

Harold J. Kushner

# Heavy Traffic Analysis of Controlled Queueing and Communication Networks

With 50 Illustrations



Springer

Harold J. Kushner  
Division of Applied Mathematics  
Brown University  
Providence, RI 02912, USA  
hjk@dam.brown.edu

*Managing Editors*

I. Karatzas  
Departments of Mathematics and Statistics  
Columbia University  
New York, NY 10027, USA

M. Yor  
CNRS, Laboratoire de Probabilités  
Université Pierre et Marie Curie  
4, Place Jussieu, Tour 56  
F-75252 Paris Cedex 05, France

---

Mathematics Subject Classification (2000): 60K25, 90Bxx, 93C70, 93E02, 93E20

---

Library of Congress Cataloging-in-Publication Data

Kushner, Harold J. (Harold Joseph), 1933–  
Heavy traffic analysis of controlled queueing and communication  
networks / Harold J. Kushner.  
p. cm. — (Applications of mathematics ; 47)  
Includes bibliographical references and index.

1. Queueing theory. I. Title. II. Series.

QA274.8 .K87 2001  
519.8'2—dc21

2001020202

Printed on acid-free paper.

© 2001 Springer Science + Business Media New York  
Originally published by Springer - Verlag New York, Inc. in 2001.

All rights reserved. This work may not be translated or copied in whole or in part without the written permission of the publisher (Springer-Verlag New York, Inc., 175 Fifth Avenue, New York, NY 10010, USA), except for brief excerpts in connection with reviews or scholarly analysis. Use in connection with any form of information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed is forbidden. The use of general descriptive names, trade names, trademarks, etc., in this publication, even if the former are not especially identified, is not to be taken as a sign that such names, as understood by the Trade Marks and Merchandise Marks Act, may accordingly be used freely by anyone.

Production managed by Michael Koy; manufacturing supervised by Erica Bresler.  
Photocomposed copy prepared from the author's  $\text{T}_{\text{E}}\text{X}$  files.

9 8 7 6 5 4 3 2 1

SPIN 10797544

ISBN 978-1-4612-6541-2 ISBN 978-1-4613-0005-2 (eBook)  
DOI 10.1007/978-1-4613-0005-2

Springer-Verlag New York Berlin Heidelberg  
A member of BertelsmannSpringer Science+Business Media GmbH

# Contents

<b>Preface</b>	<b>vii</b>
<b>1 Introduction: Models and Applications</b>	<b>1</b>
1.1 A Single Queue: Heavy Traffic Modeling . . . . .	4
1.1.1 Simple One Dimensional Models. . . . .	4
1.1.2 The Workload Form . . . . .	13
1.2 Networks . . . . .	14
1.2.1 Simple Networks . . . . .	15
1.2.2 Closed and Open/Closed Networks . . . . .	22
1.3 Server Scheduling and Assignment Problems . . . . .	23
1.3.1 Multiple Input Classes and Controlled Polling . . . . .	24
1.4 Communication and Computer Networks: The Multiplexer . . . . .	34
1.5 Controlled Admission in a Multiservice System: Formulation . . . . .	39
<b>2 Martingales and Weak Convergence</b>	<b>45</b>
2.1 Martingales and the Wiener Process . . . . .	46
2.1.1 Martingales . . . . .	46
2.1.2 The Wiener Process . . . . .	49
2.2 Stochastic Integrals . . . . .	51
2.2.1 Definition and Properties . . . . .	51
2.2.2 Itô's Lemma . . . . .	54
2.3 Poisson Measures . . . . .	56

