

Contents

Preface *xvii*

Acronyms *xix*

Part I Models for Service Systems 1

1 Introduction 3

1.1 Network Traffic Engineering: What, Why, How 3

1.2 The Art of Modeling 8

1.3 An Example: Delay Equalization 13

1.3.1 Model Setting 14

1.3.2 Analysis by Equations 15

1.3.3 Analysis by Simulation 19

1.3.4 Takeaways 21

1.4 Outline of the Book 21

1.4.1 Plan 21

1.4.2 Use 25

1.4.3 Notation 27

1.5 Further Readings 29

Problems 30

2 Service Systems and Queues 33

2.1 Service System Structure 33

2.2 Arrival and Service Processes 35

2.3 The Queue as a Service System Model 38

2.4 Queues in Equilibrium 40

2.4.1 Queues and Stationary Processes 40

2.4.2 Little's Law 45

2.5 Palm's Distributions for a Queue 49

2.6	The Traffic Process	53
2.7	Performance Metrics	56
2.7.1	Throughput	56
2.7.2	Utilization	59
2.7.3	Loss	59
2.7.4	Delay	61
2.7.5	Age of Information	62
	Summary and Takeaways	63
	Problems	65
3	Stochastic Models for Network Traffic	71
3.1	Introduction	71
3.2	The Poisson Process	72
3.2.1	Light versus Heavy Tails	78
3.2.2	Inhomogeneous Poisson Process	79
3.2.3	Poisson Process in Multidimensional Spaces	84
3.2.3.1	Displacement	89
3.2.3.2	Mapping	89
3.2.3.3	Thinning	90
3.2.3.4	Distances	91
3.2.3.5	Sums and Products on Point Processes	92
3.2.3.6	Hard Core Processes	94
3.2.4	Testing for Poisson	96
3.3	The Markovian Arrival Process	100
3.4	Renewal Processes	103
3.4.1	Residual Inter-Event Time and Renewal Paradox	108
3.4.2	Superposition of Renewal Processes	110
3.4.3	Alternating Renewal Processes	111
3.4.4	Renewal Reward Processes	113
3.5	Birth-Death Processes	115
3.6	Branching Processes	121
	Summary and Takeaways	125
	Problems	126
Part II Queues 131		
4	Single-Server Queues	133
4.1	Introduction and Notation	133
4.2	The Embedded Markov Chain Analysis of the $M/G/1$ Queue	134
4.2.1	Queue Length	136

4.2.2	Waiting Time	141
4.2.3	Busy Period and Idle Time	145
4.2.4	Remaining Service Time	148
4.2.5	Output Process	149
4.2.6	Evaluation of the Probabilities $\{a_k\}_{k \in \mathbb{Z}}$	151
4.3	The $M/G/1/K$ Queue	152
4.3.1	Exact Solution	153
4.3.2	Asymptotic Approximation for Large K	157
4.4	Numerical Evaluation of the Queue Length PDF	166
4.5	A Special Case: the $M/M/1$ Queue	168
4.6	Optimization of a Single-Server Queue	170
4.6.1	Maximization of Net Profit	171
4.6.2	Minimization of Age of Information	174
4.6.2.1	General Expression of the Average Age of Information	175
4.6.2.2	Minimization of the Age of Information for an $M/M/1$ Model	177
4.7	The $G/M/1$ Queue	178
4.8	Matrix-Geometric Queues	185
4.8.1	Quasi Birth-Death (QBD) Processes	186
4.8.2	$M/G/1$ and $G/M/1$ Structured Processes	188
4.9	A General Result on Single-Server Queues	192
	Summary and Takeaways	194
	Problems	195
5	Multi-Server Queues	199
5.1	Introduction	199
5.2	The Erlang Loss System	201
5.2.1	Insensitivity Property of the Erlang Loss System	211
5.2.2	A Finite Population Model	213
5.2.3	Non-Poisson Input Traffic	214
5.2.3.1	Wilkinson's Method	217
5.2.3.2	Fredericks' Method	218
5.2.4	Multi-Class Erlang Loss System	221
5.3	Application of the Erlang Loss Model to Cellular Radio Access Network	224
5.3.1	Cell Dimensioning under Quality of Service Constraints	225
5.3.2	Number of Handoffs in a Connection Lifetime	230
5.3.3	Blocking in a Cell with User Mobility	232
5.3.4	Trade-off between Location Updating and Paging	234
5.3.5	Dimensioning of a Cell with Two Service Classes	236
5.4	The $M/M/m$ Queue	238
5.4.1	Finite Queue Size Model	243

