ldots \end{array}\right] \end{array}. \tag{5.2.5}$$

For the Markov chain to be irreducible (the state space has a single equivalence class) 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 $k_0 + k_1 = 1$, only the two states $\{0, 1\}$ are possible in the system. (If the system starts with $i > 1$ customers, once it attains 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 PGF of the number of customers arriving during a service time,

$$K(z) = \sum_{j=0}^{\infty} k_j z^j, \quad |z| \le 1. \tag{5.2.7}$$

The results below are derived 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

$$\begin{aligned} K'(z) &= -\lambda \psi'(\lambda - \lambda z), \\ K'(1) &= -\lambda \psi'(0) \\ &= \lambda b = \rho. \end{aligned} \tag{5.2.9}$$

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 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 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. See Appendix B for the classification of states in Markov chains.

Recalling the result derived in (3.3.13), the n-step transition probabilities $P_{ij}^{(n)}$ $(i, j = 0, 1, 2, \dots)$ 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 involve mathematical techniques 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 (1963a) look at the transitions of Q_n as some first passage problems and use combinatorial methods in solving them. (See also 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 a 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 $\lim_{n\to\infty} P^n$ becoming a matrix with identical rows. Computationally, this property can be validated by obtaining 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 its 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*

$$\lim_{n\to\infty} P_{ii}^{(n)} = \frac{1}{\mu_i} = \pi_i, \quad \textit{say}.$$

(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 a Markov chain takes to return for the first time to the starting state. 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 $\pi_i = \lim_{n\to\infty} P_{ji}^{(n)}$, $i, j \in S$. 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 the discussion preceding (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

$$\begin{aligned} \boldsymbol{\pi}\mathbf{P} &= \boldsymbol{\pi} = \boldsymbol{\pi}\mathbf{P}^2, \\ \boldsymbol{\pi}\mathbf{P} &= \boldsymbol{\pi} = \boldsymbol{\pi} P^n. \end{aligned} \tag{5.2.11}$$

The last equation shows that if we use the limiting distribution as the intial 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 $\boldsymbol{\pi} = (\pi_0, \pi_1, \pi_2, \dots)$ be the limiting distribution of the imbedded chain. Using the transition probability matrix (5.2.5) in the equation $\boldsymbol{\pi P} = \boldsymbol{\pi}$ (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

$$\nu_0 \equiv 1 \quad \text{and} \quad \nu_i = \pi_i/\pi_0$$

and rewrite (5.2.14) in terms of ν_i $(i = 1, 2, \dots)$ as

$$\begin{aligned} \nu_1 &= \frac{1-k_0}{k_0}, \\ \nu_2 &= \frac{1-k_1}{k_0}\nu_1 - \frac{k_1}{k_0}, \\ &\vdots \\ \nu_j &= \frac{1-k_1}{k_0}\nu_{j-1} - \frac{k_2}{k_0}\nu_{j-2} - \cdots - \frac{k_{j-1}}{k_0}\nu_1 - \frac{k_{j-1}}{k_0}, \\ &\vdots \end{aligned} \tag{5.2.15}$$

These equations can be solved recursively to determine ν_i $(i = 1, 2, \dots)$. The limiting probabilities $(\pi_0, \pi_1, \pi_2, \dots)$ 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_is from ν_is, 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 incorporated the normalizing condition $\sum_0^\infty \pi_i = 1$. Thus we get

$$\pi_0 = (1 + \sum_{i=1}^{\infty} \nu_i)^{-1}$$

and

$$\pi_i = \frac{\nu_i}{1 + \sum_{i=1}^{\infty} \nu_i}. \tag{5.2.16}$$

Analytically, the limiting distribution $(\pi_0, \pi_1, \pi_2, \dots)$ can be determined by solving equations (5.2.14) using generating functions. Unfortunately, deriving explicit expressions for the probabilities requires inverting the resulting PGF. However, we can obtain the mean and variance of the distribution using standard techniques. Define

$$\Pi(z) = \sum_{j=0}^{\infty} \pi_j z^j \quad |z| \le 1.$$

and

$$K(z) = \sum_{j=0}^{\infty} k_j z^j \quad |z| \le 1.$$

Multiplying equations (5.2.14) with appropriate powers of z and summing, we get

$$\begin{aligned} \Pi(z) &= \pi_0 K(z) + \pi_1 K(z) + \pi_2 z K(z) + \cdots \\ &= \pi_0 K(z) + \frac{K(z)}{z}(\pi_1 z + \pi_2 z^2 + \cdots) \\ &= \pi_0 K(z) + \frac{K(z)}{z}[\Pi(z) - \pi_0]. \end{aligned}$$

Rearranging terms,

$$\Pi(z)\left[1 - \frac{K(z)}{z}\right] = \pi_0 K(z)\left[1 - \frac{1}{z}\right]$$

$$\Pi(z) = \frac{\pi_0 K(z)(z-1)}{z - K(z)}. \tag{5.2.17}$$

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

$$1 = \frac{\pi_0}{1-\rho},$$

$$\pi_0 = 1 - \rho. \tag{5.2.18}$$

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 $\sum_{j=0}^{\infty} z^j (k_{j+1} + k_{j+2} + \cdots)$ can be simplified to

$$\frac{1 - K(z)}{1 - z} = C(z), \quad \text{say.}$$

(See also the algebraic simplifications leading to (5.2.17).)

For $|z| \leq 1$,

$$|C(z)| = \left|\frac{1 - K(z)}{1 - z}\right| < 1 \quad \text{if } \rho < 1. \tag{5.2.21}$$

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 A. See Wolff (1982), who coined the acronym PASTA (Poisson arrivals see time averages). For proofs of this property, see also Cooper (1981) and Gross and Harris (1998).

Therefore, the PGF $\Pi(z)$ derived in (5.2.19) 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 those of 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

$$\begin{aligned} \pi_0 &= 1-\rho, \\ \pi_j &= (1-\rho)\int_0^\infty e^{-\lambda t}\sum_{n=0}^\infty \left[\frac{(\lambda t)^{n+j-1}}{(n+j-1)!} - \frac{(\lambda t)^{n+j}}{(n+j)!}\right] dB_n(t) \end{aligned} \tag{5.2.23}$$

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

$$\begin{aligned} L = E(Q^*) &= \Pi'(1), \\ V(Q^*) &= \Pi''(1) + \Pi'(1) - [\Pi'(1)]^2. \end{aligned} \tag{5.2.24}$$

Differentiating $\Pi(z)$ with respect to 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

$$\Pi'(1) = \frac{2K'(1)[1-K'(1)] + K''(1)}{2[1-K'(1)]}. \tag{5.2.25}$$

But note that from (5.2.9), $K'(1) = \rho$ and

$$\begin{aligned} K''(1) &= \lambda^2\psi''(0) \\ &= \lambda^2 E(S^2), \end{aligned} \tag{5.2.26}$$

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 the *Pollaczek–Khintchine formula*. Noting that ρ is the expected number in service (which is the same as the probability that 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 $\Pi''(1)$ yield

$$\begin{aligned} V(Q^*) = \rho(1-\rho) &+ \frac{\lambda^2 E(S^2)}{2(1-\rho)}\left[3 - 2\rho + \frac{\lambda^2 E(S^2)}{2(1-\rho)}\right] \\ &+ \frac{\lambda^3 E(S^3)}{3(1-\rho)} \end{aligned} \tag{5.2.29}$$

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 constant (in the queue $M/D/1$)

$$E(Q^*) = \rho + \frac{\rho^2}{2(1-\rho)} = \frac{\rho}{1-\rho}\left(1 - \frac{\rho}{2}\right). \tag{5.2.31}$$

On the other hand, when the service time distribution is Erlang with mean $1/\mu$ and scale parameter k (i.e., by writing $\lambda = \mu$ in (2.1.8)), we get $\sigma_S^2 = 1/(k\mu^2)$ 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., (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 the system time or time in system, and let T_q be the actual waiting time, both as $t \to \infty$. Let $E(T) = W$ and $E(T_q) = W_q$. Also, 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,

$$\begin{aligned} \Pi(z) &= \sum_{n=0}^{\infty} P(Q^* = n) z^n = \sum_{n=0}^{\infty} z^n \int_0^\infty e^{-\lambda t} \frac{(\lambda t)^n}{n!} dF(t) \\ &= \int_0^\infty e^{-\lambda t} \sum_{n=0}^{\infty} \frac{(\lambda t z)^n}{n!} dF(t) \\ &= \Phi(\lambda - \lambda z). \end{aligned} \tag{5.2.34}$$

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 $z = 1 - \frac{\theta}{\lambda}$;

$$\begin{aligned} \Phi(\theta) &= \frac{(1-\rho)\frac{\theta}{\lambda}\psi(\theta)}{\psi(\theta) - (\lambda - \theta)/\lambda} \\ &= \frac{(1-\rho)\theta\psi(\theta)}{\theta - \lambda[1 - \psi(\theta)]}. \end{aligned} \tag{5.2.36}$$

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 i.i.d. 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 known as the *forward recurrence time* (also known as *excess life* in the terminology of *reliability theory*). If $B(\cdot)$ is the distribution function of Z_n, it is possible to show, as $t \to \infty$, that $r_t(x)$, the density function of $R(t)$, can be given as

$$\lim_{t\to\infty} r_t(x) = \frac{1}{E[Z_n]}[1 - B(x)]. \tag{5.2.40}$$

For a more detailed description, see Chapter 8.

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 $R(t)$ (forward recurrence time) 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 example, we have

$$W = E(T) = \Phi'(0),$$
$$\sigma_T^2 = V(T) = \Phi''(0) - [\Phi'(0)]^2,$$

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. Let

$$g_i^{(n)} = P[B_i = n], \quad n = 1, 2, \ldots. \tag{5.2.45}$$

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. Consequently, $g_i^{(n)}$ is the i-fold convolution of $g_1^{(n)}$ with itself. Thus for the PGF of $g_i^{(n)}$,

$$G_i(z) = \sum_{n=i}^{n} g_i^{(n)} z^n = [G_1(z)]^i. \tag{5.2.46}$$

Noting that the busy period cannot end before the ith transition of the Markov chain, we have

$$g_i^{(i)} = k_0^{(i)},$$
$$g_i^{(n)} = \sum_{r=1}^{n-1} k_r^{(i)} g_r^{(n-i)}, \tag{5.2.47}$$

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

$$g_1^{(n)} = \sum_{r=1}^{n-1} k_r g_r^{(n-1)}, \quad n \geq 1.$$

Multiplying both sides of these equations by appropriate powers of z, we convert them into a single one in generating functions:

$$\begin{aligned} g_1^{(1)} z &= k_0 z, \\ \sum_{n=2}^{\infty} g_1^{(n)} z^n &= \sum_{n=2}^{\infty} z^n \sum_{r=1}^{n-1} g_r^{(n-1)}, \\ G_1(z) &\equiv z k_0 + z \sum_{r=1}^{\infty} k_r \sum_{n=r+1}^{\infty} z^{n-1} g_r^{(n-1)} \\ &= z \left[k_0 + \sum_{r=1}^{\infty} k_r [G_1(z)]^r \right] \\ &= z K[G_1(z)]. \end{aligned} \tag{5.2.48}$$

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 when $\rho < 1$ and determine explicit expressions for specific distributions. There are 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,

$$G'(1)[1 + \lambda\psi'(0)] = 1,$$

$$G'(1) = \frac{1}{1 + \lambda\psi'(0)}.$$

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 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

$$\boldsymbol{P} = \begin{matrix} 0 \\ 1 \\ 2 \\ \vdots \\ K-1 \end{matrix} \overset{\begin{matrix} 0 & 1 & 2 & \ldots & K-1 \end{matrix}}{\begin{bmatrix} k_0 & k_1 & k_2 & \ldots & 1-\sum_0^{k-2} k_j \\ k_0 & k_1 & k_2 & \ldots & 1-\sum_0^{k-2} k_j \\ & k_0 & k_1 & \ldots & 1-\sum_0^{k-3} k_j \\ & & & & \vdots \\ & & & & 1-k_0 \end{bmatrix}}. \tag{5.2.54}$$

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

$$\pi_j = \sum_i \pi_i P_{ij}, \quad j = 0, 1, 2, \ldots, K-1,$$

$$\sum_0^{K-1} \pi_j = 1. \tag{5.2.55}$$

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 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_is as in the infinite case for $i = 0, 1, 2, \ldots, K-1$ and determining

$$\pi_0 = \left[\sum_0^{K-1} \nu_i \right]^{-1},$$

$$\pi_i = \pi_0 \nu_i. \tag{5.2.56}$$

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 Gross and Harris (1998), pp. 231–232.

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 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

$$k_0 = 0.607; \quad k_1 = 0.303; \quad k_2 = 0.076; \quad k_3 = 0.012; \quad k_4 = 0.002.$$

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

$$\Pi(z) = \frac{(1-0.5)(z-1)e^{-0.5(1-z)}}{z - e^{-0.5(1-z)}}.$$

Instead of using $\Pi(z)$ to obtain the distribution explicitly, we use the computational method described earlier. We have

$$\begin{aligned}
\nu_1 &= \frac{1-k_0}{k_0} = 0.647,\\
\nu_2 &= \frac{1-k_1}{k_0}\nu_1 - \frac{k_1}{k_0} = 0.244,\\
\nu_3 &= 0.074,\\
\nu_4 &= 0.022,\\
\nu_5 &= 0.006,\\
\nu_6 &= 0.002,\\
\nu_7 &= 0.001,\\
\sum_{i=0}^{7} \nu_i &= 1.996,\\
\pi_0 &= \left(\sum_0^7 \nu_i\right)^{-1} = 0.501,\\
\pi_i &= \nu_i \pi_0, \quad i = 1, 2, \ldots, 7.
\end{aligned}$$

Thus we get

$$\begin{array}{llll}
\pi_0 = 0.501; & \pi_1 = 0.324; & \pi_2 = 0.122; & \pi_3 = 0.037;\\
\pi_4 = 0.011; & \pi_5 = 0.003; & \pi_6 = 0.001; & \pi_7 = 0.000.
\end{array}$$

The mean number of customers in the system as $t \to \infty$ can be determined either from 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

$$W = \frac{0.75}{0.5} = 1.5 \text{ time units.}$$

Also, for the length of a busy period B, we have

$$E(B_1) = \frac{1}{1-0.5} = 2 \text{ time units.}$$

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:

$$P(\text{no new arrivals}) = 0.5,$$
$$P(\text{one new arrival}) = 0.3,$$
$$P(\text{two new arrivals}) = 0.2.$$

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

$$k_0 = 0.5; \qquad k_1 = 0.3; \qquad k_2 = 0.2$$

with $E(\text{number of arrivals during one service period}) = 0.7 = \text{traffic intensity } \rho$. 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 (5.2.15), we get

$$\begin{array}{llll} \nu_0 = 1; & \nu_1 = 1; & \nu_2 = 0.8; & \nu_3 = 0.32; \\ \nu_4 = 0.128; & \nu_5 = 0.051; & \nu_6 = 0.021; & \nu_7 = 0.008; \\ \nu_8 = 0.003; & \nu_9 = 0.001; & \nu_{10} = 0.001. & \end{array}$$

Hence $\sum_{i=0}^{10} \nu_i = 3.333$. Since $\pi_0 = (\sum_{i=0}^{10} \nu_i)^{-1}$ and $\pi_i = \nu_i \pi_0$, we get

$$\begin{array}{llll} \pi_0 = 0.300; & \pi_1 = 0.300; & \pi_2 = 0.240; & \pi_3 = 0.096; \\ \pi_4 = 0.038; & \pi_5 = 0.015; & \pi_6 = 0.006; & \pi_7 = 0.002; \\ \pi_8 = 0.001; & \pi_9 = 0.000. & & \end{array}$$

The mean of this distribution is obtained as

$$L = E(Q^*) = 1.353.$$

Using Little's law, for the mean system time we get

$$W = \frac{1.353}{0.7} = 1.933 \text{ service time units.}$$

Note that we use the mean service time as the unit time for the purpose of determining the mean waiting time. **Answer**

Example 5.2.3. In a queueing system with RR service discipline, service is provided for a fixed amount of time, called a 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. This example illustrates the procedure for the determination of the mean total time spent in the system by a customer in receiving service. This time period is commonly called the response time in the computer science literature (Coffman and Kleinrock (1968)).

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, the 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 the 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 $\{p_j,\ j = 0, 1, 2, \dots\}$ 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 break down $W_k(j)$ in terms of the number of times the customer passes through service, which is k. Let $U_i(j)$, $i = 1, 2, \dots, k$, 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 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 of three components:

(i) the amount of time required to serve those who are ahead of the customer at the ith pass; this is $\sigma[(\frac{x}{Q}) - 1]Q$, 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

$$E[U_{i+1}|U_i = x] = \sigma\left(\frac{x}{Q} - 1\right)Q + \lambda x Q + Q, \tag{5.2.63}$$

giving

$$E[U_{i+1}] = (\lambda Q + \sigma)E(U_i) + Q(1 - \sigma), \quad i = 2, 3, \dots, k. \tag{5.2.64}$$

By successive iteration, from (5.2.64) we get

$$E[U_i] = \alpha^{i-2}E(U_2) + Q(1 - \sigma)\frac{1 - \alpha^{i-2}}{1 - \alpha}, \quad i = 2, 3, \dots, k, \tag{5.2.65}$$

where we have written $\lambda Q + \sigma = \alpha$. Also, in the first pass for $U_1(j)$, we have

$$\begin{aligned} U_1(j) = {} & \text{(time to complete the service in progress)} \\ & + \text{(total time to serve } j - 1 \text{ customers)} \\ & + \text{(one quantum of service for the arriving customer)} \\ = {} & \rho\left(\frac{Q}{2}\right) + (j - \rho)Q + Q \end{aligned}$$

$$= \left(1 - \frac{\rho}{2}\right) Q + jQ. \tag{5.2.66}$$

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

$$\begin{aligned} W_k(j) &= E(U_1) + \sum_{i=2}^{k} \left[\alpha^{i-2}(\lambda Q E(U_1) + Q(\sigma j + 1)) + Q(1-\sigma)\frac{1-\alpha^{i-2}}{1-\alpha} \right] \\ &= E(U_1) + [\lambda Q E(U_1) + Q(\sigma j + 1)]\frac{1-\alpha^{k-1}}{1-\alpha} \\ &\quad + \frac{Q(1-\sigma)}{1-\alpha}\left[(k-1) - \frac{1-\alpha^{k-1}}{1-\alpha}\right] \\ &= E(U_1) + \frac{(k-1)Q}{1-\rho} + Q\left[\lambda E(U_1) + \sigma j - \frac{\rho}{1-\rho}\right]\frac{1-\alpha^{k-1}}{1-\alpha}. \end{aligned} \tag{5.2.69}$$

In deriving (5.2.69), we have used the following simplifications:

$$\alpha = \lambda Q + \sigma; \qquad \rho = \frac{\lambda Q}{1-\sigma}; \qquad 1 - \rho = \frac{1 - \sigma - \lambda Q}{1-\sigma}.$$

Taking expectations as in (5.2.62), for the conditional mean response time we get

$$W_k = W_1 + \frac{(k-1)Q}{1-\rho} + Q\left[\lambda W_1 + \sigma L - \frac{\rho}{1-\rho}\right]\frac{1-\alpha^{k-1}}{1-\alpha}, \quad k \geq 1, \tag{5.2.70}$$

with $W_1 = (1 - \frac{\rho}{2})Q + LQ$ and L given by (5.2.61).

When the service time distribuion is exponential, $\mu e^{-\mu x} (x > 0)$, we get $\sigma = e^{-\lambda Q}$ and

$$g_i = (1 - e^{-\lambda Q})e^{-(i-1)\lambda Q}, \quad i = 1, 2, \ldots. \tag{5.2.71}$$

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 is taken out first. This is an 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 interarrival times of demands.

We want to determine the average time an item stays in the warehouse before it is disposed of.

Considering the interdemand times as the service times of the queueing system, we have here an $M/G/1$ queue with an 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

$$r(x) = \lim_{t\to\infty} r_t(x) = \frac{1}{E(S)}[1 - B(x)], \tag{5.2.72}$$

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

$$\begin{aligned} W_q &= \frac{\lambda E(S^2)}{2E(S)} \times \frac{1}{1-\rho} \cdot E(S) \\ &= \frac{\lambda E(S^2)}{2(1-\rho)}. \end{aligned} \tag{5.2.75}$$

We note here that the remainder of the service time of the customer in service gets absorbed in the busy periods initiated by the arrivals.

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 be served by a single server. Let $Z_n = t_{n+1} - t_n$, $n = 1, 2, 3, \ldots,$ be i.i.d. random variables with distribution function $A(\cdot)$ with mean a. Also, let the service time distribution be exponential with mean $1/\mu$. Note that this system has been tranditionally 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, the 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 interarrival 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

$$Q_{n+1} = \begin{cases} Q_n + 1 - X_{n+1} & \text{if } Q_n + 1 - X_{n+1} > 0, \\ 0 & \text{if } Q_n + 1 - X_{n+1} \le 0. \end{cases} \tag{5.3.2}$$

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

$$\begin{aligned} P_{ij} &= P(Q_{n+1} = j | Q_n = i) \\ &= \begin{cases} P(X_{n+1} = i - j + 1) & \text{if } j > 0, \\ P(X_{n+1} \ge i + 1) & \text{if } j = 0, \end{cases} \end{aligned}$$

giving

$$P_{ij} = b_{i-j+1}, \quad j > 0,$$
$$P_{i0} = \sum_{r=i+1}^{\infty} b_r. \tag{5.3.3}$$

The transition probability matrix takes the form

$$\boldsymbol{P} = \begin{matrix} & 0 & 1 & 2 & 3 & \dots \\ 0 & \sum_1^\infty b_r & b_0 & & & \\ 1 & \sum_2^\infty b_r & b_1 & b_0 & & \\ 2 & \sum_3^\infty b_r & b_2 & b_1 & b_0 & \dots \\ \vdots & \vdots & \vdots & \vdots & & \end{matrix}. \tag{5.3.4}$$

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

$$\phi(\theta) = \int_0^\infty e^{-\theta t} dA(t), \quad \text{Re}(\theta) > 0,$$

be the Laplace–Stieltjes transform of $A(\cdot)$. Using $\phi(\theta)$, the PGF of $\{b_j\}$ is obtained as

$$\begin{aligned} \beta(z) &= \sum_{j=0}^{\infty} b_j z^j, \quad |z| \le 1, \\ &= \int_0^\infty e^{-(\mu-\mu z)t} dA(t) \\ &= \phi(\mu - \mu z). \end{aligned}$$

Following the definitions 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 $\rho = (\text{arrival rate})/(\text{service rate})$. 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 the 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 $P_{ij}^{(n)}$ $(i, j = 0, 1, 2, \dots)$ 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 obtain its power $\mathbf{P}^n$ is likely 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

$$\pi_j = \sum_{i=0}^{\infty} \pi_i P_{ij}, \quad j = 0, 1, 2, \ldots,$$

$$\sum_{0}^{\infty} \pi_j = 1.$$

Using P_{ij}s from (5.3.4), we get

$$\begin{aligned}
\pi_0 &= \sum_{i=0}^{\infty} \pi_i \left(\sum_{r=i+1}^{\infty} b_r \right), \\
\pi_1 &= \pi_0 b_0 + \pi_1 b_1 + \pi_2 b_2 + \cdots, \\
\pi_2 &= \pi_1 b_0 + \pi_2 b_1 + \pi_3 b_2 + \cdots, \\
&\vdots \\
\pi_j &= \sum_{r=0}^{\infty} \pi_{j+r-1} b_r \quad (j \geq 1).
\end{aligned} \tag{5.3.7}$$

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.

In this case, however, (5.3.7) can be easily solved by the use of finite difference methods. 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

$$\pi_{j-1}(D - b_0 - Db_1 - D^2 b_2 - D^3 b_3 - \cdots) = 0. \tag{5.3.9}$$

Using finite difference methods, a nontrivial solution to (5.3.9) is obtained by solving its characteristic equation

$$D - b_0 - Db_1 - D^2 b_2 - \cdots = 0,$$

$$D = \sum_{j=0}^{\infty} b_j D^j,$$

$$D = \beta(D). \tag{5.3.10}$$

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:

- $\beta(0) = b_0 > 0$; $\beta(1) = \sum_0^\infty b_j = 1$; $\beta'(1) = \rho^{-1}$.
- $\beta''(z) = 2b_2 + 6b_3 z + \cdots > 0$ 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, equations $y = z$ 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 that ζ_s and 1 coincide.

Let ζ be the least positive root. We have $\zeta < 1$ if $\rho < 1$ and $\zeta = 1$ if $\rho \geq 1$. This root is used in the solution of the finite difference equation (5.3.9). (Note that our solution is in terms of probabilities that 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 $G/M/1$.

