Queueing Theory

Second Edition
Dedication

To my wife, Sue, with whom each day is fresh and new, a truly Markovian relationship.
A Path to Discovery

Theories of the known which are described by different ideas, may be equivalent in all their predictions and are hence scientifically indistinguishable. However, they are not psychologically identical when trying to move from that base into the unknown. For different views suggest different kinds of modifications which might be made. Therefore, a good scientist today might find it useful to have a wide range of viewpoints and mathematical expressions of the same theory available to him. This may be asking too much of one person. The new students should as a class have this. If every individual student follows the same current fashion in expressing and thinking about the generally understood areas, then the variety of hypotheses being generated to understand the still open problems is limited. Perhaps rightly so, ... BUT if the truth is in another direction, who will find it?

Richard P. Feynman

So spoke an honest man, the outstanding intuitionist of our age and a prime example of what may lie in store for anyone who dares to follow the beat of a different drum.

Julian Schwinger

From a special issue on Richard Feynman (who died on 15 February 1988) in Physics Today, February 1989. Feynman’s quote (slightly paraphrased here) was taken from his Nobel lecture in June 1965.

[Note: Feynman and Schwinger shared the Nobel prize with S. Tomonaga in 1965 for their work on quantum electrodynamics in the late forties. Working independently, and using radically different methods, they ended up with mathematically equivalent theories. Schwinger and Tomonaga were the “mainstreamers,” but everyone calculates using Feynman’s method to this day.]
# Contents

Preface to Second Edition .......................... xv
Preface to First Edition .......................... xvii

## 1 INTRODUCTION

1.1 Background .................................. 1
   1.1.1 Basic Formulas .......................... 1
   1.1.2 Markov Property ........................ 10
   1.1.3 Notation, Pronouns, Examples ............. 12
1.2 Distribution Functions Over Time ............. 13
   1.2.1 Exponential Distribution (Continuous Time, \( t \)) ....... 14
   1.2.2 Geometric Distribution (Discrete Time, \( n \)) .......... 17
1.3 Chapman-Kolmogorov Equations .................. 18
   1.3.1 Continuous Time ........................ 19
   1.3.2 Discrete Time ........................... 24
   1.3.3 Time-Dependent and Steady-State Solutions ...... 25
      1.3.3.1 Some Properties of Matrices .......... 25
      1.3.3.2 How a System Approaches Its Steady State . .. 27

## 2 M/M/1 QUEUE

2.1 Steady-State M/M/1-Type Loops .................. 33
   2.1.1 Time-Dependent Solution for \( N = 2 \) ............. 36
   2.1.2 Steady-State Solution for Any \( N \) ............. 38
   2.1.3 Open M/M/1 Queue (\( N \to \infty \)) .................. 41
   2.1.4 Buffer Overflow and Cell Loss for M/M/1/\( N \) Queues .... 44
   2.1.5 Load-Dependent Servers ........................ 50
   2.1.6 Departure Process ........................... 56
2.2 Relaxation Time for M/M/1/ \( /N \) Loops .......... 58
2.3 Other Transient Parameters ..................... 61
   2.3.1 Mean First-Passage Times for Queue Growth ......... 62
   2.3.2 \( k \)-Busy Period ........................... 66
      2.3.2.1 Mean Time of a Busy Period ............... 67
      2.3.2.2 Probability That Queue Will Reach Length \( k \) .... 69
      2.3.2.3 Maximum Queue Length During a Busy Period ....... 72
3 M. E. FUNCTIONS

