

---

# FOUNDATIONS OF QUEUEING THEORY

**INTERNATIONAL SERIES IN**  
**OPERATIONS RESEARCH & MANAGEMENT SCIENCE**

---

**Frederick S. Hillier, Series Editor**  
Department of Operations Research  
Stanford University  
Stanford, California

Saigal, Romesh  
The University of Michigan  
*LINEAR PROGRAMMING: A Modern Integrated Analysis*

Nagurney, Anna/ Zhang, Ding  
University of Massachusetts @ Amherst  
*PROJECTED DYNAMICAL SYSTEMS AND VARIATIONAL INEQUALITIES  
WITH APPLICATIONS*

Padberg, Manfred/ Rijal, Minendra P.  
New York University  
*LOCATION, SCHEDULING, DESIGN AND INTEGER  
PROGRAMMING*

Vanderbei, Robert J.  
Princeton University  
*LINEAR PROGRAMMING: Foundations and Extensions*

Jaiswal, N.K.  
Ministry of Defense, INDIA  
*MILITARY OPERATIONS RESEARCH: Quantitative Decision Making*

Gal, Tomas / Greenberg, Harvey J.  
FernUniversität Hagen/ University of Colorado @ Denver  
*ADVANCES IN SENSITIVITY ANALYSIS AND PARAMETRIC PROGRAMMING*

---

# FOUNDATIONS OF QUEUEING THEORY

---

**N.U. Prabhu**  
*Cornell University*  
*Ithaca, New York, USA*



SPRINGER SCIENCE+BUSINESS MEDIA, LLC

**Library of Congress Cataloging-in-Publication**

Foundations of Queueing Theory  
By N.U. Prabhu

0-7923-9962-5

A C.I.P. Catalogue record is available from the Library of Congress

---

ISBN 978-1-4613-7845-7      ISBN 978-1-4615-6205-4 (eBook)  
DOI 10.1007/978-1-4615-6205-4

Second Printing 2002.

Copyright © 1997 by Springer Science+Business Media New York  
Originally published by Kluwer Academic Publishers in 1997  
Softcover reprint of the hardcover 1st edition 1997

This printing is a digital duplication of the original edition.

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system or transmitted in any form or by any means, mechanical, photo-copying, recording, or otherwise, without the prior written permission of the publisher, with the exception of any material supplied specifically for the purpose of being entered and executed on a computer system, for exclusive use by the purchaser of the work.

*Printed on acid-free paper.*

To Vasundhara and Purnima

---

# CONTENTS

<b>PREFACE</b>	<b>xi</b>
<b>1 INTRODUCTION</b>	<b>1</b>
1.1 Description of a Queueing System	1
1.2 The Basic Model GI/G/S	3
1.3 Processes of Interest	6
1.4 The Nature of Congestion	8
1.5 Little's Formula $L = \lambda W$	9
1.6 Control of Queueing Systems	10
1.7 Historical Remarks	11
<b>2 MARKOVIAN QUEUEING SYSTEMS</b>	<b>13</b>
2.1 Introduction	13
2.2 The System $M/M/1$	14
2.3 The System $M/M/s$	17
2.4 A Design Problem	21
2.5 $M/M/s$ System with Finite Source	22
2.6 The Machine Interference Problem	23
2.7 The System $M/M/s$ with Finite Capacity	24
2.8 Loss Systems	26
2.9 Social Versus Self-Optimization	27
2.10 The System $M/M/s$ with Balking	30
2.11 The System $M/M/s$ with Reneging	32
2.12 Problems for Solution	35
<b>3 THE BUSY PERIOD, OUTPUT AND QUEUES IN SERIES</b>	<b>43</b>

3.1	Introduction	43
3.2	The Busy Period	43
3.3	The $M/M/S$ System with Last Come, First Served	50
3.4	Comparison of FCFS and LCFS	51
3.5	Time-Reversibility of Markov Processes	52
3.6	The Output Process	54
3.7	The Multi-Server System in a Series	55
3.8	Problems for Solution	56
<b>4</b>	<b>ERLANGIAN QUEUEING SYSTEMS</b>	<b>59</b>
4.1	Introduction	59
4.2	The System $M/E_k/1$	60
4.3	The System $E_k/M/1$	67
4.4	The System $M/D/1$	72
4.5	Problems for Solution	74
<b>5</b>	<b>PRIORITY SYSTEMS</b>	<b>79</b>
5.1	Description of a System with Priorities	79
5.2	Two Priority Classes with Pre-emptive Resume Discipline	82
5.3	Two Priority Classes with Head-of-Line Discipline	87
5.4	Summary of Results	91
5.5	Optimal Assignment of Priorities	91
5.6	Problems for Solution	93
<b>6</b>	<b>QUEUEING NETWORKS</b>	<b>97</b>
6.1	Introduction	97
6.2	A Markovian Network of Queues	98
6.3	Closed Networks	103
6.4	Open Networks: The Product Formula	104
6.5	Jackson Networks	111
6.6	Examples of Closed Networks; Cyclic Queues	112
6.7	Examples of Open Networks	114
6.8	Problems for Solution	118
<b>7</b>	<b>THE SYSTEM <math>M/G/1</math>; PRIORITY SYSTEMS</b>	<b>123</b>
7.1	Introduction	123

