QUEUEING THEORY

A *queue* is a waiting line (like customers waiting at a supermarket checkout counter); *queueing theory* is the mathematical theory of waiting lines. More generally, queueing theory is concerned with the mathematical modeling and analysis of systems that provide service to random demands. A queueing model is an abstract description of such a system. Typically, a queueing model represents (1) the system's physical configuration, by specifying the number and arrangement of the *servers*, which provide service to the *customers,* and (2) the stochastic (that is, probabilistic or statistical) nature of the demands, by specifying the variability in the *arrival process* and in the *service process*.

For example, in the context of computer communications, a communications channel might be a server, and the messages the customers; the (random) times at which messages request the use of the channel would be the arrival process, and the (random) lengths of service time that the messages hold the channel while being transmitted would constitute the service process. Another example is a computer system where a programmer (customer) sitting at a terminal requests access to a CPU (server) for the processing of a transaction; both the arrival time of the request for access and the amount of processing time requested are random. Then, the mathematical analysis of the models would yield formulas that presumably relate the physical and stochastic parameters to certain performance measures, such as average waiting time, server utilization, throughput, probability of buffer overflow, etc. The art of applied queueing theory is to construct a model that is simple enough so that it yields to mathematical analysis, yet contains sufficient detail so that its performance measures reflect the behavior of the real system.

Queueing theory was born in the early 1900s with the work of A. K. Erlang of the Copenhagen Telephone Company, who derived several important formulas for teletraffic engineering that today bear his name. The range of applications has grown to include not only telecommunications and computer science, but also manufacturing, air traffic control, military logistics, design of theme parks, and many other areas that involve service systems whose demands are random. Queueing theory is considered to be one of the standard methodologies (together with linear programming, simulation, etc.) of operations research and management science, and is standard fare in academic programs in industrial engineering, manufacturing engineering, etc., as well as in programs in telecommunications, computer engineering, and computer science. There are dozens of books and thousands of papers on queueing theory, and they continue to be published at an ever-increasing rate. But, despite its apparent simplicity (customers arrive, request service, and leave or wait until they get it), the subject is one of some depth and subtlety. We will illustrate this by briefly visiting some of the most important models, and describing along the way some of the obvious features and some of the subtleties.

The essence of queueing theory is that it takes into account the randomness of the arrival process and the randomness of the service process. The most common assumption about the arrival process is that the customer arrival epochs follow a *Poisson process*. One way to describe a Poisson arrival process is to imagine that time is divided into small intervals of length $\Delta\tau$. Assume that in each interval either an arrival occurs (with probability $\lambda\cdot\Delta\tau$, say, where the proportionality constant λ is the arrival rate) or it doesn't, independently of the occurrence or nonoccurrence of arrivals in the other intervals. Finally, imagine that $\Delta\tau \to 0$ (that is, take limits to pass from discrete time to continuous time). Then the arrivals are said to follow a Poisson process; and one of the properties of the Poisson process is that the times between arrivals (the *interarrival times*) are *exponentially distributed.* (A random variable X is said to be exponentially distributed if its distribution function $F_x(t)$ is given by $F_x(t)=1-e^{-\lambda t}$ for all $t \geq 0,$ where $1/\lambda$ is the average value of X. There are many textbooks that cover probability and stochastic processes. We recommend some specific ones below.)

One of the most important queueing models is the *Erlang loss model*; it assumes that the arrivals follow a Poisson process, and that the blocked customers (those who find all servers busy) are *cleared* (that is, they are denied entry into the system, so the blocked customers are lost). The fraction of arriving customers who find all the servers busy (the *probability of blocking*, or *loss probability*) is given by the famous *Erlang loss* (or *Erlang B*) *formula*,

$$B(s,a) = \frac{\frac{a^s}{s!}}{\sum_{k=0}^{s} \frac{a^k}{k!}} \quad , \tag{1}$$

where s is the number of servers and $a=\lambda\tau$ is the *offered load* in *erlangs*, where λ is the *arrival rate* and τ is the *average service time*. An important theorem is that formula (1) applies for *any* distribution of service times; this mathematically surprising and practically important result is an example of the phenomenon of *insensitivity*. Formula (1) is hard to calculate directly from its right-hand side when s and a are large, but is easy to calculate numerically using the following iterative scheme:

$$B(n,a) = \frac{aB(n-1,a)}{n + aB(n-1,a)} \quad (n = 1,2,...,s;\ B(0,a) = 1). \tag{2}$$