3.1 Properties of a Subsystem, $S$ .......................... 77

3.1.1 Mean Time to Leave $S$ .......................... 79

3.1.2 Service Time Distribution of $S$ .................. 81

3.1.3 Properties of $B$ and $V$ .......................... 83

3.1.4 Numerical Algorithm for Evaluating $b(x)$ and $R(x)$ .. 85

3.2 Matrix Exponential Distributions .................. 87

3.2.1 Commonly Used Distributions .................. 88

3.2.1.1 Erlangian Distributions .................. 88

3.2.1.2 Hyperexponential Distributions .............. 94

3.2.2 Sums of Erlangian Functions .................. 100

3.2.3 Other Examples of ME Functions .............. 104

3.2.3.1 A 4-State Hyper-Erlangian .............. 105

3.2.3.2 A Non PHase Distribution .............. 108

3.3 Distributions With Heavy Tails .............. 111

3.3.1 Subexponential Distributions .................. 111

3.3.2 Power-Tailed (PT) Distributions .............. 114

3.3.3 What Do PT Distributions Look Like? .......... 116

3.3.4 Statistical Behavior of Large Samples ........ 117

3.3.5 The Central Limit Theorem and Stable Distributions . 121

3.3.5.1 Distributions with Finite Variance ........ 121

3.3.5.2 Distributions with Infinite Variance ........ 124

3.3.5.3 Stable Distributions and Measured Averages . 128

3.3.6 Truncated Power-Tailed (TPT) Distributions .... 129

3.3.6.1 Truncation and Range of a Distribution .... 130

3.3.6.2 An ME Representation of a TPT

Distribution ...................... 133

3.3.6.3 A TPT Distribution Where $f(0) = 0$ .......... 140

3.4 Equivalent Representations ..................... 142

3.4.1 The Canonical Minimal Representation ......... 142

3.4.2 Isometric Transformations ..................... 143

3.4.2.1 Summary .................................. 149

3.4.2.2 Hierarchy of ME Functions .............. 150

3.4.3 Examples of Equivalent Representations ......... 152

3.4.4 On the Completeness of ME Distributions ......... 156

3.4.5 Setting Up Matrix Representations ............ 157

3.5 Renewal Processes and Residual Times .......... 158

3.5.1 Matrix Representations for the pdf of $Y_n$ .... 160

3.5.2 Renewal Function and Transient Renewal Processes . 164

3.5.3 Residual Times and Delayed Intervals .......... 170

3.5.3.1 Residual Vector .......................... 170

3.5.3.2 Renewal Processes ..................... 173

3.5.4 Two Illustrations of Renewal Processes ......... 175

3.5.4.1 The Poisson Process ..................... 175

3.5.4.2 Renewal Process with $E_2$ Interdeparture

Times .......................... 176
## 4 M/G/1 QUEUE

4.1 S.S. M/ME/1//N (and M/ME/1//N) Loop .......................... 185

### 4.1.1 Balance Equations ............................................ 187

### 4.1.2 Steady-State Solution ........................................ 190

### 4.1.3 Departure and Arrival Queue-Length Probabilities .......... 195

4.2 Open M/ME/1 Queue ................................................. 200

### 4.2.1 Steady-State M/ME/1 Queue .................................. 201

### 4.2.2 System Times: Pollaczek-Khinchine Formulas ............... 204

#### 4.2.2.1 Mean Queue Length ....................................... 205

#### 4.2.2.2 Queue-Length Probabilities of M/PT/1 Queues .......... 207

#### 4.2.2.3 Throughput .................................................. 208

#### 4.2.2.4 Z-Transform ................................................ 209

### 4.2.3 System Time Distribution ..................................... 211

### 4.2.4 Buffer Overflow and Customer Loss .......................... 215

### 4.2.5 Distribution of Interdeparture Times ....................... 221

4.3 M/G/1 Queue Dependence On n ..................................... 229

### 4.3.1 Residual Time as Seen by a Random Observer ............... 229

### 4.3.2 Weighted Averages of Matrix Operators ..................... 232

### 4.3.3 Waiting Time as Seen by an Arriving Customer ............. 234

### 4.3.4 System Time of an Arriving Customer ....................... 236

4.4 Relation To Standard Solution ..................................... 237

#### 4.4.1 Exponential Moments, \( \alpha_k(s) \), and Their Meaning ... 238

#### 4.4.2 Connection to Laguerre Polynomials ....................... 239

#### 4.4.3 Connection to Standard Solution ........................... 242

#### 4.4.4 M/M/X//N Approximations to M/ME/1//N Loops ............. 245

4.5 Transient Behavior of M/ME/1 Queues ............................ 250

### 4.5.1 First-Passage Processes for Queue Growth .................. 250

#### 4.5.1.1 Conditional Probabilities for Queue Growth .......... 250

#### 4.5.1.2 Mean First-Passage Time for Queue Growth .............. 255

4.5.2 Formal Procedure for Finding System Parameters ............. 263