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, readers may refer to Chaudhry (1992) and the 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)

$$L^A = E(Q^A) = \frac{\zeta}{1-\zeta}; \qquad L_q^A = \frac{\zeta^2}{1-\zeta},$$
$$V(Q^A) = \frac{\zeta}{(1-\zeta)^2}. \tag{5.3.15}$$

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 of the $M/G/1$ queue, the limiting distributions of the number of customers in the system at arrival epochs, at 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, results derived in Prabhu (1965a) 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 explicit expression for the limiting distribution $\{p_j,\ j = 0, 1, 2, \ldots\}$, when $\rho < 1$:

$$p_0 = 1 - \rho,$$
$$p_j = \rho(1-\zeta)\zeta^{j-1}, \quad j \geq 1, \tag{5.3.16}$$

From (5.3.16), the results below follow on mean queue length:

$$L = \frac{\rho}{1-\zeta}; \qquad L_q = \frac{\rho\zeta}{1-\zeta}. \tag{5.3.17}$$

As an example, consider the queue $M/M/1$. Let $A(t) = 1 - e^{-\lambda t}$ $(t \geq 0)$. Then we have

$$\phi(\theta) = \frac{\lambda}{\lambda + \theta},$$
$$\phi(\mu - \mu z) = \frac{\lambda}{\lambda + \mu - \mu z}.$$

Now the functional equation (5.3.11) takes the form

$$z = \frac{\lambda}{\lambda + \mu - \mu z},$$
$$-\mu z^2 + (\lambda + \mu)z - \lambda = 0. \tag{5.3.18}$$

This quadratic equation has two roots, 1 and $\frac{\lambda}{\mu} = \rho$. Substituting ρ in place of ζ in (5.3.16)–(5.3.17), 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 that customer's 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

$$F_q(t) = P(T_q \leq t) = 1 - \zeta e^{-\mu(1-\zeta)t},$$
$$W_q = E[T_q] = \frac{\zeta}{\mu(1-\zeta)},$$
$$V[T_q] = \frac{\zeta(2-\zeta)}{\mu^2(1-\zeta)^2}. \tag{5.3.19}$$

The time T spent by the customer in the system is obtained by adding service time to T_q. We get

$$W = E(T) = E(T_q) + \frac{1}{\mu} = \frac{1}{\mu(1-\zeta)},$$
$$V(T) = V[T_q + S] = \frac{1}{\mu^2(1-\zeta)^2}. \tag{5.3.20}$$

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.) 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,

$$\begin{aligned} H_0(z) &= \sum_{n=1}^{\infty} h_0^{(n)} z^n \\ &= \left(\sum_1^{\infty} b_k \right) z + \sum_{n=2}^{\infty} \sum_{r=1}^{n-1} h_r^{(n-1)} \left(\sum_{k=r+1}^{\infty} b_k \right). \end{aligned} \tag{5.3.26}$$

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

$$\begin{aligned}
&\beta_0 z + z\sum_{n=2}^{\infty} z^{n-1}\sum_{r=1}^{n-1}\beta_r h_r^{(n-1)} \\
&\quad = z\left[\beta_0 + \sum_{r=1}^{\infty}\beta_r \sum_{n=r+1}^{\infty} h_r^{(n-1)} z^{n-1}\right] \\
&\quad = z\left[\sum_{r=0}^{\infty}\beta_r[\eta(z)]^r\right],
\end{aligned}$$

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 $H_0'(z)$. We have

$$H_0'(z) = \frac{[1-\eta(z)][1-\eta'(z)] + [z-\eta(z)]\eta'(z)}{[1-\eta(z)]^2}$$

$$= \frac{1-\eta'(z)}{1-\eta(z)} + \frac{[z-\eta(z)]\eta'(z)}{[1-\eta(z)]^2}. \tag{5.3.28}$$

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 $z = 1$, $\eta(1)$ 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

$$\eta(1) = \zeta \quad \text{if } \rho < 1 \quad \text{and} \quad = 1 \quad \text{if } \rho \geq 1. \tag{5.3.29}$$

Consider $\eta(z) = z\beta[\eta(z)]$. We get

$$\eta'(z) = z\beta'[\eta(z)]\eta'(z) + \beta[\eta(z)],$$
$$\eta'(z)[1 - z\beta'[\eta(z)]] = \beta[\eta(z)].$$

Letting $z \to 1$ and using (5.3.29),

$$\eta'(1)[1-\beta'(\zeta)] = \beta(\zeta),$$
$$\eta'(1) = \frac{\beta(\zeta)}{1-\beta'(\zeta)}. \tag{5.3.30}$$

Substituting from (5.3.29) and (5.3.30) in (5.3.28),

$$\lim_{z\to 1} H_0'(z) = \left[1 - \frac{\beta(\zeta)}{1-\beta'(\zeta)}\right]\frac{1}{1-\zeta} + \left[(1-\zeta)\frac{\beta(\zeta)}{1-\beta'(\zeta)}\right] \times \frac{1}{(1-\zeta)^2}$$
$$= \frac{1}{1-\zeta} < \infty \quad \text{if } \rho < 1. \tag{5.3.31}$$

Similarly, we can also show that $H'_{\lim z\to 1}(z) = \infty$ when $\rho = 1$.

We 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 interarrival 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 (interarrival time) has a general distribution, the determination of the mean busy period is too complicated to be covered in this text.

The queue $G/M/s$

The imbedded Markov chain analysis of the queue $G/M/1$ can be easily extended to the multiserver 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 notation, 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 interarrival time with distribution $A(\cdot)$.

To determine transition probabilities P_{ij} $(i, j = 0, 1, 2, \dots)$, 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 interarrival period, when all s servers are busy. See the discussion under $M/M/s$ (see (4.3.1)) 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* $j < s$.

$$P_{ij} = \int_{t=0}^\infty \int_{\tau=0}^t e^{-s\mu\tau} \frac{(s\mu t)^{i-s}}{(i-s)!} s\mu \binom{s}{s-j} [1 - e^{-\mu(t-\tau)}]^{s-j} e^{-j\mu(t-\tau)} d\tau \, dA(t). \tag{5.3.35}$$

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 the P_{ij} values under cases 2 and 3, the finite difference solution given earlier for the limiting distribution needs major modifications. Interested readers are referred to Gross and Harris (1998), pp. 256–258, or Takács (1962). Considering the complexities of these procedures, the computational method developed for $G/M/1/K$ could be advantageous in this case if it is possible to work with a finite limit for the number of customers in the system.

The queue $G/M/1/K$

Consider the $G/M/1$ queue described earlier with the restriction 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 is not admitted to the system. Thus in place of (5.3.2), we have the relation

$$Q_{n+1} = \begin{cases} \min(Q_n + 1 - X_{n+1}, K) & \text{if } Q_n + 1 - X_{n+1} > 0, \\ 0 & \text{if } Q_n + 1 - X_{n+1} \leq 0. \end{cases} \tag{5.3.36}$$

Using probabilities b_j, $j = 0, 1, 2, \ldots$, defined in (5.3.1), the transition probability matrix $\mathbf{P}$ can be displayed as

$$\boldsymbol{P} = \begin{array}{c} \\ 0 \\ 1 \\ \vdots \\ K-1 \\ K \end{array} \begin{array}{c} \begin{array}{cccccc} 0 & 1 & 2 & \ldots & K-1 & K \end{array} \\ \left[\begin{array}{cccccc} \sum_1^\infty b_r & b_0 & 0 & & & \\ \sum_2^\infty b_r & b_1 & b_0 & & & \\ \vdots & \vdots & & & & \\ \sum_K^\infty b_r & b_{K-1} & b_{K-2} & \ldots & b_1 & b_0 \\ \sum_K^\infty b_r & b_{K-1} & b_{K-2} & \ldots & b_1 & b_0 \end{array}\right] \end{array}. \tag{5.3.37}$$

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 $\pi_j = \sum_{i=0}^K \pi_i P_{ij}$, we have

$$\begin{aligned} \pi_0 &= \sum_{i=0}^{K} \pi_i \left(\sum_{r=i+1}^{\infty} b_r \right), \\ \pi_1 &= \pi_0 b_0 + \pi_1 b_1 + \cdots + \pi_{K-1} b_{K-1} + \pi_K b_{K-1}, \\ \pi_2 &= \pi_1 b_0 + \pi_2 b_1 + \cdots + \pi_{K-1} b_{K-2} + \pi_K b_{K-2}, \\ &\vdots \\ \pi_{K-1} &= \pi_{K-2} b_0 + \pi_{K-1} b_1 + \pi_K b_1, \\ \pi_K &= \pi_{K-1} b_0 + \pi_K b_0. \end{aligned} \tag{5.3.38}$$

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, obtaining $\mathbf{P}^n$ 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

$$\begin{aligned} \pi_i &= \nu_i \pi_{i-1} \\ &= \nu_i \nu_{i-1} \pi_{i-2} \\ &= \nu_i \nu_{i-1} \cdots \nu_i \pi_0. \end{aligned} \tag{5.3.39}$$

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-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

$$(1 + \nu_1 + \nu_1 \nu_2 + \cdots + \nu_1 \nu_2 \cdots \nu_K)\pi_0 = 1$$

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_is as the ratio of consecutive π_is, unlike the case of (5.2.15). We do so for the reason that the π_js 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 $\boldsymbol{\pi}$ is 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 the results shown in Table 5.3.1.

Table 5.3.1. Results for Example 5.3.1.

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 $\boldsymbol{\pi} = (\pi_0, \pi_1, \pi_2, \dots)$ given by (5.3.13). We get

$$\begin{aligned}&\pi_0 = 0.584; \quad \pi_1 = 0.243; \quad \pi_2 = 0.101; \quad \pi_3 = 0.042; \\ &\pi_4 = 0.017; \quad \pi_5 = 0.007; \quad \pi_6 = 0.003; \quad \pi_7 = 0.001.\end{aligned}$$

ANSWER

Example 5.3.2. Consider the service center of Example 5.3.1 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: $K = 4$ and 7.

The distribution of the potential number of customers served during an interarrival period is Poisson with mean 1.5. We have

$$\begin{aligned}&b_0 = 0.223; \quad b_1 = 0.335; \quad b_2 = 0.251; \quad b_3 = 0.125; \\ &b_4 = 0.047; \quad b_5 = 0.015; \quad b_6 = 0.003; \quad b_7 = 0.001.\end{aligned}$$

$\boldsymbol{K = 4}$**:** Using $\nu_i = \pi_i/\pi_{i-1}$ in (5.3.38) with $K = 4$, we get

$$\begin{aligned}
\nu_4 &= b_0[1 - b_0]^{-1},\\
\nu_3 &= b_0[1 - b_1 - \nu_4 b_1],\\
\nu_2 &= b_0[1 - b_1 - \nu_3 b_2 - \nu_4 \nu_3 b_2],\\
\nu_1 &= b_1[1 - b_1 - \nu_2 b_2 - \nu_3 \nu_2 b_3 - \nu_4 \nu_3 \nu_2 b_3]^{-1}.
\end{aligned}$$

Substituting appropriate values of b_j, for $j = 0, 1, 2, 3$, we get

$$\nu_1 = 0.413; \qquad \nu_2 = 0.398; \qquad \nu_3 = 0.392; \qquad \nu_4 = 0.287.$$

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

$$\pi_0 = 0.602; \qquad \pi_1 = 0.249; \qquad \pi_2 = 0.099; \qquad \pi_3 = 0.039; \qquad \pi_4 = 0.001.$$

Answer

***K* = 7:** Looking at the structure of the ν_is, it is clear that $\nu_4, \dots, \nu_1$ determined above, in fact, yield $\nu_7, \dots, \nu_4$, when $K = 7$. Extending the equations to determine the remaining νs, viz., ν_3, ν_2, and ν_1, we get the following set of values:

$$\begin{aligned}
&\nu_7 = 0.287; \qquad \nu_6 = 0.392; \qquad \nu_5 = 0.398; \qquad \nu_4 = 0.413;\\
&\nu_3 = 0.415; \qquad \nu_2 = 0.416; \qquad \nu_1 = 0.417.
\end{aligned}$$

Converting these back to πs, we get

$$\begin{aligned}
&\pi_0 = 0.585; \qquad \pi_1 = 0.244; \qquad \pi_2 = 0.101;\\
&\pi_3 = 0.042; \qquad \pi_4 = 0.017; \qquad \pi_5 = 0.007;\\
&\pi_6 = 0.003; \qquad \pi_7 = 0.001.
\end{aligned}$$

A comparison of these values with those obtained for 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 **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 **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 merchandise in a Poisson process at the rate of 15/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 (a) if the length of service time is shortened to 3 minutes without changing the variance or (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 customers reveals that a third of them would like to have a longer wash. To satisfy their need he sets up two wash times, 5 minutes and 2.5 minutes, for the two groups. Both 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 the doctor are made at 15-minute intervals. But the amounts of time the doctor spends with her patients are mostly less than 15 minutes, although 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 workday, how often will they be able to take such breaks?

7. At a taxi stop on a busy street, the interarrival times of taxis dropping off passengers (and then being ready for a pickup) have a mean of 5 minutes with a standard deviation of 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 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) The CPU time of the job (for the first time or after I/O 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 second.
      - *Latency time*: Uniformly distributioned with mean 0.01 second.
      - *Transfer time*: Constant = 0.01 second.

   (iv) 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 an FCFS basis. Assume that the duration of service can take any of the K equal probability 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.

   (See 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 PS 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 $L \to \frac{\rho}{1-\rho}$. Set $\frac{S}{Q} = k$, where S is the service time requested by the customer.)

6

Extended Markov 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 the analytical procedures.

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 and 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 group arrivals occur in a Poisson process with rate λ and let 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 $d_k = P(X = k)$, $k = 1, 2, \ldots$, 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 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

$$\begin{aligned} \lambda p_0 &= \mu p_1, \\ (\lambda + \mu) p_n &= \lambda \sum_k d_k p_{n-k} + \mu p_{n+1}, \quad n = 1, 2, \ldots. \end{aligned} \tag{6.1.1}$$

U.N. Bhat, *An Introduction to Queueing Theory*, DOI: 10.1007/978-0-8176-4725-4_6,

Note that the first term on the right-hand side of the second equation in (6.1.1) exists only if $n - k \geq 0$.

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

$$P(z) = \sum_{n=0}^{\infty} p_n z^n; \qquad \delta(z) = \sum_{k=1}^{\infty} d_k z^k, \quad |z| \leq 1.$$

Multiplying the equations in (6.1.1) by appropriate powers of z, we have

$$\begin{aligned} \lambda p_0 &= \mu p_1, \\ (\lambda + \mu) \sum_{n=1}^{\infty} p_n z^n &= \lambda \sum_{n=1}^{\infty} z^n \sum_{k=1}^{\infty} d_k p_{n-k} + \mu \sum_{n=1}^{\infty} p_{n+1} z^n. \end{aligned} \tag{6.1.2}$$

Interchanging summations on the right-hand side of (6.1.2) and simplifying, we obtain

$$\begin{aligned} (\lambda + \mu) P(z) - \mu p_0 &= \lambda \sum_{k=1}^{\infty} d_k z^k \sum_{n=k}^{\infty} z^{n-k} p_{n-k} \\ &\quad + \mu \sum_{n=0}^{\infty} z^n p_{n+1} \\ &= \lambda \delta(z) P(z) + \frac{\mu}{z} \sum_{m=1}^{\infty} z^m p_m \\ &= \lambda \delta(z) P(z) + \frac{\mu}{z} [P(z) - p_0]. \end{aligned}$$

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 that $\lim_{z \to 1} P(z) = 1$. 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

$$\begin{aligned}\delta'(1) &= \lim_{z\to 1}\sum_{k=1}^{\infty} k d_k z^{k-1}\\ &= E(X) = d, \quad \text{say.}\end{aligned}$$

With the average group size d, we also note that $\frac{\lambda d}{\mu} = \rho$, 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 $\{d_n\}$, inverting the PGF (6.1.6) is not simple. See the 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 terms $(1-z)$ 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) simplfies 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 are 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 or not 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 and Harris (1998).

Thus the customers arrive one at a time in a Poisson process with parameter λ and are 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 $\{p_n,\ n = 0, 1, 2, \dots\}$ of the number of customers in the system as $t \to \infty$, the state balance equations can be presented as

$$\begin{aligned}
\lambda p_0 &= \mu p_K, \\
\lambda p_n &= \lambda p_{n-1} + \mu p_{n+K}, \quad n = 1, 2, \dots, K-1, \\
(\lambda + \mu) p_n &= \lambda p_{n-1} + \mu p_{n+K}, \quad n = K, K+1, \dots.
\end{aligned} \tag{6.2.1}$$

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

$$\begin{aligned}
(\lambda + \mu) P(z) &- \mu \sum_{n=0}^{K-1} p_n z^n \\
&= \lambda z \sum_{n=1}^{\infty} p_{n-1} z^{n-1} + \frac{\mu}{z^K} \sum_{n=0}^{\infty} p_{n+K} z^{n+K}
\end{aligned}$$

$$= \lambda z P(z) + \frac{\mu}{z^K}\left[P(z) - \sum_{m=0}^{K-1} p_m z^m\right].$$

Rearranging terms and multiplying by z^K, we get

$$[(\lambda+\mu)z^K - \lambda z^{K+1} - \mu]P(z) = \mu(z^K - 1)\sum_{n=0}^{K-1} p_n z^n,$$

$$P(z) = \frac{(1-z^K)\sum_{n=0}^{K-1} p_n z^n}{(\lambda/\mu)z^{K+1} - \left(\frac{\lambda}{\mu}+1\right)z^K + 1}. \qquad (6.2.2)$$

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

$$\left(\frac{\lambda}{\mu}\right)z^{K+1} - \left(\frac{\lambda}{\mu}+1\right)z^K + 1 = 0. \qquad (6.2.3)$$

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

$$\sum_{n=0}^{K-1} p_n z^n = C\frac{(\lambda/\mu)z^{K+1} - \left(\frac{\lambda}{\mu}+1\right)z^K + 1}{(z-1)(z-z_0)}. \qquad (6.2.4)$$

Substituting this result in (6.2.2), we have

$$P(z) = \frac{C(1-z^K)}{(z-1)(z-z_0)}$$
$$= \frac{C}{z_0 - z}\sum_{n=0}^{K-1} z^n. \qquad (6.2.5)$$

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 x to determine $p_n, n = 0, 1, 2, \ldots,$ explicitly as follows:

$$P(z) = \frac{z_0 - 1}{K z_0}\left(\sum_{s=0}^{K-1} z^s\right)\left(\sum_{r=0}^{\infty}\left(\frac{z}{z_0}\right)^r\right), \tag{6.2.7}$$

$$\begin{aligned} p_n &= \frac{z_0 - 1}{K z_0}\sum_{r=0}^{n}\left(\frac{1}{z_0}\right)^r, \qquad n < K, \\ &= \frac{z_0 - 1}{K z_0^{n-K+1}}\sum_{r=0}^{K-1}\left(\frac{1}{z_0}\right)^r, \quad n \geq K. \end{aligned} \tag{6.2.8}$$

Noting that

$$\sum_{r=0}^{n}\left(\frac{1}{z_0}\right)^r = \frac{1 - (\frac{1}{z_0})^{n+1}}{1 - (\frac{1}{z_0})},$$

we can present (6.2.8) as

$$\begin{aligned} p_n &= \frac{z_0^{n+1} - 1}{K z_0^{n+1}}, \quad n < K, \\ &= \frac{z_0^{K} - 1}{K z_0^{n+1}}, \quad n \geq K. \end{aligned} \tag{6.2.9}$$

As mentioned above, finding the root z_0 lying outside the unit circle of (6.2.3) is essential to 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 that are 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 The Queues $M/E_k/1$ and $E_k/M/1$

In Section 2.1 and Appendix A, we define the Erlang distribution, E_k, as the distribution of the sum of k i.i.d. exponential random variables. In Appendix B, 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 that arrivals occur in a Poisson process with rate λ. Let service be provided by a single server with service time distribution

$$f(x) = e^{-k\mu x}\frac{(k\mu x)^{k-1}k\mu}{(k-1)!}, \quad x > 0, \tag{6.3.1}$$

which has mean = $\frac{1}{\mu}$. Using the observations made in Section 2.1 and Appendix A, we note that when the service time has the Erlang distribution (6.3.1), it can be considered to be made up of k phases, each with an exponential distribution with density $k\mu e^{-k\mu x}$ $(x > 0)$, which has mean $\frac{1}{k\mu}$. 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 $\{(0,0); (1,0), (1,1), \ldots, (1,k); (2,0), (2,1), \ldots, (2,k); \ldots\}$. 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 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 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 $\frac{1}{k\mu}$. 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 given in terms of the corresponding number of phases, all we need now is a procedure to convert 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 $[\frac{n}{k}]+1$, where [ ] signifies the largest integer contained in $\frac{n}{k}$ when n is not a multiple of k and in $\frac{n}{k}$ 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

$$\begin{aligned} p_n &= \sum_{j=(n-1)k+1}^{nk} p_j^{(b)}, \quad n \geq 1, \\ p_0 &= p_0^{(b)}. \end{aligned} \tag{6.3.2}$$

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

$$\psi(\theta) = \left(\frac{k\mu}{k\mu + \theta}\right)^k, \tag{6.3.3}$$

and the PGF (5.2.7) of the number of customers arriving during a service period is now given by

$$K(z) = \left(\frac{k\mu}{k\mu + \lambda - \lambda z}\right)^k. \tag{6.3.4}$$

The resulting PGF of the limiting distribution $\{\pi_j,\ j = 0, 1, 2, \dots\}$ has the form

$$\Pi(z) = \frac{(1-\rho)(z-1)(k\mu)^k}{z(k\mu + \lambda - \lambda z)^k - (k\mu)^k}. \tag{6.3.5}$$

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 analogue $\{Q(t)\}$ 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 with the number of customers in the system as the first variable and the number of elapsed exponential interarrival phases as the second variable.* When the interarrival times are distributed as Erlang E_k, each of them may be considered to be made up of k exponentially distributed phases. Now keeping track of the number of elapsed phases in an interarrival time helps in defining the Markov process. State balance equations may be written for the bivariate process and solved using PGFs. 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 interarrival 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 real customers, 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 occur in a Poisson process in such a way that every kth customer is a real one 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 5.3 gives the limiting distribution of the number of "customers" in the $E_k/M/1$ queue. Let $\{p_n,\ n = 0, 1, 2, \dots\}$ be the limiting distribution of the number of customers in the queue $E_k/M/1$, and let $\{p_n^{(b)},\ n = 0, 1, 2, \dots\}$ be the limiting distribution of the number of customers in the $M/M^k/1$ queue. Then $\{p_n,\ n = 0, 1, 2, \dots\}$ can be determined using the relation

$$p_n = \sum_{j=nk}^{nk+k-1} p_j^{(b)}, \quad n = 1, 2, \ldots,$$

$$p_0 = \sum_{j=0}^{k-1} p_j^{(b)}. \tag{6.3.6}$$

(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 interarrival time distribution be given as

$$f(x) = e^{-k\lambda x} \frac{(k\lambda x)^{k-1} k\lambda}{(k-1)!}, \quad x > 0. \tag{6.3.7}$$

The Laplace transform of (6.3.7) takes the form

$$\phi(\theta) = \left(\frac{k\lambda}{k\lambda + \theta} \right)^k, \tag{6.3.8}$$

and the PGF $\beta(z)$ of the number of potential services during an interarrival period can be given as

$$\beta(z) = \left(\frac{k\lambda}{k\lambda + \mu - \mu z} \right)^k. \tag{6.3.9}$$

The limiting distribution of the number of customers in the system just before an arrival is obtained as

$$\pi_j = (1 - \zeta)\zeta^j, \quad j = 0, 1, 2, \ldots, \tag{6.3.10}$$

where ζ is the least positive root of the equation

$$z = \beta(z). \tag{6.3.11}$$

6.4 The Bulk Queues $M/G^K/1$ and $G^K/M/1$

In the last three sections, we have assumed that both interarrival 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 in groups is allowed. For readers interested in the continuous-time analogue 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, the server waits until the number K is reached to start the service. Note 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 $\{Q_n,\ n = 0, 1, 2, \dots\}$ 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 $\{k_j,\ j = 0, 1, 2, \dots\}$ of X_n, we have

$$k_j = P(X_n = j) = \int_0^\infty e^{-\lambda t} \frac{(\lambda t)^j}{j!} dB(t), \quad j = 0, 1, 2, \dots. \tag{6.4.1}$$

We also have the random variable relationship between Q_n and Q_{n+1},

$$Q_{n+1} = \begin{cases} Q_n + X_{n+1} - K & \text{if } Q_n > K, \\ X_{n+1} & \text{if } Q_n \le K. \end{cases} \tag{6.4.2}$$

(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 one.)

For $P_{ij} = P(Q_{n+1} = j | Q_n = i)$, (6.4.2) gives

$$\begin{aligned} P_{ij} &= \begin{cases} P(i + X_{n+1} - K = j) & \text{if } i > K, \\ P(X_{n+1} = j) & \text{if } i \le K \end{cases} \\ &= \begin{cases} k_{j-i+K} & \text{if } i > K, \\ k_j & \text{if } i \le K. \end{cases} \end{aligned} \tag{6.4.3}$$

Displaying these probabilities in matrix form, we get the transition probability matrix $\mathbf{P}$ of the imbedded Markov chain:

$$\mathbf{P} = \begin{matrix} \\ 0 \\ 1 \\ 2 \\ \vdots \\ K \\ K+1 \\ \vdots \end{matrix} \begin{matrix} \begin{matrix} 0 & 1 & \dots \end{matrix} \\ \begin{bmatrix} k_0 & k_1 & \dots \\ k_0 & k_1 & \dots \\ k_0 & k_1 & \dots \\ \vdots & \vdots & \\ k_0 & k_1 & \dots \\ & k_0 & \dots \\ & \vdots & \end{bmatrix} \end{matrix}. \tag{6.4.4}$$

Comparing (6.4.4) with (5.2.5), we note that (6.4.4) has $K + 1$ identical rows, instead of two as in (5.2.5). If we are interested in a mathematical expression for the limiting distribution of the Markov chain, we proceed in 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 can use the matrix (6.4.4) itself. Note that the elements k_j 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 interarrival 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 the reasons explained following (5.3.2), we define X_{n+1} as the number of potential departures during the $(n+1)$st interarrival period. Let $\{Q_n,\ n = 0, 1, 2, \ldots\}$ be the number of customers in the system just before the nth group arrival. Analogous to (5.3.2), we have

$$Q_{n+1} = \begin{cases} Q_n + K - X_{n+1} & \text{if } Q_n + K - X_{n+1} > 0, \\ 0 & \text{if } Q_n + K - X_{n+1} \le 0. \end{cases} \tag{6.4.5}$$

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 (6.4.5) we get

$$\begin{aligned} P_{ij} &= \begin{cases} P(i + K - X_{n+1} = j) & \text{if } j > 0, \\ P(i + K - X_{n+1} \le 0) & \text{if } j = 0 \end{cases} \\ &= \begin{cases} b_{i+K-j}, & j > 0, \\ \sum_{i+K}^{\infty} b_r, & j = 0. \end{cases} \end{aligned} \tag{6.4.6}$$

The transition probability matrix has the form

$$\mathbf{P} = \begin{array}{c} \\ 0 \\ 1 \\ \\ K \\ \\ \end{array} \begin{array}{c} \begin{array}{cccc} 0 & 1 & 2 & \cdots \end{array} \\ \left[\begin{array}{cccc} \sum_K^{\infty} b_r & b_{K-1} & b_{K-2} & \cdots \\ \sum_{K+1}^{\infty} b_r & b_K & b_{K-1} & \cdots \\ \vdots & & & \\ \sum_{2K}^{\infty} b_r & b_{2K-1} & b_{2k-2} \cdots & \\ \vdots & & \vdots & \end{array} \right] \end{array}. \tag{6.4.7}$$

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 $\sum_0^{\infty} \pi_j = 1$ along 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 solutions in bulk queueing systems that lead to algorithmic procedures, the matrix-analytic solution techniques developed by

M. F. Neuts and his associates are highly recommended. A thorough knowledge of matrix analysis is essential to understanding them. The basic references are the books by Neuts (1978, 1989). Many articles solving a wide variety of queueing problems using the Neuts method continue to appear in journals such as *Queueing Systems: Theory and Applications* and *Stochastic Models*. See also Chaudhry and Templeton (1983) as a comprehensive reference on bulk queueing systems.

