CAMBRIDGE

Cambridge University Press
978-1-107-02750-3 - Performance Modeling and Design of Computer Systems: Queueing Theory in Action
Mor Harchol-Balter
Index
More information

Index

A_S, 223, 435, 442
A_t, 435
A_y^x, 514
B, 248, 459
B_x, 514
C^2, 207
D_i, 110
$D_{\max}$, 116
N, 15
N_Q, 15
P_Q, 260
P_{block}, 261
P_{ij}, 131
R, 259, 420
S, 14
S_e, 403
S_x, 514
$S_{\overline{x}}$, 492
T, 14
T_Q, 15
W_x, 513
Z, 21
λ, 14
λ_i, 298
λ_x, 514
μ, 14
π_j, 135
ρ, 100
ρ, for multi-server system, 259, 269, 420
ρ_i, 100
ρ_x, 505
$\rho_{\overline{x}}$, 492
τ_i, 225
a_n, 242
d_n, 242
n_i, 305
$o(\delta)$, 212
p_j, 148
p_n, 242
$r(t)$, 208

Accessible states, 150
Acyclic networks, 293
Admission control, 246
Age of service, 403
All-Can-Win Theorem, 527
Aloha protocol, 195
Aperiodic, 88, 183
Aperiodic chain, 150
Arrival rate, 14
  Time-varying, 366
Arrival Theorem, 337
  Proof of, 339
Asymptotic analysis, 116

Balance equations, 170, 234
Batch system, 22
Bayes Law, 36
BCMP, 380
  FCFS servers, 381
  PLCFS servers, 384
  PS servers, 382
Bidirectional chain, 185
BigIP, 420
Bimodal, 423
Binomial distribution, 67
Birth-death process, 236
Blocking probability, 257
Bottleneck device, 119
Bottleneck Law, 110
Bounded Pareto distribution, 353, 357, 358
Burke's Theorem, 288
  Alternative proof, 290
  Proof of, 289
  Via transforms, 447
Busy period
  Expectation, 248, 406, 448
  M/G/1
    Moments, 461
    Number of jobs served during, 469
    Transform, 459, 461
    With setup time, 469
  M/M/∞, 406
  M/M/1
    Number of jobs served during, 468
    Transform, 448
  Shorts-only, 467

© in this web service Cambridge University Press www.cambridge.org

CAMBRIDGE

Cambridge University Press
978-1-107-02750-3 - Performance Modeling and Design of Computer Systems: Queueing Theory in Action
Mor Harchol-Balter
Index
More information

cμ-Rule, 516
Caching, 204
Capacity provisioning, 272, 274, 277, *see* One fast or many slow
Catalan numbers, 187
Central Limit Theorem, 61, 356
  Heuristic proof via transforms, 448
Central subsystem, 22
Chebyshev's Inequality, 91
Class-based service rates
  FCFS servers, 330, 381
  PS servers, 383
Classed Jackson network, 314
  Class-based service rates, 330
  Examples, 322, 325, 326, 329
  Limiting probabilities, 318
Closed Jackson network, 333, 345
  Limiting probabilities, 335
Closed networks, 20
  Batch system, 96
  Interactive system, 96, 345
  M/M/2, 284
  Performance of, 284, 294
  Terminal-driven system, 96
Cloud service center, 309
Communicating states, 150
Communication networks, 381
Compatible time-sharing system, 265
Competitive ratio, 424
Congestion management, 280
Connection-oriented network, 312
Constant failure rate, 350
Convergence, 79
  Convergence almost surely, 80
  Convergence in probability, 82
  Convergence with probability 1, 80
Correlation, 67
Coupon collection, 70
Covariance, 67
Coxian distribution, 380, 383, 385
CPU process lifetimes, 208, 349
CSMA/CD, 196
CTMC, *see* Markov chains
Cycle stealing, 430

D/D/1, 404
Damage caused by forest fires, 356
Database modeling, 497
Decreasing failure rate, 208, 354
Delta-step proof (δ-step), 209, 212, 213, 219
Departure process, 289, 291, 451
Device demand, 110
DFR, *see* Decreasing failure rate
Dimensionality reduction, 430
Dispatcher, 408
Distributional Little's Law, 111, 456
Doubling arrival and service rates, 5, 240, 405
Doubly stochastic matrix, 145
DTMC, *see* Markov chains