4.5.3 Properties of the \( k \)-Busy Period ............................ 263

#### 4.5.3.1 Conditional Probabilities for Queue Decrease .......... 264

#### 4.5.3.2 Mean First-Passage Times for Queue to Drop ............ 265

#### 4.5.3.3 Probability That Queue Will Reach Length \( n \) ........... 268

#### 4.5.3.4 Maximum Queue Length of a Busy Period ................. 276

### 4.5.4 Mean Time to Failure with Backup and Repair ............. 282

## 5 G/M/1 QUEUE

5.1 Steady-State Open ME/M/1 Queue ................................. 288

#### 5.1.1 Steady-State Probabilities of the G/M/1 Queue .......... 290

#### 5.1.2 Arrival and Departure Probabilities ....................... 295
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.1.3</td>
<td>Properties of Geometric Parameter $s$</td>
<td>297</td>
</tr>
<tr>
<td>5.1.4</td>
<td>Systems Where Interarrival Times Are Power-Tailed</td>
<td>308</td>
</tr>
<tr>
<td>5.1.5</td>
<td>Buffer Overflow Probabilities for the G/M/1 Queue</td>
<td>311</td>
</tr>
<tr>
<td>5.2</td>
<td>ME Representation of Departures</td>
<td>313</td>
</tr>
<tr>
<td>5.2.1</td>
<td>Arrival Time Distribution Conditioned by a Departure</td>
<td>313</td>
</tr>
<tr>
<td>5.2.2</td>
<td>Distribution of Interdeparture Times</td>
<td>318</td>
</tr>
<tr>
<td>5.3</td>
<td>ME/M/1//N and ME/M/1//N Queues</td>
<td>323</td>
</tr>
<tr>
<td>5.3.1</td>
<td>Steady-State Solution of the ME/M/1//N Queue</td>
<td>324</td>
</tr>
<tr>
<td>5.3.2</td>
<td>Arrival Probabilities and Customer Loss</td>
<td>328</td>
</tr>
<tr>
<td>5.4</td>
<td>Steady-State ME/M/C-Type Queues</td>
<td>331</td>
</tr>
<tr>
<td>5.4.1</td>
<td>Steady-State ME/M/X//N Loops</td>
<td>333</td>
</tr>
<tr>
<td>5.4.2</td>
<td>Steady-State ME/M/C Queue</td>
<td>338</td>
</tr>
<tr>
<td>5.4.3</td>
<td>Arrival and Departure Points</td>
<td>341</td>
</tr>
<tr>
<td>5.5</td>
<td>Transient Behavior of G/M/1 Queues</td>
<td>346</td>
</tr>
<tr>
<td>5.5.1</td>
<td>First-Passage Times for Queue Growth</td>
<td>346</td>
</tr>
<tr>
<td>5.5.2</td>
<td>The $k$-Busy Period</td>
<td>351</td>
</tr>
<tr>
<td>6</td>
<td>M/G/C-TYPE SYSTEMS</td>
<td>357</td>
</tr>
<tr>
<td>6.1</td>
<td>Introduction</td>
<td>357</td>
</tr>
<tr>
<td>6.2</td>
<td>Steady-State M/ME/2//N Loop</td>
<td>359</td>
</tr>
<tr>
<td>6.2.1</td>
<td>Definitions</td>
<td>359</td>
</tr>
<tr>
<td>6.2.2</td>
<td>Balance Equations</td>
<td>365</td>
</tr>
<tr>
<td>6.2.3</td>
<td>Solution of Probability Vectors</td>
<td>366</td>
</tr>
<tr>
<td>6.3</td>
<td>Steady-State M/G/C//N-Type Systems</td>
<td>371</td>
</tr>
<tr>
<td>6.3.1</td>
<td>Steady-State M/ME/C//N Loop</td>
<td>371</td>
</tr>
<tr>
<td>6.3.2</td>
<td>Alternate Representation of M/ME/C//N Systems</td>
<td>380</td>
</tr>
<tr>
<td>6.3.3</td>
<td>Generalized M/ME/C//N System</td>
<td>381</td>
</tr>
<tr>
<td>6.3.4</td>
<td>Relation to Jackson Networks</td>
<td>384</td>
</tr>
<tr>
<td>6.3.5</td>
<td>Time-Sharing Systems with Population Constraints</td>
<td>385</td>
</tr>
<tr>
<td>6.4</td>
<td>Open Generalized M/G/C Queue</td>
<td>390</td>
</tr>
<tr>
<td>6.5</td>
<td>Transient Generalized M/ME/C Queue</td>
<td>391</td>
</tr>
<tr>
<td>6.5.1</td>
<td>Queue Reduction at $S_1$ with No New Arrivals</td>
<td>391</td>
</tr>
<tr>
<td>6.5.2</td>
<td>Markov Renewal (Semi-Markov Departure)</td>
<td>399</td>
</tr>
<tr>
<td>6.5.3</td>
<td>A Little Bit of Up and Down, with Arrivals</td>
<td>403</td>
</tr>
<tr>
<td>6.5.3.1</td>
<td>First-Passage Processes for Queue Growth</td>
<td>403</td>
</tr>
<tr>
<td>6.5.3.2</td>
<td>First Passages for Queue Decrease</td>
<td>411</td>
</tr>
<tr>
<td>6.5.3.3</td>
<td>MTTF with Backup and Repair</td>
<td>417</td>
</tr>
<tr>
<td>6.6</td>
<td>Conclusions</td>
<td>419</td>
</tr>
<tr>
<td>7</td>
<td>G/G/1//N LOOP</td>
<td>421</td>
</tr>
<tr>
<td>7.1</td>
<td>Basis-Free Expression for $\Pr[X_1 &lt; X_2]$</td>
<td>421</td>
</tr>
<tr>
<td>7.2</td>
<td>Direct Products of Vector Spaces</td>
<td>424</td>
</tr>
<tr>
<td>7.2.1</td>
<td>Kronecker Products</td>
<td>425</td>
</tr>
<tr>
<td>7.2.2</td>
<td>$\Psi$ Projections onto Subspaces</td>
<td>427</td>
</tr>
<tr>
<td>7.3</td>
<td>Steady-State ME/ME/1//N Loop</td>
<td>430</td>
</tr>
<tr>
<td>7.3.1</td>
<td>Balance Equations</td>
<td>430</td>
</tr>
</tbody>
</table>
7.3.2 Steady-State Solution .......................... 435
7.3.3 Outline of an Efficient Algorithm ............... 441
7.3.4 An Example .................................. 444
7.4 A Modicum of Transient Behavior .................. 448