6.5 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, along the same lines as described in Section 6.3. Since the results of the bulk queueing analysis are given 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 details, the references mentioned in the last section and journal articles that have appeared since then on the matrix-analytic method of analysis are appropriate.

6.6 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 constitutes a natural model. Let customers arrive in a Poisson process with rate λ, let 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 $(0, b, 2b, 3b, \dots)$ 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 $\{X_n,\ n = 1, 2, \dots\}$ be the number of customers arriving during $[(n-1)b, nb]$. We have

$$k_j = P(X_n = j) = e^{-\lambda b}\frac{(\lambda b)^j}{j!}, \quad j = 0, 1, 2, \dots. \tag{6.6.1}$$

Considering the number of arrivals and service completions during $[nb, (n+1)b]$, we may write

$$Q_{n+1} = \begin{cases} Q_n + X_{n+1} - s, & Q_n > s, \\ X_{n+1}, & Q_n \le s. \end{cases} \tag{6.6.2}$$

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 (6.6.2) 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 (6.4.4) with K replaced by s, and its analysis follows along similar lines.

6.7 The Queue $M/M/1$ with Priority Disciplines

Queue disciplines that assign priority service to some customers are 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 RR, shortest-processing-time, and earliest-due-date-first disciplines are some of the examples. Since an analysis of most of the variants involves underlying processes much more complex than those we consider in this text, we shall not delve into them here. However, we will introduce the simple two-class priority model under the $M/M/1$ setting and discuss the fundamental issues of 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 *nonpreemptive*. If a customer has preemptive priority over another customer, the priority customer will preempt the nonpriority customer for service. If the priority is nonpreemptive, 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 a *head-of-the-line priority discipline*.)
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 at which 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 purposes of analysis, the 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. Under the *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 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 nonpreemptive 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 must 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 determine the limiting distribution of the state of the process ($\lambda = \lambda_1 + \lambda_2$):

$$\begin{aligned}
\lambda p_0 &= \mu_1 p_{101} + \mu_2 p_{012}, \\
(\lambda + \mu_1) p_{101} &= \lambda_1 p_0 + \mu_1 p_{201} + \mu_2 p_{112}, \\
(\lambda + \mu_2) p_{012} &= \lambda_2 p_0 + \mu_1 p_{111} + \mu_2 p_{022}, \\
(\lambda + \mu_1) p_{m01} &= \lambda_1 p_{m-1,01} + \mu_1 p_{m+1,01} + \mu_2 p_{m12}, && m \geq 1, \\
(\lambda + \mu_2) p_{0n2} &= \lambda_2 p_{0,n-1,2} + \mu_1 p_{1n1} + \mu_2 p_{0,n+1,2}, && n \geq 1, \\
(\lambda + \mu_2) p_{m12} &= \lambda_1 p_{m-1,12}, && m \geq 1, \\
(\lambda + \mu_1) p_{1n1} &= \lambda_2 p_{1,n-1,1} + \mu_1 p_{2n1} + \mu_2 p_{1,n+1,2}, && n \geq 1, \\
(\lambda + \mu_1) p_{mn1} &= \lambda_1 p_{m-1,n1} + \lambda_2 p_{m,n-1,1} \\
&\quad + \mu_1 p_{m+1,n1} + \mu_2 p_{m,n+1,2}, && m \geq 2, \quad n \geq 1, \\
(\lambda + \mu_2) p_{mn2} &= \lambda_1 p_{m-1,n2} + \lambda_2 p_{m,n-1,2}, && m \geq 1, \quad n \geq 2, \quad (6.7.1)
\end{aligned}$$

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 and Harris (1998) 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, note that the number of equations (accordingly, the number of unknowns) can become 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 must deal with 111 equations. Example 6.7.1 below is illustrative of the procedure when $K = 2$.

Without going through the analysis, we state the following conclusions that are useful in understanding the effect of priority assignment on customers as well as on 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 for a two-class system with no priorities.
2. For the queueing system, a 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 of the waiting time of customers 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 nonpreemptive 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

$$\frac{1}{\mu_2} \cdot \frac{\lambda_1}{\lambda} \left(\sum_{n=1}^{\infty} p_{0n2} \right) = \frac{\lambda_1 \lambda_2}{\lambda \mu_2^2}. \tag{6.7.2}$$

5. The derivation of the mean waiting time of the low-priority customer is much more complex. Suppose that 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 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 that when a lower-priority (class 2) customer is in service, on the arrival of a high-priority customer (class 1), the service of the class 2 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.

Because of 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 of the system. For the state balance equations, we have ($\lambda = \lambda_1 + \lambda_2$)

$$\begin{aligned}
\lambda p_0 &= \mu_1 p_{10} + \mu_2 p_{01}, \\
(\lambda + \mu_1) p_{m0} &= \lambda_1 p_{m-1,0} + \mu_1 p_{m+1,0}, && m \geq 1, \\
(\lambda + \mu_2) p_{0n} &= \lambda_2 p_{0,n-1} + \mu_1 p_{1n} + \mu_2 p_{0,n+1}, && n \geq 1, \\
(\lambda + \mu_1) p_{mn} &= \lambda_1 p_{m-1,n} + \lambda_2 p_{m,n-1} + \mu_1 p_{m+1,n}, && m, n > 0,
\end{aligned} \tag{6.7.3}$$

and $\sum p_{mn} = 1$.

Again, for the complete analytical solution of these equations, the use of probability generating functions is essential. As in the case of nonpreemptive priority, if we think of numerical solutions of these equations when the capacity limit of the system is K, the 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, compared to 111 for the nonpreemptive 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 twofold: 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 at which 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 a 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 that 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 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, the class $i-1$ completion time is the high-priority service time as seen by the class i customer. We shall not go into the analysis of such systems because of their complexity. Interested readers may refer to Jaiswal (1968) and journal articles that have appeared since then.

Example 6.7.1. 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 nonpreemptive and preemptive priority disciplines.

Nonpreemptive priority. This discipline assumes that once a nonpriority customer starts service, it is carried out to conclusion even if a priority customer arrives in the meantime. 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 as follows ($\lambda = \lambda_1 + \lambda_2$):

$$\lambda p_0 = \mu_1 p_{101} + \mu_2 p_{012}, \tag{6.7.4}$$

$$(\lambda + \mu_1) p_{101} = \lambda_1 p_0 + \mu_1 p_{201} + \mu_2 p_{112}, \tag{6.7.5}$$

$$(\lambda + \mu_2) p_{012} = \lambda_2 p_0 + \mu_1 p_{111} + \mu_2 p_{022}, \tag{6.7.6}$$

$$\mu_1 p_{201} = \lambda_1 p_{101}, \tag{6.7.7}$$

$$\mu_2 p_{022} = \lambda_2 p_{012}, \tag{6.7.8}$$

$$\mu_1 p_{111} = \lambda_2 p_{101}, \tag{6.7.9}$$

$$\mu_2 p_{112} = \lambda_1 p_{012}, \tag{6.7.10}$$

$$\sum p_{mnr} = 1.$$

Substituting from (6.7.6) and (6.7.9) in (6.7.4),

$$(\lambda + \mu_1) p_{101} = \lambda_1 p_0 + \lambda_1 p_{101} + \lambda_1 p_{012},$$

$$(\lambda_2 + \mu_1) p_{101} - \lambda_1 p_{012} = \lambda_1 p_0. \tag{6.7.11}$$

But from (6.7.4),

$$\mu_1 p_{101} + \mu_2 p_{012} = \lambda p_0.$$

Solving for p_{101} and p_{012},

$$p_{101} = \frac{\lambda_1(\lambda + \mu_2)}{\lambda_1\mu_1 + \lambda_2\mu_2 + \mu_1\mu_2} p_0 = A p_0, \quad \text{say}, \tag{6.7.12}$$

$$p_{012} = \frac{1}{\mu_2}\left[\lambda - \frac{\lambda_1\mu_1(\lambda+\mu_2)}{\lambda_1\mu_1+\lambda_2\mu_2+\mu_1\mu_2}\right]p_0 = Bp_0, \quad \text{say.} \tag{6.7.13}$$

Substituting these values in (6.7.7)–(6.7.10), we get

$$\begin{aligned} p_{201} &= \frac{\lambda_1}{\mu_1}Ap_0, \qquad & p_{111} &= \frac{\lambda_2}{\mu_1}Ap_0, \\ p_{022} &= \frac{\lambda_2}{\mu_2}Bp_0, \qquad & p_{112} &= \frac{\lambda_1}{\mu_2}Bp_0. \end{aligned} \tag{6.7.14}$$

Now p_0 is obtained from the normalizing condition $\sum p_{mnr} = 1$:

$$p_0 = \left[1 + \left(1+\frac{\lambda}{\mu_1}\right)A + \left(1+\frac{\lambda}{\mu_2}\right)B\right]^{-1}. \tag{6.7.15}$$

Summarizing, we get

$$\begin{aligned} P(0) &= P(\text{service counter idle}) = p_0 = \left[1 + \left(1+\frac{\lambda}{\mu_1}\right)A + \left(1+\frac{\lambda}{\mu_2}\right)B\right]^{-1}, \\ P(1) &= P(\text{class 1 in service}) = p_{101}+p_{201}+p_{111} = \left(1+\frac{\lambda}{\mu_1}\right)Ap_0, \\ P(2) &= P(\text{class 2 in service}) = p_{012}+p_{022}+p_{112} = \left(1+\frac{\lambda}{\mu_2}\right)Bp_0. \end{aligned} \tag{6.7.16}$$

To compare the two disciplines, let $\lambda_1 = 3$/hour, $\lambda_2 = 2$/hour, and the mean service times be 30 minutes each ($\mu_1 = \mu_2 = 2$/hour). Then we get the following results:

$$\begin{aligned} A &= \frac{3}{2}; \qquad & B &= 1, \\ p_0 &= \frac{4}{39}, \\ p_{101} &= \frac{6}{39}; \qquad & p_{201} &= \frac{9}{39}; \qquad & p_{111} &= \frac{6}{39}, \\ p_{012} &= \frac{4}{39}; \qquad & p_{022} &= \frac{4}{39}; \qquad & p_{112} &= \frac{6}{39}, \\ P(0) &= 0.103; \qquad & P(1) &= 0.538; \qquad & P(2) &= 0.359. \end{aligned}$$

Answer

Preemptive priority. Because of the preemptive nature of the discipline, we further assume that when there are two class 2 customers in the system and one of them is in service, an arriving class 1 customer will displace the one in service. This means that 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 ($\lambda = \lambda_1 + \lambda_2$):

$$\begin{aligned} \lambda p_0 &= \mu_1 p_{10} + \mu_2 p_{01}, \\ (\lambda+\mu_1)p_{10} &= \lambda_1 p_0 + \mu_1 p_{20}, \end{aligned}$$

$$\mu_1 p_{20} = \lambda_1(p_{10} + p_{11}),$$
$$(\lambda + \mu_2)p_{01} = \lambda_2 p_0 + \mu_1 p_{11} + \mu_2 p_{02},$$
$$(\lambda_1 + \mu_2)p_{02} = \lambda_2 p_{01},$$
$$(\lambda_1 + \mu_1)p_{11} = \lambda_2 p_{10} + \lambda_1(p_{01} + p_{02}),$$
$$\sum p_{mn} = 1.$$

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, we write

$$\lambda = a; \quad \mu_1 = b; \quad \mu_2 = c; \quad \lambda + \mu_1 = d; \quad \lambda_1 = e;$$
$$\lambda + \mu_2 = g; \quad \lambda_2 = h; \quad \lambda_1 + \mu_2 = j; \quad \lambda_1 + \mu_1 = k.$$

The second-to-last equation allows us to write

$$p_{02} = \frac{h}{j} p_{01}. \tag{6.7.17}$$

Eliminating p_{02} from the set of equations, we get

$$bp_{10} + cp_{01} = ap_0, \tag{6.7.18}$$
$$dp_{10} - bp_{20} = ep_0, \tag{6.7.19}$$
$$ep_{10} - bp_{20} + ep_{11} = 0, \tag{6.7.20}$$
$$-mp_{01} - bp_{11} = hp_0, \tag{6.7.21}$$
$$hp_{10} + np_{01} - kp_{11} = 0, \tag{6.7.22}$$

where we have written $c(\frac{h}{j}) - g = m$ and $e(1 + \frac{h}{j}) = n$.

Eliminating p_{11} from (6.7.20) and (6.7.21),

$$bep_{10} - b^2 p_{20} - emp_{01} = hep_0. \tag{6.7.23}$$

Eliminating p_{01} from (6.7.18) and (6.7.23),

$$(bem + bce)p_{10} - b^2 cp_{20} = (aem + che)p_0. \tag{6.7.24}$$

Eliminating p_{20} from (6.7.19) and (6.7.24),

$$b(em + ce - cd)p_{10} = e(am + ch - bc)p_0,$$

giving

$$p_{10} = \frac{e(am + ch - bc)}{b(em + ce - cd)} p_0 = Ap_0, \quad \text{say.} \tag{6.7.25}$$

Using this result in (6.7.19), we get

$$p_{20} = \frac{1}{b}(dA - e)p_0 = Bp_0, \quad \text{say.} \tag{6.7.26}$$

Substituting the values of p_{10} and p_{20} in (6.7.23),

$$\begin{aligned} p_{01} &= \frac{1}{em}[beA - b^2B - eh]p_0 \\ &= Cp_0, \quad \text{say.} \end{aligned} \tag{6.7.27}$$

Substituting this result in (6.7.21),

$$\begin{aligned} p_{11} &= \frac{-1}{b}[mC + h]p_0 \\ &= Dp_0, \quad \text{say.} \end{aligned} \tag{6.7.28}$$

Finally, going back to (6.7.17),

$$p_{02} = \frac{h}{j}Cp_0 = Ep_0, \quad \text{say.} \tag{6.7.29}$$

Using the results from (6.7.25)–(6.7.29) in the normalizing condition $\sum p_{ij} = 1$, we get

$$p_0 = (1 + A + B + C + D + E)^{-1}. \tag{6.7.30}$$

With the numerical values used in the nonpreemptive 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

$$P(0) = 0.073, \qquad P(1) = 0.849, \qquad P(2) = 0.077.$$ **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, 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.

Consider a nonpreemptive priority queue 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 $\sigma_k = \sum_{j=1}^{k} \rho_i = \rho < 1$.

Let $T_q^{(i)}$ be the waiting time of the customer. (This is the time during which 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 the customer 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 serivce 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

$$T_q^{(i)} = S_0 + \sum_{r=1}^{i} S_r + \sum_{r=1}^{i-1} S'_r. \tag{6.7.31}$$

Taking expectations,

$$W_q^{(i)} = E(S_0) + \sum_{r=1}^{i} E(S_r) + \sum_{r=1}^{i-1} E(S'_r). \tag{6.7.32}$$

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. Since the class of the customer is not known, we may use $\frac{\rho_r}{\rho}$ 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 $= \rho$. Combining these terms, we get

$$E(S_0) = \sum_{r=1}^{k} \frac{1}{\mu_r}\left(\frac{\rho_r}{\rho}\right)\rho = \sum_{r=1}^{k} \frac{\rho_r}{\mu_r}. \tag{6.7.33}$$

Let $S_r^{(j)}$ 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

$$\begin{aligned} E(S_r) &= E(n_r)E(S_r^{(j)}) \\ &= E(n_r) \cdot \frac{1}{\mu_r}. \end{aligned} \tag{6.7.34}$$

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

$$E(S_r) = \frac{\lambda_r W_q^{(r)}}{\mu_r} = \rho_r W_q^{(r)}. \tag{6.7.35}$$

We may use similar arguments for the total service time of the higher priority customers arrivng 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

$$\begin{aligned} E(S'_r) &= \frac{\lambda_r W_q^{(i)}}{\mu_r} \\ &= \rho_r W_q^{(i)}. \end{aligned} \tag{6.7.36}$$

Combining the three expressions from (6.7.33), (6.7.35), and (6.7.36), we get

$$W_q^{(i)} = E(S_0) + \sum_{r=1}^{i} \rho_r W_q^{(r)} + \sum_{r=1}^{i-1} \rho_r W_q^{(i)}, \tag{6.7.37}$$

$$[1 - \rho_i - \sigma_{i-1}] W_q^{(i)} = E(S_0) + \sum_{r=1}^{i-1} \rho_r W_q^{(r)}$$

$$W_q^{(i)} = \frac{1}{1-\sigma_i} \left[\sum_{r=1}^{i-1} \rho_r W_q^{(r)} + E(S_0) \right]. \tag{6.7.38}$$

From (6.7.37), 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

$$W_q^{(i)} = \frac{E(S_0)}{(1-\sigma_{i-1})(1-\sigma_i)}. \tag{6.7.39}$$

From (6.7.37), we get

$$(1-\sigma_{i-1}) W_q^{(i)} = \sum_{r=1}^{i} \rho_r W_q^{(r)} + E(S_0),$$

$$W_q^{(i)} = \frac{\sum_{r=1}^{i} \rho_r W_q^{(r)} + E(S_0)}{1-\sigma_{i-1}}. \tag{6.7.40}$$

Equating the right-hand sides of (6.7.39) and (6.7.40), we get

$$\sum_{r=1}^{i} \rho_r W_q^{(r)} = \frac{\sigma_i E(S_0)}{1-\sigma_i}. \tag{6.7.41}$$

Now using the assumed form of $W_q^{(i)}$ from (6.7.39) in (6.7.38) and using (6.7.41),

$$\begin{aligned} W_q^{(i+1)} &= \frac{1}{(1-\sigma_{i+1})}\left[\sum_{r=1}^{i} \rho_r W_q^{(r)} + E(S_0)\right] \\ &= \frac{1}{(1-\sigma_{i+1})}\left[\frac{\sigma_i E(S_0)}{1-\sigma_i} + E(S_0)\right] \\ &= \frac{E(S_0)}{(1-\sigma_i)(1-\sigma_{i+1})}, \end{aligned} \tag{6.7.42}$$

which shows by induction that the general form of $W_q^{(i)}$ is given by

$$W_q^{(i)} = \frac{E(S_0)}{(1-\sigma_{i-1})(1-\sigma_i)}. \tag{6.7.43}$$

Substituting from (6.7.33)

$$W_q^{(i)} = \frac{\sum_{r=1}^{k}(\frac{\rho_r}{\mu_r})}{(1-\sigma_{i-1})(1-\sigma_i)}. \tag{6.7.44}$$

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 in (6.7.43) except that $E(S_0)$ is determined as follows.

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

$$\rho_r = \lambda_r E(S_r^{(i)}).$$

Now from the arguments used in deriving (6.7.33), we get

$$\begin{aligned} E(S_0) &= \sum_{r=1}^{k} \frac{E[(S_r^{(j)}]^2}{2E(S_r^{(j)})} \cdot \frac{\lambda_r E(S_r^{(j)})}{\rho} \cdot \rho \\ &= \sum_{r=1}^{k} \frac{1}{2}\lambda_r E[(S_r^{(j)})^2], \end{aligned} \tag{6.7.45}$$

giving

$$W_q^{(i)} = \frac{\sum_{r=1}^{k} \lambda_r E[(S_r^{(j)})^2]}{2(1-\sigma_{i-1})(1-\sigma_i)}, \tag{6.7.46}$$

where we have used $S_r^{(j)}$ to denote the service time of a customer in the rth priority class.

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 (6.7.46), 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 analogue. Details of the procedure can be found in Coffman and Denning (1973).

6.8 Exercises

Note: Exercises in this chapter may require the use of computational tools.

1. In a service system groups of customers arrive and are 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 interarrival 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 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 a specific station, it would have passed through three stations, at each of which it would have spent an amount of time that is exponentially distributed with mean 3 minutes. The assembly time at the specific station is exponential with mean 5 minutes. Determine the expected number of products waiting at the station for assembly, 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 interarrival 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 6.7.1 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 nonpreemptive 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 assignments 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. Customers C_1, C_2, and C_3 call for service with rates λ_1, λ_2, and λ_3 and are serviced 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 as 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 are 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.

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 are 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 is 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 7.1.1 is a simple illustration of a network in which the sequencing of the service is shown by directional arrows between the nodes. Figure 7.1.1 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 we 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 exit 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

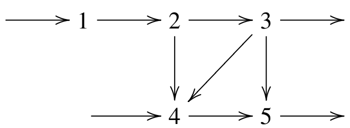

Fig. 7.1.1. Open network.

U.N. Bhat, *An Introduction to Queueing Theory*, DOI: 10.1007/978-0-8176-4725-4_7,

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 is sometimes easier to analyze, depending on its structure.)

As we shall point out later, queueing networks generally 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 by analyzing the node networks. A node network is often called the *routing chain*.

7.2 The Markovian Node Network

Consider a network of nodes $\{1, 2, \ldots, k\}$. After being served at node i, suppose a customer moves to node j with probability P_{ij} $(i, j = 1, 2, \ldots, k)$. A 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

$$\boldsymbol{P} = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ \vdots \\ k-1 \\ k \end{array} \begin{array}{c} \begin{array}{ccccc} 1 & 2 & \cdots & & k \end{array} \\ \left[\begin{array}{ccccc} 0 & P_{01} & P_{02} & \cdots & P_{0k} \\ P_{10} & P_{11} & P_{12} & \cdots & P_{1k} \\ P_{20} & P_{21} & P_{22} & \cdots & P_{2k} \\ \vdots & \vdots & & & \\ P_{k-1,0} & P_{k-1,1} & P_{k-1,2} & \cdots & P_{k-1,k} \\ P_{k0} & P_{k1} & P_{k2} & \cdots & P_{k,k} \end{array}\right] \end{array}. \tag{7.2.1}$$

For the closed network, 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. Note 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 I/O 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 $\boldsymbol{\lambda P} = \boldsymbol{\lambda}$, 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 $\boldsymbol{\lambda P} = \boldsymbol{\lambda}$ 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 $= \sum_{j=1}^{k} \nu_{ij}(\frac{1}{\mu_j})$. 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 **P** after converting state 0 to be absorbing.

The transition probability matrix (7.2.1) has the following structure when state 0 is made absorbing:

$$\boldsymbol{P} = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ \vdots \\ k \end{array} \begin{array}{c} \begin{array}{ccccc} 0 & 1 & 2 & \ldots & k \end{array} \\ \begin{bmatrix} 1 & 0 & 0 & \ldots & 0 \\ P_{10} & P_{11} & P_{12} & \ldots & P_{1k} \\ P_{20} & P_{21} & P_{22} & \ldots & P_{2k} \\ \vdots & \vdots & & & \\ P_{k0} & P_{k1} & P_{k2} & \ldots & P_{k,k} \end{bmatrix} \end{array} \tag{7.2.2}$$

$$= \begin{bmatrix} 1 & 0 \\ \boldsymbol{R} & \boldsymbol{Q} \end{bmatrix}. \tag{7.2.3}$$

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 $(\boldsymbol{I} - \boldsymbol{Q})^{-1}$, which is known as the *fundamental matrix*, give the expected number of visits of the Markov chain to the various states before it ultimately visits the absorbing state 0. Let

$$(\boldsymbol{I} - \boldsymbol{Q})^{-1} = \begin{bmatrix} \nu_{11} & \nu_{12} & \ldots & \nu_{1k} \\ \nu_{21} & \nu_{22} & \ldots & \nu_{2k} \\ \vdots & \vdots & & \vdots \\ \nu_{k1} & \nu_{k2} & \ldots & \nu_{kk} \end{bmatrix}. \tag{7.2.4}$$

Suppose that the Markov chain is initially in state i. The expected numbers of visits of the process to states $1, 2, \ldots, k$ before it is 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).

Even though we have used $1/\mu_j$ as the mean service time of a customer in node j, there is nothing to prevent us from using $1/\mu_j$ as 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):

$$\boldsymbol{P} = \begin{array}{c} \\ 0 \\ k+1 \\ 1 \\ 2 \\ \vdots \\ k \end{array} \begin{array}{c} \begin{array}{cccccc} 0 & k+1 & 1 & 2 & \dots & k \end{array} \\ \begin{bmatrix} 1 & & & & & \\ & 1 & & & & \\ P_{10} & P_{1,k+1} & P_{11} & P_{12} & \dots & P_{1k} \\ P_{20} & P_{2,k+1} & P_{21} & P_{22} & \dots & P_{2k} \\ \vdots & \vdots & \vdots & \vdots & & \vdots \\ P_{k0} & P_{k,k+1} & P_{k1} & P_{k2} & \dots & P_{k,k} \end{bmatrix} \end{array} \tag{7.2.5}$$

$$= \begin{bmatrix} \boldsymbol{I} & 0 \\ \boldsymbol{R} & \boldsymbol{Q} \end{bmatrix}. \tag{7.2.6}$$

Let f_{ij} be the probability that starting from state i $(i = 1, 2, \dots, m)$, the Markov chain ultimately is absorbed in j $(j = 0, k+1)$. Define

$$\boldsymbol{F}' = \begin{bmatrix} f_{10} & f_{20} & \cdots & f_{k0} \\ f_{1,k+1} & f_{2,k+1} & \cdots & f_{k,k+1} \end{bmatrix}. \tag{7.2.7}$$

Again appealing to the theory of finite Markov chains (Bhat and Miller (2002)), we have

$$\boldsymbol{F} = (\boldsymbol{I} - \boldsymbol{Q})^{-1}\boldsymbol{R}, \tag{7.2.8}$$

where $\boldsymbol{R}$ 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 $R = 1 - P(\text{system failure}) = \sum_{i=1}^{k} \alpha_i \sum_{j=1}^{k} \nu_{ij} P_{j0}$.