For example, it is easy to write a program that implements (2), and to verify that $B(1,\ 0.8) = 0.4444,\ B(10,\ 8) = 0.1217,\ B(100,\ 80) = 0.003992$, and $B(1000,\ 800) = 10^{-12}$. (This means, for example, that when 8 erlangs of Poisson traffic is offered to 10 servers, then about 12% of the arrivals will be blocked.) Also, it can be shown that $B(s_1+s_2,a_1+a_2)<B(s_1,\ a_1) + B(s_2,a_2)$. These examples illustrate the important fact that large systems are more efficient than small ones. The Erlang loss model is one of the fundamental models of teletraffic engineering (the "customers" are telephone calls and the "servers" are trunks, and the blocked calls are cleared from the system and thus are "lost" calls).

The *Erlang delay model* (also called M/M/s in queueing theory parlance[1]) is similar to the Erlang loss model, except that now it is assumed that the blocked customers will wait in a queue as long as necessary for a server to become available. In this model, the probability of blocking (the fraction of customers who will find all *s* servers busy and must wait in the queue) is given by the famous *Erlang delay* (or *Erlang C*) *formula*,

$$C(s,a) = \frac{\dfrac{a^s}{s!(1-\rho)}}{\displaystyle\sum_{k=0}^{s-1}\frac{a^k}{k!} + \frac{a^s}{s!(1-\rho)}} \ , \tag{3}$$

where

$$\rho = \begin{cases} \dfrac{a}{s} & if\ \ a < s \\ 1 & if\ \ a \geq s \end{cases} \tag{4}$$

The quantity *ρ* defined by (4) equals the *server utilization* (the fraction of time, on average, that a server is busy), and *C(s,a)=1* when *ρ=1.* The Erlang C formula (3) is easily calculated by combining the iteration scheme (2) with the formula

$$C(s,a) = \frac{sB(s,a)}{s - a(1 - B(s,a))} \ . \tag{5}$$

Using (5) and (2), it is easy to calculate *C(s,a)* and to compare its values with the corresponding values of *B(s,a)* computed earlier; the results are *C(1, 0.8) = 0.8, C(10,8) = 0.4092, C(100,80) = 0.01965,* and *C(1000,800) = 5.6 x* 10^{-12} . In each case the server utilization is *ρ=80%,* again showing that large systems are more efficient than small ones. Also, note that in each case, *C(s,a)> B(s,a)*. This can be explained by observing that in the Erlang B model the blocked customers are cleared from the system, whereas in the Erlang C model the blocked customers enter the system (and wait in the queue), thereby increasing the probability that future arrivals will find all the servers busy.

If the blocked customers are served in FIFO order (First In, First Out), then the probability *P(t)* that a customer will wait in the queue more than *t* before beginning service is

[1] M/M/s denotes Memoryless (exponential) distribution of interarrival times/Memoryless distribution of service times/s servers.

$$P(t) = C(s,a)e^{-(1-\rho)s\mu t} \quad , \tag{6}$$

where $\mu=1/\tau$ is the *service rate*. For example, in a 10-server system operating at 80% utilization, the fraction of customers who will wait longer than one average service time is given by the right-hand side of (6) with $s=10$, $\rho=80\%$ (and therefore $a=8$ and $C(s,a) = C(10, 8) = 0.4092$) and $t= \tau = 1/\mu$: $P(\tau)=0.05538$.