8 SEMI-MARKOV PROCESS
8.1 Introduction .................................... 454
8.1.1 Matrix Representations of Subsystems ........... 454
8.2 Markov Renewal Processes ........................ 455
8.2.1 Interdeparture Time Distributions ............... 455
8.2.2 Correlation of Departures ..................... 458
8.2.3 Laplace Transforms ........................... 462
8.3 Some Examples .................................. 462
8.3.1 Departures from Overloaded Server: Renewal Process 463
8.3.2 Markov Modulated (or Regulated) Processes .... 464
  8.3.2.1 The Underlying Generator, $\mathbf{Q}$ ............ 464
  8.3.2.2 Markov Regulated Departure Process (MRDP) ........ 467
  8.3.2.3 Markov Modulated Poisson Process (MMPP) ........... 470
  8.3.2.4 Augmented MMPP’s (AMMPP) .................... 473
  8.3.2.5 ON-OFF Models (Bursty Traffic) ............... 474
8.3.3 Merging Renewal Processes ..................... 486
8.3.4 Departures from Overloaded Multiprocessor Systems . 488
8.3.5 Departures from ME/ME/1 Queues ............... 489
  8.3.5.1 If $S_2$ Is Exponential (M/ME/1 Queues) .......... 489
  8.3.5.2 If $S_1$ Is Exponential (ME/M/1 Queues) .......... 492
  8.3.5.3 Both $S_1$ and $S_2$ Are Nonexponential .......... 496
  8.3.5.4 M/M/1//N Queues ......................... 497
8.4 MRP/M/1 Queues .................................. 498
8.4.1 Balance Equations ................................ 499
8.4.2 Some Performance Measures ..................... 502
8.4.3 The G/M/1 Queue as an Example .................. 503

9 L A Q T
9.1 Isometric Transformations ......................... 506
9.2 Linear Algebraic Formulation ...................... 508
  9.2.1 Description of a Single Server ................ 508
  9.2.2 Residual Vector and Related Properties ........... 511
9.3 Networks of Nonexponential Servers ................ 512
  9.3.1 Description of System ....................... 512
  9.3.2 Service Time Distribution .................... 516
9.4 Systems With Two Servers ........................ 519
  9.4.1 G/M/1 Queue ................................ 519
  9.4.2 Two Nonexponential Servers ..................... 520
  9.4.3 Review of Transient Behavior .................. 522
9.5 Concluding Remarks ............................. 524
Symbols 527
Abbreviations 531
Bibliography 533
Index 543
Preface to Second Edition

We have a habit in writing articles published in scientific journals to make the work as finished as possible, to cover up all the tracks, to not worry about the blind alleys or describe how we had the wrong idea first, and so on. So there isn’t any place to publish, in a dignified manner, what we actually did in order to get to do the work.


When the first edition of this book first appeared, there were few books that covered Linear Algebraic Queueing Theory (LAQT), all at a higher level. At that time I made the claim that this would become the approach of choice. Now, some 15 years later, the claim has largely been realized, particularly in problems concerning semi-Markov processes and system reliability. The prediction that because transient phenomena could now be expressed in a computationally manageable form, this subject would also become more important, seems to be slowly coming true as well. Many research papers have been published, resulting in several books containing collections of these papers, mostly on computational methods, as in for instance, [Stewart95], [Chak-Alfa97], and [LatoucheTaylor00]. The monograph by Latouche and Ramaswami (leaders in the field) [Latouche-Ram99] covers the subject well, but at a higher level. (Their title, Introduction to Matrix Analytic Methods ... could be modified to Introduction to Advanced Matrix Analytic Methods ...†.) Yet no new book at the intermediate level has emerged that takes a linear algebraic approach. That is, when possible, theorems are proven by matrix algebraic manipulation, rather than using explicit properties of matrices or probabilistic arguments. Therefore, an updated version of the original is needed.