For a finite time reliability analysis of a Markovian network, readers are referred to Bhat and Kavi (1987).

7.3 Queues in Series

The simplest open queueing network structure is that in which 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 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 that the waiting rooms feeding customers to these stations have infinite capacity.

To illustrate the behavior of a series of queues, we 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), $n_1, n_2 = 0, 1, 2, \ldots$. 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 7.3.1.

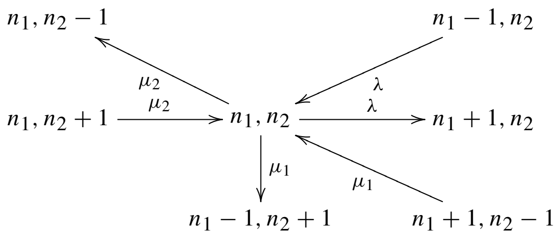

Fig. 7.3.1. Transition diagram.

Writing the corresponding state balance equations, we get

$$\begin{aligned}
\lambda p_{00} &= \mu_2 p_{01}, \\
(\lambda + \mu_2) p_{0n_2} &= \mu_1 p_{1,n_2-1} + \mu_2 p_{0,n_2+1}, \\
(\lambda + \mu_1) p_{n_1,0} &= \mu_2 p_{n_1,1} + \lambda p_{n_1-1,0}, \\
(\lambda + \mu_1 + \mu_2) p_{n_1 n_2} &= \mu_1 p_{n_1+1,n_2-1} + \mu_2 p_{n_1,n_2+1} \\
&\quad + \lambda p_{n_1-1,n_2}, \quad n_1, n_2 > 0.
\end{aligned} \tag{7.3.2}$$

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 equations (7.3.2). Now using the normalizing condition

$$\sum_{n_1=0}^{\infty}\sum_{n_2=0}^{\infty} p_{n_1n_2} = 1,$$

we get

$$p_{00} = (1-\rho_1)(1-\rho_2).$$

Hence

$$p_{n_1n_2} = (1-\rho_1)(1-\rho_2)\rho_1^{n_1}\rho_2^{n_2}. \tag{7.3.4}$$

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

$$p_{n_1n_2\cdots n_k} = \Pi_{i=1}^{k}(1-\rho_i)\rho_i^{n_i} \tag{7.3.5}$$

(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 $M/M/s$-type queue was also Poisson with the same rate as the arrival process, as $t \to \infty$. In the queueing network literature this property is 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 a 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:

$$\lambda p_{n_1n_2} = \mu_2 p_{n_1,n_2+1}, \qquad n_1, n_2 = 0, 1, 2, \ldots,$$

$$\mu_1 p_{n_1 n_2} = \lambda p_{n_1 - 1, n_2}, \qquad n_1 = 1, 2, \ldots; \quad n_2 = 0, 1, 2, \ldots,$$
$$\mu_2 p_{n_1, n_2} = \mu_1 p_{n_1+1, n_2-1}, \qquad n_1 = 0, 1, 2, \ldots; \quad n_2 = 1, 2, \ldots. \tag{7.3.6}$$

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, 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 (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_{0n_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) $M \to M$, (2) local balance, and (3) the product-form solution. A general approach to the analysis of these systems is provided in Section 7.5.

7.4 Queues with Blocking

The analysis becomes 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 with an example.

Example 7.4.1. A machine repair has two stages, and there are two repairman 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 the machine that has completed the first stage of repair. 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 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 follows:

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 7.4.1.

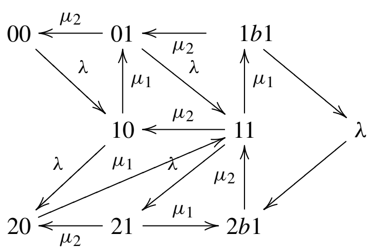

Fig. 7.4.1. Transition diagram.

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:

$$\lambda p_{00} = \mu_2 p_{01},$$
$$(\lambda + \mu_2) p_{01} = \mu_1 p_{10} + \mu_2 p_{1b,1},$$

$$
\begin{aligned}
(\lambda + \mu_1)p_{10} &= \lambda p_{00} + \mu_2 p_{11},\\
(\lambda + \mu_1 + \mu_2)p_{11} &= \lambda p_{01} + \mu_2 p_{2b,1} + \mu_1 p_{20},\\
(\lambda + \mu_2)p_{1b,1} &= \mu_1 p_{11},\\
\mu_1 p_{20} &= \lambda p_{10} + \mu_2 p_{21},\\
(\mu_1 + \mu_2)p_{21} &= \lambda p_{11},\\
\mu_2 p_{2b,1} &= \lambda p_{1b,1} + \mu_1 p_{21}.
\end{aligned}
$$

Solving these equations with appropriate substitutions, we get

$$
\begin{array}{lll}
p_{01} = \dfrac{\lambda}{\mu_2} p_{00}; & p_{10} = A p_{00}; & p_{11} = B p_{00},\\
p_{1b,1} = C p_{00}; & p_{21} = D p_{00}; & p_{20} = E p_{00},\\
p_{2b,1} = F p_{00},
\end{array}
$$

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

$$
p_{00} = \left[1 + \frac{\lambda}{\mu_2} + A + B + C + D + E + F\right]^{-1}.
$$

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 rule used in the example above—will change the transition diagram significantly, requiring rewriting of the equations and reworking of 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 cite Perros (1994).

7.5 Open Jackson Networks

Suppose in the Markovian node network of Section 7.2, each node represents an $M/M/s$ 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 $M/M/s$ 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 that 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 let 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

$$p_{n_1 n_2 \cdots n_k} = P(Q_1 = n_1, Q_2 = n_2, \ldots, Q_k = n_k). \tag{7.5.1}$$

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 $p_{n_1 n_2 \cdots n_k}$ of (7.5.1), Jackson has shown that

$$p_{n_1 n_2 \cdots n_k} = p_1(n_1) p_2(n_2) \cdots p_k(n_k), \tag{7.5.2}$$

where

$$p_i(r) = \begin{cases} p_i(0) \dfrac{(\gamma_i/\mu_i)^r}{r!}, & r = 0, 1, 2, \ldots, s_i, \\ p_i(0) \dfrac{(\gamma_i/\mu_i)^r}{s_i! s_i^{r-s_i}}, & r = s_i, s_i + 1, \ldots, \end{cases} \tag{7.5.3}$$

and

$$\gamma_i = \lambda_i + \sum_j \alpha_{ji} \gamma_j, \quad i = 1, 2, \ldots, k. \tag{7.5.4}$$

Given λ_i and α_{ij} $(i, j = 1, 2, \ldots, k)$, the quantity γ_i can be determined from (7.5.4). Note 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

$$\sum_{n_1} \sum_{n_2} \cdots \sum_{n_k} p_{n_1 n_2 \cdots n_k} = 1.$$

The structure of the distribution $p_i(r)$ in (7.5.3) is similar to the limiting distribution of the queue $M/M/s_i$ 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 feedforward transitions, we could apply Burke's (1956) result on the departure process and conclude that as $t \to \infty$, the feedforward 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 interarrival time distribution is a mixture of exponentials (i.e., hyperexponential). Thus from the limiting distribution $p_{n_1 n_2 \cdots n_k}$ of (7.5.3), which is the product of limiting distributions of $M/M/s_i$ queueing systems, the only conclusion we can draw is that in the limit, the Jackson network behaves as if it is a series of $M/M/s_i$ 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 (see 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 the state balance equations as follows:

$$\begin{aligned}
(\lambda_1 + \lambda_2) p_{00} &= \mu_1 \alpha_{10} p_{10} + \mu_2 \alpha_{20} p_{01}, \\
(\lambda_1 + \lambda_2 + \mu_1) p_{10} &= \lambda_1 p_{00} + \mu_2 \alpha_{21} p_{01} + \mu_1 \alpha_{10} p_{20}, \\
(\lambda_1 + \lambda_2 + \mu_2) p_{01} &= \lambda_2 p_{00} + \mu_1 \alpha_{12} p_{10} + \mu_2 \alpha_{20} p_{02}, \\
(\lambda_1 + \lambda_2 + \mu_1 + \mu_2) p_{11} &= \lambda_1 p_{01} + \lambda_2 p_{10} \\
&\quad + \mu_1 \alpha_{10} p_{21} + \mu_2 \alpha_{20} p_{12} \\
&\quad + \mu_1 \alpha_{12} p_{20} + \mu_2 \alpha_{21} p_{02}, \\
&\ \vdots \\
(\lambda_1 + \lambda_2 + \mu_1 + \mu_2) p_{n_1 n_2} &= \lambda_1 p_{n_1 - 1, n_2} + \lambda_2 p_{n_1, n_2 - 1} \\
&\quad + \mu_1 \alpha_{10} p_{n_1 + 1, n_2} + \mu_2 \alpha_{20} p_{n_1, n_2 + 1} \\
&\quad + \mu_1 \alpha_{12} p_{n_1 + 1, n_2 - 1} + \mu_2 \alpha_{21} p_{n_1 - 1, n_2 + 1}, \\
n_1, n_2 &> 0. \qquad (7.5.5)
\end{aligned}$$

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),

$$\gamma_1 = \frac{\lambda_1 + \lambda_2 \alpha_{21}}{1 - \alpha_{12} \alpha_{21}},$$

$$\gamma_2 = \frac{\lambda_2 + \lambda_1 \alpha_{12}}{1 - \alpha_{12}\alpha_{21}}. \tag{7.5.7}$$

Write $\rho_i = \frac{\gamma_i}{\mu_i}$, $i = 1, 2$. Suppose that a trial solution is

$$p_{n_1 n_2} = C \rho_1^{n_1} \rho_2^{n_2}. \tag{7.5.8}$$

Verifying that (7.5.8) is, in fact, the correct solution to the state balance equation (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 and Harris (1998).

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 moving within the network.

Following the same reasoning as in open networks with k nodes, and the ith node supporting s_i servers ($i = 1, 2, \ldots$), the limiting distribution $p_{n_1 n_2 \cdots n_k} = P(Q_1 = n_1, Q_2 = n_2, \ldots, Q_k = n_k)$ can be obtained in the product form as

$$p_{n_1 n_2 \cdots n_k} = C \Pi_{i=1}^{k} \frac{\rho_i^{n_i}}{a_i(n_i)}, \tag{7.6.1}$$

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 $\sum_{n_1 n_2 \cdots n_k} p_{n_1 n_2} \cdots n_k = 1$. We note here that the term "product form" is used only for 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 ρ_i, $i = 1, 2, \ldots, k$, we note that since the total traffic is known, only $k - 1$ equations are independent. Hence we start by setting one of the ρ_is equal to 1.

The determination of $C \equiv C(Q)$ is not a simple problem. We have

$$C^{-1}(Q) = [C(Q)]^{-1} = \sum_{n_1+n_2+\cdots+n_k=Q} \Pi_{i=1}^{k} \frac{\rho_i^{n_i}}{a_i(n_i)}, \tag{7.6.4}$$

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.) Thus determining $C^{-1}(Q)$ directly from (7.6.4) is easy only for small values of Q and k, even with the help of a computer. One of the earliest algorithms to compute $G(Q) = C^{-1}(Q)$ systematically was 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

$$g_m(n) = \sum_{n_1+n_2+\cdots+n_k=n} \Pi_{i=1}^{m} f_i(n_i) \tag{7.6.6}$$

and $g_k(Q) = G(Q)$ ($m = k$ and $n = Q$). We may write

$$\begin{aligned} g_m(n) &= \sum_{r=0}^{n} \left[\sum_{n_1+n_2+\cdots+n_{m-1}+r=n} \Pi_{i=1}^{m} f_i(n_i) \right] \\ &= \sum_{r=0}^{n} f_m(r) \left[\sum_{n_1+n_2+\cdots+n_{m-1}=n-r} \Pi_{i=1}^{m-1} f_i(n_i) \right] \\ &= \sum_{r=0}^{n} f_m(r) g_{m-1}(n-r), \quad n = 0, 1, 2, \ldots, Q. \end{aligned} \tag{7.6.7}$$

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 12.

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, 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 and Harris (1998).

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 (see 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 notation 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

$$p_{n_1,n_2,\ldots,n_k} = \frac{1}{G(Q)} \frac{\mu_1^{Q-n_1}}{\mu_2^{n_2}\mu_3^{n_3}\cdots\mu_k^{n_k}} \tag{7.7.4}$$

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 that 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 $\sum_{n_1} p_{n_1,Q-n_1} = 1$,

$$\frac{1}{G(Q)}\sum_{n_1=0}^{Q}\left(\frac{\mu_1}{\mu_2}\right)^{Q-n_1} = 1,$$

$$G(Q) = \frac{1-\left(\frac{\mu_1}{\mu_2}\right)^{Q+1}}{1-\frac{\mu_1}{\mu_2}}.$$

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 Operational Laws for Performance Analysis

Computer systems lend themselves easily for modeling as queueing networks. They require measures and relationships additional to what we have discussed so far in their performance analysis. We introduce a few of them here preparatory to the modeling and analysis in applications to be illustrated in Chapter 12. For elaborations and extensions of these performance measures and relationships, commonly known as *operational laws* in the computer science literature, the readers may refer to the articles and books cited here as well as elsewhere in the text. Our discussion follows the 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 simple central server network that includes a central processing unit (CPU) and k I/O stations (devices). A job begins with the CPU and continues with zero or more I/O services. Assuming that a new job enters the system as soon as an active job completes service we make this a closed network. With the assumption of exponential distributions for service, we have a Markovian network in which conditions for state balance exist.

Assume that the system operation has been observed long enough to be able to estimate the following quantities. Let T be the observation period and B_i be the time that device i has been busy providing service. Also, let C_{ij} be the number of times a job requests service at device j immediately after computing 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 (note that the original notation is used for the estimates):

$$\text{utilization } U_i = \rho_i = \frac{B_i}{T},$$
$$\text{effective output rate } \gamma_i = \frac{C_i}{T},$$
$$\text{routing probabilities } \alpha_{ij} = \frac{C_{ij}}{C_i}.$$

Note that we are assuming input and output flow balance in these expressions. Since the CPU is where the job is initiated and completed, using a subscript 0 to indicate its status, we have the job-flow balance equations,

$$\gamma_0 = \sum_{i=1}^{k} \gamma_i \alpha_{i0}. \tag{7.8.1}$$

To be consistent with the computer science literature we use the term "*response time*" instead of "system time" (i.e., waiting + service). The response time R_i at device i can be estimated as (the total amount of time accumulated by a device)/(the number of services completed at the device). If Q_i represents the number at the device waiting for or being served in the long run, using Little's law ($L = \lambda W$), for each device we get

$$E(Q_i) = \gamma_i E(R_i). \tag{7.8.2}$$

In Markovian networks, job flow is balanced and therefore γ_i can be identified as the device *throughput*. These quantitites also give us the *visit ratios*, which are the mean number of service requests per job for a device relative to the mean number of jobs coming into the system. The visit ratio V_i for device i can be defined as

$$V_i = \frac{\gamma_i}{\gamma_0}$$

and estimated as C_i/C_0. The relation

$$\gamma_i = V_i \gamma_0 \tag{7.8.3}$$

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 (7.8.3) into (7.8.1), we obtain the visit ratio equations

$$V_0 = 1, \qquad V_j = \alpha_{0j} + \sum_{i=1}^{k} V_i \alpha_{ij}, \quad j = 1, 2, \ldots, k. \tag{7.8.4}$$

The system response time R is obtained by pooling the response times of all devices. From Little's law, writing $E(Q) = \sum_{i=1}^{k} E(Q_i)$, we have

$$E(R) = E(Q)/\gamma_0$$

Using (7.8.2) and (7.8.3), we get

$$E(R) = \sum_{i=1}^{i} V_i E(R_i), \tag{7.8.5}$$

which is known as the *general response time law*. This result is valid even when the network is not Markovian.

If M computers are sending jobs to the terminal, for the terminal user, its think time Z (preparing to submit a job), together with the response time, forms the think–wait cycle. When the job flow is balanced, we have

$$M = [E(Z) + E(R)]\gamma_0,$$

giving

$$E(R) = \frac{M}{\gamma_0} - E(Z). \tag{7.8.6}$$

This relationship is known as the *interactive response time law*.

Denning and Buzen (1978) includes several illustrative examples of these relationships. (See also Jain (1991) for further elaborations on these laws.)

Another important analysis procedure used in applications is the *mean value analysis*. Ordinarily, to determine the response times in networks, one has to obtain 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 the 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 as $t \to \infty$. The recursion established by Reiser and Lavenberg is the relationship

$$E[R_j(N)] = E(S_j)\{1 + E[Q_j(N - 1)]\}.$$

The derivation of this relationship is omitted because of its complexity. A numerical illustration of its use in the mean value analysis is provided in Chapter 12. The readers should to note that in the numerical illustration, the notation is simplified by dropping the expected value operator E.

7.9 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 the 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* mentioned in the last section is one such example. In the category of approximations, we have 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 (1998).

The paper by Baskett et al. (1975) on 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 exponential (e.g., Coxian) and service disciplines other than 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), Kelly (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.

7.10 Exercises

Note: Exercises in this chapter may require the use of computational tools.

1. Solve Example 7.4.1 with a 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 ongoing 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 machines with each mechanic.
   Solve the problem under service disciplines in the 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 $n_1 = 0, 1, 2, \dots$ and $n_2 = 0, 1, 2, \dots, m, m+1$. 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), μ_2; I/O(2), μ_3; file server, μ_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:

$$N = 40; \quad \lambda = 0.01/\text{second}; \quad \beta_1 = 0.6; \quad \beta_3 = 0.4,$$
$$\frac{1}{\mu_1} = 0.8 \text{ second}; \quad \frac{1}{\mu_2} = 0.3 \text{ second}; \quad \frac{1}{\mu_3} = 0.6 \text{ second}; \quad \frac{1}{\mu_4} = 1 \text{ second}$$

(Krakowiak (1988)).

6. Consider the following motor vehicle registration process with four stations (reception, clerk 1, clerk 2, cashier):
   (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 average 10 minutes, the amount of time having an exponential distribution.
   (c) *Clerk* 2 handles standard in-state renewal applications and takes on 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 minute.

   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 $C_N \to C_{N-1} \to \cdots \to C_2 \to C_1$. 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. (See Bhat et al. (1975); see also Bhat (1984), Section 13.5.)

8

Renewal Process Models

Up until now, we have used only Markov models for analyzing queueing systems. As indicated in several places, Markov models are not general enough to provide a complete analysis of the systems discussed. For instance, the queue length process $\{Q(t),\ t \in T\}$ in the queue $M/G/1$ is Markovian only at departure points and the process $\{Q(t),\ t \in T\}$ in $G/M/1$ is Markovian only at arrival points. Even though we can use a Markov process analysis technique to analyze these processes 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 $G/M/1$, explicit results are very difficult to obtain. In this chapter, we provide an alternative approach based on renewal processes. Since only very brief mention has been made of the renewal process earlier in Section 3.2, we start with an introduction to the basics of the process in the next section.

8.1 Renewal Process

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

$$P(Z_i \leq x) = F(x). \tag{8.1.1}$$

The process consisting of the sequence of random variables $(Z_1, Z_2, \dots)$ is known as a *renewal process*. Let $N(t)$ be the number of events occurring in $(0, t]$. This 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 interarrival times form a sequence of

U.N. Bhat, *An Introduction to Queueing Theory*, DOI: 10.1007/978-0-8176-4725-4_8,

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 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 $S_n = Z_1 + Z_2 + \cdots + Z_n$. 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

$$\phi(\theta) = \int_0^\infty e^{-\theta x} dF(x), \quad \mathrm{Re}(\theta) > 0, \tag{8.1.2}$$

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

$$\int_0^\infty e^{-\theta x} dF_n(x) = [\phi(\theta)]^n. \tag{8.1.3}$$

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

$$P_n(t) = P[N(t) = n]. \tag{8.1.4}$$

Consider two events

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

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

$$P_n(t) = F_n(t) - F_{n+1}(t). \tag{8.1.5}$$

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 (8.1.5), it is easy to show that

$$\begin{aligned} U(t) = E[N(t)] &= \sum_{n=1}^{\infty} n P_n(t) \\ &= \sum_{n=1}^{\infty} F_n(t) \end{aligned} \tag{8.1.6}$$

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:

$$
\begin{aligned}
u(t) &= \lim_{\Delta t \to 0} \frac{P(\text{renewal event occurs in } (t, t+\Delta t])}{\Delta t} \\
&= \sum_{r=1}^{\infty} \lim_{\Delta t \to 0} \frac{P(r\text{th renewal occurs in } (t, t+\Delta t])}{\Delta t} \\
&= \sum_{r=1}^{\infty} \lim_{\Delta t \to 0} \frac{f_r(t)\Delta t + o(\Delta t)}{\Delta t} \\
&= \sum_{r=1}^{\infty} f_r(t) = U'(t), \qquad (8.1.7)
\end{aligned}
$$

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

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

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

$$u^*(\theta) = \theta U^*(\theta). \qquad (8.1.8)$$

Referring back to (8.1.2) and (8.1.3) and using (8.1.8), we get

$$U^*(\theta) = \frac{1}{\theta} \sum_{n=1}^{\infty} [\phi(\theta)]^n, \qquad (8.1.9)$$

$$u^*(\theta) = \sum_{n=1}^{\infty} [\phi(\theta)]^n. \qquad (8.1.10)$$

From (8.1.9) and (8.1.10), we get

$$U^*(\theta) = \frac{\phi(\theta)}{\theta[1-\phi(\theta)]}, \qquad (8.1.11)$$

$$u^*(\theta) = \frac{\phi(\theta)}{1-\phi(\theta)}. \qquad (8.1.12)$$

Rearranging terms in (8.1.11), we get

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

which on inversion gives

$$U(t) = F(t) + \int_0^t U(t-\tau)dF(\tau). \tag{8.1.13}$$

Similarly, from (8.1.12) we can get

$$u(t) = f(t) + \int_0^t u(t-\tau)f(\tau)d\tau. \tag{8.1.14}$$

The integral equation (8.1.13) is known as the *renewal equation*, which in its general form can be written as

$$Z(t) = h(t) + \int_0^t Z(t-\tau)dF(\tau), \tag{8.1.15}$$

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

$$Z(t) = h(t) + \int_0^t h(t-\tau)dU(\tau). \tag{8.1.16}$$

The significance of (8.1.16) 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. $$U(t+\Delta) - U(t) \to \frac{\Delta}{R} \quad \text{as } t \to \infty, \tag{8.1.17}$$
   $$u(t) \to \frac{1}{R} \quad \text{as } t \to \infty, \tag{8.1.18}$$
   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
   $$\int_0^t h(t-\tau)dU(\tau) \to \frac{1}{R}\int_0^\infty h(t)dt \quad \text{as } t \to \infty. \tag{8.1.19}$$
   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 (8.1.17)–(8.1.19) can be easily understood if we note that $U(t)$ is the expected number of renewals in $(0, t]$. The implication of (8.1.17) 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 (8.1.18) follows directly from (8.1.17) and gives the rate of occurrence of the renewal. The result (8.1.19) follows directly by taking limits ($t \to \infty$) in (8.1.16). Since $h(t)$ is Riemann integrable it tends to 0 as $t \to \infty$ and the contribution of $dU(t)$ in the integral in (8.1.16) is $\frac{1}{R}$, which is the rate of occurrence of the renewal. Thus we get

$$\lim_{t \to \infty} Z(t) = \frac{1}{R} \int_0^{\infty} h(t)dt. \tag{8.1.20}$$

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 let $s_t(x)$ and $r_t(x)$ be their probability density functions, respectively. Using renewal arguments, we can write

$$s_t(x) = u(t-x)[1 - F(x)], \quad 0 < x < t, \tag{8.1.21}$$

and

$$r_t(x) = f_1(t+x) + \int_{\tau=0}^{t} u(\tau) f(t - \tau + x)d\tau, \tag{8.1.22}$$

where $f_1(x)$ is the density function of the initial renewal period. As $t \to \infty$, (8.1.21) and (8.1.22) yield

$$\begin{aligned} \lim_{t \to \infty} s_t(x) &= \frac{1 - F(x)}{R}, \\ \lim_{t \to \infty} r_t(x) &= \frac{1 - F(x)}{R}. \end{aligned} \tag{8.1.23}$$

Taking expected values, we get

$$\int_0^{\infty} x \left[\frac{1 - F(x)}{R} \right] dx = \frac{E[Z^2]}{2R}, \tag{8.1.24}$$

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

8.2 Renewal Process Models for Queueing Systems

In order to apply these results to determine the properties of the underlying processes of queueing systems, we proceed as follows. Suppose $Q(t)$, the number of customers in the system is the process of interest. 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 0 during that period. Probabilistically, this can be defined as

$${}^0P_{ij}(t) = P[Q(t) = j, Q(\tau) \neq 0,\ 0 < \tau < t|Q(0) = i]. \tag{8.2.1}$$

Note that ${}^0P_{ij}(t)$ is the probability of transition of the process $\{Q(t)t \in T\}$ within a busy period. Noting further that the busy cycle is the renewal period for the queue length process, we can write

$$P_{ij}(t) = {}^0P_{ij}(t) + \int_0^t {}^0P_{0j}(t-\tau)dU(\tau). \tag{8.2.2}$$

The two terms on the right-hand side of this equation give the probabilities of the 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 (8.1.16). Therefore, using the key renewal theorem, we get

$$\lim_{t\to\infty} P_{ij}(t) = \frac{1}{R}\int_0^\infty {}^0P_{0j}(t)dt, \tag{8.2.3}$$

where $R = E[\text{busy cycle}]$.

