QUEUEING NOTATION

ROBERT B. COOPER
Department of Computer Science and Engineering, Florida Atlantic University, Boca Raton, Florida

Queueing theory is concerned with the mathematical analysis of systems that provide service to random demands. This theory has many application areas (e.g., industrial engineering, electrical engineering, computer science, telecommunications, and operations management). Consequently, the terminology and notation, while often intuitively sensible, are not always consistent. The mathematical theory is focused on simple (but meaningful) mathematical models that can be described in the precise terminology required for mathematical analysis. In this article, we give an overview of the notation and terminology used in describing the standard basic queueing models.

In the basic queueing model, "customers" arrive (to request service), wait (if necessary) for a "server" to become available to provide the required service, and then leave. Thus the model consists of three components: (i) the (stochastic) arrival process, (ii) the (stochastic) service requirement, and (iii) the physical configuration of the servers and their operating rules. The objective of the theory is to understand the relationship between these components and the behavior (performance measures) of the system.

In a now-classic 1953 paper, D.G. Kendall [1] proposed the following notation to describe a queueing model: $a/b/c$, where a describes the input (arrival) process, b describes the service process, and c is the number of servers. Implicit in this description is the assumption that each arriving customer is assigned to a server immediately if one is available and holds that server for the length of time required (the "service time"); if a server is not immediately available, the customer waits in an infinite-capacity queue (or waiting room). Furthermore, the servers are all identical, and the waiting customers are served according to the FIFO (first in, first out) "queue discipline" (although the queue discipline might have no effect on the values of any particular performance measure, and therefore its omission from the explicit notation might be irrelevant).

Kendall's notation is now widely used to describe queueing models. But, clearly, there are an infinite number of possible physical configurations, operating protocols, queue disciplines, and so on, with the consequence that sometimes authors extend this notation to include those variations; that is, they treat Kendall's scheme as an algorithm that can be decoded to describe a particular model, rather than as a descriptive name of that model. For example, $a/b/c/n$ might be used to indicate that the model has a capacity of n customers, or that there are n different sources that generate the arrivals (see ***Finite Population Models—Single Station Queues***), and so on. It has become commonplace to extend Kendall's original three-symbol description to longer strings, such as $a/b/c/d/e$, where d is some measure of system capacity and e is the queue discipline. (However, sometimes even longer strings are used; clearly, this can get out of hand.) Besides FIFO, other queue disciplines include LIFO (last in, first out), LIFO-PR (last in, first out, preemptive-resume), SIRO (service in random order), SPTF (shortest processing time first), SRPT (shortest remaining processing time first), RR (round robin), PS (processor sharing), VAC (server vacations), and others. If, in any particular case, one wants to introduce Kendall-like notation to describe a particular model, then the shorthand notation should be described upfront, rather than assuming that the reader can decode its meaning; similarly, the reader should be careful in interpreting the meaning of any such notation without first ascertaining exactly what the model is. With this caveat, we describe Kendall's notation.

Wiley Encyclopedia of Operations Research and Management Science, edited by James J. Cochran

Here are the conventions:

G	*g*eneral (no particular assumption)
GI	*g*eneral *i*ndependent (the random variables in question are mutually independent and identically distributed)
M	*M*arkov or memoryless (exponential random variables)
D	*d*eterministic (the same constant value for each realization of the random variable)
E_k	k-phase *E*rlangian (sum of k independent, identical, exponentially distributed "phases")
PH	*ph*ase-type (sum of a (possibly random) number of independent, exponentially distributed phases)
MAP	*M*arkovian *a*rrival *p*rocess
$BMAP$	*b*atch *M*arkovian *a*rrival *p*rocess

Then, for example, a typical (and most important) model is $M/G/1$ (see ***The M/G/1 Queue***). This describes a model with Poisson input (M denotes independent, identically distributed, exponential interarrival times), *G*eneral service times, and 1 server. It is usually assumed implicitly that the service times are independent of each other (hence, $M/GI/1$ would be more appropriate, but is rarely used, although renewal-process input is often denoted explicitly, as in $GI/M/1$), and independent of the arrival process. It is also implicitly assumed here that there is infinite queueing capacity (so that no customer is ever "cleared" from the system because of lack of waiting space) and that the customers are served in FIFO order (although, depending on the performance measure of interest, the queue discipline might have no effect). The fact that the service times here are *G*eneral means that the formulas that describe the behavior of $M/G/1$ are to be evaluated using the distribution function of the service times as if it were a parameter. (Thus, $M/M/1$ and $M/D/1$ are important special cases.) For $M/G/1$, the celebrated Pollaczek–Khintchine formula relates the mean waiting time to the utilization of the server (typically denoted by ρ) and to the mean and variance of the distribution of service times.