This second edition, in addition to making many corrections and improvements, is larger by a third than the first edition. The increase in size reflects the growing recognition of the importance of processes that generate unboundedly large variances or long-range autocorrelation, as seen in CPU times, file sizes, telecommunications traffic, finance, and insurance claims. Thus an extensive amount of material has been added to Chapter 3 describing a broad set of ME functions. In particular there is an entire section on power-tail (PT), or Pareto distributions (they form a proper subset of the heavy-tailed distributions), including a section showing how they can be represented by ME distributions within the Markovian structure, even though they have infinite moments. They are then used in Chapters 4 and 5 to study queues with

†A definition of elementary, or introductory is: “that which the author understands.” Advanced means: “that which the author is not sure about,” whereas intermediate is: “that which the author figured out while writing the book.”
PT service times, and to see how PT renewal processes affect system times.

A new Chapter 8 has been added that covers Semi-Markov processes (SMP), an important topic that is used extensively in queueing models of performance from system reliability to telecommunications systems to performance of computer clusters to inventory problems in operations research. We first give a formal mathematical description of the properties of SMPs and the related Markov Renewal Processes (MRP). Several detailed examples are then presented, each with a different state-space construction. We then look at some \textit{ON-OFF} models used in modeling telecommunications traffic.

The old Chapter 8 is now Chapter 9, and includes a new section on how to deal with networks of nonexponential servers.

\section*{Acknowledgments}

When I decided to write a second edition in 2000 I realized that the original text, written in DITROFF, would have to be translated to $\LaTeX$. Lucky for me that my friend, Dr. Michael Greiner, willingly took on the task of writing the translator and overseeing its execution by Michael Schneider. Without their efforts I might still be doing the translation by hand. I must thank my friends at Technical University of Munich, Prof. Eike Jessen and Dr. Manfred Jobmann, for their longtime support and encouragement from the time I spent my sabbatical year at TUM in 1994. Manfred has carefully read the original and now the final version and has found more errata than I can afford to pay at $1 per error. Don Costello invited me to give a series of lectures and then encouraged me to write the second edition with expanded coverage of heavy-tailed distributions. Thanks to former Dean Amir Fagri and the School of Engineering at UCONN, and my department chair, Reda Ammar, for providing funding and a sabbatical so I could work on the book and hire Justin Besiglio and Robert Sheahan to produce many of the figures herein. In the last two years, Robert and Feng Zhang have generated the rest of the graphs and helped with the formatting of the book. I don’t know how I can show proper appreciation of their extraordinary efforts. Thanks to my former students, Jisung Woo, Steve Thompson, Marwan Sleiman, Sarah Tasneem, Cindy Siriwong, Gehan Verasinghe and the many students who have taken my LAQT course but whose names have slipped from my grasp, for sharing in the proofreading. Special thanks to my former students and present collaborators, Hans-Peter Schwefel, Pierre Fiorini, Ahmed Mohamed, and Imad Antonios for their invaluable input. Prof. S\o ren Asmussen provided valuable suggestions on making tighter definitions, in particular on defining heavy-tailed distributions and their subsets. I also want to thank Peter Kihl for spending so much time editing the entire book, as well as other suggestions for improving the text. Any errors that remain are mine alone. My thanks to Springer-Verlag for their offer to publish the book, and thanks most of all to my wife, Sue, for persevering through it all.

Storrs, CT, April 2008

Lester Lipsky
Preface to First Edition

“Necessity is the mother of invention” is a misleading proverb. “Necessity is the mother of temporary fixups” is much nearer to the truth. The basis of the growth of modern invention is science, and science is largely the outgrowth of intellectual curiosity.

Alfred North Whitehead

At least 50 worthwhile books on queueing theory have been written in the last 35 years. Two or three times as many books have been published in which queueing theory and Markov chains play an important part. Most of these books, even the older ones, are still useful for understanding at least some part of the subject. Why, then, should yet another book be published? The answer, simply, is that there is no book (or even collection of papers) that covers intermediate queueing theory using what I call “Linear Algebraic Queueing Theory” (LAQT). There are in fact only two books which use a linear algebraic approach, both by Marcel Neuts [Neuts81] and [Neuts89], and both of them are written for experts in the field. I waited five years for someone to write a book that could be used for a first or second course in the subject (never do anything if someone else is going to do it), but to no avail. So in 1988 I started to write it myself.