In the case of the queue $M/G/1$, R has been obtained in (5.2.52) as

$$R = \frac{1}{\lambda(1-\rho)}. \tag{8.2.4}$$

In the case of queue $G/M/1$ from (5.3.32), we have

$$R = \frac{1}{\lambda(1-\zeta)}, \tag{8.2.5}$$

where $1/\lambda$ is the mean interarrival time. Note that ζ is the least positive root of the functional equation (5.3.24).

The determination of the transition probabilities ${}^0P_{0j}(t)$ is complicated in both these queueing systems, and we do not present it here. Interested readers may refer to Bhat (1968) for these results as well as a complete analysis of the queue length processes in the two queueing systems. See also Takács (1962).

A *semi-Markov process* or a *Markov renewal process* is defined by incorporating renewal process concepts 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).

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 to easy applications. But often real-world systems are so complex and so general that simple Markov and renewal process models do not represent them well. Computer and communication systems, which have had a major role in advancing technology in the last three decades, require queueing models that go well beyond those we have seen 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.

9.1 The General Queue $G/G/1$

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 $(n = 0, 1, 2, \dots)$, and we let the interarrival times $T_n = t_{n+1} - t_n$ be i.i.d. random variables with distribution function $A(\cdot)$. Let the service time of the nth customer be S_n, and let $\{S_n,\ n = 1, 2, \dots\}$ be i.i.d. random variables with distribution function $B(\cdot)$. We represent the means and variances of these random variables as follows:

$$\begin{aligned} E(T_n) &= \frac{1}{\lambda}, \qquad E(S_n) = \frac{1}{\mu}, \\ V(T_n) &= \sigma_A^2, \qquad V(S_n) = \sigma_B^2. \end{aligned} \tag{9.1.1}$$

Note that for the means of interarrival times and service times, we have used the same notation of $\frac{1}{\lambda}$ and $\frac{1}{\mu}$ as in $M/M/1$ queues, but with a broader interpretation so as to make comparisons simple. We also define the traffic intensity $\rho = \frac{\lambda}{\mu}$ as before.

Let W_n $(n = 1, 2, \dots)$ be the waiting time of the nth customer and $W(t)$ be 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, $\{W_n\}$ $(n = 1, 2, \dots)$ are the actual waiting times of the arrivals at

U.N. Bhat, *An Introduction to Queueing Theory*, DOI: 10.1007/978-0-8176-4725-4_9,

$(t_1, t_2, \ldots)$, and the process W_n is a subset of the $W(t)$ process. These are graphically illustrated in Figure 9.1.1.

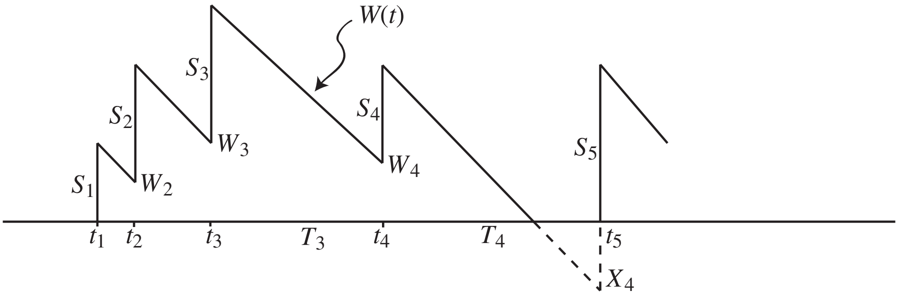

Fig. 9.1.1. Waiting time processes $\{W(t)\}$ and $\{W_n\}$.

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 $W_n + S_n$ becomes 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 that $W_{n+1} = 0$, 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

$$X_n = -\min(0, W_n + S_n - T_n). \tag{9.1.5}$$

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 this notation, (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). See also Gross and Harris (1998) 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 and 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, say I, that 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

$$E(X^2) = E[U^2] + 2E[W]E[U],$$

$$E[W] = \frac{E[X^2] - E[U^2]}{2E(U)}. \tag{9.1.14}$$

Defining $E(X^2)$ in a manner similar to (9.1.11), we may write $E(X^2) = a_0 E(I^2)$. From (9.1.10), 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

$$\begin{aligned} E(W) &= \frac{a_0 E(I^2)}{2[-a_0 E(I)]} - \frac{\sigma_A^2 + \sigma_B^2 + \frac{1}{\lambda^2}(1-\rho)^2}{2\left(\frac{1}{\mu} - \frac{1}{\lambda}\right)} \\ &= \frac{\lambda^2(\sigma_A^2 + \sigma_B^2) + (1-\rho)^2}{2\lambda(1-\rho)} - \frac{E(I^2)}{2E(I)}. \end{aligned} \tag{9.1.16}$$

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

$$\begin{aligned} E(W) &= \frac{\lambda^2(\sigma_A^2 + \sigma_B^2)}{2\lambda(1-\rho)} + \frac{1}{2\lambda}(1-\rho) - \frac{[E(I)]^2}{2E(I)} \\ &\leq \frac{\lambda^2(\sigma_A^2 + \sigma_B^2)}{2\lambda(1-\rho)} + \frac{1-\rho}{2\lambda} - \frac{E(I)}{2}, \end{aligned}$$

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 $G/G/s$, even getting the bounds for $E(W)$ becomes 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.

See also Suzuki and Yoshida (1970) and the discussion in Gross and Harris (1998).

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 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 notation, 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 give a plausibility argument by tracking the arrivals and departures in a queue.

Consider a queueing system in which customers arrive with rate λ. In Chapter 8, we described how a queueing process can be considered a renewal process with a 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.

Consider a busy cycle of 10σ units of time. Assume that arrivals and departures $(C_1, C_2, C_3, C_4, C_5)$ occur at σ unit epochs as shown in Table 9.2.1. Note that the 10σ point is the start of the next busy cycle.

Table 9.2.1. Arrivals 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

Let $L(\text{BC})$, $W(\text{BC})$, and $\lambda(\text{BC})$ be, respectively, the average number in the system, average time in system for a customer, and average rate of arrival in the busy cycle considered here. We get

$$L(\text{BC}) = \frac{16}{10}.$$

The amounts of times the four customers have spent in the system are

$$C_1\colon 4; \qquad C_2\colon 4; \qquad C_3\colon 5; \quad \text{and} \quad C_4\colon 3,$$

for a total of 16σ units of time. Thus we get

$$W(\text{BC}) = \frac{16\sigma}{4} = 4\sigma \text{ units.}$$

The arrival rate is obtained as

$$\lambda(\text{BC}) = \frac{4}{10} \text{ per } \sigma \text{ unit,}$$

verifying the relationship

$$L(\text{BC}) = \lambda(\text{BC}) \cdot W(\text{BC}).$$

A similar relationship can be verified for $L_q(\text{BC})$, $W_q(\text{BC})$, and $\lambda(\text{BC})$ as well. Thus, in general, we have the relationships

$$L = \lambda W; \qquad L_q = \lambda W_q.$$

(See also Jewell (1967).)

9.3 Approximations

Architectural models are exact; purely 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/probability 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 involves mainly simplifying 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 by 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 interarrival time or service time). Then the best usable 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 approximation through simpler distribution models are the use of Erlangian and hyperexponential 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 that 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 must evaluate constantly the tradeoff 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 error analysis, experimentation, and simulation.

We have already seen one example of the process of approximating the result in Section 9.1. There, 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 that is simpler for analysis while retaining as much of the original properties as possible. An example with 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)

$$\begin{aligned} W_2 &\cong U_1, \\ W_3 &\cong U_1 + U_2, \end{aligned}$$

$$\vdots$$

$$W_{n+1} = \sum_{r=1}^{n} U_r = U^{(n)}, \quad \text{say.} \tag{9.3.2}$$

$U^{(n)}$, $n = 1, 2, \ldots$, 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 the 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.) But

$$\begin{aligned} \frac{\sigma^2}{2\alpha} &= \frac{1}{2}\left(\frac{V(U)}{-E(U)}\right) \\ &= \frac{1}{2}\left[\frac{V(T) + V(S)}{E(T) - E(S)}\right]. \end{aligned} \tag{9.3.5}$$

Referring back to (9.1.19), we 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 the $G/G/s$ 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 and Harris (1998).)

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 interarrival time $sE(T) = \frac{s}{\lambda}$:

$$E(W) \cong \frac{sV(T) + V(S)}{2[sE(T) - E(S)]}$$
$$= \frac{V(T) + sV(S/s)}{2[E(T) - (1/s)E(S)]}. \tag{9.3.7}$$

Clearly, (9.3.7) is larger than (9.3.6) by $(s - 1)V(S/s)$, 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 $F(x, t; y, s) = P(X(s) \leq y|X(t) = s), t < s$. 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):

$$\lim_{\Delta t \to 0} \frac{1}{\Delta t} \int_{|y-x|\leq\delta} (y - x) d_y F(x, t; y, t + \Delta t) = a(x, t), \tag{9.4.2}$$
$$\lim_{\Delta t \to 0} \frac{1}{\Delta t} \int_{|y-x|\leq\delta} (y - x)^2 d_y F(x, t; y, t + \Delta t) = b(x, t) > 0. \tag{9.4.3}$$

Applying these properties in the derivation of the 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 (1965b) 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

$$\alpha(t)\Delta t \cong E[X(t + \Delta t) - X(t)|X(t)],$$

$$\sigma^2(t)\Delta t \cong V[X(t+\Delta t) - X(t)|X(t)]. \tag{9.4.5}$$

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 9.1.1. For a small interval of time Δt, the changes occurring in $W(t)$ are

$$\begin{aligned} W(t+\Delta t) - W(t) &= -\Delta t \quad &&\text{with probability } 1 - \lambda\Delta t + o(\Delta t), \\ W(t+\Delta t) - W(t) &= S - \Delta t \quad &&\text{with probability } \lambda\Delta t + o(\Delta t). \end{aligned} \tag{9.4.6}$$

In the statements of (9.4.6), we have used the following assumptions:

1. 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.)
2. The value of $W(t)$ decreases at a unit rate with time.
3. 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 $\Delta t \to 0$,

$$\alpha = \lambda E(S) - 1 \tag{9.4.7}$$

$$\begin{aligned} \sigma^2(t)\Delta t &\cong (\Delta t)^2[1 - \lambda\Delta t + o(\Delta t)] + E[(S-\Delta t)^2][\lambda\Delta t + o(\Delta t)] \\ &\quad - [\lambda E(S)\Delta t - \Delta t + o(\Delta t)]^2 \\ &= \lambda E(S^2)\Delta t - [\lambda E(S) - 1]^2(\Delta t)^2 + o(\Delta t), \end{aligned}$$

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

$$\frac{\partial f(x,t)}{\partial t} = -\alpha\frac{\partial f(x,t)}{\partial x} + \frac{\sigma^2}{2}\frac{\partial^2 f}{\partial x^2} \tag{9.4.9}$$

with conditions

$$f(x,t|x_0) \geq 0,$$

$$\int_0^\infty f(x,t|x_0)dx = \frac{\lambda}{\mu},$$
$$\lim_{t\to 0} f(x,t|x_0) = 0, \qquad x \neq x_0,$$
$$f(x,t|x_0) = 0 \quad \text{for } x \leq 0 \quad \text{and} \quad t \geq 0. \tag{9.4.10}$$

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 much literature on the use of 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 referred to Glynn (1990).

9.5 Fluid Approximation

The fluid approximation of a queueing process ignores the randomness in the arrival and service processes and extends the early investigations of road traffic into queueing problems. Starting with an engineering approach, the approximation procedure developed by Newell (1971) 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 and 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 $Q(0) \gg 1$, $Q(t)$ will gradually decrease until it reaches zero. On the other hand, if $\lambda > \mu$, $Q(t)$ will grow progressively larger and larger and will go to infinity as $t \to \infty$.

The interesting problem is in modeling the rush hour traffic. Let $\lambda(t)$ be time dependent with 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 $\lambda(t_0) = \mu$, $\lambda(t_2) = \mu$, and $Q(t_3) = 0$. This means that during this rush hour, $Q(t) = 0$ at t_0 and t_3. The β of (9.5.2) is obtained as

$$\beta = -\frac{1}{2}\frac{d^2\lambda(t)}{dt^2}\bigg|_{t=t_1}. \tag{9.5.3}$$

At t_0 and t_2, since $\lambda(t_0) = \lambda(t_2) = \mu$, we have

$$\mu = \lambda(t_1) - \beta(t_0 - t_1)^2,$$
$$\mu = \lambda(t_1) - \beta(t_2 - t_1)^2.$$

Hence we get

$$t_0 = t_1 - \left[\frac{\lambda(t_1) - \mu}{\beta}\right]^{1/2},$$
$$t_2 = t_1 + \left[\frac{\lambda(t_1) - \mu}{\beta}\right]^{1/2}. \tag{9.5.4}$$

Note that the positive square root in the first expression and the negative square root in the second expression lead to inadmissible results.

Now

$$\begin{aligned}\lambda(t) - \mu &= \lambda(t_1) - \beta(t - t_1)^2 \\ &\quad - [\lambda(t_1) - \beta(t_0 - t_1)^2] \\ &= \beta[(t_0 - t_1)^2 - (t - t_1)^2].\end{aligned}$$

Also, by a similar argument, we get

$$\lambda(t) - \mu = \beta[(t_2 - t_1)^2 - (t - t_1)^2].$$

Note that $\lambda(t) - \mu$ is a quadratic in t with zeros at $t = t_0$ and $t = t_2$. Hence we may write

$$\lambda(t) - \mu = A(t - t_0)(t - t_2). \tag{9.5.5}$$

Taking the second derivative of both sides of (9.5.5) and using (9.5.3), we get

$$A = -\beta,$$

giving

$$\lambda(t) - \mu = \beta(t - t_0)(t_2 - t). \tag{9.5.6}$$

Since $Q(t_0) = Q(t_3) = 0$, we have

$$\begin{aligned} Q(t) = A(t) - D(t) &= \int_{t_0}^{t} [\lambda(\tau) - \mu] d\tau, \quad t_0 < t < t_3, \\ &= \int_{t_0}^{t} \beta(\tau - t_0)(t_2 - \tau) d\tau \\ &= \beta(t - t_0)^2 \left[\frac{t_2 - t_0}{2} - \frac{t - t_0}{3} \right]. \end{aligned} \tag{9.5.7}$$

This gives

$$Q(t_2) = \frac{\beta}{6}(t_2 - t_0)^3.$$

But from (9.5.4), we have

$$t_2 - t_0 = 2\left[\frac{\lambda(t_1) - \mu}{\beta}\right]^{1/2}. \tag{9.5.8}$$

Hence

$$Q(t_2) = \frac{4}{3}\beta^{-\frac{1}{2}}[\lambda(t_1) - \mu)]^{3/2}, \tag{9.5.9}$$

showing that the maximal queue length is proportional to the $\frac{3}{2}$ power of oversaturation $\lambda(t_1) - \mu$.

Noting that the right-hand side of (9.5.7) is a cubic in t, we may write

$$\frac{t_2 - t_0}{2} = \frac{t - t_0}{3},$$

allowing us to express t_3, where $Q(t) = 0$, as

$$\begin{aligned} t_3 &= t_0 + \frac{3}{2}(t_2 - t_0), \\ \frac{t_3 - t_0}{3} &= \frac{t_2 - t_0}{2}. \end{aligned} \tag{9.5.10}$$

Substituting this result back in (9.5.7), we get

$$Q(t) = \frac{\beta}{3}(t - t_0)^2(t_3 - t). \tag{9.5.11}$$

Now the total delay between t_0 and t_3 can be given as

$$= \int_{t_0}^{t_3} Q(\tau)d\tau$$
$$= \frac{\beta}{36}(t_3 - t_0)^4. \tag{9.5.12}$$

Using (9.5.10) and (9.5.8), this can be simplified to write

$$\text{total delay} = \frac{9[\lambda(t_1) - \mu]^2}{4\beta}. \tag{9.5.13}$$

The results derived here come out of purely deterministic assumptions on system elements. Also, the results are greatly dependent on the state of the system at specific time points. For these reasons, the usefulness of such results is limited. One way of addressing these problems is to use stochastic differential equations to represent 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 an introductory-level book. 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 a better knowledge of topics pertaining to general queues, approximations, limit theorems, etc., are encouraged to seek out more recent issues of research journals and books on such topics.

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 a random phenomenon. Thus to use 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 shall discuss methods of parameter estimation appropriate to various data collection procedures.

We have not discussed data collection and analysis procedures in this text simply because there are several books on this topic in the literature on statistics. Statistical inference procedures are also well established, and a chapter on statistical inference 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 of interarrival 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 interarrival times can be obtained, then the parameters of the distribution of the interarrival times can be estimated using standard statistical procedures. On the other hand, if the information 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.

U.N. Bhat, *An Introduction to Queueing Theory*, DOI: 10.1007/978-0-8176-4725-4_10,

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 the 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 the method of maximum likelihood. 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. In spite of its simplicity, a major drawback of this procedure is that the desirable properties of the estimators are either difficult to establish or do not exist in order to 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, therefore, it would be necessary to obtain the properties of the estimators themselves (such as asymptotic normality, minimum variance, etc.).

In order to avoid the problems associated with the method of moments, the preferred procedure of estimation is the method of maximum likelihood (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, 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), and
3. invariance (the m.l. estimator of a function of the parameter is the corresponding function of the m.l. estimator).

However, m.l. estimation is not perfect either. The estimate obtained by this procedure can be biased.

As indicated earlier, if random samples of interarrival 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 interarrival 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 that 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 that can be modeled as birth-and-death processes. Therefore, for 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 that 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 intial 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 that 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

$$P(\text{an upward jump, i.e., an arrival}) = \frac{\lambda}{\lambda + \mu},$$
$$P(\text{a downward jump, i.e., a departure}) = \frac{\mu}{\lambda + \mu}.$$

   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, the probability element to be associated with that event is $e^{-\lambda x_\ell}$.
6. Because of the Markovian nature of the process, the intervals of time representing the interevent times as identified in items 1 and 3–5 above are independent of each other and also of the events identified in item 2, and the nature of jumps is

independent of all other events. Thus the sample path is made up of independent realizations of various random variables, which can be used to construct a likelihood function for the m.l. estimation.

For the purpose of deriving the m.l. estimator, we define

$$
\begin{aligned}
n_{ae} &= \text{number of arrivals to an empty system,}\\
n_{ab} &= \text{number of arrivals to a busy system,}\\
m &= \text{number of departures from the system,}\\
x_i &= \text{intervals of time spent in state } i \text{ when the system is busy}\\
&\quad\ (i = 0, 1, 2, \ldots, (n_{ab} + m)),\\
x_j &= \text{intervals of time the system has been empty} \quad (j = 0, 1, 2, \ldots, n_{ae}),\\
x_\ell &= \text{the very last interval terminating in } T,\\
n &= n_{ab} + n_{ae},\\
T_b &= \sum x_i + x_\ell,\\
T - T_b &= \sum x_j + x_\ell.
\end{aligned}
$$

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

$$f(\lambda, \mu) = C'\lambda^n \mu^m e^{-\lambda T} e^{-\mu T_b}. \qquad (10.2.3)$$

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)

$$\hat{\lambda}_{\text{crude}} = \frac{n}{T}; \qquad \hat{\mu}_{\text{crude}} = \frac{m}{T_b}. \qquad (10.2.4)$$

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:

$$\begin{aligned} \hat{\lambda} &= (\hat{\mu} - \hat{\lambda})(n + n_0 - \hat{\lambda}T), \\ \hat{\lambda} &= (\hat{\lambda} - \hat{\mu})(m - n_0 - \hat{\mu}T_b). \end{aligned} \qquad (10.2.5)$$

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

$$\begin{aligned} \hat{\rho} &= (1 - \hat{\rho})(n + n_0 - \hat{\mu}\hat{\rho}T), \\ \hat{\rho} &= (\hat{\rho} - 1)(m - n_0 - \hat{\mu}T_b). \end{aligned} \qquad (10.2.6)$$

These equations give

$$\begin{aligned} \hat{\mu} &= \frac{n + m}{\hat{\rho}T + T_b}, \\ \hat{\lambda} &= \frac{(n + m)\hat{\rho}}{\hat{\rho}T + T_b}. \end{aligned} \qquad (10.2.7)$$

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

$$\frac{\hat{\rho} - (n + n_0)(1 - \hat{\rho})}{\hat{\rho} - (m - n_0)(\hat{\rho} - 1)} = -\frac{\hat{\rho}T}{T_b}, \qquad (10.2.8)$$

which gives a quadratic equation in $\hat{\rho}$,

$$f(\hat{\rho}) = T(m - n_0 - 1)\hat{\rho}^2 - [(m - n_0)T + (n + n_0 + 1)T_b]\hat{\rho} + (n + n_0)T_b = 0. \quad (10.2.9)$$

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 $m - n_0 - 1$ with $m - n_0$ and $n + n_0 + 1$ with $n + n_0$ in (10.2.9). The corresponding quadratic equation

$$f^*(\hat{\rho}) = T(m - n_0)\hat{\rho}^2 - [(m - n_0)T + (m + n_0)T_b]\hat{\rho} + (n + n_0)T_b = 0 \quad (10.2.10)$$

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) into (10.2.7), we get

$$\hat{\lambda}_{\text{approx}} \cong \frac{n+n_0}{T}, \qquad \hat{\mu}_{\text{approx}} \cong \frac{m-n_0}{T_b} \tag{10.2.12}$$

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., $\ll 1$) 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 $(T_b) = 25$ minutes.

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)

$$\hat{\lambda}_{\text{crude}} = \frac{75}{30} = 2.5,$$
$$\hat{\mu}_{\text{crude}} = \frac{70}{25} = 2.8.$$

Evaluating the admissible root in (0, 1) of (10.2.9), we get

$$\hat{\rho}_{\text{exact}} = 0.827,$$

from which, after substituting back in (10.2.7), we get

$$\hat{\lambda}_{\text{exact}} = 2.407; \qquad \hat{\mu}_{\text{exact}} = 2.911.$$

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

$$\hat{\lambda}_{\text{approx}} = 2.567, \qquad \hat{\mu}_{\text{approx}} = 2.720.$$

Table 10.2.1 summarizes these results.

Table 10.2.1. 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 $(n = 0, 1, 2, \dots)$ and service parameters μ_n $(n = 1, 2, 3, \dots)$, ignoring the information on the initial state ($\hat{\lambda}_n$ and $\hat{\mu}_n$ are the corresponding estimates), we get

$$\hat{\lambda}_n = \frac{\text{number of arrivals when the process is in state } n}{\text{total time the process is in state } n},$$
$$\hat{\mu}_n = \frac{\text{number of departures when the process is in state } n}{\text{total time the process is in state } n}. \tag{10.2.14}$$

(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 interarrival and service time distributions in an $M/G/1$ queue. Recalling definitions and notation 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

$$k_j = \int_0^\infty e^{-\lambda t} \frac{(\lambda t)^j}{j!} dB(t), \quad j = 0, 1, 2, \dots. \tag{10.3.2}$$

Now $\boldsymbol{\theta}$ includes arrival rate λ and the parameters of the distribution $B(\cdot)$.

Suppose that the sampling plan is to observe the queueing system until N departures have occurred and to 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 we let n_{ij} be the number of transitions of observed values of Q_n from state i to state j $(i, j = 0, 1, 2, \dots)$. Let $(r_0, r_1, \dots, r_N)$ be the observed values of Q_n $(n = 0, 1, 2, \dots, N)$. Using the observed values of the sample path, the likelihood function may be written as (ignoring the distribution of r_0)

$$f(\boldsymbol{\theta}) = \Pi_{n=1}^{N} P(Q_n = r_n | Q_{n-1} = r_{n-1}).$$

Taking logarithms,

$$\ln f(\boldsymbol{\theta}) = \sum_{n=1}^{N} \ln P(Q_n = r_n | Q_{n-1} = r_{n-1}). \tag{10.3.3}$$

Expressing the transition probabilities in terms of the k_js defined in (10.3.2) and using the transition counts n_{ij}, (10.3.3) simplifies to

$$\ln f(\boldsymbol{\theta}) = \sum_{j=0}^{\infty} (n_{0j} + n_{1j}) \ln k_j + \sum_{i=2}^{\infty} \sum_{j=i-1}^{\infty} n_{ij} \ln k_{j-i+1}. \tag{10.3.4}$$

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 for the $G/M/1$ queue as well. But the expressions are a little 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 $\{X_n,\ n = 1, 2, \dots\}$ are 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, $n = 0, 1, \dots, N$, we get a corresponding sample of $\{X_n\}$ suitable for use as a random sample. Now the product of corresponding density elements gives the likelihood function. But 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 (5.3.2), when rearranged, yields the relation

$$\begin{aligned} X_{n+1} &= Q_n + 1 - Q_{n+1} \quad &\text{when } Q_{n+1} > 0, \\ &\geq Q_n + 1 \quad &\text{when } Q_{n+1} = 0, \end{aligned} \tag{10.3.6}$$

which does not provide a complete sample on $\{X_n\}$ since when $Q_{n+1} = 0$, X_{n+1} is available only as being 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 obtain 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 interarrival and service times in estimation is originally due to Cox (1965).

