Queueing Theory

U. Narayan Bhat
Southern Methodist University, Dallas, TX, USA 75275

November 4, 2015

Keywords: birth and death process; Markov chain; Markov process; matrix analytic method; queue length; steady state probability; stochastic process; transition probability; waiting line.

Queueing is essential to manage congestion in traffic of any type in the modern technological world. This does not mean it is a new phenomenon. More than one hundred years ago, recognizing its importance to telephone traffic, Danish mathematician A. K. Erlang (1909) showed for the first time how probability theory can be used to provide a mathematical model for telephone conversations. From then on, slowly in the first three decades, moderately in the next three decades, and tremendously in the last four decades, the probabilistic approach to modeling queueing phenomena has grown and contributed significantly to the technological progress. For a historical perspective of the growth of queueing theory see Chapter 1 of Bhat (2008).

Queueing theory describes probabilistically and mathematically the interaction between the arrival process of customers and the service provided to them in order to manage the system in an efficient manner. The term customer is used in a generic sense representing a unit, human or otherwise, demanding service. The unit providing service is known as the server. Some examples of a queueing system are: a communication system with voice and data traffic demanding transmission; a manufacturing system with several work stations; patients arriving at a doctor's office; vehicles requiring service; and so on.

Since the arrival process and service are random phenomena we start with a probabilistic model (also known as a stochastic model) of a queueing system. If we analyze such models using mathematical techniques we can derive its properties that can be used in understanding its behavior and managing it for its efficient use.

In order to build a probabilistic model, first we describe the arrival process (called the input process) using probability distributions. For example, the arrival of customers could be in a Poisson process; i.e. the number of customers

arriving in a set period of time has a Poisson distribution. Its parameter, say λ, gives the mean number of customers arriving during a unit time. The known distribution now identifies the arrival process. The amount of service provided by the facility is represented by a random variable since it could be random. The distribution of the random variable identifies the service process. When we talk about service we have to take into consideration the mode of service such as service provided with several servers, service provided in a network of servers, etc. Also we must include factors such as queue discipline (e.g., first come, first served (FCFS), also known as first in, first out (FIFO); last come, first served (LCFS or LIFO); group service; priority service; etc). Another factor that complicates the model is the system capacity, which may be finite or infinite.

Because of the multitude of factors involved in a queueing system, we use a three or four element symbolic representation in discussing various types of systems. The basic structure of the representation is to use symbols or numbers for the three elements: input/service/number of servers. When the system capacity is finite an additional element is added. The commonly used symbols for distributions are: M for Poisson or exponential, E_k for Erlangian with k phases (gamma distribution with an integer scale parameter k), D for deterministic, and G for a general (also GI for general independent) or an unspecified distribution. Thus M/G/1 represents a Poisson arrival, general service, and single server system, and M/G/1/N has the same description as above with a capacity restriction of N customers in the system.

When the arrival process is represented by a random variable with an index parameter t, define $A(t)$ as the number of customers arriving and $D(t)$ the number of customers leaving the system during a time period $(0, t)$. Let the number of customers in the system at time t be $Q(t)$. Then $Q(t) = A(t) - D(t)$. In order to manage the system efficiently one has to understand how the process $Q(t)$ behaves over time. Note that all $A(t), D(t)$, and $Q(t)$ are stochastic processes (which are sequences of random variables indexed by the time parameter t.) Since the total number of customers leaving the system at t is dependent on the number customers arriving during that time, the mode of their arrival (e.g. there may be time periods with no customers in the system, commonly called idle periods), the service mechanism, queue discipline (when some customers get preferred treatment) and other factors that affect the operation of the system (e.g. service breakdowns), to analyze $Q(t)$, all these factors need to be taken into account in the model.

