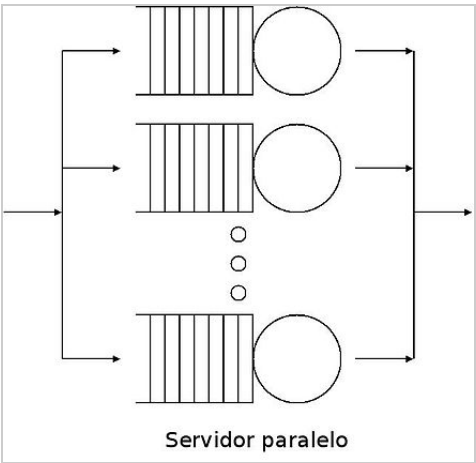
Queueing theory

From Wikipedia, the free encyclopedia

Queueing theory is the mathematical study of waiting lines, or queues.[1] A queueing model is constructed so that queue lengths and waiting time can be predicted.[1] Queueing theory is generally considered a branch of operations research because the results are often used when making business decisions about the resources needed to provide a service.

Queueing theory has its origins in research by Agner Krarup Erlang when he created models to describe the Copenhagen telephone exchange.[1] The ideas have since seen applications including telecommunication, traffic engineering, computing[2] and, particularly in industrial engineering, in the design of factories, shops, offices and hospitals, as well as in project management.[3][4]



Queue networks are systems in which single queues are connected by a routing network. In this image, servers are represented by circles, queues by a series of rectangles and the routing network by arrows. In the study of queue networks one typically tries to obtain the equilibrium distribution of the network, although in many applications the study of the transient state is fundamental.

Contents

Spelling

The spelling "queueing" over "queuing" is typically encountered in the academic research field. In fact, one of the flagship journals of the profession is named *Queueing Systems*.

Single queueing nodes

Single queueing nodes are usually described using Kendall's notation in the form *A*/*S*/*C* where *A* describes the time between arrivals to the queue, *S* the size of jobs and *C* the number of servers at the node.[5][6] Many theorems in queueing theory can be proved by reducing queues to mathematical systems known as Markov chains, first described by Andrey Markov in his 1906 paper.[7]

Agner Krarup Erlang, a Danish engineer who worked for the Copenhagen Telephone Exchange, published the first paper on what would now be called queueing theory in 1909.[8][9][10] He modeled the number of telephone calls arriving at an exchange by a Poisson process and solved the M/D/1 queue in 1917 and M/D/k queueing model in 1920.[11] In Kendall's notation:

- M stands for Markov or memoryless and means arrivals occur according to a Poisson process
- D stands for deterministic and means jobs arriving at the queue require a fixed amount of service
- k describes the number of servers at the queueing node ($k = 1, 2,...$). If there are more jobs at the node than there are servers then jobs will queue and wait for service

The M/M/1 queue is a simple model where a single server serves jobs that arrive according to a Poisson process and have exponentially distributed service requirements. In an M/G/1 queue the G stands for general and indicates an arbitrary probability distribution. The M/G/1 model was solved by Felix Pollaczek in 1930,[12] a solution later recast in probabilistic terms by Aleksandr Khinchin and now known as the Pollaczek–Khinchine formula.[11][13]

After the 1940s queueing theory became an area of research interest to mathematicians.[13] In 1953 David George Kendall solved the GI/M/k queue[14] and introduced the modern notation for queues, now known as Kendall's notation. In 1957 Pollaczek studied the GI/G/1 using an integral equation.[15] John Kingman gave a formula for the mean waiting time in a G/G/1 queue: Kingman's formula.[16]

The matrix geometric method and matrix analytic methods have allowed queues with phase-type distributed inter-arrival and service time distributions to be considered.[17]

Problems such as performance metrics for the M/G/k queue remain an open problem.[11][13]