The reason that LAQT should become familiar to novices as well as to those who are already knowledgeable in intermediate and advanced queueing theory is that any problems that can be cast into a matrix-vector format can easily be adapted to make use of the high-speed parallel and vector processors available today. Also, many problems in queueing theory that traditionally are solved by unrelated mathematical techniques can now be solved in a consistent integrated fashion. This allows for better physical insight. But, most important, many system performance measures that are normally ignored because of their computational and formulational difficulties can be dealt with easily in LAQT. Some examples are: properties of the busy period, departure processes, first-passage times, residual times, distinctions between what an observer sees and what a customer sees, and compound processes in general. Each of these topics is treated here without requiring prior knowledge of the reader. This book makes the following claim. “Any problem that can be solved for exponential servers can somehow be extended to treat nonexponential servers.” Of course, it remains to be seen whether the future will vindicate this optimism.

Many decisions had to be made before this book could be written. First, who is the intended audience? There are a half a dozen disciplines that claim queueing theory as one of their “bread-and-butter” techniques. Applied probability, computer science, electrical engineering, management science, operations research, systems engineering, and even physics lay claim to various
parts of this subject as their own, each with its own terminology. Because I
dabble in all these fields, I decided to try to write a generic book that could
be understood by all. The terms used are defined in relation to customers ar-
riving at, being served by, and departing from subsystems, from the different
viewpoints of the customer and of an outside (sometimes random) observer.
The mental image one gets is of humans being served by mechanical objects,
while being observed by other human beings.

Another decision to be made was the level at which to present the ma-
terial, namely, as a first or second course in queueing theory, as a reference
book for practitioners, or as a monograph for would-be researchers in the
field. Once again, I decided to try to aim for all. There is no reason why
this material cannot be taught to mathematically mature college seniors or
new graduate students who have already had courses in linear algebra and
probability theory, but have not necessarily had any queueing theory. Unfor-
tunately this would have required that the first two chapters be expanded to
more than twice their present size without ever mentioning LAQT. There are
already many books available that give an excellent introduction to queueing
theory. Therefore I opted for either a first course, where the student already
has had some background in Markov processes and elementary queueing the-
ory, or a second course. For instance, many students in computer science and
electrical engineering take a course in applied probability covering material
such as that in Chapters 7 and 8 of Trivedi’s book [Trivedi82]. Alternately,
many courses in performance modeling (e.g., courses using [Molloy89] or
[Lazowskaetal84]) are adequate to serve as an introduction to this book.

We assume that the reader is already familiar with matrix theory. However,
except for such elementary formulas as that defining matrix multiplication,
we do not expect the student to have any particular theorem at his or her
fingertips. Therefore background information is introduced as needed. There
is no special section put aside for reviewing linear algebra. We assume the
same about the reader’s knowledge of integral and differential calculus (in
particular, Taylor’s series and l’Hospital’s rule) and elementary probability
theory. For those whose mathematics is a bit rusty, we recommend that an
elementary text in each of these areas be kept handy. But worry not; for all
the mathematical content, this is not a rigorous text. It is a “why and how
to” book. Whenever we would like a matrix to have a particular property, we
assume it is so, whether or not we can prove it.

The material is rather densely packed, so several readings and rereadings
may be necessary for the less experienced queueing theorist, particularly be-
cause there are numerous definitions in the text, and definitions do not usually
stick in one’s mind without some effort. This problem is reduced somewhat
by the book’s layout. We are inclined to introduce an idea in one chapter,
and then use it again in a subsequent section, but in a more intricate way. We
have done our best to give explicit reference to material previously discussed.
For Instructors and Practitioners

One might say that the “father” of LAQT is Victor Wallace, who in the 1960s introduced the concept of Quasi Birth-Death (QBD) processes and proved that there exists a matrix geometric solution for a large class of such systems, including the open $G/G/C$ queue [Wallace69]. His presentation, although motivated by queueing theory [Wallace72], was couched in terms of abstract Markov chains, and so was acknowledged, but was not picked up as a practical way of dealing mathematically, conceptually, or computationally with specific problems in elementary or intermediate queueing theory.