E_2 distribution, 364
E_2/M/1, 365
Embedded DTMC, 286, 451, 479
Empirical measurements, 350
  Human wealth, 356
  IP flow durations, 356
  Natural disasters, 356
  Phone call durations, 356
  Web file sizes, 355
  Wireless session times, 356
Ensemble average, 84, 86 200
Equilibrium distribution, 396, 403
  Distribution, 407
  Transform, 448
Equivalence between Geometric and Exponential, 209, 211
Ergodic, 88, 164
Ergodic theorem of Markov chains, 164, 178
Ergodicity, 148
Erlang-2, 423
Erlang-B formula, 257
Erlang-C formula, 261
Erlang-k distribution, 360
Exceptional first service, 280
Excess of random variable, 396
  Distribution, 407
  Expectation, 401
  Transform, 448
Excess of service, 396, 402
Expected number of visits to state, 157
Expected time to k failures, 185
Exponential distribution, 206, 350
Exponential phases, 360

Failure rate function, 208, 222
Failures and repair, 407
Fairness, 356, 475
  All-Can-Win Theorem, 527
  Fair scheduling, 487, 526
Feedback in network, 307, 308
Fibonacci sequence, 205
Financial application, 185
Finite buffer space, 282
First-Come-First-Served (FCFS), 478
Flow time, 14
Forced Flow Law, 106

© in this web service Cambridge University Press www.cambridge.org

CAMBRIDGE

Cambridge University Press
978-1-107-02750-3 - Performance Modeling and Design of Computer Systems: Queueing Theory in Action
Mor Harchol-Balter
Index
More information

Forward chain, 286
Fractional moments, 70
Frequency-division multiplexing, 241, 263

G/G/k, 428
  Stability, 428
Gambler’s ruin, 160
General service times, 383
Generalized Erlang, 360
Generating random variables
  Accept/Reject method, 72
  Inverse-Transform method, 70
Generator matrix, 368
Geometric distribution, 69, 209
Google’s PageRank algorithm, 190, 191
  Dead end, 192
  Implementation, 195
  Spider trap, 193
Grabbing multiple servers simultaneously, 308, 309

H_2 distribution, 361
Hair salon, 282
Heavy-tail property, 354, 358
Heterogeneous servers, 266, 268
High-variability job sizes, 408
Honeypot, 223
HTTP request scheduling, 356
HTTP request times, 355
Human wealth, 356
Hyperexponential distribution, 361, 377, 378
  Decreasing failure rate, 362, 378
  Degenerate, 362
Hypoexponential, 360

Immediate dispatching, 408
Increasing failure rate, 208
Increasing number of servers, 276
Independent increments, 214
Indicator random variables, 56
Infinite variance, 354
Insensitivity results, 257, 278
Inspection Paradox, 395, 402
Interactive closed system, 21
Interarrival time, 14
IP flow durations, 356
Irreducible, 88, 183
Irreducible chain, 150

Jackson network, 297, 305
  Acyclic, 293
  Arrival process into server, 299
  Classed, *see* Classed Jackson network
  Example, 306
  Limiting probabilities, 305
  Local balance, 301
  Product form, 304
  Total arrival rate into server, 298
  With load-dependent service rates, 344
  With M/M/k queues, 344
Job
  Age, 349, 403
  Lifetime, 349
  Remaining lifetime, 349
  Size, 14, 349
Job migration, 349
  Criterion, 355
Join-Shortest-Queue, *see* Task assignment policy