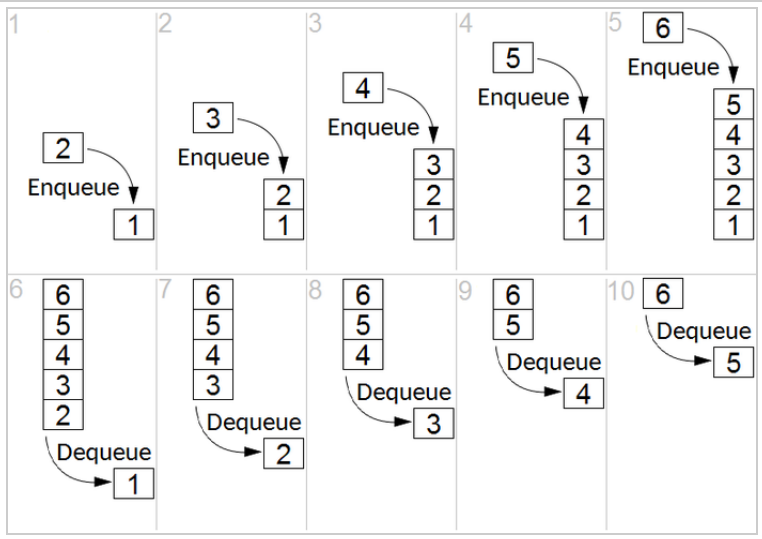
Service disciplines

Various scheduling policies can be used at queuing nodes:

First in first out
: This principle states that customers are served one at a time and that the customer that has been waiting the longest is served first.[18]

Last in first out
: This principle also serves customers one at a time, but the customer with the shortest waiting time will be served first.[18] Also known as a stack.

Processor sharing
: Service capacity is shared equally between customers.[18]



First in first out (FIFO) queue example.

Priority
: Customers with high priority are served first.[18] Priority queues can be of two types, non-preemptive (where a job in service cannot be interrupted) and preemptive (where a job in service can be interrupted by a higher-priority job). No work is lost in either model.[19]

Shortest job first

The next job to be served is the one with the smallest size

Preemptive shortest job first

The next job to be served is the one with the original smallest size[20]

Shortest remaining processing time

The next job to serve is the one with the smallest remaining processing requirement.[21]

Service facility

- Single server: customers line up and there is only one server
- Parallel servers: customers line up and there are several servers
- Tandem queue: there are many counters and customers can decide going where to queue

Customer's behavior of waiting

- Balking: customers deciding not to join the queue if it is too long
- Jockeying: customers switch between queues if they think they will get served faster by doing so
- Reneging: customers leave the queue if they have waited too long for service

Queueing networks

Networks of queues are systems in which a number of queues are connected by customer routing. When a customer is serviced at one node it can join another node and queue for service, or leave the network. For a network of m the state of the system can be described by an m–dimensional vector $(x_1,x_2,...,x_m)$ where x_i represents the number of customers at each node.

The first significant results in this area were Jackson networks,[22][23] for which an efficient product-form stationary distribution exists and the mean value analysis[24] which allows average metrics such as throughput and sojourn times to be computed.[25] If the total number of customers in the network remains constant the network is called a closed network and has also been shown to have a product–form stationary distribution in the Gordon–Newell theorem.[26] This result was extended to the BCMP network[27] where a network with very general service time, regimes and customer routing is shown to also exhibit a product-form stationary distribution. The normalizing constant can be calculated with the Buzen's algorithm, proposed in 1973.[28]

Networks of customers have also been investigated, Kelly networks where customers of different classes experience different priority levels at different service nodes.[29] Another type of network are G-networks first proposed by Erol Gelenbe in 1993:[30] these networks do not assume exponential time distributions like the classic Jackson Network.