The first researcher actually to take this viewpoint in solving problems specific to queueing theory was Marcel Neuts, who in the mid-1970s introduced PHase distributions [Neuts75] and showed that they had matrix representations which could be manipulated algebraically, while operating on state vectors corresponding to the queue length probabilities (one vector for each value of $n$, the queue length). He strongly argued that a matrix formulation could more easily be handled by computers than could integration or differentiation [Neuts81]. Also, since so many problems seemed to have a recursive solution, algorithms for their numerical evaluation became straightforward. However, he and his students concentrated most of their efforts attacking hitherto unsolved problems, and thus remained too abstract to be appreciated by the practical users (as I was then) of queueing theory. It seemed as though this was just another one of the many techniques one might use to solve a small set of problems.

This researcher became interested in the subject in the late 1970s in studying the problem of what happens to a subnetwork of exponential servers when the number of customers who can be active simultaneously is restricted. My students and I soon realized that if the subsystem was restricted to one active customer, then that subsystem was equivalent to a single server with a nonexponential (Coxian, or Kendall [Kendall64], or RLT, or matrix exponential) distribution. Then, after John Carroll reduced the balance equations from second-order to first-order difference equations [Carroll79], we independently, and virtually simultaneously with Neuts, found the explicit matrix geometric solution to $M/G/1$ and $G/M/C$ queues. The two papers appeared back-to-back in the May-June 1982 issue of Operations Research [CarrollLipvdL82], [Neuts82]. I consider this to be the true beginning of LAQT, for then it became clear that many seemingly diverse problems could be solved using one technique and one viewpoint.

It is interesting to realize that the basis for LAQT was established by Erlang himself [Erlang17] when he represented a single server by a series of exponential stages, but linear algebra was not in vogue at the turn of this century, so queueing theory had to be developed entirely within the framework of what is called “modern analysis.” The “method of stages” is really a part of LAQT, distorted so it could fit into the classical view, whereas D. R. Cox’s work in the 1950s [Cox55], showing in effect that “every pdf can be approximated arbitrarily closely by a function whose Laplace transform can be written as the ratio of two polynomials (RLT functions)” is really the basis
for claiming that there exists a linear algebraic formulation of every problem which can be formulated otherwise.

You might question whether LAQT really is a peer to the standard variety of queueing theory. Well, for decades now, it has been standard technique in various areas of applied mathematics to replace differential operators on a solution function by an equivalent linear operator on a vector in Hilbert space. In fact, the pair of representations of quantum theory, Werner Heisenberg’s matrix mechanics and Erwin Schrödinger’s wave mechanics, is the prime example of this duality. The proof by John von Neumann that they are mathematically equivalent is closely related to Cox’s completeness statement in extending A. K. Erlang’s method of stages to include all functions with rational Laplace transforms [Cox55]. Fortunately for physics, linear algebra was a known quantity by the 1920s, so the two viewpoints grew together and have become so intertwined that the typical quantum practitioner switches from one to the other and back again with little difficulty. A similar statement can be made about linear control theory. Both of those disciplines deal with functions of complex variables, even though what is actually observed must be real. If physicists can talk about the charm of quarks, which can never be seen outside their nuclear home, and electrical engineers can have imaginary currents, surely our customers should be allowed to travel with negative probabilities and complex service times from one phase to another, as long as they remain inside one subsystem or another, and as long as all observable entities are real.

The reader should avoid mapping this material onto already familiar techniques, at least until Chapter 4 has been covered. By then you will see the power and elegance of this methodology, as well as its usefulness, and be able to “switch back and forth without difficulty.” Furthermore, because most solutions are in terms of matrix operations rather than integrals, or roots of equations, highly efficient algorithms for both single and parallel computer systems can easily be written. There are several mathematical tool kits readily available (e.g., MATLAB, Mathematica, Maple) that execute matrix equations directly.

Organization

The book is laid out by chapter in order of increasing complexity of structure. There is more than enough material for a two-semester course, but a one-semester first course or a one-semester second course can easily be fashioned.

In Chapter 1 we make a quick survey of those topics normally connected to Markov chains. Chapter 2 starts out as a continuation of Chapter 1 by using the Chapman-Kolmogorov equations to set up the M/M/1 queue. But we soon switch to the simpler and intuitively more satisfying view associated with steady-state transition diagrams. Every queueing system is made up of two subsystems, each of which contains one exponential server. In Chapter 3 we show that by adding structure to a subsystem we give it a nonexponential (called Matrix Exponential, ME) service time distribution. In Chapter 4 we
combine the ideas of the two previous chapters to study the M/G/1 queue (i.e., one nonexponential and one exponential subsystem). As long as our system is closed (finite population of customers), there is no difference between an M/G/1//N loop and a G/M/1//N loop. But if the population is increased unboundedly, one or the other server will saturate. So, if the nonexponential server is the faster one, we have the open M/G/1 queue as given in Chapter 4. However, in Chapter 5 we assume that the exponential server is faster, and derive the properties of an open G/M/1 queue.