7.2	The Waiting Time in $M/G/1$	124
7.3	The Sojourn Time and the Queue Length	129
7.4	The Service Interval	132
7.5	The $M/G/1$ System with Exceptional Service	133
7.6	The Busy Period in $M/G/1$	137
7.7	Completion Times in Priority Systems	141
7.8	Low Priority Waiting Time	145
7.9	Problems for Solution	146
<b>8</b>	<b>THE SYSTEM <math>GI/G/1</math>; IMBEDDED MARKOV CHAINS</b>	<b>149</b>
8.1	Imbedded Markov Chains	149
8.2	The System $GI/G/1$	150
8.3	The Wiener-Hopf Technique; Examples	154
8.4	Set-up Times; Server Vacations	161
8.5	The Queue Length and Waiting Time in $GI/M/1$	166
8.6	The Queue Length in $M/G/1$	170
8.7	Time Sharing Systems	173
8.8	The $M/M/1$ System with RR Discipline	174
8.9	Problems for Solution	177
<b>A</b>	<b>APPENDIX</b>	<b>181</b>
A.1	The Poisson Process	181
A.2	Renewal Theory	184
A.3	The Birth-And-Death Process	187
A.4	Markov Processes with a Countable State Space	190
A.5	Markov Chains	191
A.6	Two Theorems on Functional Equations	196
A.7	Review Problems in Probability and Stochastic Processes	197
<b>B</b>	<b>BIBLIOGRAPHY</b>	<b>199</b>
	<b>INDEX</b>	<b>203</b>

---

# PREFACE

Over the last twenty years several books on queueing theory have appeared, treating it at a level suitable for an undergraduate course of study. These books were adequate for their purpose for a while, but with rapid advances in the subject area their treatment has become a little outdated. This is because current research has shown the need to pay more attention to the basic concepts and techniques of queueing theory. These include the busy period, imbedded chains, regeneration points, Wiener-Hopf technique, time-reversibility, output, vector Markov processes, remaining workload and completion times. The use of these concepts and techniques considerably simplifies the analysis of models involving last come, first served queue discipline, priorities, networks, set-up times and server vacations. These form the foundations of queueing theory, developed by D.G. Kendall in his pioneering work in 1951-1954 and by other authors during the two decades that followed. Of course there are also techniques based on random walks, martingales and point processes, but these are beyond the scope of an undergraduate level text.

The present book deals with the foundations of queueing theory and is intended as a text for an undergraduate course on queueing theory. I have not attempted to merely chronicle the results of queueing theory as historically derived, but established them by my own approach to the subject. However, I have avoided the "monograph" style of citing the author and year of publication of each and every result. For convenience of reference the main results are stated in the form of theorems.

A pre-requisite for the book is an undergraduate course on stochastic processes. The earlier part of the text uses results from the Poisson process, renewal theory, birth-and-death processes and Markov chains. As the presentation progresses, relatively advanced concepts such as time-reversibility, vector Markov processes, Wiener-Hopf technique and regenerative sets are introduced, with strong motivation from the queueing models considered. Applied mathematical tools used in the text include generating functions, Laplace transforms, Laplace-Stieltjes transforms and Fourier transforms. It is expected that the instructor will provide a brief review of this material.

Chapters 1-6 adequately cover the material for a one-term course at the senior undergraduate level. For a graduate level course the instructor can skip Chapters 1-2 and concentrate on Chapters 3-8.

The following is a brief summary of the contents of the book. In the introductory chapter the different features of a queueing system are described. We mainly follow Kendall's characterization of the queueing model in terms of input, queue discipline and service mechanism, but add a new feature, namely, cost structure, which is important in the formulation of optimality problems we consider. Kendall's notation  $GI/G/s$  is retained, in spite of a tendency in current literature to change it to  $G/G/S$ . The traffic intensity is defined as the ratio of the workload submitted to the system and the maximum service that the system is capable of providing. In addition to the queue length and waiting time, we introduce the notions of remaining workload and busy time. To illustrate the nature of congestion we consider a deterministic model. A partial proof of Little's formula  $L = \lambda W$  is given, but we have preferred to derive the mean queue length and mean waiting time separately for most of the models considered.