Example of M/M/1

Birth and Death process

- A/B/C

  A:distribution of arrival time

  B:distribution of service time

  C:the number of parallel servers

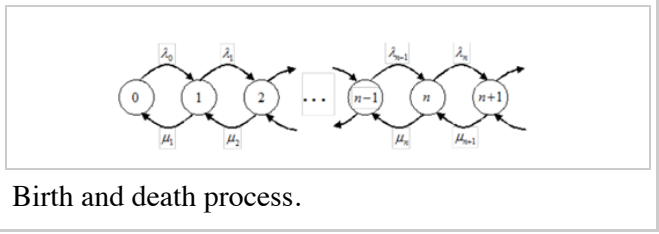
A system of inter-arrival time and service time showed exponential distribution, we denoted M.

λ: the average arrival rate

μ: the average service rate of a single service

P : the probability of n customers in system

n :the number of people in system



Birth and death process.

- Let E represent the number of times of entering state n, and L represent the number of times of leaving state n. We have $|E - L| \in \{0, 1\}$. When the system arrives at steady state, which means t, we have, therefore arrival rate=removed rate.
- Balance equation

situation 0: $\mu_1 P_1 = \lambda_0 P_0$

situation 1: $\lambda_0 P_0 + \mu_2 P_2 = (\lambda_1 + \mu_1) P_1$

situation n: $\lambda_{n-1} P_{n-1} + \mu_{n+1} P_{n+1} = (\lambda_n + \mu_n) P_n$

By balance equation, $P_1 = \frac{\lambda_0}{\mu_1} P_0 \quad P_2 = \frac{\lambda_1}{\mu_2} P_1 + \frac{1}{\mu_2}(\mu_1 P_1 - \lambda_0 P_0) = \frac{\lambda_1}{\mu_2} P_1 = \frac{\lambda_1 \lambda_0}{\mu_2 \mu_1} P_0$

By mathematical induction, $P_n = \frac{\lambda_{n-1} \lambda_{n-2} \cdots \lambda_0}{\mu_n \mu_{n-1} \cdots \mu_1} P_0 = P_0 \prod_{i=0}^{n-1} \frac{\lambda_i}{\mu_{i+1}}$

Because $\sum_{n=0}^{\infty} P_n = P_0 + P_0 \sum_{n=1}^{\infty} \prod_{i=0}^{n-1} \frac{\lambda_i}{\mu_{i+1}} = 1$

we get $P_0 = \frac{1}{1 + \sum_{n=1}^{\infty} \prod_{i=0}^{n-1} \frac{\lambda_i}{\mu_{i+1}}}$

Routing algorithms

In discrete time networks where there is a constraint on which service nodes can be active at any time, the max-weight scheduling algorithm chooses a service policy to give optimal throughput in the case that each job visits only a single service node. In the more general case where jobs can visit more than one node, backpressure routing gives optimal throughput.

A network scheduler must choose a queuing algorithm, which affects the characteristics of the larger network.

Mean field limits

Mean field models consider the limiting behaviour of the empirical measure (proportion of queues in different states) as the number of queues (*m* above) goes to infinity. The impact of other queues on any given queue in the network is approximated by a differential equation. The deterministic model converges to the same stationary distribution as the original model.[31]

Fluid limits

Fluid models are continuous deterministic analogs of queueing networks obtained by taking the limit when the process is scaled in time and space, allowing heterogeneous objects. This scaled trajectory converges to a deterministic equation which allows the stability of the system to be proven. It is known that a queueing network can be stable, but have an unstable fluid limit.[32]

Heavy traffic/diffusion approximations

In a system with high occupancy rates (utilisation near 1) a heavy traffic approximation can be used to approximate the queueing length process by a reflected Brownian motion,[33] Ornstein–Uhlenbeck process or more general diffusion process.[34] The number of dimensions of the RBM is equal to the number of queueing nodes and the diffusion is restricted to the non-negative orthant.

