# Preface

This book covers fundamental concepts in queueing and loss models, as well as simulation methods, with their application to modeling and analysis of computer systems and communication networks. Modeling and analysis are essential components in the process of designing and dimensioning a computer system or network. The aim of the book is to provide the reader with state-of-the-art analytical and computational tools to evaluate the performance and operating characteristics of today's computer systems and communication networks. The theory and methodologies discussed in this book may be applied to other fields such as manufacturing, operations research, and industrial engineering.

This volume is an updated and expanded version of an earlier book by the first author entitled *Modeling and Analysis: An Introduction to System Performance Evaluation Methodology* published by Addison-Wesley in 1978. The present book includes more advanced and recent materials on queueing and traffic models in the context of both computing systems and networks. For instance, the generalized Erlang and Engset loss models and loss network theory are discussed for the first time in a cohesive manner. The book also features an up-to-date treatment of simulation methodologies for systems and networks, including an introduction to network simulation packages and recent developments on random number generation.

## ORGANIZATION OF THE BOOK

Chapter 1 provides an overview of the scope of the performance evaluation methodologies to be covered in the book. We start with a discussion of the significant role stochastic modeling has played in science and engineering, followed by a brief review of successful "modeling and analysis" efforts in the development of computing systems and communication networks. The art of performance modeling is illustrated by means of several examples taken from today's computing systems and networks. State-of-the-art practice in performance evaluation is reviewed and the strengths and limitations of these techniques are discussed.

The remainder of the book is organized into four parts. Part I, "Basic Queueing and Loss Models", covers the principles of queueing and loss analysis and teletraffic theory. Chapter 2 starts with Little's formula, the most fundamental and useful formula in queueing analysis, followed by a review of Poisson processes, which leads to formulation of birth-and-death (BD) queueing models. M/M/1, M/M/$\infty$, and other Markovian queueing models are derived as special cases of the BD queueing models. Chapter 3 introduces the classical loss models studied by Erlang and Engset in telephone engineering a century ago, but they still capture the essence of connection-oriented network services, be they over wired, wireless, or optical channels. Chapter 4 begins with various ways of representing general distributions of either service time or interarrival time, followed by discussion of the embedded

Markov chain technique due to Kendall, as applied to queues in which either the arrival process or the service process is non-Markovian, i.e., M/G/1 and G/M/$m$ queueing models. Chapter 5 discusses a class of queueing models whose departure process is shown to be Poisson when the arrival process is Poisson. This important class of models is derived by using the notions of quasi-reversibility and symmetric queues introduced by Kelly. Loss servers (as in the Erlang and Engset loss models), M/G/$\infty$ and M/G/1 with PS (processor-sharing) scheduling, and M/G/1 with LCFS (last-come, first-served) with PR (preemptive resume) scheduling fall in this class.

In Part II, “Queueing and Loss Networks”, we provide a cohesive treatment of queueing network models, loss network models, and computational algorithms associated with these models. Chapter 6 starts with the classical Jackson network that has the so-called product-form solution for the joint queue distribution, and then introduces a more general model, often referred to as the *BCMP* (Baskett–Chandy–Muntz–Palacios) model. Multiple classes of customers and general routing behavior are also incorporated into these generalized Jackson-type networks. Chapter 7 addresses the loss system counterparts of the results presented in Chapter 6; it can also be viewed as a generalization of the classical loss models discussed in Chapter 3. Chapter 8 provides a comprehensive treatment of efficient computational algorithms to numerically evaluate performance measures associated with queueing and loss networks.

Part III, “Advanced Queueing Models”, consists of several advanced topics. Chapter 9 begins with a discussion of some conservation laws due to Kleinrock. Then, priority queues and the so-called server vacation models are discussed, the latter being suited for modeling scheduling algorithms involving polling and token passing. Chapter 10 discusses the classical theory of the G/G/1 queue due to Lindley, followed by the more recent work of Neuts and others on phase-type (PH) distributions and their application to a numerically treatable subset of G/G/1 queues, known as PH/PH/1 queues. The notion of quasi-birth-and-death (QBD) is also discussed.

Queueing theory is generally formulated in the continuous-time setting. However, in such systems as time-slotted and synchronous systems, discrete-time modeling is more appropriate. In Chapter 11, we present Geo/Geo/1 (the discrete-time analog of M/M/1), Geo/G/1, discrete-time M/G/$\infty$, and discrete-time G/G/1 models.

In Chapter 12 we treat two different types of traffic model in which interarrival times are not independent: one is the class of Markovian traffic models, the other is the class of long-range dependent (LRD) traffic models. Empirical studies reported in the literature have suggested that some Internet and LAN traffic exhibit the self-similar property, and the theory of fractional Brownian motion (fBM) developed by Mandelbrot provides a mathematical characterization of such non-Markovian traffic.