In the analysis of a queueing system the stochastic process $W(t)$ representing the waiting time of a customer to get served, and the random variable, say B, representing the busy period (the amount of time the system is continuously busy at a stretch) are also used. The objective of the analysis is to get the distributional properties of the stochastic processes $Q(t)$ and $W(t)$ and the random variable B for use in decision making. Analyzing stochastic processes

in finite time t often becomes very complex. When the constituent elements such as arrival and service are not time-dependent we can derive the distributions of the limit random variables $Q = \lim_{t\to\infty} Q(t)$ and $W = \lim_{t\to\infty} W(t)$ when they exist. The ratio arrival rate/service rate is commonly known as the traffic intensity of the queueing system (say ρ). The property $\rho < 1$ is generally the requirement for the existence of the limit distributions of the stochastic processes $Q(t)$ and $W(t)$, when they are time-independent. The behavioral performance measures of interest in a queueing system are transition probability distributions of $Q(t)$ and $W(t)$, probability distributions of Q, W, and B, and their means and variances.

In addition to the behavioral problems of underlying stochastic processes mentioned above, we are also interested in inference problems such as estimation and tests of hypotheses regarding basic parameters and performance measures, and optimization problems for assistance in decision making. An introduction to these topics and the necessary references may be found in Bhat (2008).

In order to provide an illustration of the behavioral analysis of a queueing system we consider below a system with Poisson arrivals, exponential service, and single server, symbolically known as an M/M/1 queue. This is the simplest and the most used system in applications.

Let customers arrive in a Poisson process with rate λ. This means that the number $A(t)$ of the customers arriving in $(0, t)$ has a Poisson distribution

$$P[A(t) = j] = e^{-\lambda t}\frac{(\lambda t)^j}{j!}, \quad j = 0, 1, 2, ...$$

It also means that the interarrival times have an exponential distribution with probability density $a(x) = \lambda e^{-\lambda x}(x > 0)$. We assume the service times to have an exponential distribution with probability density $b(x) = \mu e^{-\mu x}(x > 0)$. With these assumptions we have E[inter-arrival time] $= (1/\lambda) = 1/$arrival rate and E[service time] $= (1/\mu) = 1/$service rate. The ratio of arrival rate to service rate is the traffic intensity . Note that we have assumed the processes to be time-independent.

Let $Q(t)$ be the number of customers in the system at time t and its transition probability distribution be defined as

$$P_{ij} = P[Q(t) = j|Q(0) = i]$$

Because of the Poisson arrival process and the exponential service distribution, $Q(t)$ can be modeled as a birth and death process (a class of stochastic processes with major properties (a) probability of more than one state change during an infinitesimal interval of time is close to zero; (b) the rate of change in a unit time is constant and (c) changes occurring in non-overlapping intervals of time are independent of each other) governed by the following difference-differential

equations.

$$\begin{aligned} P_{i0}(t) &= -\lambda P_{i0}(t) + \mu P_{i1}(t) \\ P_{in}(t) &= -(\lambda+\mu)P_{in}(t) + \lambda P_{i,n-1}(t) \\ &\quad +\mu P_{i,n+1}(t) \qquad n = 1, 2, ... \end{aligned} \tag{1}$$

with $P_{in}(0) = 1$ when $n = i$ and $= 0$ otherwise. Solving these equations to obtain $P_{in}(t)$ is not very simple. Readers may refer to Gross et al (2008) and its earlier editions for their solutions.

When $\rho < 1$, the limit $P_{ij}(t) = p_j$ exists and is independent of the initial state i. It is known as the steady state probability. It can be obtained easily from the following equations that result by letting $t \to \infty$ in the above set of difference-differential equations.

$$\begin{aligned} \lambda p_0 &= \mu p_1 \\ (\lambda+\mu)p_n &= \lambda p_{n-1} + \mu p_{n+1} \qquad n = 1, 2, ... \end{aligned} \tag{2}$$

along with $\sum_{n=0}^{\infty} p_n = 1$. We get $p_0 = 1-\rho, p_n = (1-\rho)\rho^n (n = 0, 1, 2, ...,)$. The mean $E(Q)$ and variance $V(Q)$ of this distribution can be obtained as $E(Q) = \rho/(1-\rho)$ and $V(Q) = \rho/(1-\rho)^2$.