Formula (3) (or (5)) predicts how many customers (more precisely, what fraction of arriving customers) will have to wait. Formulas (6) and (7) below predict how long the customers will have to wait; that is, if w denotes the average waiting time, then

$$w = C(s,a)\frac{1}{1-\rho}\frac{\tau}{s} \; . \tag{7}$$

Significantly, although (6) is based on the assumption of FIFO service, (7) remains correct even when service is not FIFO, for example, LIFO (L= Last) or SIRO (Service In Random Order). This is true because interchanging statistically identical customers waiting in the queue does not change the number of customers or the amount of work waiting to be served, so average waiting time remains the same.

It is important to note that the Erlang delay model does *not* have the insensitivity property enjoyed by its Erlang loss counterpart; the Erlang C formula is derived under the assumption that the service times (like the interarrival times) are exponentially distributed. If the service times are not exponentially distributed, then results corresponding to (3), (6) and (7) are difficult to obtain, except in one very important case, the single-server ($s=1$) queue. This fundamental model is often referred to as M/G/1 (the G denotes general distribution of service times). When $s=1$ the analogue of formula (7) is the celebrated *Pollaczek-Khintchine* formula,

$$w = \frac{\rho\tau}{2(1-\rho)}(1+\frac{\sigma^2}{\tau^2}) \; , \tag{8}$$

where σ^2 is the variance (a measure of variability or "spread") of the service times. For example, when service times are exponential, then $\sigma^2=\tau^2$ and (8) coincides with (7) (when $s=1$). When service times are constant, then $\sigma^2=0$ and (8) shows that, all other things being equal, average waiting times are twice as long when service times are exponential as when they are constant. Remarkably, when $s=1$ formula (3) remains valid for all service-time distributions.

These examples were chosen to illustrate the richness of queueing theory: simple models accurately describe real systems and often yield surprising insights. There is much more to this useful and mathematically interesting subject.

REFERENCES

Heyman & Sobel, 1982, Wolff, 1989, and Ross, l993 provide good treatments of background material in probability and stochastic processes, together with material that relates directly to queueing theory. Cooper, l981 is a textbook on queueing theory, with some emphasis on models useful in teletraffic engineering; and Cooper, l990 is a survey with an updated list of references. Kleinrock, l976 is a classic textbook that emphasizes computer applications, and Bertsekas & Gallager, l992 has a similar focus but addresses more modern technology. Kulkarni, l995 is a recent textbook with many examples chosen from computer science and engineering, and Takagi, l991-1993 is an encyclopedic compendium with a comprehensive bibliography.

1976 Kleinrock, L. *Queueing Systems, Vol. II, Computer Applications*. New York: Wiley.

1981 Cooper, R.B. *Introduction to Queueing Theory*, 2nd ed. New York: North-Holland (Elsevier). Reprinted 1990. Washington DC: CEEPress, The George Washington University.

1982 Heyman, D.P. and M.J. Sobel. *Stochastic Models in Operations Research, Vol. I, Stochastic Processes and Operating Characteristics*. New York: McGraw-Hill.

1989 Wolff, R.W. *Stochastic Modeling and the Theory of Queues*. Englewood Cliffs, NJ: Prentice Hall.

1990 Cooper, R.B. “Queueing Theory”. In D.P. Heyman and M.J. Sobel, eds. *Stochastic Models,* Chap. 10, pp.469-518. Amsterdam: North-Holland (Elsevier).

1991, 1993 Takagi, H. *Queueing Analysis,* Vol. 1 (1991), Vols. 2,3 (1993). Amsterdam: North-Holland (Elsevier).

1992 Bertsekas, D., and R. Gallager. *Data Networks*, 2nd ed. Englewood Cliffs, NJ: Prentice Hall.

1993 Ross, S.M. *Introduction to Probability Models*, 5th ed. San Diego: Academic Press

1995 Kulkarni, V.G. *Modeling and Analysis of Stochastic Systems*. New York: Chapman & Hall.

ROBERT B. COOPER

Published in *ENCYCLOPEDIA OF COMPUTER SCIENCE,* Fourth Edition (Anthony Ralston, Edwin D. Reilly, David Hemmendinger, eds.), Groves Dictionaries, Inc., 2000, pp. 1496-1498