M/M/c queue

From Wikipedia, the free encyclopedia

In queueing theory, a discipline within the mathematical theory of probability, the **M/M/c queue** (or **Erlang–C model**[1]:495) is a multi-server queueing model.[2] In Kendall's notation it describes a system where arrivals form a single queue and are governed by a Poisson process, there are c servers and job service times are exponentially distributed.[3] It is a generalisation of the M/M/1 queue which considers only a single server. The model with infinitely many servers is the M/M/∞ queue.

Contents

Model definition

An M/M/c queue is a stochastic process whose state space is the set {0, 1, 2, 3, ...} where the value corresponds to the number of customers in the system, including any currently in service.

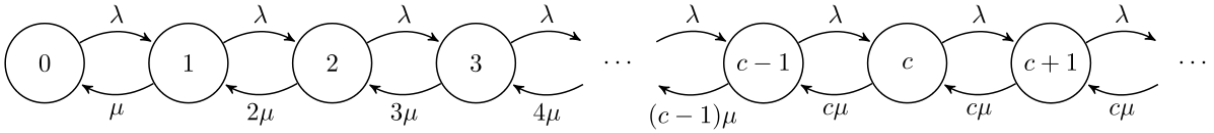
- Arrivals occur at rate λ according to a Poisson process and move the process from state i to i+1.
- Service times have an exponential distribution with parameter μ. If there are less than c jobs, some of the servers will be idle. If there are more than c jobs, the jobs queue in a buffer.
- The buffer is of infinite size, so there is no limit on the number of customers it can contain.

The model can be described as a continuous time Markov chain with transition rate matrix

$$Q = \begin{pmatrix} -\lambda & \lambda & & & & & & & \\ \mu & -(\mu+\lambda) & \lambda & & & & & & \\ & 2\mu & -(2\mu+\lambda) & \lambda & & & & & \\ & & 3\mu & -(3\mu+\lambda) & \lambda & & & & \\ & & & & \ddots & & & & \\ & & & & c\mu & -(c\mu+\lambda) & \lambda & & \\ & & & & & c\mu & -(c\mu+\lambda) & \lambda & \\ & & & & & & c\mu & -(c\mu+\lambda) & \lambda \\ & & & & & & & \ddots & \end{pmatrix}$$

on the state space $\{0, 1, 2, 3, ...\}$. The model is a type of birth–death process. We write $\varrho = \lambda/(c\ \mu)$ for the server utilization and require $\varrho < 1$ for the queue to be stable. ϱ represents the average proportion of time which each of the servers is occupied (assuming jobs finding more than one vacant server choose their servers randomly).

The state space diagram for this chain is as below.



Stationary analysis

Number of customers in the system

If the traffic intensity is greater than one then the queue will grow without bound but if server utilization $\rho = \dfrac{\lambda}{c\mu} < 1$ then the system has a stationary distribution with probability mass function[4][5]

$$\pi_0 = \left[\left(\sum_{k=0}^{c-1} \frac{(c\rho)^k}{k!} \right) + \frac{(c\rho)^c}{c!} \frac{1}{1-\rho} \right]^{-1}$$

$$\pi_k = \begin{cases} \pi_0 \dfrac{(c\rho)^k}{k!}, & \text{if } 0 < k < c \\ \pi_0 \dfrac{\rho^k c^c}{c!}, & \text{if } c \leq k \end{cases}$$

where π_k is the probability that the system contains k customers.

The probability that an arriving customer is forced to join the queue (all servers are occupied) is given by

$$\mathrm{C}\left(c, \lambda/\mu\right) = \frac{\left(\frac{(c\rho)^c}{c!}\right)\left(\frac{1}{1-\rho}\right)}{\sum_{k=0}^{c-1} \frac{(c\rho)^k}{k!} + \left(\frac{(c\rho)^c}{c!}\right)\left(\frac{1}{1-\rho}\right)} = \frac{1}{1 + (1-\rho)\left(\frac{c!}{(c\rho)^c}\right)\sum_{k=0}^{c-1} \frac{(c\rho)^k}{k!}}$$

which is referred to as Erlang's C formula and is often denoted $C(c, \lambda/\mu)$ or $E_{2,c}(\lambda/\mu)$.[4] The average number of customers in the system (in service and in the queue) is given by[6]

$$\frac{\rho}{1-\rho} C(c, \lambda/\mu) + c\rho.$$

Busy period of server

The busy period of the M/M/c queue can either refer to

- full busy period: the time period between an arrival which finds $c-1$ customers in the system until a departure which leaves the system with $c-1$ customers
- partial busy period: the time period between an arrival which finds the system empty until a departure which leaves the system again empty.[7]

