

Contents

<i>Notation</i>	xi
<i>Introduction</i>	xiii
Part I The Single Queue	1
1 Queues and Their Simulations, Birth and Death Queues	3
1.1 The Single Queue	3
1.2 Simulation of a Queue	6
1.3 Birth and Death Queues	7
1.4 Historical Notes, Sources and Extensions	13
Exercises	13
2 The M/G/1 Queue	16
2.1 Little’s Law	16
2.2 Work Conservation	21
2.3 Some Renewal Theory	23
2.4 Length Biasing	25
2.5 Stationary Point Processes and Palm Measure	27
2.6 PASTA — Poisson Arrivals See Time Averages	28
2.7 M/G/1 Average Waiting Times	31
2.8 Busy Periods	33
2.9 Supplementary Material: Embedded Markov Chains	34
2.10 Sources	39
Exercises	39
3 Scheduling	42
3.1 Batch Scheduling	43
3.2 Value of Information	45
3.3 Stream Scheduling	46
3.4 Optimality of SRPT	47

vi	<i>Contents</i>	
3.5	Mean and Variance for Non-Preemptive, Non-Predictive Scheduling	49
3.6	M/G/1 Priority Scheduling	50
3.7	M/G/1 under Preemptive LCFS and under PS	52
3.8	Smith's Rule, Klimov's Control of M/G/1, and the $c\mu$ Rule.	53
3.9	Bandit Processes and the Gittins Index	53
3.10	Sources	55
	Exercises	56
	Part II Approximations of the Single Queue	59
4	The G/G/1 Queue	61
4.1	Loynes Construction for Stability of G/G/1 Queues	61
4.2	G/G/1 and the Random Walk	64
4.3	Bounds for G/G/1 Waiting Times	67
4.4	Sources	68
	Exercises	69
5	The Basic Probability Functional Limit Theorems	71
5.1	Convergence of Stochastic Processes	71
5.2	Functional Limit Theorems for Random Walks	73
5.3	Functional Limit Theorems for Renewal Processes	79
5.4	Sources	81
	Exercises	81
6	Scaling of G/G/1 and G/G/∞	83
6.1	Primitives and Dynamics of G/G/1	83
6.2	Scaling Time, Space, and Initial Conditions	85
6.3	Skorohod's Reflection Mapping for the Single-Server Queue	86
6.4	Fluid Scaling and Fluid Limits of G/G/1	88
6.5	Diffusion Scale Limits for Underloaded and Overloaded G/G/1	90
6.6	Diffusion Scaling of the G/G/1 Queue in Heavy Traffic	93
6.7	Approximations of the G/G/1 Queue	95
6.8	Another Approach: Strong Approximation	97
6.9	Diffusion Limits for G/G/ ∞ in Heavy Traffic	98
6.10	Sources	100
	Exercises	101
7	Diffusions and Brownian Processes	102
7.1	Diffusion Processes, Brownian Motion, and Reflected Brownian Motion	102
7.2	Diffusions and Birth and Death Processes	105

<i>Contents</i>		vii
7.3	Approximation of the G/G/1 Queue	107
7.4	Two Sided Regulation of Brownian Motion	107
7.5	Optimal Control of a Manufacturing System	109
7.6	Oblique Reflection and Multivariate RBM	111
7.7	Supplement: Calculations for Brownian Motion and Derived Diffusions	112
7.8	Sources	120
	Exercises	120
Part III Queueing Networks		123
8	Product-Form Queueing Networks	125
8.1	The Classic Jackson Network	125
8.2	Reversibility and Detailed Balance Equations	127
8.3	Partial Balance and Stationary Distribution of Jackson Networks	129
8.4	Time Reversal and the Arrival Theorem	131
8.5	Sojourn Times in Jackson Networks	133
8.6	Closed Jackson Networks	134
8.7	Kelly Networks	135
8.8	Symmetric Queues and Insensitivity	138
8.9	Multi-Class Kelly-Type Networks	140
8.10	Sources	140
	Exercises	141
9	Generalized Jackson Networks	143
9.1	Primitives and Dynamics	143
9.2	Traffic Equations and Stability	144
9.3	Centering and Skorohod Oblique Reflection	146
9.4	Fluid Limits	147
9.5	Diffusion Limits	149
9.6	Stability of Generalized Jackson Networks	153
9.7	Sources	154
	Exercises	154
Part IV Fluid Models of Multi-Class Queueing Networks		157
10	Multi-Class Queueing Networks, Instability, and Markov Rep- resentations	159
10.1	Multi-Class Queueing Networks	159
10.2	Some Unstable MCQN	161