The waiting time of an arriving customer, when the queue discipline is FCFS, is the total amount of time required to serve the customers who are already in the system and this total time has an Erlangian distribution. Let us denote it as T_q (we use T as the random variable representing the total time the customer is in the system, also known as total workload.) Accordingly we get

$$\begin{aligned} P[T_q \le t] &= 1 - \rho + \int_0^t \sum_{n=1}^{\infty} p_n e^{-\mu t} \frac{\mu^n t^n}{(n-1)!} dt \\ &= 1 - \rho e^{-\mu(1-\rho)t} \\ E[T_q] &= \rho/\mu(1-\rho) \text{ and } E[T] = 1/\mu(1-\rho). \end{aligned}$$

Let $E(Q) = L$ and $E(T) = W$. Looking at the above results we can see that $L = \lambda W$ showing, how L and W are related in this system. This property is known as Little's Law and it holds in more complex systems under certain general conditions. Another property is the exponential nature of the limiting waiting time distribution shown above which holds in more general queueing systems as well.

The derivation of the distribution of the busy period B is more complicated even in this simple system. We may refer the reader to Gross et al (2008) for its derivation.

When systems include more complicated features, more advanced techniques are employed to analyze resulting stochastic models. For instance, generalizing the

concepts introduced earlier, the birth and death process can be used to model a wide class of queueing systems. Using the terminology of birth for the arrival and death for the departure of customers, when there are n customers in the system, let λ_n be the rate of birth and μ_n be the rate of dealth. The birth and dealth process is governed by the three properties identified earlier. Writing $P_{in}(t) \equiv P_n(t)$ for ease of notation, the transition probabilities of the stochastic process $Q(t)$ can be shown to satisfy equations similar to (1), with λ and μ replaced by λ_n and μ_n respectively. Denoting the row vector $[P_0(t), P_1(t), ...]$ as $\boldsymbol{P(t)}$, the row vector of their derivatives as $\boldsymbol{P'(t)}$, and the transition rate matrix (also called the generator matrix)

$$\boldsymbol{A} = \begin{bmatrix} -\lambda_0 & \lambda_0 & & & \\ \mu_1 & -(\lambda_1+\mu_1) & \lambda_1 & & \\ & \mu_2 & -(\lambda_2+\mu_2) & \lambda_2 & \\ & & \ddots & & \end{bmatrix} \tag{3}$$

the generalized Eq. (1) can be written as

$$\boldsymbol{P^{'}(t) = P(t)A} \tag{4}$$

As $t \to \infty$ and when $P_n(t)$ exists, writing $P_n(t)_{\lim t\to\infty} = p_n$ and denoting the row vector of $\{p_n, n = 0, 1, 2, ...\}$ as $\boldsymbol{p}$, Eq. (4) becomes

$$0 = \boldsymbol{p\ A} \tag{5}$$

[Eq. (2) in the M/M/1 case.]
In queueing models of a wide variety of systems from computer, communication, manufacturing, and other areas of applications, birth and death process models can be used with generator matrix $\boldsymbol{A}$ of appropriate structures, sometimes with submatrices taking the place of matrix elements. The problem now consists of solving equations of the type (4) and (5).

In two classical papers Kendall (1951, 1953) introduced the imbedded Markov chain technique for the analysis of queueing systems when their arrival processes are not Poisson and/or the service time distributions are not exponential. In an M/G/1 queue the queue length process $Q(t)$ is not a Markov process for all t. But if we observe the process only at departure points, say $t_0, t_1, t_2, ..., \{Q(t_n) n = 0, 1, 2, ..., \}$ is an imbedded markov chain, the transition probablility matrix of which can be fully described when the arrival rate and the service time distribution are known. In a GI/M/1 queue, to get an imbedded Markov chain, we have to observe the system at arrival points. Then knowing the distribution of inter-arrival times and the rate of service we can specify the elements of the corresponding transition probability matrix. The analysis of the queueing systems then follow using standard techniques for the analysis of Markov chains.

