AN INTRODUCTION TO QUEUEING THEORY AND MATRIX-ANALYTIC METHODS

An Introduction to Queueing Theory and Matrix-Analytic Methods

by

L. BREUER
University of Trier, Germany

and

D. BAUM
University of Trier, Germany

Springer

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

ISBN 1-4020-3630-2 (HB)
ISBN 978-1-4020-3630-9 (HB)
ISBN 1-4020-3631-0 (e-book)
ISBN 978-1-4020-3631-6 (e-book)

Published by Springer,
P.O. Box 17, 3300 AA Dordrecht, The Netherlands.

www.springeronline.com

Printed on acid-free paper

Printed in the Netherlands.

Contents

List of Figures ix
Foreword xi

1. QUEUES: THE ART OF MODELLING 1

Part I Markovian Methods

2. MARKOV CHAINS AND QUEUES IN DISCRETE TIME 11
   1 Definition 11
   2 Classification of States 15
   3 Stationary Distributions 20
   4 Restricted Markov Chains 27
   5 Conditions for Positive Recurrence 29
   6 The M/M/1 queue in discrete time 31

3. HOMOGENEOUS MARKOV PROCESSES ON DISCRETE STATE SPACES 39
   1 Definition 39
   2 Stationary Distribution 46

4. MARKOVIAN QUEUES IN CONTINUOUS TIME 51
   1 The M/M/1 Queue 51
   2 Skip–Free Markov Processes 54
   3 The M/M/∞ Queue 55
   4 The M/M/k Queue 56
   5 The M/M/k/k Queue 58
   6 The M/M/k/k+c/N Queue 59

5. MARKOVIAN QUEUEING NETWORKS 63
1 Balance Equations and Reversibility Properties 65
2 Jackson and Gordon-Newell Networks 80
3 Symmetric Service Disciplines 99

Part II Semi–Markovian Methods

6. RENEWAL THEORY 113
1 Renewal Processes 113
2 Renewal Function and Renewal Equations 116
3 Renewal Theorems 118
4 Residual Life Times and Stationary Renewal Processes 124
5 Renewal Reward Processes 130

7. MARKOV RENEWAL THEORY 135
1 Regenerative Processes 135
2 Semi–Markov Processes 138
3 Semi–regenerative Processes 144

8. SEMI–MARKOVIAN QUEUES 147
1 The GI/M/1 Queue 147
2 The M/G/1 Queue 155
3 The GI/M/m Queue 160

Part III Matrix–Analytic Methods

9. PHASE–TYPE DISTRIBUTIONS 169
1 Motivation 169
2 Definition and Examples 171
3 Moments 176
4 Closure Properties 178

10. MARKOVIAN ARRIVAL PROCESSES 185
1 The PH renewal process 185
2 From PH renewal processes to MAPs 187
3 From MAPs to BMAPs 188
4 Distribution of the Number of Arrivals 190
5 Expected Number of Arrivals 192

11. THE GI/PH/1 QUEUE 197
1 The Embedded Markov Chain 198
2 Stationary Distribution at Arrival Instants 199
3 Ergodicity of the Embedded Markov Chain 204
4 Asymptotic Distribution of the System Process 208

12. THE BMAP/G/1 QUEUE 213
1 The Embedded Markov Chain 214
2 The Matrix G 215
3 Stationary Distribution at Service Completions 216
4 Asymptotic Distribution of the System Process 218
5 Stability Conditions 224

13. DISCRETE TIME APPROACHES 229
1 Discrete Phase–Type Distributions 229
2 BMAPs in Discrete Time 232
3 Blockwise Skip–Free Markov Chains 234
4 The PH/PH/1 Queue in Discrete Time 236

14. SPATIAL MARKOVIAN ARRIVAL PROCESSES 239
1 Arrivals in Space 240
2 Properties of Spatial MAPs 245

15. APPENDIX 253
1 Conditional Expectations and Probabilities 253
2 Extension Theorems 256
3 Transforms 258
4 Gershgorin's Circle Theorem 260

References 263

Index 269

List of Figures

1.1	Single server queue	2
1.2	Multi–server queue	2
1.3	Total system time	6
3.1	Typical path	40
3.2	Poisson process	43
4.1	M/M/1 queue	51
4.2	Transition rates for the M/M/1 queue	52
4.3	A skip–free Markov process	54
4.4	M/M/k queue	57
4.5	A closed computer network	60
5.1	Open Queueing Network	64
5.2	Modified Network	91
5.3	Central Server Model	96
5.4	Modified Central Server Model	98
5.5	FCFS Order	103
5.6	Cox Distribution	103
5.7	LCFS Order	104
5.8	Simple model of a computer pool	108
6.1	Random variables of a renewal process	113
7.1	Typical path of a semi-Markov process	139
8.1	Fix point as intersection with diagonal	150
9.1	Erlang distribution	174
9.2	Generalized Erlang distribution	175
9.3	Hyper–exponential distribution	175
9.4	Cox distribution	176

9.5 Convolution of two PH distributions 179
9.6 Mixture of two PH distributions 180
9.7 Superposition of two PH distributions 182

Foreword

The present textbook contains the records of a two–semester course on queueing theory, including an introduction to matrix–analytic methods. This course comprises four hours of lectures and two hours of exercises per week and has been taught at the University of Trier, Germany, for about ten years in sequence. The course is directed to last year undergraduate and first year graduate students of applied probability and computer science, who have already completed an introduction to probability theory. Its purpose is to present material that is close enough to concrete queueing models and their applications, while providing a sound mathematical foundation for the analysis of these. Thus the goal of the present book is two–fold.