Write[8][9] T_k = min(t: k jobs in the system at time 0^+ and $k-1$ jobs in the system at time t) and $\eta_k(s)$ for the Laplace–Stieltjes transform of the distribution of T_k. Then[8]

1. For $k > c$, T_k has the same distribution as T_c.
2. For $k = c$,

$$\eta_c(s) = \frac{c\mu}{k\mu + s + \lambda - \lambda\eta_c(s)}.$$

1. For $k < c$,

$$\eta_k(s) = \frac{k\mu}{k\mu + s + \lambda - \lambda\eta_{k+1}(s)}.$$

Response time

The response time is the total amount of time a customer spends in both the queue and in service. The average response time is the same for all work conserving service disciplines and is[6]

$$\frac{C(c, \lambda/\mu)}{c\mu - \lambda} + \frac{1}{\mu}.$$

Customers in first-come, first-served discipline

The customer either experiences an immediate exponential service, or must wait for k customers to be served before their own service, thus experiencing an Erlang distribution with shape parameter $k + 1$.[10]

Customers in processor sharing discipline

In a processor sharing queue the service capacity of the queue is split equally between the jobs in the queue. In the M/M/c queue this means that when there are c or fewer jobs in the system, each job is serviced at rate μ. However, when there are more than c jobs in the system the service rate of each job decreases and is $\frac{c\mu}{n}$ where n is the number of jobs in the system. This means that arrivals after a job of interest can impact the service time of the job

of interest. The Laplace–Stieltjes transform of the response time distribution has been shown to be a solution to a Volterra integral equation from which moments can be computed.[11] An approximation has been offered for the response time distribution.[12][13]

Finite capacity

In an M/M/c/K queue (sometimes known as the Erlang–A model[1]:495) only K customers can queue at any one time (including those in service[4]). Any further arrivals to the queue are considered "lost". We assume that $K \geq c$. The model has transition rate matrix

$$Q = \begin{pmatrix} -\lambda & \lambda & & & & & & & & \\ \mu & -(\mu+\lambda) & \lambda & & & & & & & \\ & 2\mu & -(2\mu+\lambda) & \lambda & & & & & & \\ & & 3\mu & -(3\mu+\lambda) & \lambda & & & & & \\ & & & & \ddots & & & & & \\ & & & & c\mu & -(c\mu+\lambda) & \lambda & & & \\ & & & & & c\mu & -(c\mu+\lambda) & \lambda & & \\ & & & & & & & \ddots & & \\ & & & & & & & c\mu & -(c\mu) \end{pmatrix}$$

on the state space $\{0, 1, 2, ..., c, ..., K\}$. In the case where $c = K$, the M/M/c/c queue is also known as the Erlang–B model.[1]:495

Transient analysis

See Takács for a transient solution[14] and Stadje for busy period results.[15]

Stationary analysis

Stationary probabilities are given by[16]

$$\pi_0 = \left[\sum_{k=0}^{c} \frac{\lambda^k}{\mu^k k!} + \frac{\lambda^c}{\mu^c c!} \sum_{k=c+1}^{K} \frac{\lambda^{k-c}}{\mu^{k-c} c^{k-c}} \right]^{-1}$$

$$\pi_k = \begin{cases} \frac{(\lambda/\mu)^k}{k!}\pi_0 & \text{for } k = 1, 2, \ldots, c \\ \frac{(\lambda/\mu)^k}{c^{k-c}c!}\pi_0 & \text{for } k = c+1, \ldots, K. \end{cases}$$

The average number of customers in the system is[16]

$$\frac{\lambda}{\mu} + \pi_0 \frac{\rho(c\rho)^c}{(1-\rho)^2 c!}$$

and number of average response time for a customer[16]

$$\frac{1}{\mu} + \pi_0 \frac{\rho(c\rho)^c}{\lambda(1-\rho)^2 c!}.$$

Heavy traffic limits

Writing $X(t)$ for the number of customers in the system at time t, it can be shown that under three different conditions the process

$$\hat{X}_n(t) = \frac{X(nt) - \mathbb{E}(X(nt))}{\sqrt{n}}$$

converges to a diffusion process.[1]:490

1. Fix μ and c, increase λ and scale by $n = 1/(1-\varrho)^2$.
2. Fix μ and ϱ, increase λ and c, and scale by $n = c$.
3. Fix as a constant β where

$$\beta = (1-\rho)\sqrt{s}$$

and increase λ and c using the scale $n = c$ or $n = 1/(1-\varrho)^2$. This case is called the Halfin–Whitt regime.[17]