2.3.1	The Poisson Process and Poisson Random Measures . . . . .	56
2.3.2	Martingale Decomposition of a Jump Process . . . . .	58
2.4	The Doob–Meyer Process . . . . .	60
2.4.1	Introduction and Itô’s Formula . . . . .	60
2.4.2	The Doob–Meyer Process for a Jump Process . . . . .	62
2.5	Weak Convergence . . . . .	63
2.5.1	Motivation and Examples . . . . .	63
2.5.2	Basic Theorems of Weak Convergence . . . . .	65
2.5.3	The Function Space $D(\mathbb{R}^k; 0, \infty)$ . . . . .	68
2.5.4	The Function Spaces $D(S; 0, T)$ and $D(S; 0, \infty)$ . . . . .	69
2.5.5	Truncated Processes . . . . .	70
2.6	The Time Transformation Method . . . . .	72
2.7	Measure-Valued Processes . . . . .	77
2.8	Tightness and Convergence to Wiener Processes . . . . .	79
2.8.1	Criteria for Weak Convergence to Martingales . . . . .	79
2.8.2	Tightness and Wiener Process Limits . . . . .	82
<b>3</b>	<b>Stochastic Differential Equations: Controlled and Uncontrolled</b>	<b>89</b>
3.1	Stochastic Differential Equations and Diffusion Processes . . . . .	91
3.1.1	Definitions: Uncontrolled SDEs . . . . .	91
3.1.2	Controlled Diffusions . . . . .	92
3.1.3	Construction of a Strong Solution . . . . .	94
3.1.4	The Girsanov Transformation . . . . .	96
3.2	Relaxed Controls . . . . .	97
3.2.1	Existence of a Deterministic Optimal Control . . . . .	98
3.2.2	The Topology on the Space of Relaxed Controls . . . . .	101
3.2.3	Stochastic Relaxed Controls . . . . .	102
3.3	Impulsive and Singular Control Problems . . . . .	105
3.4	Reflected Diffusions . . . . .	108
3.4.1	Introduction and Examples . . . . .	108
3.4.2	Reflected Brownian Motion and Related Models . . . . .	111
3.5	The Skorohod Problem . . . . .	118
3.5.1	Definitions and Lipschitz Conditions . . . . .	118
3.5.2	Reflected Stochastic Differential Equations . . . . .	124
3.5.3	Impulsive and Singular Control Problems . . . . .	129
3.6	Tightness for the Skorohod Problem . . . . .	130
3.7	Reflected Jump–Diffusion Processes . . . . .	134
3.8	Approximations of Optimal Controls . . . . .	136
<b>4</b>	<b>Invariant Measures and the Ergodic Problem</b>	<b>141</b>
4.1	Convergence to Invariant Measures . . . . .	142
4.2	Properties of Solutions . . . . .	145

4.3	The Process Model . . . . .	148
4.4	Optimal Feedback Controls . . . . .	155
4.5	A Maximum Principle . . . . .	161
4.6	Optimality Over All Admissible Controls . . . . .	166
4.7	Functional Occupation Measures . . . . .	168
4.8	Approximating Controls . . . . .	176
<b>5</b>	<b>The Single-Processor Problem</b>	<b>179</b>
5.1	A Canonical Single-Server Problem . . . . .	182
5.1.1	The Basic Model . . . . .	182
5.1.2	Multiple Arrival Streams of Different Rates . . . . .	190
5.1.3	The Fictitious Services Model . . . . .	193
5.2	Tightness of the Reflection Terms . . . . .	195
5.3	The Workload Processes . . . . .	200
5.3.1	The Workload Equation for the Basic Model . . . . .	200
5.3.2	Several Input Streams: Different Work Requirements . . . . .	202
5.3.3	Asymptotic Relations Between the Queue Size and Workload . . . . .	203
5.4	Extensions . . . . .	207
5.4.1	Batch Arrivals and Services . . . . .	207
5.4.2	Many Servers . . . . .	210
5.5	Correlated Processes: Bursty Arrivals . . . . .	212
5.6	Processor Interruptions, Priorities, and Vacations . . . . .	217
5.6.1	Short and Frequent Vacations . . . . .	217
5.6.2	Priorities . . . . .	220
5.6.3	Short and Infrequent Vacations . . . . .	222
5.6.4	Long but Infrequent Vacations . . . . .	223
5.7	An Alternative Scaling: Fast Arrivals and Services . . . . .	227
<b>6</b>	<b>Uncontrolled Networks</b>	<b>229</b>
6.1	The Heavy Traffic Limit Theorem . . . . .	231
6.1.1	Independent Routing . . . . .	231
6.1.2	Martingale Difference Intervals . . . . .	237
6.1.3	Correlated Routing . . . . .	242
6.1.4	A General Skorohod Problem Model . . . . .	244
6.2	Extensions . . . . .	245
6.2.1	Closed Networks . . . . .	245
6.2.2	Nonbottleneck Processors . . . . .	248
6.2.3	Batch Arrivals, Multiple Servers, Multicast and Fork-Join Queues . . . . .	249
6.3	Blocking . . . . .	253
6.4	The Workload Formulation and Priorities . . . . .	257
6.4.1	Queued Workload . . . . .	257
6.4.2	A Multiclass Feedforward System. . . . .	258