Chapter 13 is devoted to fluid models and their applications. While the diffusion process approximation characterizes the queue process in terms of the first and second moments of the arrival and departure processes, the fluid approximation uses only the first moment. It is equivalent to treating the movement of customers (e.g., packets) as the flow of a fluid, ignoring random fluctuations associated with the irregularity of interarrival and service times. We will find fluid models useful in the analysis of statistical multiplexer models.

The analytic solution of G/G/1 based on Lindley's theory presented in Chapter 10 requires spectrum decomposition, which in practice may present great difficulty except for special cases. Alternatives to the numerical solution technique via the PH/PH/1 formulation are some bounding and approximation techniques, which are often extremely simple and insightful. In Chapter 14 we discuss exponentially tight bounds on both the waiting time distribution and the mean waiting time. The latter argument can be generalized to the G/G/$m$ queue as well. Heavy traffic approximation and diffusion process approximation are closely related. Both require only the first two moments of service time and interarrival time distributions. Also covered in this chapter are diffusion approximation of a queueing network, reduced load approximation in a loss network, and the concept of effective bandwidth.

The last topic of Part III is time-dependent solutions of some queueing models. The time-dependent solution is generally hard to come by: a closed-form solution for the time-dependent queue distribution of the M/M/1, the simplest queueing model, involves modified Bessel functions! In Chapter 15 we present several different methods for transient analysis of Markovian models—the matrix representation method, the spectral expansion method, the eigenvector method, and the Laplace transform method. These solution methods can be shown to be mathematically equivalent or closely related to each other. We also show that the diffusion and fluid approximations often lend themselves to simpler transient analysis.

Part IV, "Simulation Modeling and Analysis", focuses on implementation methodologies for discrete-event simulators, and statistical tools for the design and analysis of simulation experiments. In Chapter 16, basic concepts of discrete-event simulation are discussed, including the formulation of simulation models and random number generation. Then, simulation languages and environments for more complex simulation tasks are discussed. We discuss two popular packages for network simulation: a commercial package OPNET and the public-domain software ns-2. Then three case studies—a single server queue, a queueing network, and a loss network—are presented to illustrate how discrete-event simulators can be developed.

Chapter 17 discusses how a simulation experiment ought to be designed and how its output data should be analyzed. Basic concepts such as significance level and confidence interval are reviewed. The chapter also discusses some useful practices in analyzing a simulation run. Finally, the chapter discusses a very important but seldom practiced technique in simulation experiments, i.e., how to minimize the simulation run time while maintaining the accuracy of the simulation estimate at a required level. Efficient simulation is very critical when we model rare events such as buffer overflow or packet loss. Various variance reduction techniques as well as the importance sampling method are discussed.

Appendix A provides a brief exposition of concepts in number theory that are needed to follow the theory behind the various random number generation methods discussed in Section 16.5.

## SUGGESTED COURSE PLANS

We assume that the readers or students are familiar with *calculus, linear algebra (or matrix theory), and statistics* at the undergraduate level, and *probability, random variables, and random processes* at the first-year graduate level. In Appendix A, we provide a brief review of concepts in *number theory*, which are pertinent to the topic of random number generation discussed in Chapter 17.

The book is suitable as a graduate-level textbook, but should also be of interest to researchers concerned with the performance evaluation of computer systems and networks. As a text the book can be used for a two-semester course sequence on system performance evaluation. The first semester could cover the fundamental material in Chapters 2 through 8, as well as Chapter 16 on simulation methodology. The second semester could then cover a selection of material of the advanced topics from Chapters 9 through 15, as well Chapter 17 on simulation experiments and data analysis. Alternatively, portions of the book can be covered in a one-semester course. For example, a one-semester course emphasizing computational methods could comprise Chapters 2 through 8, with selected material from Chapters 9 through 15. A course emphasizing simulation methods could comprise Chapters 2 through 7, with selected material from Chapters 16 and 17.

At Princeton, the materials in Chapters 2 through 8 and some selected topics of Chapters 9 through 15 have been taught as a part of a graduate course "ELE 531: Communication Networks". Syllabi for these courses are available at `www.princeton.edu/kobayashi`. The students are expected to have taken "ELE525: Random Processes for Information Systems". Syllabi for these courses are available at `www.princeton.edu/kobayashi`. At George Mason University, draft copies of the book were used in a graduate course "ECE 642: Design and Analysis of Computer Networks", and the course syllabus can be found at `ece.gmu.edu/~bmark`.

Princeton, New Jersey<br>
Fairfax, Virginia<br>
March 2008

Hisashi Kobayashi<br>
Brian L. Mark