Kendall notation, 236, 253
Kleinrock’s independence assumption, 381

Laplace transform, 433
  Conditioning, 441
  Linearity, 439
  Moments, 436, 439
  Sum of random number of random variables, 444
  Two-sided, 447
Last-Come-First-Served (LCFS), 478
Law of large numbers
  Strong, 84
  Weak, 83
Least-Work-Left, *see* Task assignment policy
Limited processor-sharing, 498
Limiting distribution, 136, 140
Limiting probabilities as rates, 168
Limiting probability, 134, 135, 136, 140, 242
  As seen by arrival, 242
  As seen by departure, 242
Little’s Law, 95, 98
  Distributional, 111, 456
  For closed systems, 96, 101, 112
  For open systems, 95
  For red jobs, 101
  For waiting time, 100
Little-o, $o(\delta)$, 211
Load, 18
  Resource requirement, R, 259, 273
  System utilization, ρ, 259, 269
Load balancer, 408
Load balancing, 349
Load balancing versus unbalancing, 267, 268, 415, 432
Load in multi-server system, 273
Load-dependent service rates, 344
Local balance, 301, 318, 333, 386

© in this web service Cambridge University Press www.cambridge.org

CAMBRIDGE

Cambridge University Press
978-1-107-02750-3 - Performance Modeling and Design of Computer Systems: Queueing Theory in Action
Mor Harchol-Balter
Index
More information

LocalDirector, 420
Log-log plot, 351

$M^*/E_2^*/1$, 372
M_t, 366
M_t/M/1, 366
M/E_2/1, 364
M/E_k/1, 398
M/H_2/1, 364, 398
M/H_2/2, 378
  With setup time, 378
M/BP/1/PS, 394
M/Cox/1/PS, 385
M/D/1, 398, 404
M/G/∞, 278
  Insensitivity, 278
M/G/1, 395
  Busy period, 459, 461
  Different job types, 405
  Failures and repair, 407
  Laplace transform of response time, 450, 488
  Low utilization, 404
  Mean time in queue, 397
  Priority queue, *see* Priority queue
  Special busy periods, 462
  Stability, 419
  Tagged-job argument, 397
  Time in queue, 404
  Variability in service time, 405
  Variance time in queue, 404, 456
  With setup time
    Mean, 465
    Transform, 464
  Z-transform of number in queue, 456
  Z-transform of number in system, 450
M/G/1/PS, 385, 394
  Ages of jobs, 484
M/G/2, 418
  Stability, 418
M/G/k, 413, 428
  Inaccuracy of approximations, 413
  Lee-Longton approximation, 413
M/M/∞, 266, 271
  Busy period, 406
  With setup time, 468
M/M/1, 236
  Busy period, 448
  Busy period mean, 248
  Departure process, 289, 291
  Finite capacity, 246
  Number in queue, 246
  Number in system, mean, 239
  Number in system, variance, 239
  Response time, distribution, 248, 443, 447
  Response time, mean, 239
  Simulation of, 246
  Threshold queue, 249
  With feedback loop, 307
  With setup time, 280, 465
M/M/1/N, 246
M/M/1/PS, 384
M/M/2
  Heterogeneous servers, 266, 268
  Transform analysis, 449, 456
  With setup time, 378
M/M/2/3, 447
  Example, 265
M/M/k, 253, 258
  Capacity provisioning, 272
  Departure process, 289
  Distribution time in queue, 277
  Expected number busy, 259, 262
  Expected number in queue, 262
  Increasing number of servers, 276
  Mean time in queue, 270
  Mean time in queue given delayed, 270
  Probability of queueing, P_Q, 261
  Resource requirement, R, 259
  System utilization, ρ, 259, 269
  Transform analysis, 449, 456
M/M/k/k, 253, 255
  Blocking probability, 256
  Erlang-B formula, 257
  Insensitivity result, 257
M/PH/1, 377
Markov chains
  Accessible states, 150
  Aperiodic, 150, 183
  Balance equations, 170, 234
  Communicating states, 150
  Continuous-time (CTMC), 130, 225
    View 1, 226
    View 2, 227
  Converting CTMC to DTMC, 229, 234, 235
  Discrete-time (DTMC), 129, 130
  Ergodic, 164
  Ergodic theorem, 164, 178
  Ergodicity, 148
  Expected number of visits to state, 157
  Finite-state, 131, 138, 189
  Gambler's ruin, 160
  Infinite-state, 139
  Irreducible, 150, 183
  Limiting distribution, 135, 136, 139, 140
  Limiting distribution equals stationary distribution, 136, 140