Let $a(u; \boldsymbol{\theta})$ and $b(v; \boldsymbol{\phi})$ be the interarrival time and service time densities, respectively, with θ and ϕ 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 $u_1, u_2, \ldots, u_{N_A}$ and $v_1, v_2, \ldots, v_n$ 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

$$f(\boldsymbol{\theta}, \boldsymbol{\phi}) = [\Pi_{i=1}^{N_A} a(u_i; \boldsymbol{\theta})][\Pi_{j=1}^{n} b(v_j; \boldsymbol{\phi})][1 - A(x_n; \boldsymbol{\theta})]. \tag{10.4.1}$$

Since the factor $[1 - A(x_n; \boldsymbol{\theta})]$ causes difficulty in obtaining simple estimates, consider the alternative approximate likelihood function, sometimes called the *conditional likelihood function*, obtained by dropping the last term in (10.4.1):

$$f_c(\boldsymbol{\theta}, \boldsymbol{\phi}) = [\Pi_{i=1}^{N_A} a(u_i; \boldsymbol{\theta})][\Pi_{j=1}^{n} b(v_j; \boldsymbol{\phi})]. \tag{10.4.2}$$

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 interdeparture 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 interdeparture 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 that it is busy. Also, when it is busy the interdeparture interval is the service time itself.

With this information, it is not difficult to write

$$C(t) = \frac{\lambda}{\mu}B(t) + \left(1 - \frac{\lambda}{\mu}\right)\int_0^t B(t-x)\lambda e^{-\lambda x}dx. \tag{10.5.1}$$

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 interarrival 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

$$\begin{aligned} C(t) &= 0, & t &< \frac{1}{\mu}, \\ &= \frac{\lambda}{\mu}, & t &= \frac{1}{\mu}, \\ &= \frac{\lambda}{\mu} + \left(1 - \frac{\lambda}{\mu}\right)\left(1 - e^{-\lambda(t-1/\mu)}\right), & t &> \frac{1}{\mu}. \end{aligned} \tag{10.5.2}$$

Because of the nonzero probability associated with $t = 1/\mu$, we may use the minimum observed interdeparture time as the estimate of $1/\mu$. (Even without the help of (10.5.2), this is the best conclusion because the length of the service time is the minimum of the interdeparture 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 interdeparture 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 their arrival and departure epochs. The time in system (waiting + service time) has the Laplace–Stieltjes transform given by (5.2.36). Its moments can be determined by differentiation and by 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. For details, readers are referred to Gross and Harris (1998), p. 320.

When 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 parametric form itself? From Appendix A, we know that the Erlangian family of distributions for different values of k have a 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 to the exponential distribution, they are easy for analysis as models for interarrival 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 to 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 the 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 the properties of $\psi(\theta)$, we get $\mu_1^s = -\psi'(0)$ and $\mu_2^s = \psi''(0)$. Thus

$$K''(1) = \lambda^2[\sigma^2 + (\mu_1^s)^2] = \lambda^2\sigma^2 + [K'(1)]^2, \tag{10.5.5}$$

which leads to

$$\sigma^2 = \lambda^{-2}[K''(1) - (K'(1))^2].$$

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 $\hat{C} \ll 1$, use Erlang; when $\hat{C} \gg 1$, 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 the 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 unwieldly. 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 its applications, over the years researchers in industrial labs 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, 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 occur when there is some prior information on parameter values of the process or when the goodness of fit of a distribution form for the interarrival time or the service time must 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, the 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, Harishchandra 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), 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. The notation used here is from Section 10.2. 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. (See also 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 $H_0\colon \boldsymbol{\theta} = \boldsymbol{\theta}_0$, where $\boldsymbol{\theta}$ is the vector of parameters, with $\boldsymbol{\theta}^0$ as the set of desired values, against a suitable alternative, say $H_1\colon \boldsymbol{\theta} = \boldsymbol{\theta}_1$. 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 appropriate 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 their study, they use the fact that the queue length process $\{Q_n,\ n = 0, 1, 2, \dots\}$, representing 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 as follows: after observing Q_n, 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}, 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. For the systems $M/M/1$, $M/E_k/1$, $E_k/M/s$, and $M/M/s/s$ 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 $t_0, t_1, \dots$ 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 (the 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 that may result because of fluctuations in the sample path of the process.

The first set of limits is determined using the limiting distribution of $\{Q_n,\ n = 0, 1, 2, \dots\}$. 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

$$c_u = \min\{k|P(Q^* \geq k) \leq \alpha_u\},$$

$$c_\ell = \max\{k | P(Q^* \le k) \le \alpha_\ell\}. \tag{10.7.2}$$

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 interarrival 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 $\{Q_n\}$ reaches 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 is $\le \beta_u$. This can be stated as

$$d_u = \min\{n | (1 - k_0)^n \le \beta_u\} \tag{10.7.3}$$

When $\{Q_n\}$ reaches c_ℓ, we do not conclude that the traffic intensity $< \rho_0$ unless it stays at or lower than 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 is $\le \beta_\ell$. This can be stated as

$$d_\ell = \min\{n | k_0^n \le \beta_\ell\}. \tag{10.7.4}$$

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 interarrival 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 $(c_u, c_\ell; d_u, d_\ell)$ 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 steps 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 the book by Gross and Harris (1998), starting from its first edition in 1974. Generally, it seems that 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 by Gross and Harris. For a comprehensive discussion of these and other topics in inference on queueing systems, readers may consult Bhat et al. (1997), which also includes an extensive bibliography.

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 telephone 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 behavorial 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 volume 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 these models. 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.

In the next three sections, we introduce the three modes of decision making: using (1) 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

U.N. Bhat, *An Introduction to Queueing Theory*, DOI: 10.1007/978-0-8176-4725-4_11,

discussion will be minimal because using performance measures in decision making is a natural process in probability modeling, and real-world applications of queueing theory do not make extensive use of design and control procedures.

11.2 Performance Measures for Decision Making

The first half of the 20th century was the formative period for 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 the 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) and 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 tollbooths.

An additional aid to decision making developed in the last 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 shall not go into it in detail in this text. An introduction to simulation and some examples are provided in Chapter 12. See also the books on the subject by Law and Kelton (1991) and 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 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

$$\mu = \lambda + \sqrt{\frac{C}{D}}. \tag{11.3.4}$$

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, 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 are used to 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, 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

$$\begin{aligned} W_q &= \left(\frac{\lambda}{s}\right) b_2 \Big/ 2\left(1 - \frac{\lambda b_1}{s}\right) \\ &= \frac{\lambda b_2}{2(s - \lambda b_1)}. \end{aligned} \tag{11.3.5}$$

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 values 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 variable number of servers depending on whether the queue length reaches a given number N. When the queue length reaches N, if the number of servers is $s + i$ $(i \geq 0)$, an additional server is added. This procedure is continued until the number of servers reaches a maximum of S. On the other hand, if the queue length drops below $N - 1$ when the number of servers is $s + i$, one server is removed from service. Again, this procedure is continued until the minimum number s of servers is reached. 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:

$$C(i) = \text{the customer cost rate incurred when there are } i \text{ customers in the system,}$$
$$r_i = \text{cost rate incurred when the service rate } \mu_i \text{ is being used.}$$

The general policy stated by Crabill is as follows: If $C(i)$ is nondecreasing and if $\to \infty$ as $i \to \infty$, $0 < \mu_1 < \mu_2 < \cdots < \mu_k$, $0 \leq r_1 < r_2 < \cdots < r_k$, $\lambda < \mu_k$, 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 $0 = d_1 \leq d_2 \leq d_3 \leq \cdots \leq d_k \leq d_{k+1} = \infty$ 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 startup 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. For a more detailed discussion of problems, see Gross and Harris (1998).

Numerous 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), and Crabill et al. (1977) and to the special issue of the journal *Queueing Systems* edited by Stidham (1995). These articles provide extensive, though sometimes overlapping, bibliographies.

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 usable 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 the reader's interest.

12

Modeling and Analysis Using Computational Tools[1]

Most of the results from Chapters 4–6 can be used for the numerical analysis of queueing systems using standard techniques from applied mathematics with the help of software packages provided by numerical analysis tools such as MATLAB and Mathematica. Some of the significant mathematical techniques that may be needed are matrix multiplications; solving of difference, differential, or integral equations; and root-finding algorithms. Software specific to queueing systems is also available for these analyses. See, for instance, the QTS software accompanying the book by Gross and Harris (1998) the use of Mathematica by Hastings (2006).

In this chapter, we present two special algorithms, *mean value analysis* and the *convolution algorithm*, for the analysis of closed queueing networks, as well as an introduction to simulation techniques that are widely used in analyzing queueing systems in general. To illustrate special algorithms, we use simplifying assumptions that also show how they provide practical solutions to systems that are intractable or whose behaviors cannot be easily modeled using simple probability distributions.

As a note to the reader, we point out that the notation used in this chapter is slightly modified from that used in Chapter 7. For instance, since the service times used are deterministic, no expected values are used in the analysis and the notation representing random variables is also used for the expected values.

12.1 Mean Value Analysis

Mean value analysis (MVA) applies to closed queueing networks and provides their performance in mean values. Also, MVA can be used only if a queueing network has a product-form solution. We will limit ourselves to simple service centers with a fixed limit on the queue size and a single class of customers (or jobs).

Before we introduce MVA, consider the network in Figure 12.1.1, representing a computer system with a single central processing unit (CPU) and several I/O devices

[1] This chapter is authored by Professors Robert Akl and Krishna Kavi, Department of Computer Science and Engineering, University of North Texas, Denton, TX 76203, USA.

U.N. Bhat, *An Introduction to Queueing Theory*, DOI: 10.1007/978-0-8176-4725-4_12,

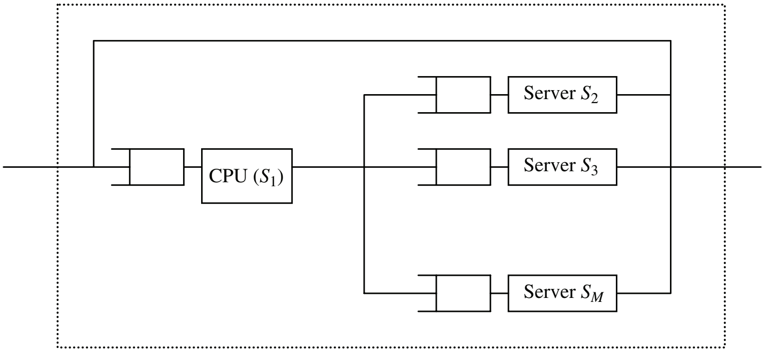

Fig. 12.1.1. A computer system model.

(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 the 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 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 the jth service center. The number of visits to the CPU is given by $1 + \sum_{j=2}^{j=M} V_j$.

These formulations are based on the "operational laws" (Denning and Buzen (1978)) introduced in Chapter 7, which 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 in which a new job arrives soon after a completed job leaves the system. Such models are used to represent time-sharing computer systems in which 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 the think time of a user sitting at a terminal.

Reiser and Lavenberg (1980) showed that the mean response time for service at the jth service station in a closed network with N jobs is given by

$$R_j(N) = (1/\mu_j) * [1 + Q_j(N-1)], \tag{12.1.1}$$

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 one being serviced) given by $Q_j(N-1)$ and must wait for these jobs to be serviced. Note that this formulation assumes that the service distribution is exponential. The response time shown in (12.1.1) 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 will use operational laws that specify that

$$R(N) = \sum_{j=1}^{M} R_j(N)V_j. \tag{12.1.2}$$

Here V_j is the number of visits that a job makes to the jth service center.

Using Little's law, we can obtain the mean throughput rate and mean number of jobs at service station j. In the 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

$$Q_j(N) = X(N) * R_j * V_j. \tag{12.1.3}$$

Example 12.1.1. Consider a computer system with a CPU (C) and 3 file servers (labeled $F1$, $F2$, and $F3$) that can perform file reads and writes. Let us assume that each job vists $F1$ 10 times, $F2$ 20 times, and $F3$ 30 times. After each visit to a file server, the job comes back to the CPU. (Thus the number of visits that each program makes to the 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. $N = 1$:

$$\begin{aligned}
R_C(1) &= (1/\mu_C)[1 + Q_C(0)] = 1 * [1 + 0] = 1,\\
R_{F1}(1) &= (1/\mu_{F1})[1 + Q_{F1}(0)] = 2 * [1 + 0] = 2,\\
R_{F2}(1) &= (1/\mu_{F2})[1 + Q_{F2}(0)] = 3 * [1 + 0] = 3,\\
R_{F3}(1) &= 1/\mu_{F3})[1 + Q_{F3}(0)] = 4 * [1 + 0] = 4.
\end{aligned}$$

System response time:

$$\begin{aligned}
R(1) &= R_C(1) * V_C + R_{F1}(1) * V_{F1} + R_{F2}(1) * V_{F2} + R_{F3}(1) * V_{F3}\\
&= 1 * 61 + 2 * 10 + 3 * 20 + 4 * 30 = 261.
\end{aligned}$$

Queue lengths at each service station are computed as follows:

$$\begin{aligned}
Q_j(N) &= [N/R(N)] * Rj(N) * Vj,\\
Q_C(1) &= [1R(1)] * R_C(1) * V_C = (1/261) * 1 * 61 = 0.234,\\
Q_{F1}(1) &= [1/R(1)] * R_{F1}(1) * V_{F1} = (1/261) * 2 * 10 = 0.077,\\
Q_{F2}(1) &= [1/R(1)] * R_{F2}(1) * V_{F2} = (1/261) * 3 * 20 = 0.230,\\
Q_{F3}(1) &= [1/R(1)] * R_{F3}(1) * V_{F3} = (1/261) * 4 * 30 = 0.460,
\end{aligned}$$

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:

$$\begin{aligned} R(2) &= R_C(2) * V_C + R_{F1}(2) * V_{F1} + R_{F2}(2) * V_{F2} + R_{F3}(2) * V_{F3} \\ &= 1.234 * 61 + 2.154 * 10 + 3.69 * 20 + 5.84 * 30 = 345.814. \end{aligned}$$

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 a higher number of jobs N in the system. Table 12.1.1 shows some values.

Table 12.1.1. Mean response times and queue lengths.

N	R	Q_C	Q_{F1}	Q_{F2}	Q_{F3}
1	261.000	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	2.385
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	0.948	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

The following pseudoalgorithm 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 in the response time computation. For delay centers, we use $R_j(N) = 1/\mu_j$.

MVA can be used for multiple classes of customers. In this case, we iterate the MVA for each class of customers. In other words, we find the average queue lengths iteratively for each customer class.

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. We consider it beyond the scope of this book for discussion.

12.2 The Convolution Algorithm

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 will 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 $p_{n,n_2,\ldots,n_M}$ when there are n_j jobs at service station j (including the job being serviced) and $\{n_1, n_2, \ldots, n_M\}$ denotes the state of the system. Note that the system state represents an element of the set defined as

$$\overrightarrow{N} = \left\{ \{n_1, n_2, \ldots, n_m\} \left| \sum_{j=1}^{j=M} n_j = N \right| \right\}.$$

In this chapter, we will restrict ourselves to closed networks of queues and provide a technique that can be implemented as a computer program.

For systems in which the service time per job is independent of the queue lengths (load-independent service), we can use the following result (Gordon and Newell (1967)):

$$p_{n,n_2,\ldots,n_M} = \frac{d_1^{n_1} d_2^{n_2} \cdots d_M^{n_M}}{G(N)}, \tag{12.2.1}$$

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 the 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 Chapter 7, is based on the observation

$$\sum_{\vec{N}} \prod_{j=1}^{M} (d_j)^{n_j} = \sum_{\vec{N}|n_M=0} \prod_{j=1}^{M} (d_j)^{n_j} + \sum_{\vec{N}|n_M>0} \prod_{j=1}^{M} (d_j)^{n_j}, \tag{12.2.2}$$

where the summation is over the set of all possible states $\{n_1, n_2, \ldots, n_M\}$, 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 with the summation over the set of all possible vectors $\{n_1, n_2, \ldots, n_M\}$, such that $\sum_{j=1}^{M} n_j = N - 1$. Note that since there is at least one customer at service station M, we factor d_M out. The summation now deals with a system with one less customer. Thus we have

$$\sum_{\vec{N}} \prod_{j=1}^{M} (d_j)^{n_j} = \sum_{\vec{N}|n_M=0} \prod_{j=1}^{M} (d_j)^{n_j} + d_M \sum_{\overrightarrow{N-1}} \prod_{j=1}^{M} (d_j)^{n_j}. \tag{12.2.3}$$

If we use $g(n, m)$ for $\sum_{\vec{n}} \prod_{j=1}^{m} (d_j)^{n_j}$ then the normalizing constant $G(N)$ is given by $g(N, M)$. But as we have seen,

$$g(n, m) = g(n, m-1) + d_m * g(n-1, m), \tag{12.2.4}$$

with initial conditions

$$\begin{aligned} g(j, 0) &= 0 \quad \text{for } j = 1, 2, \ldots, n, \\ g(0, k) &= 1 \quad \text{for } k = 1, 2, \ldots, m. \end{aligned}$$

(12.2.4) 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 (12.2.1).

Example 12.2.1. Let us use the same example (Example 12.1.1) as 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:

$$d_1 = d_{\text{CPU}} = 1 * 61 = 61,$$
$$d_2 = d_{F1} = 2 * 10 = 20,$$
$$d_3 = d_{F2} = 3 * 20 = 60,$$
$$d_4 = d_{F3} = 4 * 30 = 120.$$

Table 12.2.1 shows the $g(n, m)$ values for $n = 0, 1, \ldots, 10$ and $m = 1, 2, 3, 4$.

Table 12.2.1. Iterative 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	3721	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	$3.14274E+12$	$4.67516E+12$	$3.32810E+13$	$1.70613E+15$
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	$7.13343E+17$	$1.06131E+18$	$1.03207E+19$	$3.00989E+21$

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 (12.2.1). For example, the probability that all 10 customers are waiting at the CPU is given by

$$p_{10,0,0,0} = \frac{61^{10}}{3.00989 \times 10^{21}} = 2.027 \times 10^{-10}.$$

As can be seen from this example, using total service demands d_j 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 a 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.

12.2.1 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. Note that the convolution algorithm not only computes $G(N)$ but also computes several intermediate values including $G(N - i)$.

Queue lengths. The probability that there are k or more jobs at service station j is given by

$$P(n_j \geq k) = \sum_{\vec{N}|n_j \geq k} \frac{d_1^{n_1} d_2^{n_2} \cdots d_M^{n_M}}{G(N)} = d_j^k \frac{G(N-k)}{G(N)}. \tag{12.2.5}$$

Note that if we use a scaled value $y_j = (\frac{d_j}{k})$ when computing $G(N)$, we will use y_j in the above equation.

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. 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 the queue lengths.

Utilizations. 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

$$U_{\text{CPU}} = 0.508, \qquad U_{F1} = 0.167, \qquad U_{F2} = 0.50, \qquad U_{F3} = 1.00.$$

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 $\gamma_j = \frac{U_j}{1/\mu_j}$. Since closed queueing networks are based on forced flows, the system throughput is given by $\frac{\gamma_j}{V_j} = \frac{U_j}{d_j}$. For the above 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.

12.3 Simulation

As systems modeled as stochastic processes or queueing systems become complex and dynamic, analytical or numerical solutions 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.

A key to good simulations is the quality of the random-number generators used. 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 be small. The linear congruential (LC) method is a widely used technique for generating numbers. In this method, the next random number r_n is generated using the current random number r_{n-1} in the equation

$$r_n = (a * r_{n-1} + c) \text{ modulo } m,$$

where a and c are nonnegative constants. In order to produce m different numbers, the following conditions must hold:

- 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). Because of the growing interest in computer security using cryptography, which requires the generation of random keys, several new techniques have arisen 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) in order 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; 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
```

In order to develop a simulator for a queueing system (e.g., $M/M/1$), we can select one of 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 interarrival 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, shortest job first) can be applied to the same population.

It is also necessary to decide on a termination test based on either a total number of jobs processed by the simulation or on 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 the first-in, first-out (FIFO) discipline, it is only necessary to use the time of arrival, the time when a service is initiated, and the 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. Deadlines 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 earliest-deadline-first scheduling, it is necessary to sort the waiting list of jobs by their deadlines. To implement shortest-job-first 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 the 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.

$M/M/s$ 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; and 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 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 the behavior of a modeled system. The behaviors of each modeled component, the connections among the components (how a job moves from one component to another), and the way in which a waiting queue of jobs is processed must be coded into the simulation. In the next section, we provide a basic introduction to MATLAB and how it can be used to model queueing systems.

12.4 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

[2] The MathWorks, Inc., 3 Apple Hill Drive, Natick, MA 01760-2098, USA, http://www.mathworks.com/products/matlab/.

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 modeling, hypothesis testing, and probability distributions that may be very useful when simulating queues.

The MATLAB program that follows[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) interarrival 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,
```

[3] Available online from http://www.stat.auckland.ac.nz/∼stat320/.

```
if nextarrival < nextdeparture,
    now= nextarrival;
    currentlength= currentlength + 1;
    disp(sprintf('Arrival at :  %f (current length %d)', now, currentlength));
    nextarrival= now + (-1/lambda)*log(rand);
    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 the following:

$$
\begin{aligned}
\text{Arrival at : } 0.102314 &\quad \text{(current length 1).}\\
\text{Departure at : } 0.601800 &\quad \text{(current length 0).}\\
\text{Arrival at : } 3.031791 &\quad \text{(current length 1).}\\
\text{Departure at : } 3.146866 &\quad \text{(current length 0).}\\
\text{Arrival at : } 4.474956 &\quad \text{(current length 1).}\\
\text{Arrival at : } 5.018319 &\quad \text{(current length 2).}\\
\text{Departure at : } 5.259194 &\quad \text{(current length 1).}\\
&\vdots
\end{aligned}
$$

Each time it 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 arrival 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));

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] Simulating $M/G/1$ and $M/G/\infty$

```
function [jumptimes, systsize, systtime] = simmg1(tmax, lambda)
% SIMMG1 simulate a M/G/1 queueing system.  Poisson arrivals
```

[4] Written by R. Gaigalas and I. Kaj and available online from http://www.mathworks.com/matlabcentral/fileexchange/loadFile.do?objectId=2494.