5.4.2	Resource Sharing versus Isolation	244
5.5	Infinite Server Queues	247
5.5.1	Analysis of Message Propagation in a Linear Network	252
	Summary and Takeaways	257
	Problems	258
6	Priorities and Scheduling	265
6.1	Introduction	265
6.2	Conservation Law	268
6.3	<i>M/G/1</i> Priority Queueing	272
6.3.1	Non-FCFS Queueing Disciplines	273
6.3.2	Head-of-Line (HOL) Priorities	276
6.3.3	Preempt-Resume Priorities	283
6.3.4	Shortest Job First	284
6.3.5	Shortest Remaining Processing Time	286
6.3.6	The μ C Rule	288
6.4	Processor Sharing	289
6.4.1	The <i>M/G/1</i> Processor Sharing Model	290
6.4.2	Generalized Processor Sharing	293
6.4.3	Weighted Fair Queueing	298
6.4.4	Credit-Based Scheduling	302
6.4.5	Deficit Round Robin Scheduling	306
6.4.6	Least Attained Service Scheduling	308
6.5	Miscellaneous Scheduling	312
6.5.1	Scheduling on a Radio Link	312
6.5.1.1	Proportional Fairness	312
6.5.1.2	Multi-rate Orthogonal Multiplexing	313
6.5.2	Job Dispatching	318
6.6	Optimal Scheduling	324
6.6.1	Anticipative Systems	325
6.6.2	Server-Sharing, Nonanticipative Systems	325
6.6.3	Non-Server-Sharing, Nonanticipative Systems	326
	Summary and Takeaways	327
	Problems	327
7	Queueing Networks	331
7.1	Structure of a Queueing Network and Notation	331
7.2	Open Queueing Networks	332
7.2.1	Optimization of Network Capacities	345
7.2.2	Optimal Routing	347
7.2.3	Braess Paradox	350

7.3	Closed Queueing Networks	355
7.3.1	Arrivals See Time Averages (ASTA)	358
7.3.2	Buzen's Algorithm for the Computation of the Normalization Constant	359
7.3.3	Mean Value Analysis	360
7.4	Loss Networks	369
7.4.1	Erlang Fixed-Point Approximation	373
7.4.2	Alternate Routing	378
7.5	Stability of Queueing Networks	381
7.5.1	Definition of Stability	385
7.5.2	Turning a Stochastic Discrete Queueing Network into a Deterministic Fluid Network	387
7.6	Further Readings	390
	Appendix	391
	Summary and Takeaways	394
	Problems	394
8	Bounds and Approximations	399
8.1	Introduction	399
8.2	Bounds for the $G/G/1$ Queue	401
8.2.1	Mean Value Analysis	404
8.2.2	Output Process	406
8.2.3	Upper and Lower Bounds of the Mean Waiting Time	407
8.2.4	Upper Bound of the Waiting Time Probability Distribution	409
8.3	Bounds for the $G/G/m$ Queue	412
8.4	Approximate Analysis of Isolated G/G Queues	416
8.4.1	Approximations from Bounds	416
8.4.2	Approximation of the Arrival or Service Process	417
8.4.3	Reflected Brownian Motion Approximation	418
8.4.4	Heavy-traffic Approximation	423
8.5	Approximate Analysis of a Network of $G/G/1$ Queues	426
8.5.1	Superposition of Flows	427
8.5.2	Flow Through a Queue	428
8.5.3	Bernoulli Splitting of a Flow	428
8.5.4	Putting Pieces Together: The Decomposition Method	429
8.5.5	Bottleneck Approximation for Closed Queueing Networks	442
8.6	Fluid Models	443
8.6.1	Deterministic Fluid Model	444
8.6.2	From Fluid to Diffusion Model	452
8.6.3	Stochastic Fluid Model	456
8.6.4	Steady-State Analysis	459

8.6.4.1	Infinite Buffer Size ($K = \infty$)	462
8.6.4.2	Loss Probability	463
8.6.5	First Passage Times	466
8.6.6	Application of the Stochastic Fluid Model to a Multiplexer with ON-OFF Traffic Sources	468
	Summary and Takeaways	471
	Problems	472

Part III Networked Systems and Protocols 477

9	Multiple Access	479
9.1	Introduction	479
9.2	Slotted ALOHA	482
9.2.1	Analysis of the Naïve Slotted ALOHA	483
9.2.2	Finite Population Slotted ALOHA	487
9.2.3	Stabilized Slotted ALOHA	494
9.3	Pure ALOHA with Variable Packet Times	499
9.4	Carrier Sense Multiple Access (CSMA)	504
9.4.1	Features of the CSMA Protocol	505
9.4.1.1	Clear Channel Assessment	505
9.4.1.2	Persistence Policy	506
9.4.1.3	Retransmission Policy	507
9.4.2	Finite Population Model of CSMA	509
9.4.3	Multi-Packet Reception CSMA	513
9.4.3.1	Multi-Packet Reception 1-Persistent CSMA with Poisson Traffic	515
9.4.3.2	Multi-Packet Reception Nonpersistent CSMA with Poisson Traffic	519
9.4.4	Stability of CSMA	523
9.4.5	Delay Analysis of Stabilized CSMA	531
9.5	Analysis of the WiFi MAC Protocol	534
9.5.1	Outline of the IEEE 802.11 DCF Protocol	534
9.5.2	Model of CSMA/CA	538
9.5.2.1	The Back-off Process	540
9.5.2.2	Virtual Slot Time	543
9.5.2.3	Saturation Throughput	545
9.5.2.4	Service Times of IEEE 802.11 DCF	549
9.5.2.5	Correlation between Service Times	554
9.5.3	Optimization of Back-off Parameters	556
9.5.3.1	Maximization of Throughput	556
9.5.3.2	Minimization of Service Time Jitter	561