See also

- Spectral expansion solution

References

1. Gautam, Natarajan (2012). *Analysis of Queues: Methods and Applications*. CRC Press. ISBN 9781439806586.
2. Harrison, Peter; Patel, Naresh M. (1992). *Performance Modelling of Communication Networks and Computer Architectures*. Addison–Wesley. p. 173.
3. 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).
4. Kleinrock, Leonard (1975). *Queueing Systems Volume 1: Theory*. pp. 101–103, 404. ISBN 0471491101.
5. Bolch, G.; Greiner, S.; de Meer, H.; Trivedi, K. S. (1998). "Single Station Queueing Systems". *Queueing Networks and Markov Chains*. pp. 209–262. ISBN 0471193666. doi:10.1002/0471200581.ch6 (https://doi.org/10.1002%2F0471200581.ch6).
6. Barbeau, Michel; Kranakis, Evangelos (2007). *Principles of Ad-hoc Networking*. John Wiley & Sons. p. 42. ISBN 0470032901.
7. Artalejo, J. R.; Lopez-Herrero, M. J. (2001). "Analysis of the Busy Period for the M/M/c Queue: An Algorithmic Approach". *Journal of Applied Probability*. **38** (1): 209–222. JSTOR 3215752 (https://www.jstor.org/stable/3215752).
8. Omahen, K.; Marathe, V. (1978). "Analysis and Applications of the Delay Cycle for the M/M/c Queueing System" (https://docs.lib.purdue.edu/cgi/viewcontent.cgi?article=1110&context=cstech). *Journal of the ACM*. **25** (2): 283. doi:10.1145/322063.322072 (https://doi.org/10.1145%2F322063.322072).
9. Daley, D. J.; Servi, L. D. (1998). "Idle and busy periods in stable M / M / k queues". *Journal of Applied Probability*. **35** (4): 950. doi:10.1239/jap/1032438390 (https://doi.org/10.1239%2Fjap%2F1032438390).

10. Iversen, Villy B. (June 20, 2001). "ITU/ITC Teletraffic Engineering Handbook" (http://www.itu.int/ITU-D/study_groups/SGP_1998-2002/SG2/StudyQuestions/Question_16/RapporteursGroupDocs/teletraffic.pdf) (PDF). Retrieved August 7, 2012.
11. Braband, J. (1994). "Waiting time distributions for M/M/N processor sharing queues". *Communications in Statistics. Stochastic Models*. **10** (3): 533–548. doi:10.1080/15326349408807309 (https://doi.org/10.1080%2F15326349408807309).
12. Braband, J. (1995). "Waiting time distributions for closed M/M/N processor sharing queues". *Queueing Systems*. **19** (3): 331–344. doi:10.1007/BF01150417 (https://doi.org/10.1007%2FBF01150417).
13. Braband, Jens; Schassberger, Rolf (21–23 September 1993). B. Walke and O. Spaniol, ed. "Random Quantum Allocation: A New Approach to Waiting Time Distributions for M/M/N Processor Sharing Queues". Aachen: Springer: 130–142. ISBN 3540572015.
14. Takács, L. (1962). *Introduction to the Theory of Queues*. London: Oxford University Press. pp. 12–21.
15. Stadje, W. (1995). "The busy periods of some queueing systems". *Stochastic Processes and their Applications*. **55**: 159–167. doi:10.1016/0304-4149(94)00032-O (https://doi.org/10.1016%2F0304-4149%2894%2900032-O).
16. Allen, Arnold O. (1990). *Probability, Statistics, and Queueing Theory: With Computer Science Applications*. Gulf Professional Publishing. pp. 679–680. ISBN 0120510510.
17. Halfin, Shlomo; Whitt, Ward (1981). "Heavy-Traffic Limits for Queues with Many Exponential Servers" (http://www.columbia.edu/~ww2040/Halfin81.pdf) (PDF). *Operations Research*. **29** (3): 567–588. JSTOR 170115 (https://www.jstor.org/stable/170115). doi:10.1287/opre.29.3.567 (https://doi.org/10.1287%2Fopre.29.3.567).

Retrieved from "https://en.wikipedia.org/w/index.php?title=M/M/c_queue&oldid=783834004"