Another important example is $M/G/s/s$ (see ***The M/G/s/s Queue***). Here, the second s specifies that the capacity of the system is s, which is the same as the number of servers; that is, in this model any customer who finds all servers busy does not wait, but is immediately cleared from the system. In a sense, this is not a queueing model, because there is no queue (unless one considers the customers in service to be part of the queue, which is a common convention). Some representations for the case of Poisson input and $n > 0$ waiting positions might be $M/G/s/n, M/G/s/s+n$, or $M/G/s+n$, all indicating s servers and n waiting positions, for a total system capacity of $s+n$, and, in the infinite-capacity case, the default notation $M/G/s$. Another example is $M^x/G/1$, say, in which the customers arrive in batches of size X (which may be a random variable), and the arrival times of the batches follow a Poisson process (see ***Batch Arrivals and Service—Single Station Queues***). The point is that Kendall's notation provides a convenient shorthand for describing queueing models, but the reader should be alert to avoid incorrect interpretations. With this background, most of the basic queueing models are now completely specified.

The notations *PH, MAP*, and *BMAP* refer to generalizations of the method of phases, where the key idea is to model random time intervals as composed of a (possibly random) number of exponentially distributed phases and then to exploit the simplifications arising from the resulting Markovian structure (see ***Matrix Analytic Method: Overview and History***).

The important model $M/G/s/s$ (s servers and 0 waiting positions) is often called the *Erlang B model* or the *Erlang loss model* (because customers who find all servers busy are cleared from the system and thus are lost). The formula that describes the probability of loss (the fraction of arriving customers who are cleared from the system) is typically called the *Erlang B formula* or *Erlang's first formula*. Likewise, the important model $M/M/s$ (or, $M/M/s/\infty$, s servers and infinite waiting capacity) is typically called the *Erlang C model* or the *Erlang delay model*,

and the formula that describes the probability that an arriving customer will be forced to wait until a server becomes available is often called the *Erlang C formula* or *Erlang's second formula* (see ***The M/M/s Queue***). The Erlang B and Erlang C formulas were first published in 1917 by A.K. Erlang. The context was telephone traffic (teletraffic) theory and engineering, and the terminology reflects this. The "B" in Erlang B refers to Blocking; a call is blocked if it finds all servers busy (but beware, there is not a universally accepted definition of "blocked"), in which case "lost calls are cleared." In the early teletraffic literature, the $M/G/\infty$ model is called *lost calls held*; this can be interpreted to say that the calls leave the system at the same rate whether they are in service or are waiting in the queue. (Aficionados should see Brockmeyer *et al*. [2] for reprints of Erlang's papers and interesting historical and biographical material. Syski [3] gives an invaluable authoritative and comprehensive treatment of teletraffic theory prior to 1960.) This model is now often called *Erlang A* (for Abandonment), generalizations of which can be used to describe systems (like the ubiquitous call center) in which customers renege or depart from the queue before entering service. In this context, the Kendall notation $a/b/s + M$, for example, would indicate that there are s servers and an infinite-capacity waiting room from which customers abandon after waiting an exponentially (M) distributed length of time (but beware, this could be interpreted to describe an s-server system with a waiting room of capacity M).

Two of the most important theorems in queueing theory are Little's Law [4] (see ***Little's Law and Related Results***) and PASTA [5] (see ***PASTA and Related Results***). Little's Law expresses the equality $L = \lambda W$, where L stands for expected number of customers in the "system," W stands for expected time spent by a customer in the system, and λ is the rate at which the customers enter the system. Originally, L denoted queue *L*ength, W denoted *W*aiting time in queue, and λ denoted the customer arrival rate, but these definitions are very flexible as long as one is consistent in the definition of "system." Little's Law is notable for its great generality; all that is required of L, λ, and W is, essentially, their existence.

PASTA is an acronym for Poisson Arrivals See Time Averages. In essence, it expresses the fact that if the customers arrive according to a Poisson process, then the states of the system that they observe at (just prior to) their arrival epochs (in the queueing theory vernacular, "epoch" means "instant") has the same distribution as the states observed at random epochs or averaged over time. PASTA is notable for its utility in solving queueing models (one can shift one's viewpoint as is convenient) and for its counterintuitive subtlety.

In summary, queueing theory is vast and, because of its many various application areas, its notation, while usually intuitive, is not always consistent across application areas. Finally, it is interesting and amusing to observe that almost all books and most technical journals use the spelling "queuEing," which makes it the only (ordinary) word in the English language with five successive vowels, whereas virtually all spell-checkers and nontechnical editors (in the United States) omit the second E.

REFERENCES

1. Kendall DG. Stochastic processes occurring in the theory of queues and their analysis by the method of the imbedded Markov chain. Ann Math Stat 1953;24(3):338–354.
2. Brockmeyer E, Halstrøm HL, Jensen A. The life and works of A.K. Erlang. Transactions 2. Copenhagen: Danish Academy of Technical Sciences; 1948. pp. 1–277.
3. Syski R. Introduction to congestion theory in telephone systems. 2nd ed. Amsterdam: Elsevier Science Publishers B.V. 1986 (1st ed. Edinburgh: Oliver & Boyd; 1960.)
4. Little JDC. A proof for the queuing formula: L = λW. Oper Res 1961;9(3):383–387.
5. Wolff RW. Poisson arrivals see time averages. Oper Res 1982;30(2):223–231.