On the one hand, students who are mainly interested in applications easily feel bored by elaborate mathematical questions in the theory of stochastic processes. The presentation of the mathematical foundations in our courses is chosen to cover only the necessary results, which are needed for a solid foundation of the methods of queueing analysis. Further, students oriented towards applications expect to have a justification for their mathematical efforts in terms of immediate use in queueing analysis. This is the main reason why we have decided to introduce new mathematical concepts only when they will be used in the immediate sequel.

On the other hand, students of applied probability do not want any heuristic derivations just for the sake of yielding fast results for the model at hand. They want to see the close connections between queueing theory and the theory of stochastic processes. For them, a systematic introduction to the necessary concepts of Markov renewal theory is indispensable. Further, they are not interested in any technical details of queueing applications, but want to see the reflection of the mathematical concepts in the queueing model as purely as possible.

A prominent part of the book will be devoted to matrix–analytic methods. This is a collection of approaches which extend the applicability of Markov renewal methods to queueing theory by introducing a finite number of auxiliary states. For the embedded Markov chains this leads to transition matrices in block form having the same structure as the classical models. With a few modifications they can be analyzed in the same way.

Matrix–analytic methods have become quite popular in queueing theory during the last twenty years. The intention to include these in a students' introduction to queueing theory has been the main motivation for the authors to write the present book. Its aim is a presentation of the most important matrix–analytic concepts like phase–type distributions, Markovian arrival processes, the GI/PH/1 and BMAP/G/1 queues as well as QBDs and discrete time approaches. This is the content of part III of this book.

As an introductory course for students it is necessary to provide the required results from Markov renewal theory before. This is done in part I, which contains Markovian theory, and part II which combines the concepts of part I with renewal theory in order to obtain a foundation for Markov renewal theory. Certainly only few students would like to acquire this theoretical body without some motivating applications in classical queueing theory. These are introduced as soon as the necessary theoretical background is provided.

The book is organized as follows. The first chapter gives a short overview of the diverse application areas for queueing theory and defines queues and their system processes (number of users in the system). The appendix sections in chapter 15 provide an easy reference to some basic concepts of analysis and probability theory.

For the simple Markovian queueing models (in discrete and continuous time) it suffices to give a short introduction to Markov chains and processes, and then present an analysis of some queueing examples. This is done in chapters 2 through 4. Chapter 5 gives an introduction to the analysis of simple queueing networks, in particular Jackson and Gordon–Newell networks as well as BCMP networks. This concludes the first part of the book, which deals with Markovian methods exclusively.

The second part is devoted to semi–Markovian methods. In chapter 6 the most important results of renewal theory are provided. Chapter 7 contains a short introduction to Markov renewal theory. This will be necessary for the analysis of the classical semi–Markovian queues (namely the GI/M/1 and M/G/1 systems), which is presented in chapter 8.

More recent approaches which are usually subsumed under the term "matrix–analytic methods" are presented in the third part of the book. In chapters

9 and 10 the basic concepts of phase–type distributions and Markovian arrival processes are introduced. The matrix–analytic analogues to the GI/M/1 and M/G/1 queues, namely the GI/PH/1 and BMAP/G/1 systems are analyzed in chapters 11 and 12. Chapter 13 gives a short overview on discrete time analogues. Further blockwise skip–free Markov chains, also known as QBD processes, are analyzed, with an application to the PH/PH/1 queue in discrete time. Finally, in chapter 14 a generalization of BMAPs towards spatial Markovian arrival processes is presented.

Of course, most of the more classical material can be found in existing textbooks on stochastic processes. For example, Çinlar [25] and Ross [75] still contain, in our view, the most systematic treatment of semi–Markovian queues. Also of great value, mostly for the theory of Markov chains and processes, are the courses on stochastic processes by Karlin and Taylor [46, 47]. Further important results may be found in Doob [31], Asmussen [5], and Nelson [61]. The material on queueing networks can be found in Mitrani [60], Kelly [48], and Kleinrock [50]. Monographs on matrix–analytic methods are the pioneering books by Neuts [65, 66], and Latouche and Ramaswami [52]. For discrete time methods the overview paper by Alfa [2] was helpful.

However, some aspects of standard presentation have been changed in order to alleviate the mathematical burden for the students. The stationary regime for Markov chains has been introduced as an asymptotic mean over time in order to avoid the introduction of periodicity of states. The definition of Markov processes in chapter 3 is much closer to the derivation of immediate results. It is not necessary to derive the standard path properties in lengthy preliminary analyses, since these are already included in the definition. Nevertheless, the close connection between the phenomena observed in queueing systems and the definition given in our textbook is immediately clear to the student.

The introduction of renewal theory has been postponed to the second part of the book in order to show a variety of queueing application of a purely Markovian nature first. The drawback that a proof for asymptotic behaviour of Markov processes must be deferred appears bearable for an average student. The proof of Blackwell's theorem, and thus also for the equivalent key renewal theorem, has been omitted as it is too technical for a student presentation in the authors' opinion. The same holds for proofs regarding the necessity of the stability condition for the queues GI/PH/1 and BMAP/G/1. Only proofs for sufficiency have been included because they are easily based on the classical Foster criteria.

At the end of each chapter there will be a collection of exercises, some of them representing necessary auxiliary results to complete the proofs presented in

the lectures. Additional material is given as exercises, too, e.g. examples of computer networks or certain special queueing system.

The book is written according to the actual scripts of the lecture courses given at the University of Trier, Germany. It is intended not only to collect material which can be used for an introductory course on queueing theory, but to propose the scripts of the lectures themselves. The book contains exactly as much material as the authors (as lecturers) could present in two semesters. Thus a lecturer using this textbook does not need to choose and reassemble the material for a course from sources which must be shortened because there is no time to treat them completely. This entails saving the work of reformulating notations and checking dependencies. For a course of only one semester we propose to teach parts I and II of this book, leaving out sections 5.3 and 8.3.