© in this web service Cambridge University Press www.cambridge.org

CAMBRIDGE

Cambridge University Press
978-1-107-02750-3 - Performance Modeling and Design of Computer Systems: Queueing Theory in Action
Mor Harchol-Balter
Index
More information

Limiting probabilities, 134, 135, 136, 139, 140
Limiting probabilities as rates, 168
Markovian property, 130, 225
n-step transition probabilities, 133
Null recurrent, 162
Periodic, 171
Positive recurrent, 162, 183
Powers of **P**, 133
Random walk, 160
Recurrent chain, 161
Recurrent state, 156, 157
Recurrent versus transient, 188
Semi-Markov process, 406
Solution via generating functions, 201, 205
Stationary distribution, 136, 140
Stationary equations, 136
Stationary property, 131
Steady state, 137
Summary theorem, 165
Symmetric random walk, 163
Time average, 166
Time average versus ensemble average, 88, 148
Time between visits to state, 153, 164, 406
Time to empty, 204
Time until leave state, 226
Time-reversibility equations, 171, 254
Time-reversible, 170, 254
Transient chain, 159, 161
Transient state, 156, 157
Transition probability matrix, 131
Markov-modulated Poisson process, 366
Markov's Inequality, 91
Markovian property, 130, 225
Matching moments of distribution, 361, 363
Matrix-analytic method, 359, 366
Generator matrix, 368
M/PH/1, 377
Time-varying load, 377
Max of Exponentials, 223, 224
Mean value analysis (MVA), 337, 340
Memoryless, 207, 222
Memoryless distribution, 209
Method of phases, 359
Method of stages, 359
Migrating jobs, 349
Migrating old jobs, 355
Minimum of Exponentials, 212
Modification analysis, 114, 118, 124
Multiple resources at once, 309
Multiprogramming level (MPL), 21, 23
For Processor-Sharing, 497

Network Dispatcher, 420
Network of PS servers, 391, 393
Network of workstations, 349, 383
Network with two job types, 326
Non-Markovian arrival process, 366
Normal approximation, 68
Null recurrent, 162

One fast or many slow, 7, 263, 431
Open networks, 16
Open versus closed systems, 123, 247, 267
Operational laws, 93
Asymptotic analysis, 116
Bottleneck Law, 110
Combining operational laws, 107, 112
Forced Flow Law, 106
Little's Law, 95, 98
For closed systems, 101, 112
For mean slowdown, 113
For red jobs, 101
For waiting time, 100
Modification analysis, 118, 124
Response Time Law, 103
Utilization Law, 100

Packet-routing network, 312
Parallel jobs, 308
Pareto distribution, 352, 355, 415
PASTA, 242, 243, 338
Performance metrics, 473
Performance-per-Watt, 458
Periodic chains, 171, 189
Phase-type distribution, 359, 362
Phone call durations, 356
Poisson
Number of arrivals during S, 223
Poisson approximation to Binomial, 67
Poisson Arrivals See Time Averages (PASTA), 242, 243, 398
Application to simulations, 244
Poisson process, 213, 222
Definition 1, 215
Definition 2, 215
Definition 3, 217
Independent increments, 214
Merging processes, 218
Number of arrivals during service, 223, 435
Poisson splitting, 218
Stationary increments, 215
Uniformity, 221
Pollaczek-Khinchin (P-K) formula, 404
Positive correlation, 67
Positive recurrent, 88, 162, 183
Power laws in Internet, 355

© in this web service Cambridge University Press www.cambridge.org

CAMBRIDGE

Cambridge University Press
978-1-107-02750-3 - Performance Modeling and Design of Computer Systems: Queueing Theory in Action
Mor Harchol-Balter
Index
More information