```
% 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
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
```

12.5 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 turns 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 the cross-street, once a signal turns green it will stay green. Assume that cars arrive at the intersection as a Poisson process with an arrival rate of λ. If there are no cars in the left-turn lane, no turn signal appears.

   You need to generate a random number indicating how many cars arrive at the intersection from each direction and if a car is requesting a 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 until 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 of 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, poststore. A memory processor (MP) executes preload and poststore phases, providing access to memory-resident data. An execute processor (EP) provides service during the execute phase. New threads are enabled as some threads complete their execution and supply data to waiting threads (modeled as Synch service). Consider the queueing network in Figure 12.5.1 as a model of the system.

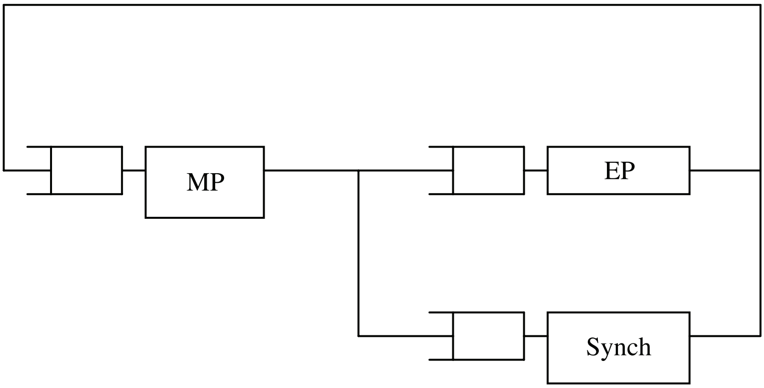

Fig. 12.5.1. A multithreaded computer system model.

The service time of the MP is based on the average number of load and store instructions, while the service time of the EP is based on the average number of nonmemory 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 a 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, Synch = 6. (The time unit is one instruction cycle, and using a 1GHz processor, the time unit is a nanosecond.) Using $N = 10, 20, 50$ threads, perform MVA for this exercise.

4. Using the 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 ensure 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 earliest deadline first (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 the 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.

   Write a program to simulate EDF scheduling. A job in a real-time system, τ_i, is defined as $\tau_i = (r_i, e_i, D_i)$, where r_i is its arrival time, e_i is its estimated average execution time, and D_i is its deadline. You should also maintain a dynamic deadline d_i with an initial value $r_i + D_i$, which tracks the absolute time before the deadline expires. In other words, D_i is the relative deadline of the job with respect to the arrival time and d_i is the absolute (wall clock) deadline.

   Figure 12.5.2 shows the relationship among the various parameters.

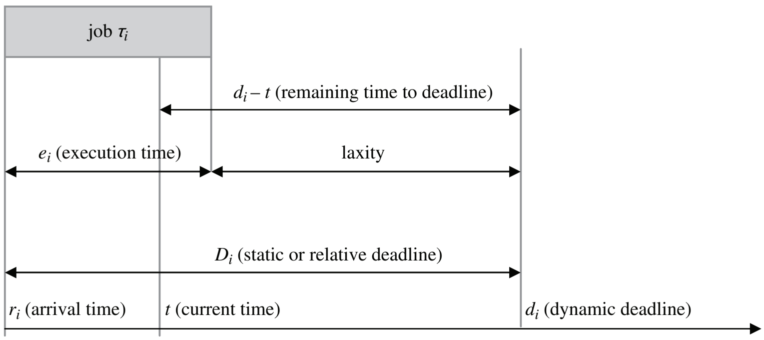

Fig. 12.5.2. Relationship among parameters.

   For your simulations, generate a fixed number (N) of jobs with randomly generated arrivals, execution times, and deadlines. Assume that jobs are mutually independent. Each simulation is terminated when the predetermined experimental time T has expired.

   Investigate the sensitivity of the various task parameters on the success rates of EDF. Use random distributions available in MATLAB to generate the necessary parameters for tasks.

   (a) Generate five jobs ($N = 5$) that arrive at the same time 0 and have the same deadline. Schedule the jobs based on FIFO. What is the success ratio using EDF with FIFO? What is the average response time for the completed jobs?

(b) Another commonly used scheduling method to increase throughput of a system is known as shortest job first (SJF). As the name implies, tasks are scheduled based on their execution times—schedule jobs with progressively increasing execution times. As in the previous exercise, create jobs using an arrival time and service time. Schedule the jobs based on SJF. Determine the success ratio and the response time for EDF with SJF.
(c) Repeat parts (a) and (b) for five jobs with different deadlines.
(d) Repeat parts (a) and (b) for five jobs with different deadlines and different arrival times.

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 the smallest laxity is scheduled first. Repeat the experiment of Exercise 5 with LLF and compare its success ratio with that of EDF.

APPENDICES

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 those given in Section 2.1 of Chapter 2 and related distributions that are often used in applications. Distributions other than those mentioned here that are sometimes used in queueing theory can be found in standard statistics texts.

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

$$h(t) = \frac{f(t)}{1 - F(t)}. \tag{A.1.1}$$

Note 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 the term *instantaneous failure rate*, or simply the *failure rate*.

When $f(x)$ is exponential, i.e., equal to $\lambda e^{-\lambda x}$ $(x > 0)$, 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 i.i.d.

(c) For purposes of modeling, the implications of (a) and (b) above and the memoryless property of the exponential distribution to the Poisson process described in Section 2.1 are the following:

- Events occurring in nonoverlapping intervals of time are independent of each other.

U.N. Bhat, *An Introduction to Queueing Theory*, DOI: 10.1007/978-0-8176-4725-4_13,

- There is a constant λ such that the probabilities of occurrence of events in a small interval of length Δt are given as follows:
  - $P\{$number of events of occurring in

$$(t, t+\Delta t] = 0\} = 1 - \lambda\Delta t + o(\Delta t).$$

  - $P\{$number of events occurring in

$$(t, t+\Delta t] = 1\} = \lambda\Delta t + o(\Delta t).$$

  - $P\{$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 $\Delta t \to 0$.

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

$$P(X_1 < X_2) = \int_{x=0}^{\infty} P(X_1 < X_2 | X_2 = x) f_2(x) dx,$$

where we have written $f_2(x)$ for the density function of the random variable X_2. We get

$$\begin{aligned} P(X_1 < X_2) &= \int_{x=0}^{\infty} P(X_1 < x) f_2(x) dx \\ &= \int_{x=0}^{\infty} (1 - e^{-\lambda_1 x}) \lambda_2 e^{-\lambda_2 x} dx \\ &= \frac{\lambda_1}{\lambda_1 + \lambda_2}. \end{aligned} \tag{A.1.2}$$

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 $t \geq 0$, we have

$$P[X_1(t) = n_1] = e^{-\lambda_1 t} \frac{(\lambda_1 t)^{n_1}}{n_1!},$$

$$P[X_2(t) = n_2] = e^{-\lambda_2 t} \frac{(\lambda_2 t)^{n_2}}{n_2!}.$$

Using these results, we can show that $X(t)$ is also Poisson for $t \geq 0$:

$$\begin{aligned} P[X(t) = n] &= \sum_{n_2=0}^{n} P[X_1(t) = n - n_2] P[X_2(t) = n_2] \\ &= e^{-(\lambda_1+\lambda_2)t} \frac{[(\lambda_1 + \lambda_2)t]^n}{n!}. \end{aligned} \tag{A.1.3}$$

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 $t_1 < t_2 < t_3 < \cdots < t_n$ in the interval $[0, T]$. Then the random variables $t_1, t_2, \ldots, t_n$ have the same distribution as the nth-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

$$f_{t_1,t_2,\ldots,t_n}(x_1, x_2, \ldots, x_n) = \frac{n!}{T^n}, \quad 0 \le x_1 \le x_2 \le \cdots \le x_n \le T. \tag{A.1.4}$$

A.2 Variants of the Poisson Process

The Poisson process assumes that the events occur one at a time. However, 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

$$\Pr(G_n = j) = g_j.$$

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., the number of claims $X(t)$ and the total amount of claims $Y(t)$ in insurance-risk theory). Then we get

$$\Pr[X(t) = n,\ Y(t) \le x] = e^{-\lambda t}\frac{(\lambda t)^n}{n!}H_n(x) \qquad \text{(A.2.2)}$$

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 occurring in $(0, t]$, can be given as

$$P_n(t) = P[X(t) = n] = \int_0^{\infty} e^{-\lambda t}\frac{(\lambda t)^n}{n!} dL(\lambda). \qquad \text{(A.2.3)}$$

When the range of Λ is other than $(0, \infty)$, a suitable range is to be used for integration. From (A.2.3), we get

$$E[X(t)] = tE[\Lambda].$$

For instance, when

$$dL(\lambda) = e^{-\mu\lambda}\frac{\mu^k \lambda^{k-1}}{(k-1)!} d\lambda \quad (0 \le \lambda < \infty),$$

we get

$$P_n(t) = \binom{n+k-1}{k-1}\left(\frac{\mu}{t+\mu}\right)^k \left(\frac{t}{t+\mu}\right)^n, \quad n = 0, 1, 2, \ldots, \qquad \text{(A.2.4)}$$

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 (HE) Distribution

Let random variables $\{Z_1, Z_2, \dots\}$ be distributed as

$$
\begin{aligned}
F(x) &= 1 - \sum_{i=1}^{K} p_i e^{-\lambda_i x}, && 0 \le x < \infty, \\
\lambda_i &> 0 && \text{for all } i \text{ for which } p_i > 0; \\
1 \ge p_i &\ge 0, && \sum_1^K p_i = 1. \qquad \text{(A.3.1)}
\end{aligned}
$$

We get

$$E(Z_n) = \sum_1^K \left(\frac{p_i}{\lambda_i}\right)$$

and

$$\psi(\theta) = \sum_1^K p_i \left(\frac{\lambda_i}{\theta + \lambda_i}\right).$$

Also,

$$E[Z_n^2] = \sum_{i=1}^{k} \frac{2p_i}{\lambda_i^2} \quad \text{and} \quad \mathrm{CV}(Z_n) = \left[\frac{2\sum_{i=1}^{K} \frac{p_i}{\lambda_i^2}}{\left(\sum \frac{p_i}{\lambda_i}\right)^2} - 1\right]^{1/2}.$$

This distribution is generated if events fall into identifiable classes and an event belonging to class i generates with probability p_i an interoccurrence time that is exponential with mean $\frac{1}{\lambda_i}$. Depending on the values of p_i and λ_i and the possible values of i, a wide variety of distributions can be generated.

To retain the same mean $\frac{1}{\lambda}$, the following form of the HE distribution can be used:

$$
\begin{aligned}
F(x) &= 1 - \sum_{i=1}^{K} p_i e^{-K p_i \lambda x} \quad (x \ge 0), \qquad \text{(A.3.2)} \\
\lambda &> 0, \quad 1 \ge p_i \ge 0, \quad \sum_1^K p_i = 1.
\end{aligned}
$$

Then

$$
\begin{aligned}
E[Z_n] &= \frac{1}{\lambda}, \\
E[Z_n^2] &= \frac{2}{(K\lambda)^2} \sum_{i=1}^{K} \left(\frac{1}{p_i}\right),
\end{aligned}
$$

$$\mathrm{CV}[Z_n] = \left[\left(\sum_{i=1}^{K} \frac{1}{p_i} \right) \frac{2}{K^2} - 1 \right]^{1/2}.$$

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 $\{Z_1, Z_2, \dots\}$ be distributed as

$$\begin{aligned} F(x) &= \int_0^x e^{-\lambda y} \frac{\lambda^k y^{k-1}}{(k-1)!} dy, \quad 0 \le x \le \infty, \quad \lambda > 0, \\ &= 1 - \sum_{r=0}^{k-1} e^{-\lambda x} \frac{(\lambda x)^r}{r!}. \end{aligned} \tag{A.4.1}$$

We get

$$\begin{aligned} E[Z_n] &= \frac{k}{\lambda}, \\ \psi(\theta) &= \left(\frac{\lambda}{\theta + \lambda} \right)^k, \end{aligned}$$

and

$$E[Z_n^2] = \left(1 + \frac{1}{k}\right) \frac{k^2}{\lambda^2} \quad \text{and} \quad \mathrm{CV}[Z_n] = \frac{1}{\sqrt{k}}.$$

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 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*. Let

$$F(x) = \int_0^x \sum_{i=1}^{K} p_i e^{-k_i \lambda y} \frac{(k_i \lambda)^{k_i} y^{k_i - 1}}{(k_i - 1)!} dy, \quad 0 \le x < \infty, \quad \lambda > 0. \tag{A.5.1}$$

We get

$$E[Z_n] = \frac{1}{\lambda},$$

$$\psi(\theta) = \sum_{i=1}^{K} p_i \left(\frac{k_i \lambda}{\theta + k_i \lambda}\right)^{k_i},$$

and

$$E[Z_n^2] = \frac{1}{\lambda^2} \sum_{i=1}^{K} p_i \left(1 + \frac{1}{k_i}\right),$$

$$\mathrm{CV}(Z_n) = \left[\sum_{i=1}^{K} \frac{p_i}{k_i}\right]^{1/2}.$$

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 K has also been found satisfactory. (See Luchak (1956).)

(b) Both λ and k varying.

$$F(x) = \int_0^x \sum_{i=1}^{K} p_i e^{-k_i \lambda_i y} \frac{(k_i \lambda_i)^{k_i} y^{k_i - 1}}{(k_i - 1)!} dy. \tag{A.5.2}$$

This general form admits both the HE and Erlang distribution as special cases:

$$\text{HE:} \quad k_i = 1 \quad \text{for } i = 1, 2, \ldots, K,$$
$$\text{mixed Erlang:} \quad \lambda_i = \lambda \quad \text{for } i = 1, 2, \ldots, K.$$

Assuming the coefficient of variation (CV) to be a measure providing an adequate representation of the model, the Erlang (with CV ≤ 1) and HE (with CV ≥ 1) distributions offer a wide spectrum of choice for model selection. In the Erlang model, the CV is decreased by increasing the value of the parameter k. In the HE model with $K = 2$, CV is increased by moving p_1 and p_2 away from $\frac{1}{2}$.

A.6 Coxian Distributions; Phase-Type Distribution

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 distribution, which has the Laplace transform

$$\psi(\theta) = \Pi_{i=1}^{K} \left(\frac{\lambda_i}{\theta + \lambda_i}\right). \tag{A.6.1}$$

The corresponding distribution can be thought of as the distribution of the time that a process takes to pass through K phases $(X_1, X_2, \ldots, X_K)$, where X_i has an exponential distribution with mean $\frac{1}{\lambda_i}$. It is obtained as the convolution of K exponentials with parameters $\lambda_1, \lambda_2, \ldots, \lambda_K$. Another large subset of Coxian distributions is the phase-type distribution introduced by Neuts (1975, 1989), 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 Section B.4 of Appendix B.

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.

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 interarrival 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 interoccurrence times of events. Define

$$p_k = P(Z_n = k), \quad n = 1, 2, \ldots, \quad k = 0, 1, 2, \ldots,$$

and

$$\phi(z) = \sum_k p_k z^k, \quad |z| \leq 1.$$

Three discrete distributions that are analogues 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 let 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

$$\phi(z) = \frac{\alpha z}{1-(1-\alpha)z}.$$

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,

$$p_k(n) = \binom{n}{k}\alpha^k(1-\alpha)^{n-k}, \quad k = 0, 1, 2, \ldots, n. \tag{A.8.2}$$

We get

$$E[X(n\sigma)] = n\alpha \quad \text{and} \quad V[X(n\sigma)] = n\alpha(1-\alpha)$$

and

$$\phi_n(z) = \sum_{k=0}^{n} p_k(n)z^k = (1-\alpha+\alpha z)^n.$$

Clearly, (A.8.2) is the discrete analogue 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*: Now 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 the binomial probability given in (A.8.2). Since the event has to occur at the nth time point, we have

$$\begin{aligned} p_k^{(n)} &= \binom{n-1}{k-1}\alpha^{k-1}(1-\alpha)^{n-k}\alpha \\ &= \binom{n-1}{k-1}\alpha^{k}(1-\alpha)^{n-k}, \quad n = k, k+1, \ldots . \end{aligned} \tag{A.8.3}$$

We get

$$E[Z] = \frac{k}{\alpha} \quad \text{and} \quad V[Z] = \frac{k(1-\alpha)}{\alpha^2}$$

and

$$\phi(z) = \left[\frac{\alpha z}{1-(1-\alpha)z}\right]^k .$$

As in the Erlang, this distribution is generated by counting every kth event as an effective event for the queueing system.

B

Markov Process

This appendix builds on the basic concepts introduced in Section 3.3 of the main text and provides additional background material for further work on modeling and analysis of queueing systems. The notation used in the material given here is consistent with that used in Chapter 3.

B.1 Kolmogorov Equations

Let $\{X(t),\ t \in T\}$ 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. They are the *forward Kolmogorov equations* and *backward Kolmogorow equations*. Forward Kolmogorov equations are the ones commonly used in applications because of their convenient structure, even though the backward equations are considered to be more fundamental due to the nature of the 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 as

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

Set $s = \Delta t$; then

$$P_{ij}(t+\Delta t) = \sum_{k \in S} P_{ik}(t) P_{kj}(\Delta t).$$

Subtracting $P_{ij}(t)$ from both sides of the equation and dividing by Δt,

$$\frac{P_{ij}(t+\Delta t) - P_{ij}(t)}{\Delta t} = \sum_{k \neq j} \frac{P_{ik}(t) P_{kj}(\Delta t)}{\Delta t}$$

U.N. Bhat, *An Introduction to Queueing Theory*, DOI: 10.1007/978-0-8176-4725-4_14,

$$+ P_{ij}(t)\frac{P_{jj}(\Delta t) - 1}{\Delta t}.$$

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). Equations (B.1.2) for $i,\ j \in S$ are known as *forward Kolmogorov equations*.

In matrix notation, we can write them as

$$\mathbf{P}'(t) = \mathbf{P}(t)\mathbf{A}. \tag{B.1.3}$$

The transition probability $P_{ij}(t)$ can be determined by solving these differential equations along with the boundary condition $\mathbf{P}(0) = I$.

Backward Kolmorov equations can be obtained in a similar manner by starting with the relation

$$P_{ij}(\Delta t + t) = \sum_{k \in S} P_{ik}(\Delta t) P_{kj}(t).$$

The corresponding matrix equation can be given as

$$\mathbf{P}'(t) = \mathbf{A}\mathbf{P}(t).$$

Formally, the solution for both sets of equations can be given as

$$\mathbf{P(t)} = e^{\mathbf{A}t} = \mathbf{I} + \sum_{n=1}^{\infty} \mathbf{A}^n \frac{t^n}{n!}. \tag{B.1.4}$$

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 interoccurrence times are exponential. 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) $P\{\text{number of events occurring in } (t, t + \Delta t] = 0\} = 1 - \lambda\Delta t + o(\Delta t)$,
   (b) $P\{\text{number of events occurring in } (t, t + \Delta t] = 1\} = \lambda\Delta t + o(\Delta t)$,
   (c) $P\{\text{number of events occurring in } (t, t + \Delta t] > 1\} = o(\Delta t)$,
   where $o(\Delta t)$ is such that $\frac{o(\Delta t)}{\Delta t} \to 0$ as $\Delta t \to 0$.

Using the notation and equations developed for Markov processes, in this context, we have (see (3.3.15) and (3.3.16))

$$P'_{ij}(0) = \lambda, \qquad P'_{ii}(0) = -\lambda,$$

resulting in a 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 $\mathbf{P}(t) = (P_0(t), P_1(t), \dots)$ and $\mathbf{P}'(t) = (P'_0(t), P'_1(t), \dots)$, 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 the equations in (B.1.2) are not always easily determined. For the Poisson process, because of the bidiagonal structure of **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 **A** is finite and diagonalizable, 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 **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 that λ_{ii} is also the sum of all the instantaneous transition rates out of state i.

This allows us to interpret $\frac{1}{\lambda_{ii}}$ 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 $\frac{1}{\lambda_{ii}}$.

For a proof of this result and for a discussion of special forms of Markov processes used in stochastic modeling, readers are referred to Bhat and Miller (2002) and the advanced books cited there.

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 that ceases to operate when Q_n reaches a 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 that 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 be absorbed in M. On the other hand, the Markov chain of the second system will remain in the class 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* (also known as *persistent*) 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 value of the *recurrence time*, which is the mean time the process takes to return to the same state. 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 positive 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. Here we list 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 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).

B.4 Phase-Type Distributions

In Appendix A, we postponed the description of a phase-type (PH) distribution because it required results from Markov processes. For illustration, we use the generalized Erlang distribution, one of the simpler PH distributions, given by the Laplace transform $\psi(\theta)$ of (A.6.1):

$$\psi(\theta) = \Pi_{i=1}^{k}\left(\frac{\lambda_i}{\theta + \lambda_i}\right). \tag{B.4.1}$$

The generalized Erlang distribution can be generated as the distribution of the total time a process takes to traverse k phases, with phase i lasting a duration that has an exponential distribution with mean $\frac{1}{\lambda_i}$. Recalling the properties of the Markov process, we can identify it as a Markov process with states $\{1, 2, \ldots, k, k+1\}$ of which state $k+1$ is absorbing. The generator matrix of the process can be given as

$$\mathbf{A} = \left[\begin{array}{cccccc|c} -\lambda_1 & \lambda_1 & & & \cdot\cdot & & 0 \\ & -\lambda_2 & \lambda_2 & & \cdot\cdot & & 0 \\ & & -\lambda_3 & \lambda_3 & \cdot & 0 & \\ & & & & \cdot & -\lambda_k & \lambda_k \\ \hline 0 & 0 & 0 & 0 & \cdot & \ldots & 0 \end{array}\right]$$

$$= \begin{bmatrix} \mathbf{T} & \mathbf{T}^0 \\ \mathbf{0} & 0 \end{bmatrix}, \tag{B.4.2}$$

where $\mathbf{T}$ $(m \times m)$ and $\mathbf{T}^0 (m \times 1)$ are submatrices.

Let $P_i(t)$ be the probability that the process is in state i at time t. Note that, because $k + 1$ is an absorbing state, ultimately the process will come to reside in that state. Therefore, $\sum_{i=1}^{k} P_i(t)$ is the probability that the process is in one of the transient states $\{1, 2, \ldots, k\}$ at time t. Let Y_k be the time the process takes to traverse all the k states. Then

$$P(Y_k > t) = \sum_{i=1}^{k} P_i(t).$$

Hence

$$\begin{aligned} P(Y_k \leq t) &= 1 - \sum_{i=1}^{k} P_i(t) \\ &= P_{k+1}(t). \end{aligned} \tag{B.4.3}$$

For the Markov process with generator matrix $\mathbf{A}$, we can write the forward Kolmogorov equations as

$$\begin{aligned} P_1'(t) &= -\lambda_1 P_1(t), \\ P_i'(t) &= -\lambda_i P_i(t) + \lambda_{i-1} P_i(t), \quad 1 < i \leq k, \\ P_{k+1}'(t) &= \lambda_k P_k(t), \end{aligned} \tag{B.4.4}$$

with $P_1(0) = 1$ and $P_i(0) = 0$ for $i > 1$. Define the Laplace transform

$$\phi_i(\theta) = \int_0^\infty e^{-\theta t} P_i(t) dt.$$

Taking transforms of (B.4.4), we get

$$\begin{aligned} -1 + \theta\phi_1(\theta) &= -\lambda_1\phi_1(\theta), \\ \theta\phi_i(\theta) &= -\lambda_i\phi_i(\theta) + \lambda_{i-1}\phi_{i-1}(\theta), \quad 1 < i \leq k. \end{aligned}$$

Solving these equations recursively, we get

$$\phi_k(\theta) = \left(\frac{1}{\theta + \lambda_k}\right) \Pi_{i=1}^{k-1} \left(\frac{\lambda_i}{\theta + \lambda_i}\right). \tag{B.4.5}$$

If $f_k(y)$ is the probability density of Y_k, it is easy to see that $P'_{k+1}(t)$ is, in fact, $f_k(y)$. Thus from the last equation in (B.4.4), and (B.4.5), we have the Laplace transform of Y_k as

$$\int_0^\infty e^{-\theta y} f_k(y)dy = \Pi_{i=1}^k \left(\frac{\lambda_i}{\theta + \lambda_i}\right),$$

which is the same as (B.4.1).

Referring back to the general form of the solution to the forward Kolmogorov equations given by (B.4.4), the distribution function of Y_k can be given as

$$F(t) = 1 - \boldsymbol{\alpha} \exp(\mathbf{T}t)\mathbf{e} \quad \text{for } t \geq 0, \tag{B.4.6}$$

where $\boldsymbol{\alpha} = (1, 0, 0, \ldots, 0)$ and $\mathbf{e}' = (1, 1, \ldots, 1)$.

Generalizing this structure, Neuts (1989) has defined the PH distribution as the time until absorption in a finite Markov process with generator matrix

$$\mathbf{A} = \begin{bmatrix} \mathbf{T} & \mathbf{T}^0 \\ \mathbf{0} & 0 \end{bmatrix},$$

where the $m \times m$ matrix $\mathbf{T}$ satisfies $T_{ii} < 0$ for $1 \leq i \leq k$ and $T_{ij} \geq 0$ for $i \neq j$. Also, $\mathbf{Te} + \mathbf{T}^0 = \mathbf{0}$. The initial probability vector $(\boldsymbol{\alpha}, \alpha_{k+1})$ is such that $\boldsymbol{\alpha} e + \alpha_{k+1} = 1$. States $\{1, 2, \ldots, k\}$ are transient and $k + 1$ is absorbing. A large number of PH distributions can be generated by using different structures for $\mathbf{T}$. For the properties of PH distribution and its use in queueing theory, readers are referred to Neuts (1978, 1989).

The simplest PH distribution is the Erlang, in which the λ_i, $i = 1, 2, \ldots, k$ of the generator matrix (B.4.2) are a constant, say λ. The underlying random variable represents the time taken by the process to traverse k phases of service, each with an exponential distribution with mean $\frac{1}{\lambda}$.

C

Results from Mathematics

In this appendix,[1] we present useful results from several areas of mathematics that have been used in the book.

C.1 Riemann–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

$$a = x_0 \leq x_1 \leq x_2 \leq \cdots \leq x_{n-1} \leq x_n = b.$$

For any partition $\boldsymbol{P}$ of the closed interval $[a, b]$, we define

$$\begin{aligned} N_1 &= \text{least upper bound } f(x), \qquad \text{where } x \in [x_{i-1}, x_i], \\ n_i &= \text{greatest lower bound } f(x) \quad \text{where } x \in [x_{i-1}, x_i], \\ \Delta\phi_i &= \phi(x_i) - \phi(x_{i-1}), \\ U(\boldsymbol{P}, f, \phi) &= \sum_{i=1}^{n} N_i \Delta\phi_i, \\ L(\boldsymbol{P}, f, \phi) &= \sum_{i=1}^{n} n_i \Delta\phi_i. \end{aligned}$$

Then

$$\int_a^{b^-} f(x) d\phi(x) = \text{greatest lower bound } U(\boldsymbol{P}, f, \phi),$$

[1] Reprinted with permission from U. N. Bhat and G. K. Miller, *Elements of Applied Stochastic Processes*, 3rd ed., Wiley, New York, 2002.

U.N. Bhat, *An Introduction to Queueing Theory*, DOI: 10.1007/978-0-8176-4725-4_15,

$$\int_{a^-}^{b} f(x)d\phi(x) = \text{least upper bound } L(\boldsymbol{P}, f, \phi),$$

where the greatest lower bound and least upper bound are taken over all partitions of $[a, b]$.

We say that $f(x)$ is Riemann–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 partiions 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 C.2.1. 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 C.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 C.2.2. 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.

Properties C.2.1 below 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.

Properties 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

$$g(t) = \begin{cases} f(t-\xi) & \text{for } t > \xi, \\ 0 & \text{for } t \le \xi, \end{cases}$$

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 $g(t) = \frac{d^n[f(t)]}{dt^n} = f^{(n)}(t)$, then

$$L\{g(t)\} = s^n\phi(s) - s^{n-1}f(0) - s^{n-2}f^{(1)}(0) - \cdots - sf^{n-2}(0) - f^{(n-1)}(0).$$

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

$$\lim_{s\to\infty} \phi(s) = 0,$$
$$\lim_{t\to 0} f(t) = \lim_{s\to\infty} s\phi(s),$$
$$\lim_{t\to\infty} f(t) = \lim_{s\to 0} s\phi(s).$$