Until the 1970s researchers analyzed stochastic processes underlying queueing system models using difference, differential and integral equations with the help of transforms; combinatorial relationships; and recursive solution techniques. As queueing models in applications such as computer-communication and manufacturing systems became more complex the available methods became inadequate. The need for analyzing such processes was answered with the development of the matrix analytic approach initiated by Neuts (1981) on generalized forms of matrix structures show in (3) and the transition probability matrices of the imbedded Markov chains of M/G/1 and GI/M/1, and advanced by many researchers since then. See Latouche and Ramaswami (1999).

The literature on queueing theory is vast and it is impossible to cover all facets of the analysis of queueing systems using various modeling and sophisticated mathematical techniques in a short article in an encyclopedia. One such system is the queueing network in which customers are served in various types of networks of service nodes. It has spawned a major area of research and applications relevant to systems such as computer, communication, manufacturing and transportation. Retrial, polling and vacation models are some of the models used in the analysis of individual or networks of queues.

As mentioned earlier, analysis of queueing systems involves the study of interaction between arrival and service processes under various queueing structures and queueing disciplines. These complexities generate a large number of problems in practice. For instance, if the arrival rate is larger than service rate, the system becomes unstable. Limit theorems on appropriate distributions provide the behavior of such congested systems. The properties of the tail probabilities of queue length and waiting time distributions describe the influence the basic distributions of arrival and service processes have on the system behavior. Also the complexities of the systems may require the use of mathematically more sophisticated stochastic processes such as diffusion processes.

The following references provide the basic understanding of the subject at two levels: Bhat (2008) for those who have a background only in probability and statistics at an introductory level and Gross and Harris (1998) or Gross et al. (2008) for those who have some background in stochastic processes. Bibliographies given in these books and specialized books such as Buzacott and Shanthikumar (1993) and Chen and Yao (2001) provide a good selection of references on the subject. The major source for articles in queueing theory is the journal *Queueing Systems*. Other sources are major operations research journals such as *European Journal of Operations Research*, *Management Science*, *Operations Research*, and *Stochastic Models*; and other journals in application areas such as electrical and communications engineering, industrial engineering, and manufacturing.

References

[1] Bhat, U. N. (2008). *An Introduction to Queueing Theory*, Birkhauser, Boston.

[2] Buzacott, J. A. and Shanthikumar, J. G. (1993), *Stochastic Models of Manufacturing Systems*, Prentice Hall, Upper Saddle River, NJ.

[3] Chen, H. and Yao, D. D. (2001), *Fundamentals of Queueing Networks*, Springer, New York.

[4] Erlang, A. K. (1909), The theory of probabilities and telephone conversations *Nyt Tidsskrift fur Matematik B*, **20**, 33.

[5] Gross, D. and Harris, C. M. (1998), *Fundamentals of Queueing Theory*, 3rd Ed., Wiley, New York.

[6] Gross, D., Shortle, J. F., Thompson, J. M. and Harris, C. M. (2008), *Fundamentals of Queueing Theory*, 4th Ed., Wiley, New York.

[7] Kendall, D. G. (1951), "Some problems in the theory of queues" *J. Royal Statist. Soc. B*, **13**, 151-185.

[8] Kendall, D. G. (1953), "Stochastic processes occurring in the theory of queues and their analysis by the method of imbedded Markov chains", *Ann. Math. Statist.*, **24**, 338-354.

[9] Latouche, G. and Ramaswami, V. (1999), *Introduction to Matrix Analytic Methods in Stochastic Modeling*, ASA-SIAM Series on Statistics and Applied Probability, Philadelphia, PA.

[10] Neuts, M. F. (1981), *Matrix-Geometric Solutions in Stochastic Models. An Algorithmic Approach*, The Johns Hopkins University Press, Baltimore, MD.

Acknowledgement : Based on an article from Lovric, Miodrag (2011), International Encyclopedia of Statistical Science. Heidelberg: Springer Science + Business Media, LLC.