Power management, 457
  Dynamic power management, 111
  Performance-per-Watt, 466
  Policies
    DelayedOff, 469
    ON/IDLE, 458, 465
    ON/OFF, 458, 466, 469
  Power allocation, 125
  Server farm, 467
  Setup cost, 458
Power-law distribution, 352
Preemptive-resume, 482
Pricing for queues, 279
Priority queue, 499
  Non-preemptive, 500, 502
  Preemptive, 500, 508
Probability
  Alternative definition of expectation, 69
  Bayes Law, 36, 68
  Bernoulli distribution, 38
  Binomial distribution, 39
  Conditional independence, 68
  Conditional probability, 33
  Conditional random variables, 49
  Conditionally independent events, 34
  Conditioning, 53
  Covariance, 67
  Expectation, 44
  Expectation of product, 48, 65
  Expectation of quotient, 65
  Expected time to k failures, 185
  Exponential distribution, 42
  Geometric distribution, 39
  Independent events, 34
  Independent random variables, 48
  Indicator random variables, 56
  Law of Total Probability, 35, 53
  Linear transformation property, 60
  Linearity of Expectation, 54
  Markov's Inequality, 91
  Mutually exclusive events, 32
  Normal distribution, 57, 68
  Pareto distribution, 43
  Poisson distribution, 40
  Random variable
    Continuous, 37
    Discrete, 37
  Sum of Geometric number of Exponentials, 223
  Sum of random number of random variables, 62, 187
  Variance, 46
  Variance of sum, 56, 65
Process migration, 349
Processor with failures, 205
Processor-Sharing, 380
Product form, 304, 311, 318, 330, 333, 380, 392, 394
Program analysis example, 132
  Limiting distribution, 145

Quality of Service, 67
Quick versus slow customers, 329

Raising matrix to power, 133
Random, *see* Task assignment policy, 478
Random walk, 160
Randomized chess, 186
Recurrent chain, 161
Recurrent state, 156, 157
Recurrent versus transient, 188
Relationship between closed and open systems, 25
Reliability
  Max of Exponentials, 223, 224
  Min of Exponentials, 212
Remaining service requirement, 518
Remote execution, 349
Renewal process, 167, 399
Renewal theorem, 167
Renewal-Reward theorem, 400
Renewal-Reward theory, 399, 406
Repair facility, 330
Repair facility example, 131, 133, 138
Rerouting IP flows, 356
Residence time, 510, 519
Residual distribution, 396
Residue classes, 173, 189
Resource requirement, R, 273
Response time, 14
  95th percentile, 277
Response time in closed system, 22
Response Time Law for Closed Systems, 103
Reverse chain, 285, 286
Reverse process, 286
Round-Robin, *see* Task assignment policy
Routing
  Class-based, 311, 326
Routing probability, 297

Sample mean, 357
Sample path, 80
Sample variance, 357
Scheduling
  Comparison of policies, 524
  Fairness, *see* Fairness
  First-Come-First-Served (FCFS), 10
  Foreground-Background (FB), 490

© in this web service Cambridge University Press www.cambridge.org

CAMBRIDGE

Cambridge University Press
978-1-107-02750-3 - Performance Modeling and Design of Computer Systems: Queueing Theory in Action
Mor Harchol-Balter
Index
More information

Response time, 494
Transform, 497
Last-Come-First-Served (LCFS), 10
Transform, 481
Least-Attained-Service (LAS), 491
Limited Processor-Sharing, 497
Minimizing mean slowdown, 125
Non-preemptive, 410, 478, 479, 481
Preemptive, 482
Preemptive-LCFS (PLCFS), 488
Response time, 489
Transform, 497
Preemptive-Shortest-Job-First (PSJF), 512
Transform, 514
Processor-Sharing (PS), 382, 483
Response time, 486
Random, 10, 478
RS policy, 113
Shortest-Job-First (SJF), 499, 505
Shortest-Remaining-Processing-Time (SRPT), 26, 113, 356, 519
SRPT beats FB, 523
Starvation, 356, 475
Work-conserving policy, 474
Scheduling web requests, 356, 499
Scherr's thesis, 265
Security application, 223
Semi-Markov process, 406
Server farm, 408
FCFS servers, 410
Optimal design, 424
PS and FCFS servers, 430
PS servers, 391, 419, 430
Task assignment, *see* Task assignment in server farms
Service rate, 14
Service requirement, 14
Setup cost, 11, 458
Setup time, 280, 378, 465, 468
Simulation
Generating random variables, 70
M/BP/1, 357
M/M/1, 78
Pareto, 358
Sample mean, 357
Sample variance, 357
Time average versus ensemble average, 90
Size-Interval-Task-Assignment (SITA), *see* Task assignment policy
Slotted Aloha protocol, 196
Slowdown, 26, 474, 481
Sojourn time, 14