In Chapter 6 two or more customers can independently be active at once in one subsystem, the M/G/C system. This increases the complexity of the mathematics required, as well as the computational complexity and sizes of matrices. But it also enormously increases the range of problems that can be solved, the so-called “generalized M/G/C systems.” In Chapter 7 we revert to one active customer per subsystem, but now both subsystems have structure, and we are dealing with a G/G/1//N loop. This leads to a different increase in complexity, requiring a direct product of vector spaces, which we must first discuss before actually finding the steady-state solution.

Finally, in Chapter 8 we try to give a linear algebraic formulation that does not depend upon a physical interpretation of individual states. As such, it acts as a review of the book.

The chapters are all structured in more or less the same way, with obvious deviations because of the material. First we find the closed steady-state solution. Then we “open” the loop by increasing the customer population unboundedly. Then we look at certain specialized topics (e.g., load-dependent servers, renewal theory, comparison with other methods). Finally we explore the transient behavior of the appropriate queue.

A one-semester first course would cover Chapter 1 and the steady-state parts of Chapters 2, 3, 4, and 5. Depending on the background of the students, the instructor might add some descriptive material to Chapters 1 and 2.

Assuming that students have already had a course in queueing theory, but not one that covered LAQT, a one-semester second course would skim through Chapter 1 and the first part of Chapter 2. But then Section 2.3 must be covered in earnest, as must the first part of Chapter 3. Except for the material on residual times, which must be covered, Section 3.5 can be omitted. Most of Chapters 4 and 5 should be covered, but the instructor can skip Chapter 6 if desired and go directly to Chapter 7. However, Chapter 6 is potentially of great practical importance, therefore the instructor may prefer to skip Chapter 7 instead. Chapter 9 can be put in or left out, as per taste.

A two-semester course can be given that combines the two one-semester courses in the order just described, or one can go sequentially from beginning to end, skipping those topics which seem inappropriate. However, one cannot study Section 6.5, for example, without first covering the related material in Chapters 2, 4, and 5.
Acknowledgments

I would like to thank Professor Howard Sholl and BECAT of the University of Connecticut for continued support in the technical creation of this book. In particular, Anthony Guzzi has rewritten DITROFF\textsuperscript{†} so it actually does what it is supposed to do (at least on my workstation). His devotion to my needs has been beyond the call. Also, I thank John Marshall and Sue Zajac* for keeping the system up (most of the time) and the secretaries (Jean, Sue, Ruth, Sandi, and Sherry) for keeping me up (most of the time). To my former students, now collaborators, Apple van de Liefvoort (University of Missouri-KC), Aby Tehranipour (Eastern Michigan University), and Yiping Ding (now at BGS Systems, Inc.), I give thanks for technical advice in the various chapters where they are experts. Thanks to Seva nanda Adari, Jinzhu (Jim) Chen, and Houzhong Yan for reading the first draft and pointing out how ideas could be made clearer. I thank the students who were in my class in the fall of 1990 (Somnath Deb, Sharad Garg, Rudi Hackenberg, Chengdong Lu, Jim Moriarty, Carolyn Pe, and Cien Xu), who used the second version of this book and searched for errors of content. Siddhartha Roy and Dilip Tagare meticulously went through the final draft, searching for errors of all kinds (and they found many). Dilip and Ed Bigos were also responsible for generating most of the graphs. To Professors Joseph Macek (University of Tennessee); George Nagy (Rensselaer Polytechnic Institute); Don Costello and Sharad Seth (University of Nebraska); Don Towsley (University of Massachusetts); Victor Wallace (University of Kansas); and Arnie Russek (my thesis advisor), Jim Galligan, and Krishna Pattipati (all of the University of Connecticut), thanks for useful critical comments. I must also thank Macmillan Publishing Company’s Ed Maura for taking the initiative in inviting me to do the book, John Griffin who oversaw the project from cover to cover, the unknown proofreader who went through the text with the devotion of a mother combing her daughter’s hair, and especially Leo Malek who was determined that this would be a good-looking book. Thanks to Erikson/Dillon Art services for creating such fine figures and page layout, and finally, Janet Pecorelli and the American Mathematical Society’s printing service in Providence, RI, for producing such high-quality galley proofs at short notice.

Storrs, CT, November 1991

Lester Lipsky

\textsuperscript{†}DITROFF is a registered trademark of AT&T.

*As of May 4, 1996, now known as Sue Marie Lipsky.