6.5	Fluid Scaling and Limits . . . . .	264
6.5.1	A Single-Class Network . . . . .	264
6.6	An Alternative Scaling: Fast Arrivals and Services . . . . .	267
<b>7</b>	<b>Uncontrolled Networks, Continued</b>	<b>269</b>
7.1	A Manufacturing System . . . . .	270
7.1.1	The Basic Model . . . . .	270
7.1.2	Multiple Types of Final Products I . . . . .	272
7.1.3	Multiple Types of Final Products: II . . . . .	277
7.2	Shared Buffers . . . . .	279
7.3	Process Interruptions and Vacations . . . . .	282
7.3.1	Short Vacations . . . . .	283
7.3.2	Long But Infrequent Interruptions . . . . .	285
7.4	Correlated Service Times . . . . .	290
7.4.1	A Simple Feedforward System . . . . .	291
7.4.2	Random Routing in a Feedforward System . . . . .	292
7.4.3	A Feedback System with Correlated Service Times . . . . .	294
7.5	Processor Sharing . . . . .	298
<b>8</b>	<b>State Dependence</b>	<b>305</b>
8.1	Marginal State Dependence . . . . .	307
8.2	Poisson-Type Input and Service Process . . . . .	316
8.2.1	The Basic Model . . . . .	316
8.2.2	A Generalization: Relationships with Section 1 . . . . .	321
8.2.3	Vacations . . . . .	322
8.3	Self-Service with Fast Arrivals . . . . .	323
8.4	Discontinuous Dynamics . . . . .	326
8.5	An Application to the Multiplexer-Buffer System . . . . .	329
8.5.1	Introduction and Problem Description . . . . .	329
8.5.2	Convergence . . . . .	332
8.6	Balking, Withdrawing, and Retrials . . . . .	337
<b>9</b>	<b>Bounded Controls</b>	<b>341</b>
9.1	Discounted Cost . . . . .	343
9.1.1	A Canonical Model and Assumptions . . . . .	343
9.1.2	Controls Appearing Linearly . . . . .	350
9.1.3	Controls Appearing Nonlinearly . . . . .	352
9.1.4	Extensions . . . . .	356
9.2	The Ergodic Cost Problem . . . . .	359
9.3	Examples . . . . .	363
9.3.1	Service Rate and Assignment Controls . . . . .	363
9.3.2	The Multiplexer Problem: Convergence . . . . .	364
9.4	Data for the Multiplexer Problem . . . . .	368

<b>10 Singular Controls</b>	<b>375</b>
10.1 A Canonical Model . . . . .	377
10.2 Vacations . . . . .	384
10.3 Controlled Admission . . . . .	391
10.3.1 Introduction: The Basic System . . . . .	391
10.3.2 Upper Limit to the Bandwidth for the BE-Sharing Customers . . . . .	396
10.4 Rerouting and Singular Control . . . . .	398
10.4.1 Rerouting with Penalty . . . . .	398
10.4.2 Rerouting Only When a Queue Is Full . . . . .	401
<b>11 Polling and Control of Polling</b>	<b>405</b>
11.1 An Averaging Principle for the Individual Queues . . . . .	408
11.1.1 Continuous Switching Policies . . . . .	408
11.1.2 Lower Thresholds . . . . .	417
11.2 Rate of Switching and Nonzero Switching Time . . . . .	420
11.2.1 Switching Rate . . . . .	420
11.2.2 Nonzero Switching Time . . . . .	422
11.3 Gated Polling . . . . .	423
11.4 The Control Problem: Switching Cost and Time . . . . .	424
11.4.1 Switching Cost . . . . .	424
11.4.2 Nonzero Switching Time . . . . .	427
11.5 Controlled Polling with Interruptions . . . . .	428
11.6 Weak Convergence . . . . .	431
11.7 The Limit Control Problem . . . . .	434
11.8 Extensions and Comments . . . . .	439
11.8.1 Minimizing the Total Expected Workload. . . . .	439
11.8.2 No Vacations: The Asymptotic Optimality of the $c\mu$ -rule . . . . .	441
11.9 Relaxed Poisson Measures . . . . .	442
11.9.1 A Difficulty with Controlled Jumps . . . . .	442
11.9.2 The Relaxed Poisson Measure . . . . .	446
<b>12 Assignment and Scheduling:</b>	
<b>Many Classes and Processors</b>	<b>453</b>
12.1 Many Input Classes and a Single Processor . . . . .	455
12.2 A One-Stage, Two-Class and Two-Processor Problem . . . . .	459
12.3 Many Input Classes and Servers: One Stage . . . . .	466
12.4 Feedforward Networks . . . . .	476
<b>References</b>	<b>485</b>
<b>Symbol Index</b>	<b>504</b>
<b>Index</b>	<b>508</b>