Solving recurrences via generating functions, 201
Square-root staffing, 274, 277
Squared coefficient of variation, C^2, 65, 207, 359
Stability, 26
M/G/2, 418
Minimum number of servers needed, 273
Single queue, 15, 419
Starvation, 475, 496
Stationary distribution, 136, 140
Stationary equations, 136
Stationary increments, 215
Stationary property, 131
Statistical multiplexing, 241
Steady state, 137
Stirling's approximation, 155
Stochastic process, 130
Stopping time, 187
Strong Law of Large Numbers, 84, 167
Sum of Geometric number of Exponentials, 223
With transforms, 446
Sum of random number of random variables, 62, 187
With transforms, 444
Supercomputer center, 308
SURGE, 355
Symmetric random walk, 163, 187, 188
SYNC, 356
System utilization, ρ, 259, 269

Tagged-job argument, 396
Tail of response time, 475
Tandem queue, 291
With finite buffers, 282
Task assignment in server farms, 9, 408
Dynamic, 412
Load balancing, 267
Load balancing versus unbalancing, 267, 268
Static, 412
Task assignment policy, 408
Central queue, 411
Central-Queue-SRPT, 425
Comparison, 423
Cycle stealing, 430
Hybrid, 428
IMD, 427
Join-Shortest-Queue (JSQ), 429
For PS servers, 422
Least-Work-Left (LWL), 412, 431
M/G/k, 411, 431
OPT-0, 423
Random policy, 410

© in this web service Cambridge University Press www.cambridge.org

CAMBRIDGE

Cambridge University Press
978-1-107-02750-3 - Performance Modeling and Design of Computer Systems: Queueing Theory in Action
Mor Harchol-Balter
Index
More information

Task assignment policy (*cont.*)
  Round-Robin policy, 410
  SITA versus LWL, 416
  Size-Interval-Task-Assignment (SITA), 414
    Optimal size cutoffs, 414
  TAGS, 428
  Under low variability, 419
TCP flow scheduling, 356
Think time, 21
Threshold queue, 146, 249
Throughput, 17
Throughput in closed system, 23
Throughput of device, 18
Throwing away servers, 267
Time average, 84, 86, 166
Time average versus ensemble average, 88, 148
Time between visits to state, 153, 164, 406
Time in system, 14
Time-reversibility, 170, 288, 294
Time-reversibility equations, 171, 254
Time to empty, 204
Time until k consecutive failures, 70
Time-sharing, 380
Time-sharing CPU, 382
Time-varying arrival rate, 366
Time-varying load, 377
Transient chain, 159, 161
Transient state, 156, 157
Transition probability matrix, 131
Transmission time, 382
Turnaround time, 14

Umbrella problem example, 132, 133, 139
Unbounded queue example, 142
UNIX process lifetime, 349
  Distribution, 350
  Empirical measurements, 350
Utilization, 17, 100
  Derivation of, 100
Utilization Law, 19, 100
Utilization of device, 18
Utilization of multi-server system, 273

Vacation of server, 467
Variability in service time, 377, 378, 403, 405
Variance, 46

Waiting time, 510
Wald's equation, 187
Walk on undirected weighted graph, 186
Weak Law of Large Numbers, 83, 91
Web file sizes, 355
Web server farm, 419
Weibull distribution, 423, 524, 525
What-if questions, 93, 114
Wireless session times, 356
Work in system, 474
Work-conserving policy, 474
Worst-case analysis, 424

Z-transform, 201, 434
  Conditioning, 441
  For solving recurrences, 201
  Linearity, 440
  Moments, 437
  Sum of random number of random variables, 444

© in this web service Cambridge University Press www.cambridge.org