A convenient way to start the study of queueing systems is with the queue length in  $M/M/s$ , which is the familiar birth-and-death process. This is done in Chapter 2. The steady state distributions of the queue length and waiting time are derived for  $s = 1$  and  $s > 1$ . We also consider the cases where the customers come from a finite source, the system has finite waiting space capacity, customers balk or renege. These classical models have proved to be of importance in recent applications. In these systems (except for renegeing) only the queue length is considered. We also treat two design problems, one involving the number of servers and the other involving the capacity of the system.

Relatively advanced topics in  $M/M/s$  are treated in Chapter 3. We establish I.J. Good's analogy between the  $M/M/1$  busy period and a branching process. This analogy is extended to derive the joint distribution of the length of the busy period and the number of customers served. The system busy period in  $M/M/s$  is also considered. The concept of busy period is used to characterize the waiting time in  $M/M/s$  with last come, first served queue discipline (and the low priority waiting time in Chapter 5 and completion times in Chapter 7). Time-reversibility of the  $M/M/s$  queue length process is established and used to characterize its output. The results are then applied to the multi-server system in a series.

A.K. Erlang's analysis of  $M/E_k/1$  is based on the method of phases. However, in Chapter 4 we use a different approach for  $M/E_k/1$  as well as  $E_k/M/1$ . Thus, in  $M/E_k/1$  we study the bivariate Markov process  $\{Q(t), R(t)\}$ , where  $Q(t)$  is the queue length and  $R(t)$  is the number of remaining phases of the customer being served. This yields more information concerning the system (such as the distribution of the remaining service time) and also motivates the study of vector Markov processes in the priority systems of Chapter 5 and queueing networks of Chapter 6. The system  $M/D/1$  is studied as a limiting case of  $M/E_k/1$ .

In Chapter 5 we study Markovian priority systems. The steady state distributions of the queue lengths and waiting times are derived. The optimality of the shortest processing times is established.

Queueing networks are discussed in Chapter 6, with Markovian networks as the main focus of study. A unified approach is used for closed and open networks to derive the steady state distribution of the queue length. For open networks the product formula is established, and the notion of quasi-reversibility is used to characterize external departures. Jackson networks are treated as a special case of open Markovian networks. Several examples of open and closed networks are discussed.

In Chapter 7 we study queueing systems with Poisson arrivals. For the standard system  $M/G/1$  explicit expressions are derived for the steady state distributions of the waiting time and the queue length. The basic tool used is renewal theory. We study the spent and remaining service times of the customer at the counter and explain the paradox concerning the service interval. Results for the busy period are obtained by extending the branching process analogy used in Chapter 3. We also consider a variant of the  $M/G/1$  system in which the customer who initiates a busy period has a service time different from those of other customers. The results of this system are applied to priority systems with Poisson arrivals. Our approach is based on the notion of completion times. The steady state distribution of low priority customer's waiting time is expressed as a weighted sum of two other distributions.

In the final Chapter 8 of the book we consider the system  $GI/G/1$  using imbedded Markov chains. D.V. Lindley's analysis of the system is extended to cover waiting times as well as idle times. This makes the application of the Wiener-Hopf technique probabilistically more meaningful. Apart from the standard  $M/G/1$  system and its variant with exceptional service times (treated in Chapter 7) we also consider the  $M/G/1$  system with additional workloads imposed on the server at the beginning of each busy period. These results are applied

to the  $M/G/1$  system with set-up times and server vacations. The product form property which has attracted recent attention is explained. The imbedded chain analysis is carried out for the queue length in  $GI/M/1$  and  $M/G/1$ . We also consider time-sharing systems.

At the end of each chapter there are problems for solution. These are at a level of difficulty compatible with those in a typical course in operations research.

The Appendix contains a brief review of the results of the elementary theory of stochastic processes: the Poisson process, renewal theory, birth-and-death process and (the more general) Markov processes with a countable state space, and Markov chains. We also establish the uniqueness of solution for two functional equations that arise in queueing theory.

Although we have considered some elementary optimization problems, it has not been possible to include a more complete treatment of design and control problems arising in queueing theory. The overall constraint on the size of the book has also meant that numerical computations and simulation studies also be left out of the scope of the presentation. These important topics deserve to be included in an advanced text on queueing theory, but the present book does not belong to this category.

Thanks are due to Sharon Hobbie for her efficient typing of the manuscript and to Gary Folven, OR/MS Editor of Kluwer Academic Publishers, for his encouragement.

Ithaca, New York  
April 1997

N.U. Prabhu