9.5.4	Fairness of CSMA/CA	565
9.6	Further Readings	570
	Appendix	572
	Summary and Takeaways	573
	Problems	575
10	Congestion Control	579
10.1	Introduction	579
10.2	Congestion Control Architecture in the Internet	583
10.3	Evolution of Congestion Control in the Internet	587
10.3.1	TCP Reno	588
10.3.1.1	TCP Congestion Control Operations	589
10.3.1.2	NewReno	593
10.3.1.3	TCP Congestion Control with SACK	594
10.3.1.4	Congestion Window Validation	595
10.3.2	TCP CUBIC	596
10.3.3	TCP Vegas	598
10.3.4	Data Center TCP (DCTCP)	601
10.3.4.1	Marking at the Switch	602
10.3.4.2	ECN-Echo at the Receiver	603
10.3.4.3	Controller at the Sender	603
10.3.5	Bottleneck Bandwidth and RTT (BBR)	604
10.3.5.1	Delivery Rate Estimate	607
10.3.5.2	StartUp and Drain	608
10.3.5.3	ProbeBW	609
10.3.5.4	ProbeRTT	610
10.3.5.5	Pseudo-code of BBR Algorithm	610
10.4	Traffic Engineering with TCP	611
10.5	Fluid Model of a Single TCP Connection Congestion Control	614
10.5.1	Classic TCP with Fixed Capacity Bottleneck Link	615
10.5.2	Classic TCP with Variable Capacity Bottleneck Link	617
10.5.2.1	Discretization of the Evolution Equations	625
10.5.2.2	Accuracy of the Fluid Approximation of TCP	627
10.5.3	Application to Wireless Links	630
10.5.3.1	Random Capacity	630
10.5.3.2	TCP over Cellular Link	632
10.6	Fluid Model of Multiple TCP Connections Congestion Control	635
10.6.1	Negligible Buffering at the Bottleneck	635
10.6.2	Classic TCP with Drop Tail Buffer at the Bottleneck	637
10.6.3	Classic TCP with AQM at the Bottleneck	638
10.6.4	Data Center TCP with FIFO Buffer at the Bottleneck	639

10.7	Fairness and Congestion Control	642
10.8	Network Utility Maximization (NUM)	645
10.9	Challenges to TCP	652
10.9.1	Fat-Long Pipes	653
10.9.2	Wireless Channels	655
10.9.3	Bufferbloat	656
10.9.4	Interaction with Applications	658
	Appendix	659
	Summary and Takeaways	664
	Problems	665
11	Quality-of-Service Guarantees	669
11.1	Introduction	669
11.2	Deterministic Service Guarantees	670
11.2.1	Arrival Curves	673
11.2.2	Service Curves	677
11.2.3	Performance Bounds	681
11.2.4	Regulators	683
11.2.5	Network Calculus	688
11.2.5.1	Single Node Analysis	689
11.2.5.2	End-to-End Analysis	692
11.3	Stochastic Service Guarantees	703
11.3.1	Multiplexing with Marginal Buffer Size	703
11.3.2	Multiplexing with Non-Negligible Buffer Size	711
11.3.3	Effective Bandwidth	714
11.3.3.1	Definition of the Effective Bandwidth	714
11.3.3.2	Properties of the Effective Bandwidth	715
11.3.3.3	Effective Bandwidth of a Markov Source	716
11.3.4	Network Analysis and Dimensioning	721
11.4	Further Readings	727
	Appendix	728
	Summary and Takeaways	732
	Problems	733
A	Refresher of Probability, Random Variables, and Stochastic Processes	735
A.1	Probability	735
A.2	Random Variables	737
A.3	Transforms of Probability Distribution Functions	739
A.4	Inequalities and Limit Theorems	744
A.4.1	Markov Inequality	744

A.4.2	Chebychev Inequality	745
A.4.3	Jensen Inequality	746
A.4.4	Chernov Bound	746
A.4.5	Union Bound	747
A.4.6	Central Limit Theorem (CLT)	747
A.5	Stochastic Processes	748
A.6	Markov Chains	749
A.6.1	Classification of States	750
A.6.2	Recurrence	751
A.6.3	Visits to a State	754
A.6.4	Asymptotic Behavior and Steady State	756
A.6.5	Absorbing Markov Chains	762
A.6.6	Continuous-Time Markov Processes	763
A.6.7	Sojourn Times in Process States	765
A.6.8	Reversibility	766
A.6.9	Uniformization	768
A.7	Wiener Process (Brownian Motion)	769
A.7.1	Wiener Process with an Absorbing Barrier	771
A.7.2	Wiener Process with a Reflecting Barrier	772
References		775
Index		789