See also

- Ehrenfest model
- Erlang unit
- Industrial engineering
- Network simulation
- Poisson distribution
- Project Production Management
- Queue area
- Queueing delay
- Queue management system
- *Queueing Systems* – a journal of queueing theory
- Random early detection
- Renewal theory
- Throughput
- Scheduling (computing)
- Traffic jam
- Traffic generation model
- Flow network

References

1. Sundarapandian, V. (2009). "7. Queueing Theory". *Probability, Statistics and Queueing Theory*. PHI Learning. ISBN 8120338448.
2. Lawrence W. Dowdy, Virgilio A.F. Almeida, Daniel A. Menasce. "Performance by Design: Computer Capacity Planning by Example" (http://www.cs.gmu.edu/~menasce/perfbyd/).
3. Schlechter, Kira (March 2, 2009). "Hershey Medical Center to open redesigned emergency room" (http://www.pennlive.com/midstate/index.ssf/2009/03/hershey_med_to_open_redesigned.html). *The*

Patriot-News.
4. Mayhew, Les; Smith, David (December 2006). *Using queuing theory to analyse completion times in accident and emergency departments in the light of the Government 4-hour target* (http://www.cass.city.ac.uk/media/stories/story_96_105659_69284.html). Cass Business School. ISBN 978-1-905752-06-5. Retrieved 2008-05-20.
5. Tijms, H.C, *Algorithmic Analysis of Queues", Chapter 9 in A First Course in Stochastic Models, Wiley, Chichester, 2003*
6. Kendall, D. G. (1953). "Stochastic Processes Occurring in the Theory of Queues and their Analysis by the Method of the Imbedded Markov Chain" (http://projecteuclid.org/euclid.aoms/1177728975). *The Annals of Mathematical Statistics*. **24** (3): 338. JSTOR 2236285 (https://www.jstor.org/stable/2236285). doi:10.1214/aoms/1177728975 (https://doi.org/10.1214%2Faoms%2F1177728975).
7. A.A. Markov, Extension of the law of large numbers to dependent quantities, Izvestiia Fiz.-Matem. Obsch. Kazan Univ., (2nd Ser.), 15(1906), pp. 135–156 [Also [37], pp. 339–361].
8. "Agner Krarup Erlang (1878 - 1929) | plus.maths.org" (http://pass.maths.org.uk/issue2/erlang/index.html). Pass.maths.org.uk. Retrieved 2013-04-22.
9. Asmussen, S. R.; Boxma, O. J. (2009). "Editorial introduction". *Queueing Systems*. **63**: 1. doi:10.1007/s11134-009-9151-8 (https://doi.org/10.1007%2Fs11134-009-9151-8).
10. Erlang, Agner Krarup (1909). "The theory of probabilities and telephone conversations" (https://web.archive.org/web/20111001212934/http://oldwww.com.dtu.dk/teletraffic/erlangbook/pps131-137.pdf) (PDF). *Nyt Tidsskrift for Matematik B*. **20**: 33–39. Archived from the original (http://oldwww.com.dtu.dk/teletraffic/erlangbook/pps131-137.pdf) (PDF) on 2011-10-01.
11. Kingman, J. F. C. (2009). "The first Erlang century—and the next". *Queueing Systems*. **63**: 3–4. doi:10.1007/s11134-009-9147-4 (https://doi.org/10.1007%2Fs11134-009-9147-4).
12. Pollaczek, F., Ueber eine Aufgabe der Wahrscheinlichkeitstheorie, Math. Z. 1930
13. Whittle, P. (2002). "Applied Probability in Great Britain". *Operations Research*. **50**: 227–177. JSTOR 3088474 (https://www.jstor.org/stable/3088474). doi:10.1287/opre.50.1.227.17792 (https://doi.org/10.1287%2Fopre.50.1.227.17792).
14. Kendall, D.G.:Stochastic processes occurring in the theory of queues and their analysis by the method of the imbedded Markov chain, Ann. Math. Stat. 1953
15. Pollaczek, F., Problèmes Stochastiques posés par le phénomène de formation d'une queue
16. Kingman, J. F. C.; Atiyah (October 1961). "The single server queue in heavy traffic". *Mathematical Proceedings of the Cambridge Philosophical Society*. **57** (4): 902. JSTOR 2984229 (https://www.jstor.org/stable/2984229). doi:10.1017/S0305004100036094 (https://doi.org/10.1017%2FS0305004100036094).
17. Ramaswami, V. (1988). "A stable recursion for the steady state vector in markov chains of m/g/1 type". *Communications in Statistics. Stochastic Models*. **4**: 183–188. doi:10.1080/15326348808807077 (https://doi.org/10.1080%2F15326348808807077).
18. Penttinen A., *Chapter 8 – Queueing Systems*, Lecture Notes: S-38.145 - Introduction to Teletraffic Theory.
19. Harchol-Balter, M. (2012). "Scheduling: Non-Preemptive, Size-Based Policies". *Performance Modeling and Design of Computer Systems*. p. 499. ISBN 9781139226424. doi:10.1017/CBO9781139226424.039 (https://doi.org/10.1017%2FCBO9781139226424.039).
20. Harchol-Balter, M. (2012). "Scheduling: Preemptive, Size-Based Policies". *Performance Modeling and Design of Computer Systems*. p. 508. ISBN 9781139226424. doi:10.1017/CBO9781139226424.040 (https://doi.org/10.1017%2FCBO9781139226424.040).
21. Harchol-Balter, M. (2012). "Scheduling: SRPT and Fairness". *Performance Modeling and Design of Computer Systems*. p. 518. ISBN 9781139226424. doi:10.1017/CBO9781139226424.041 (https://doi.org/10.1017%2FCBO9781139226424.041).
22. Jackson, J. R. (1957). "Networks of Waiting Lines". *Operations Research*. **5** (4): 518–521. JSTOR 167249 (https://www.jstor.org/stable/167249). doi:10.1287/opre.5.4.518 (https://doi.org/10.1287%2Fopre.5.4.518).