10.3	Policy Driven Markovian Description of MCQN	165
10.4	Stability and Ergodicity of Uncountable Markov Processes	168
10.5	Sources	173
	Exercises	174
11	Stability of MCQN via Fluid Limits	175
11.1	System Equations, Fluid Limit Model, Fluid Equations and Fluid Solutions	175
11.2	Stability via Fluid Models	181
11.3	Some Fluid Stability Proofs	185
11.4	Piecewise Linear Lyapunov Functions and Global Stability	191
11.5	Sources	194
	Exercises	194
12	Processing Networks and Maximum Pressure Policies	197
12.1	A More General Processing System	198
12.2	Maximum Pressure Policies	201
12.3	Rate Stability Proof via the Fluid Model	204
12.4	Further Stability Results under Maximum Pressure Policy	206
12.5	Applications	209
12.6	Sources	215
	Exercises	215
13	Processing Networks with Infinite Virtual Queues	219
13.1	Introduction, Motivation, and Dynamics	219
13.2	Some Countable Markovian MCQN-IVQ	221
13.3	Fluid Models of Processing Networks with IVQs	226
13.4	Static Production Planning Problem and Maximum Pressure Policies	228
13.5	An Illustrative Example: The Push-Pull System	230
13.6	Sources	233
	Exercises	234
14	Optimal Control of Transient Networks	236
14.1	The Semiconductor Wafer Fabrication Industry	237
14.2	The Finite Horizon Problem Formulation	238
14.3	Formulation of the Fluid Optimization Problem	240
14.4	Brief Summary of Properties of SCLP and its Solution	242
14.5	Examination of the Optimal Fluid Solution	245
14.6	Modeling Deviations from the Fluid as Networks with IVQs	247
14.7	Asymptotic Optimality of the Two-Phase Procedure	250
14.8	An Illustrative Example	252
14.9	Implementation and Model Predictive Control	253

<i>Contents</i>	<i>ix</i>
14.10 Sources	255
Exercises	255
Part V Diffusion Scaled Balanced Heavy Traffic	257
15 Join the Shortest Queue in Parallel Servers	259
15.1 Exact Analysis of Markovian Join the Shortest Queue	259
15.2 Variability and Resource Pooling	262
15.3 Diffusion Approximation and State Space Collapse	264
15.4 Threshold Policies for Routing to Parallel Servers	266
15.5 A Note about Diffusion Limits for MCQN	266
15.6 Sources	267
Exercises	268
16 Control in Balanced Heavy Traffic	270
16.1 MCQN in Balanced Heavy Traffic	270
16.2 Brownian Control Problems	273
16.3 The Criss-Cross Network	276
16.4 Sequencing for a Two-Station Closed Queueing Network	281
16.5 Admission Control and Sequencing for a Two-Station MCQN	287
16.6 Admission Control and Sequencing in Multi-Station MCQN	295
16.7 Asymptotic Optimality of MCQN Controls	300
16.8 Sources	301
Exercises	302
17 MCQN with Discretionary Routing	305
17.1 The General Balanced Heavy Traffic Control Problem	305
17.2 A Simple Network with Routing and Sequencing	306
17.3 The Network of Laws and Louth	309
17.4 Further Examples of Pathwise Minimization	312
17.5 Routing and Sequencing with General Cuts	315
17.6 Sources	318
Exercises	319
Part VI Many-Server Systems	321
18 Infinite Servers Revisited	323
18.1 Sequential Empirical Processes and the Kiefer Process	323
18.2 Stochastic System Equations for Infinite Servers	325
18.3 Fluid Approximation of Infinite Server Queues	326
18.4 Diffusion Scale Approximation of Infinite Server Queues	327

x	<i>Contents</i>	
18.5	Sources	329
	Exercises	329
19	Asymptotics under Halfin–Whitt Regime	330
19.1	Three Heavy Traffic Limits	330
19.2	M/M/s in Halfin–Whitt Regime	331
19.3	Fluid and Diffusion Limits for G/G/s under Halfin–Whitt Regime	336
19.4	sources	345
	Exercises	345
20	Many Servers with Abandonment	346
20.1	Fluid Approximation of G/G/n+G	346
20.2	The M/M/n+M System under Halfin–Whitt Regime	350
20.3	The M/M/n+G System	352
20.4	Sources	358
	Exercises	358
21	Load Balancing in the Supermarket Model	360
21.1	Join Shortest of d Policy	361
21.2	Join the Shortest Queue under Halfin–Whitt Regime	373
21.3	Approaching JSQ: Shortest of $d(n)$ and Join Idle Queue	381
21.4	Sources	382
	Exercises	383
22	Parallel Servers with Skill-Based Routing	385
22.1	Parallel Skill-Based Service under FCFS	386
22.2	Infinite Bipartite Matching under FCFS	390
22.3	A FCFS Ride-Sharing Model	401
22.4	A Design Heuristic for General Parallel Skill-Based Service	404
22.5	Queue and Idleness Ratio Routing	406
22.6	Sources	409
	Exercises	409
	<i>References</i>	413
	<i>Index</i>	427