(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 that $f_2(t)$ is a null function and $f(t) = f_1(t) + f_2(t)$; then by Properties C.2.1(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 C.2.1 (an Abelian theorem). *If for some nonnegative number* ξ (≥ 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 C.2.2 (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* ξ (≥ 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 C.3.1. 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.

Properties 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 $A(z) = zB^{(1)}(z) = z\frac{dB(z)}{dz}$.
(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) $Q(z) = \frac{[1-P(z)]}{1-z}$,
(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 C.3.1 (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 C.3.2 (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 C.3.3. *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):

$$\sum_{n=0}^{\infty} a_n = a \quad \textit{iff} \qquad \lim_{z\to 1^-} A(z) = a,$$

$$\lim_{m\to\infty}\left(\frac{1}{m}\sum_{n=0}^{m} a_n\right) = c \quad \textit{iff} \quad \lim_{z\to 1^-}\,[(1-z)A(z)] = c.$$

(Beightler et al. (1961), Feller (1968), Hardy (1949), Whittaker and Watson (1962).)

References

J. Abate and H. Dubner (1968), A new method of generating power series expansions of functions, *SIAM J. Numer. Anal.*, **5**, 102–112.

J. Abate, H. Dubner, and S. B. Weinberg (1968), Queueing analysis for the BIM 2314 disk storage facility, *J. Assoc. Comput. Mach.*, **15**, 577–589.

J. Abate and W. Whitt (1992), The Fourier-series method for inverting transforms of probability distributions, *Queueing Systems*, **10**, 5–88.

S. K. Acharya (1999), On normal approximation for maximum likelihood estimation from single server queues, *Queueing Systems*, **31**, 207–216.

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.

A. S. Alfa (2003), Vacation models in discrete time, *Queueing Systems*, **44**, 5–30.

A. O. Allen (1990), *Probability, Statistics, and Queueing Theory with Computer Science Applications*, 2nd ed., Academic Press, Boston.

C. Armero (1994), Bayesian inference in Markovian queues, *Queueing Systems*, **15**, 419–426.

C. Armero and D. Conesa (2000), Prediction in Markovian bulk arrival queues, *Queueing Systems*, **34**, 327–350.

N. T. J. Bailey (1952), Study of queues and appointment systems in out-patient departments with special reference to waiting times, *J. Roy. Statist. Soc.* B, **14**, 185–199.

N. T. J. Bailey (1954), A continuous time treatment of a simple queue using generating functions 1, *J. Roy. Statist. Soc.* B, **16**, 288–291.

N. T. J. Bailey (1964), *The Elements of Stochastic Processes with Applications to the Natural Sciences*, Wiley, New York.

K. R. Balachandran (1973), Control policies for a single server system, *Management Sci.*, **19**, 1013–1018.

I. V. Basawa, U. N. Bhat, and J. Zhou (2008), Parameter estimation in queueing systems using partial information, *Statist. Probab. Lett.*, to appear.

I. V. Basawa and N. U. Prabhu (1981), Estimation in single server queues, *Nav. Res. Logist. Quart.*, **28**, 475–487.

I. V. Basawa and N. U. Prabhu (1988), Large sample inference from single server queues, *Queueing Systems*, **3**, 289–304.

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. Assoc. Comput. Mach.*, **22**, 248–260.

M. Bäuerle (2002), Optimal control of queueing networks: An aproach via fluid models, *Adv. Appl. Probab.*, **34**, 313–328.

C. S. Beightler, L. G. Mitten, and G. L. Nemhauser (1961), A short table of Z-transforms and generating functions, *Oper. Res.*, **9**, 547–578.

F. Benson and D. R. Cox (1951, 1952), The productivity of machines requiring attention at random intervals, *J. Roy. Statist. Soc.* B, **13**, 65–82; **14**, 200–210.

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, Vol. 2, Springer-Verlag, New York.

U. N. Bhat (1969), Sixty years of queueing theory, *Management Sci.*, **15**, B280–B294.

U. N. Bhat (1984), *Elements of Applied Stochastic Processes*, 2nd ed., Wiley, New York.

U. N. Bhat (1987), A sequential technique for the control of traffic intensity in Markovian queues, *Ann. Oper. Res.*, **8**, 151–164.

U. N. Bhat (2003), Parameter estimation in $M/G/1$ and $GI/M/1$ queues using queue length data, in S. K. Srinivasan and A. Vijayakumar, eds., *Stochastic Point Processes*, Narosa, New Delhi, 96–107.

U. N. Bhat and I. V. Basawa (2002), Maximum likelihood estimation in queueing systems, in N. Balakrishnan, ed., *Advances on Methodological and Applied Aspects of Probability and Statistics*, Taylor ans Francis, New York, 13–29.

U. N. Bhat and K. M. Kavi (1987), Reliability models for computer systems: An overview including dataflow graphs, Part 1, *Sadhana*, **11**, 167–186; Part 2, in N. Viswanadham, ed., *Reliability and Fault-Tolerant Issues in Realtime Systems*, Indian Academy of Sciences, Bangalore, India.

U. N. Bhat and G. K. Miller (2002), *Elements of Applied Stochastic Processes*, 3rd ed., Wiley, New York.

U. N. Bhat, G. K. Miller, and S. S. Rao (1997), Statistical analysis of queueing systems, in J. H. Dshalalow, ed., *Frontiers in Queueing*, CRC Press, New York, Chapter 13, 351–393.

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.

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.

P. Billingsley (1961), *Statistical Inference for Markov Processes*, University of Chicago Press, Chicago.

A. Birnbaum (1954), Statistical methods for Poisson processes and exponential populations, *J. Amer. Statist. Assoc.*, **49**, 254–266.

G. Boole with J. F. Moulton, ed. (1970), *A Treatise on the Calculus of Finite Differences*, 5th ed., Chelsea, New York.

O. J. Boxma and H. Takagi (1992), Editorial introduction, *Queueing Systems*, **11** (special issue on polling models), 1–5.

G. Brigham (1955), On a congestion problem in an aircraft factory, *Oper. Res.*, **3**, 412–428.

E. Brockmeyer, H. L. Halstrøm, and A. Jensen (1960), *The Life and Works of A. K. Erlang*, Applied Mathematics and Computing Machinery Series, Vol. 6, Acta Polytechnica Scandinavica, Copenhagen.

P. J. Burke (1956), The output of a queueing system, *Oper. Res.*, **4**, 699–714.

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.

J. A. Buzacott and J. G. Shanthikumar (1992), Editorial introduction, *Queueing Systems*, **12** (special issue on manufacturing systems), 1–2.

J. A. Buzacott and J. G. Shanthikumar (1993), *Stochastic Models of Manufacturing Systems*, Prentice–Hall, Upper Saddle River, NJ.

J. P. Buzen (1973), Computational algorithms for closed queueing networks with exponential servers, *Comm. Assoc. Comput. Mach.*, **16**, 527–531.

K. M. Chandy (1972), The analysis and solutions for general queueing networks, in *Proceedings of the* 6*th Annual Princeton Conference on Information Science and Systems*, Princeton University, Princeton, NJ.

M. L. Chaudhry (1992), Editorial introduction, *Queueing Systems*, **10** (special issue on numerical computations in queues), 1–3.

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.

M. L. Chaudhry and J. G. C. Templeton (1983), *A First Course in Bulk Queues*, Wiley, New York.

A. B. Clarke (1957), Maximum likelihood estimates in a simple queue, *Ann. Math. Statist.*, **28**, 1036–1040.

A. Cobham (1954), Priority assignment in waiting line problems, *Oper. Res.*, **2**, 70–76; correction, *Oper. Res.*, **3**, 547.

E. G. Coffman, Jr. and P. J. Denning (1973), *Operating Systems Theory*, Prentice–Hall, Englewood Cliffs, NJ.

E. G. Coffman, Jr. and M. Hofri (1986), Queueing models of secondary storage devices, *Queueing Systems*, **1**, 129–168.

E. G. Coffman, Jr. and L. Kleinrock (1968), Feedback queueing models for time shared systems, *J. Assoc. Comput. Mach.*, **15**, 549–576.

J. W. Cohen (1969), *The Single Server Queue*, North-Holland, London.

W. J. Conover (1971), *Practical Nonparametric Statistics*, Wiley, New York.

R. W. Conway, W. L. Maxwell, and L. W. Miller (1967), *Theory of Scheduling*, Addison–Wesley, Reading, MA.

R. B. Cooper (1981), *Introduction to Queueing Theory*, North-Holland, New York.

P. J. Courtois (1977), *Decomposability: Queueing and Computer Science Applications*, Academic Press, New York.

D. R. Cox (1955), The analysis of non-Markovian stochastic processes by the inclusion of supplementary variables, *Proc. Cambridge Philos. Soc.*, **51**, 433–441.

D. R. Cox (1962), *Renewal Theory*, Melthuen, London.

D. R. Cox (1965), Some problems of statistical analysis connected with congestion, in W. L. Smith and W. B. Wilkinson, eds., *Proceedings of the Symposium on Congestion Theory*, University of North Carolina Press, Chapel Hill, NC, 289–316.

D. R. Cox and P. A. W. Lewis (1966), *The Statistical Analysis of Series of Events*, Methuen, London.

T. B. Crabill (1972), Optimal control of a service facility with variable exponential service times and constant arrival rate, *Management Sci.*, **18**, 560–566.

T. B. Crabill, D. Gross, and M. J. Magazine (1977), A classified bibliography of research on optimal design and control of queues, *Oper. Res.*, **25**, 219–232.

P. J. Denning and J. P. Buzen (1978), The operational analysis of queueing network models, *Comput. Surv.*, **10**-3, 225–261.

R. L. Disney and P. C. Kiessler (1987), *Traffic Processes in Queueing Networks*, Johns Hopkins University Press, Baltimore.

B. T. Doshi (1986), Queueing systems with vacations: A survey, *Queueing Systems*, **1**, 29–66.

B. T. Doshi and D. Yao (1995), Editorial introduction, *Queueing Systems*, **20** (special issue on telecommunication systems), 1–5.

J. H. Dshalalow (1997), Queueing systems with state dependent parameters, in J. H. Dshalalow, ed., *Frontiers in Queueing*, CRC Press, New York, Chapter 4, 61–116.

H. Dubner and J. Abate (1968), Numerical inversion of Laplace transforms by relating them to the finite Fourier cosine transform, *J. Assoc. Comput. Mach.*, **15**, 115–123.

L. C. Edie (1954), Traffic delays at toll booths, *J. Oper. Res. Soc. Amer.*, **2**, 107–138.

A. K. Erlang (1917), Solution of some problems in the theory of probabilities of significance in automatic telephone exchanges, *Elektroteknikeren*, **13**, 5.

G. Falin (1990), A survey of retrial queues, *Queueing Systems*, **7**, 127–168.

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

T. C. Fry (1928), *Probability and Its Engineering Uses*, Van Nostrand, Princeton, NJ.

S. H. Fuller (1980), Performance evaluation, in H. S. Stone, ed., *Introduction to Computer Architecture*, 2nd ed., Science Research Associates, Chicago, 527–590.

D. P. Gaver, Jr. (1968), Diffusion approximations and models for certain congestion problems, *J. Appl. Probab.*, **5**, 607–623.

E. Gelenbe and G. Pujolle (1998), *Introduction to Queueing Networks*, Wiley, New York.

P. W. Glynn (1990), Diffusion approximations, in D. P. Heyman and M. J. Sobol, eds., *Stochastic Models*, Vol. 2, Elsevier Science, Amsterdam, Chapter 4, 145–198.

W. J. Gordon and G. F. Newell (1967), Closed queueing systems with exponential servers, *Oper. Res.*, **15**, 254–265.

T. L. Goyal and C. M. Harris (1972), Maximum likelihood estimation for queues with state dependent service, *Sankhya* A, **34**, 65–80.

D. Gross and C. M. Harris (1998), *Fundamentals of Queueing Theory*, 3rd ed., Wiley, New York.

F. A. Haight (1957), Queueing with balking, *Biometrika*, **44**, 360–369.

F. A. Haight (1967), *Handbook of the Poisson Distribution*, Wiley, New York.

G. Hardy (1949), *Divergent Series*, Oxford University Press, Oxford, UK.

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*, **50**, 144–148.

K. J. Hastings (2006), *Introduction to the Mathematics of Operations Research with Mathematica*, 2nd ed., Chapman and Hall, New York.

D. P. Heyman (1968), Optimal operating policies for $M/G/1$ queueing systems, *Oper. Res.*, **16**, 362–382.

F. B. Hildebrand (1968), *Finite Difference Equations and Simulations*, Prentice–Hall, Englewood Cliffs, NJ.

F. S. Hillier (1963), Economic models for industrial waiting line problems, *Management Sci.*, **10**, 119–130.

F. S. Hillier and G. J. Lieberman (1986), *Introduction to Operations Research*, 4th ed., Holden–Day, Oakland, CA.

T. Hirayama, S. J. Hong, and M. M. Krunz (2004), A new approach to analysis of polling systems, *Queueing Systems*, **48**, 135–158.

D. L. Iglehart and W. Whitt (1970), Multiple channel queues in heavy traffic I, II, *Adv. Appl. Probab.*, **2**, 150–177, 355–369.

J. R. Jackson (1957), Networks of waiting lines, *Oper. Res.*, **5**, 518–521.

J. R. Jackson (1963), Jobshop-like queueing systems, *Management Sci.*, **10**, 131–142.

R. R. P. Jackson (1954), Queueing processes with phase type service, *Oper. Res. Quart.*, **5**, 109–120.

R. R. P. Jackson (1956), Random queueing processes with phase type service, *J. Roy. Statist. Soc.* B, **18**, 129–132.

A. A. Jagers and E. A. van Doorn (1986), On the continued Erlang loss function, *Oper. Res. Lett.*, **5**, 43–46.

R. Jain (1991), *The Art of Computer Systems Performance Analysis*, Wiley, New York.

N. K. Jaiswal (1968), *Priority Queues*, Academic Press, New York.

W. S. Jewell (1967), A simple proof of $L = \lambda W$, *Oper. Res.*, **15**, 1109–1116.

N. L. Johnson and S. Kotz (1969), *Distributions in Statistics: Discrete Distributions*, Wiley, New York.

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.

S. Karlin and H. M. Taylor (1975), *A First Course in Stochastic Processes*, 2nd ed., Academic Press, New York.

F. P. Kelly (1979), *Reversibility and Stochastic Networks*, Wiley, New York.

D. G. Kendall (1951), Some problems in the theory of queues, *J. Roy. Statist. Soc.* B, **13**, 151–185.

D. G. Kendall (1953), Stochastic processes occurring in the theory of queues and their analysis by the method imbedded Markov chains, *Ann. Math. Statist.*, **24**, 338–354.

J. F. C. Kingman (1962a), Some inequalities for the $GI/G/1$ queue, *Biometrika*, **49**, 315–324.

J. F. C. Kingman (1962b), On queues in heavy traffic, *J. Roy. Statist. Soc.* B, **24**, 383–392.

J. F. C. Kingman (1965), The heavy traffic approximation in the theory of queues, in W. L. Smith and W. B. Wilkinson, eds., *Proceedings of the Symposium on Congestion Theory*, University of North Carolina Press, Chapel Hill, NC.

J. F. C. Kingman (1969), Markov population processes, *J. Appl. Probab.*, **6**, 1–18.

L. Kleinrock (1975), *Queueing Systems, Vol.* I: *Theory*, Wiley, New York.

L. Kleinrock (1976), *Queueing Systems Vol.* II: *Computer Applications*, Wiley, New York.

E. Koenigsberg (1958), Cyclic queues, *Oper. Res. Quart.*, **9**, 22–35.

T. Konstantopoulos (1998), Editorial introduction, *Queueing Systems*, **28** (special issue on mathematical and probabilistic methods in communication networks), 1–5.

J. A. Koziol and A. F. Nemec (1979), On a Cramer–von Mises type statistic for testing bivariate independence, *Canad. J. Statist.*, **7**, 43–52.

S. Krakowiak (1988), *Principles of Operating Systems*, MIT Press, Cambridge, MA (translated by D. Beeson).

V. G. Kulkarni (1997), Fluid models for single buffer systems, in J. H. Dshalalow, ed., *Frontiers in Queueing*, CRC Press, New York, Chapter 11, 321–338.

V. G. Kulkarni and H. M. Liang (1997), Retrial queues revisited, in J. H. Dshalalow, ed., *Frontiers in Queueing*, CRC Press, New York, Chapter 2, 19–34.

S. S. Lavenberg, ed. (1983), *Computer Performance Modeling Handbook*, Academic Press, New York.

A. M. Law and W. D. Kelton (1991), *Simulation Modeling and Analysis*, 2nd ed., McGraw–Hill, New York.

W. Ledermann and G. E. Reuter (1956), Spectral theory for the differential equations of simple birth and death processes, *Philos. Trans. Roy. Soc. London Ser.* A, **246**, 321–369.

P. A. W. Lewis (1972), *Stochastic Point Processes*, Wiley, New York, 1–54.

H. W. Lilliefors (1966), Some confidence intervals for queues, *Oper. Res.*, **14**, 723–727.

D. V. Lindley (1952), The theory of queues with a single server, *Proc. Cambridge Philos. Soc.* **48**, 277–289.

J. D. C. Little (1961), A proof for the queueing formula: $L = \lambda W$, *Oper. Res.* **9**, 383–387.

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, *Oper. Res.*, **4**, 711–732.

J. McKinney (1969), Survey of analytical time sharing models, *Comput. Surv.*, **1**, 105–116.

W. G. Marchal (1978), Some simpler bounds on the mean queueing time, *Oper. Res.*, **26**, 1083–1088.

K. T. Marshall (1968), Some inequalities in queueing, *Oper. Res.*, **16**, 651–665.

D. R. Miller (1981), Computation of steady state probabilities for $M/M/1$ priority queues, *Oper. Res.*, **29**, 945–958.

G. K. Miller (1996), *Estimation for Renewal Processes with Unobservable Interarrival Times*, doctoral dissertation, Southern Methodist University, Dallas; available through ProQuest/University Microfilms (http://il.proquest.com/brand/umi.shtml).

G. K. Miller (1999), Maximum likelihood estimation for Erlang integer parameter, *Statist. Probab. Lett.*, **43**, 335–341.

G. K. Miller and U. N. Bhat (1997), Estimation for renewal processes with unobservable gamma or Erlang interarrival times, *J. Statist. Planning Inference*, **61**, 355–372.

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. Math. Sci.*, **1**, 1–11.

D. Mitra and I. Mitrani (1991), Editorial introduction, *Queueing Systems*, **9** (special issue on communication systems), 1–4.

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.

J. J. Moder and C. R. Phillips, Jr. (1962), Queueing with fixed and variable channels, *Oper. Res.*, **10**, 218–231.

E. C. Molina (1927), Application of the theory of probability to telephone trunking problems, *Bell Systems Tech. J.*, **6**, 461–494.

M. K. Molloy (1989), *Fundamentals of Performance Modeling*, Macmillan, New York.

S. C. Moore (1975), Approximating the behavior of non-stationary single server queues, *Oper. Res.*, **23**, 1011–1032.

P. M. Morse (1958), *Queues, Inventories, and Maintenance*, Wiley, New York.

R. R. Muntz (1973), Poisson departure processes and queueing networks, in *Proceedings of the 7th Annual Princeton Conference on Information Science and Systems*, Princeton University Press, Princeton, NJ.

M. F. Neuts (1966), The single server queue with Poisson input and semi-Markov service times, *J. Appl. Probab.*, **3**, 202–230.

M. F. Neuts (1967), A general class of bulk queues with Poisson input, *Ann. Math. Statist.*, **68**, 759–770.

M. F. Neuts (1975), Probability distributions of phase type, in *Liber Amicorum Professor Emeritus H. Florin*, Department of Mathematics, University of Louvain, Louvain, Belgium, 173–206.

M. F. Neuts (1978), Markov chains with applications in queueing theory, which have a matrix geometric invariant probability vector, *Adv. Appl. Probab.*, **10**, 185–212.

M. F. Neuts (1989), *Structured Stochastic Matrices of $M/G/1$ Type and Their Applications*, Marcel Dekker, New York.

G. F. Newell (1968), Queues with time depedent arrival rates I, II, *J. Appl. Probab.*, **5**, 436–451, 579–606.

G. F. Newell (1971), *Applications of Queueing Theory*, Chapman and Hall, London.

C. Palm (1947), The distribution of repairmen in servicing automatic machines, *Industritidningen Norden*, **75**, 75–80, 90–94, 119–123 (in Swedish).

R. D. Pedersen and J. C. Shah (1972), Multiserver queue storage requirements with unpacked messages, *IEEE Trans. Comm.*, June, 462–465.

H. Perros (1994), *Queueing Networks with Blocking*, Oxford University Press, New York.

S. M. Pitts (1994), Nonparametric estimation of the stationary waiting time distribution for the $GI/G/1$ queue, *Ann. Statist.*, **22**, 1428–1446.

F. Pollaczek (1934), Über das Warteproblem, *Math. Z.*, **38**, 492–537.

F. Pollaczek (1965), Concerning an analytical method for the treatment of queueing problems, in W. L. Smith and W. B. Wilkinson, eds., *Proceedings of the Symposium on Congestion Theory*, University of North Carolina Press, Chapel Hill, NC, 1–42.

N. U. Prabhu (1960), Some results for the queue with Poisson arrivals, *J. Roy. Statist. Soc.* B, **22**, 104–107.

N. U. Prabhu (1965a), *Queues and Inventories*, Wiley, New York.

N. U. Prabhu (1965b), *Stochastic Processes*, Macmillan, New York.

N. U. Prabhu (1987), A bibliography of books and survey papers on queueing systems: Theory and applications, *Queueing Systems*, **2**, 393–398.

N. U. Prabhu (1997), *Foundations of Queueing Theory*, Kluwer Academic Publishers, Boston.

N. U. Prabhu (1998), *Stochastic Storage Processes*, Springer-Verlag, New York.

N. U. Prabhu and U. N. Bhat (1963a), Some first passage problems and their application to queues, *Sankhya* A, **25**, 281–292.

N. U. Prabhu and U. N. Bhat (1963b), Further results for the queue with Poisson arrivals, *Oper. Res.*, **11**, 380–386.

J. Putter (1955), The treatment of ties in some nonparametric tests, *Ann. Math. Statist.*, **26**, 368–386.

V. Ramaswami (1990), From the matrix-geometric to the matrix-exponential, *Queueing Systems*, **6**, 229–260.

V. Ramaswami (2001), The surprising reach of the matrix analytic approach, in A. Krishnamoorthy, N. Raju, and V. Ramaswami, eds., *Advances in Probability and Stochastic Processes*, Notable Publications, Neshanic Station, NJ, 167–177.

R. H. Randles and D. A. Wolfe (1979), *Introduction to the Theory of Nonparametric Statistics*, Wiley, New York.

S. S. Rao and U. N. Bhat (1991), A sequential test for a denumerable Markov chain and an application to queues, *J. Math. Phys. Sci.*, **25**, 521–527.

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.

E. Reich (1965), Departure processes, in W. L. Smith and W. B. Wilkinson, eds., *Proceedings of the Symposium on Congestion Theory*, University of North Carolina Press, Chapel Hill, NC, 439–457.

M. Reiser and S. S. Lavenberg (1980), Mean-value analysis of closed multichain queueing networks, *J. Assoc. Comput. Mach.*, **27**-2, 313–322.

W. Rudin (1964), *Principles of Mathematical Analysis*, McGraw–Hill, New York.

T. L. Saaty (1961), *Elements of Queueing Theory and Applications*, McGraw–Hill, New York.

T. L. Saaty (1966), Seven more years of queueing theory: A lament and a bibliography, *Nav. Res. Logist. Quart.*, **13**, 447–476.

C. H. Sauer and K. M. Chandy (1981), *Computer Systems Performance Modeling*, Prentice–Hall, Englewood Cliffs, NJ.

T. J. Schriber (1991), *Introduction to Simulation*, Wiley, New York.

B. Sengupta (1989), Markov processes whose steady state distribution is matrix-exponential with an application to $GI/PH/1$, *Adv. Appl. Probab.*, **22**, 159–180.

V. P. Singh (1970), Two-server Markovian queues with balking heterogeneous vs. homogeneous servers, *Oper. Res.*, **18**, 145–159.

V. P. Singh (1971), Markovian queues with three heterogeneous servers, *AIIE Trans.*, **III**-1 (March), 45–48.

D. R. Smith and W. Whitt (1981), Resource sharing for efficiency in traffic systems, *Bell Systems Tech. J.*, **60**, 39–65.

M. J. Sobel (1974), Optimal operation of queues, in A. B. Clarke, ed., *Mathematical Methods in Queueing Theory: Proceedings of a Conference at Western Michigan University*, Springer-Verlag, New York.

W. J. Stewart (1994), *Introduction to the Numerical Solution of Markov Chains*, Princeton University Press, Princeton, NJ.

S. Stidham, Jr. (1995), Editorial introduction, *Queueing Systems*, **21** (special issue on optimal design and control of queueing systems), 239–243.

S. Stidham and N. U. Prabhu (1974), Optimal control of queueing systems, in A. B. Clarke, ed., *Mathematical Methods in Queueing Theory: Proceedings of a Conference at Western Michigan University*, Springer-Verlag, New York.

T. Suzuki and Y. Yoshida (1970), Inequalities for many server queue and other queues, *J. Oper. Res. Soc. Japan*, **13**, 59–77.

R. Syski (1960), *Introduction to Congestion Theory in Telephone Systems*, Oliver and Boyd, Edinburgh.

L. Takács (1962), *Introduction to the Theory of Queues*, Oxford University Press, New York.

L. Takács (1967), *Combinatorial Methods in the Theory of Stochastic Processes*, Wiley, New York.

H. Takagi (1997), Queueing analysis of polling models: Progress in 1990–1994, in J. H. Dshalow, ed., *Frontiers in Queueing*, CRC Press, New York, Chapter 5, 119–146.

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.

D. Thiruvaiyaru and I. V. Basawa (1996), Maximum likelihood estimation for queueing networks, in B. L. S. Prakasa Rao and B. R. Bhat, *Stochastic Processes and Statistical Inference*, New Age International Publications, New Delhi, 132–149.

D. Thiruvaiyaru, I. V. Basawa, and U. N. Bhat (1991), Estimation for a class of simple queueing networks, *Queueing Systems*, **9**, 301–312.

K. S. Trivedi (2002), *Probability and Statistics with Reliability, Queueing, and Computer Science Applications*, 2nd ed., Wiley, New York.

W. N. Venables and B. D. Ripley (2002), *Modern Applied Statistics with S*, 4th ed., Springer-Verlag, New York.

P. T. Whittaker and G. N. Watson (1962), *A Course of Modern Analysis*, 4th ed., Cambridge University Press, Cambridge, UK.

W. Whitt (2000), An overview of Brownian and non-Brownian FCLTs for the single server queue, *Queueing Systems*, **36**, 39–70.

P. Whittle (1967), Nonlinear migration processes, *Bull. Internat. Statist. Inst.*, **42**, 642–646.

P. Whittle (1968), Equilibrium distributions for an open migration process, *J. Appl. Probab.*, **5**, 567–571.

D. V. Widder (1946), *The Laplace Transform*, Princeton University Press, Princeton, NJ.

R. W. Wolff (1965), Problems of statistical inference for birth and death queueing models, *Oper. Res.*, **13**, 343–357.

R. W. Wolff (1982), Poisson arrivals see time averages, *Oper. Res.*, **30**, 223–231.

T. Yang and J. G. C. Templeton (1987), A survey of retrial queues, *Queueing Systems*, **2**, 201–233.

S. F. Yashkov (1987), Processor showing queues: Some progress in analysis, *Queueing Systems* **2**, 1–17.

Index