23. Jackson, James R. (Oct 1963). "Jobshop-like Queueing Systems". *Management Science*. **10** (1): 131–142. JSTOR 2627213 (https://www.jstor.org/stable/2627213). doi:10.1287/mnsc.1040.0268 (https://doi.org/10.1287%2Fmnsc.1040.0268).
24. Reiser, M.; Lavenberg, S. S. (1980). "Mean-Value Analysis of Closed Multichain Queuing Networks". *Journal of the ACM*. **27** (2): 313. doi:10.1145/322186.322195 (https://doi.org/10.1145%2F322186.322195).
25. Van Dijk, N. M. (1993). "On the arrival theorem for communication networks". *Computer Networks and ISDN Systems*. **25** (10): 1135–2013. doi:10.1016/0169-7552(93)90073-D (https://doi.org/10.1016%2F0169-7552%2893%2990073-D).
26. Gordon, W. J.; Newell, G. F. (1967). "Closed Queuing Systems with Exponential Servers". *Operations Research*. **15** (2): 254. JSTOR 168557 (https://www.jstor.org/stable/168557). doi:10.1287/opre.15.2.254 (https://doi.org/10.1287%2Fopre.15.2.254).
27. Baskett, F.; Chandy, K. Mani; Muntz, R.R.; Palacios, F.G. (1975). "Open, closed and mixed networks of queues with different classes of customers". *Journal of the ACM*. **22** (2): 248–260. doi:10.1145/321879.321887 (https://doi.org/10.1145%2F321879.321887).
28. Buzen, J. P. (1973). "Computational algorithms for closed queueing networks with exponential servers" (http://www-unix.ecs.umass.edu/~krishna/ece673/buzen.pdf) (PDF). *Communications of the ACM*. **16** (9): 527. doi:10.1145/362342.362345 (https://doi.org/10.1145%2F362342.362345).
29. Kelly, F. P. (1975). "Networks of Queues with Customers of Different Types". *Journal of Applied Probability*. **12** (3): 542–554. JSTOR 3212869 (https://www.jstor.org/stable/3212869). doi:10.2307/3212869 (https://doi.org/10.2307%2F3212869).
30. Gelenbe, Erol (Sep 1993). "G-Networks with Triggered Customer Movement". *Journal of Applied Probability*. **30** (3): 742–748. JSTOR 3214781 (https://www.jstor.org/stable/3214781). doi:10.2307/3214781 (https://doi.org/10.2307%2F3214781).
31. Bobbio, A.; Gribaudo, M.; Telek, M. S. (2008). "Analysis of Large Scale Interacting Systems by Mean Field Method". *2008 Fifth International Conference on Quantitative Evaluation of Systems*. p. 215. ISBN 978-0-7695-3360-5. doi:10.1109/QEST.2008.47 (https://doi.org/10.1109%2FQEST.2008.47).
32. Bramson, M. (1999). "A stable queueing network with unstable fluid model". *The Annals of Applied Probability*. **9** (3): 818. JSTOR 2667284 (https://www.jstor.org/stable/2667284). doi:10.1214/aoap/1029962815 (https://doi.org/10.1214%2Faoap%2F1029962815).
33. Chen, H.; Whitt, W. (1993). "Diffusion approximations for open queueing networks with service interruptions". *Queueing Systems*. **13** (4): 335. doi:10.1007/BF01149260 (https://doi.org/10.1007%2FBF01149260).
34. Yamada, K. (1995). "Diffusion Approximation for Open State-Dependent Queueing Networks in the Heavy Traffic Situation". *The Annals of Applied Probability*. **5** (4): 958. JSTOR 2245101 (https://www.jstor.org/stable/2245101). doi:10.1214/aoap/1177004602 (https://doi.org/10.1214%2Faoap%2F1177004602).

Further reading

- Gross, Donald; Carl M. Harris (1998). *Fundamentals of Queueing Theory*. Wiley. ISBN 0-471-32812-X. Online (https://books.google.com/books?id=K3lQGeCtAJgC)
- Deitel, Harvey M. (1984) [1982]. *An introduction to operating systems* (http://portal.acm.org/citation.cfm?id=79046&dl=GUIDE&coll=GUIDE) (revisited first ed.). Addison-Wesley. p. 673. ISBN 0-201-14502-2. chap.15, pp. 380–412
- Lazowska, Edward D.; John Zahorjan; G. Scott Graham; Kenneth C. Sevcik (1984). *Quantitative System Performance: Computer System Analysis Using Queueing Network Models* (http://www.cs.washington.edu/homes/lazowska/qsp/). Prentice-Hall, Inc. ISBN 0-13-746975-6.
- Zukerman, Moshe. *Introduction to Queueing Theory and Stochastic Teletraffic Models* (http://www.ee.cityu.edu.hk/~zukerman/classnotes.pdf) (PDF).

External links

- Queueing theory calculator (http://www.supositorio.com/rcalc/rcalclite.htm)
- Teknomo's Queueing theory tutorial and calculators (http://people.revoledu.com/kardi/tutorial/Queuing/index.html)
- Virtamo's Queueing Theory Course (http://www.netlab.tkk.fi/opetus/s383143/kalvot/english.shtml)
- Myron Hlynka's Queueing Theory Page (http://web2.uwindsor.ca/math/hlynka/queue.html)
- Queueing Theory Basics (http://www.eventhelix.com/RealtimeMantra/CongestionControl/queueing_theory.htm)
- A free online tool to solve some classical queueing systems (http://queueing-systems.ens-lyon.fr/)
- What You Hate Most About Waiting in Line: (It's not the length of the wait.) (http://www.slate.com/articles/business/operations/2012/06/queueing_theory_what_people_hate_most_about_waiting_in_line_.html), by Seth Stevenson, *Slate*, 2012 – popular introduction