Reviewed are fundamental principles of queuing in terms that apply to computing systems.

After laying a foundation of a minimum number of definitions, the author provides a working familiarity with extended principles and applications to system performance estimation through the use of worked out examples.

Elements of queuing theory for system design

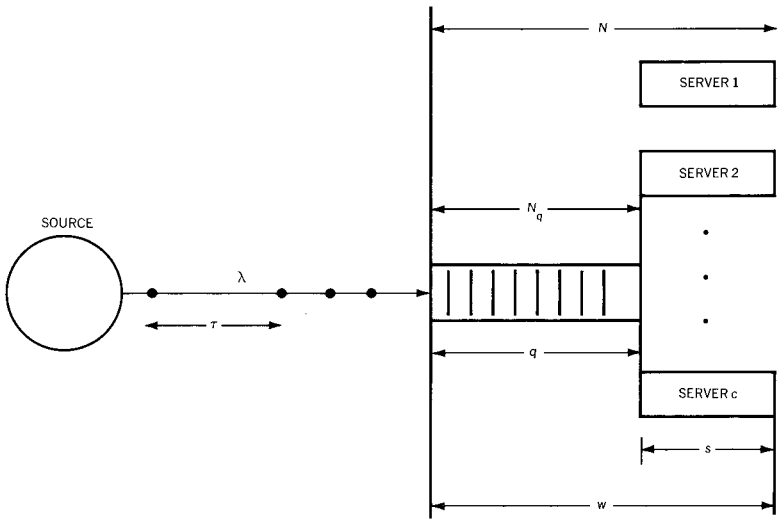
by A. O. Allen

Queuing theory, which was originally developed by the Danish mathematician A. K. Erlang for designing telephone systems, has been adapted and extended to become a useful tool in computer system design and analysis. Queuing theory can be used to predict such computer performance measures as the following: waiting time to use an on-line terminal; estimates of buffer storage requirements at message switching centers; and estimates of the effect of assigning priorities in an inquiry system. These and other topics are discussed in this paper.

A *queue* is a waiting line, and, in communications and computer engineering, *queuing theory* is the study of waiting-line and line-serving phenomena. One speaks of a queue of inquiries waiting to be processed by an on-line system. Such a queue may create still other queues. That queue may further cause a waiting line of Input/Output (I/O) requests to form, which may, in turn, generate a queue of channel requests, and so on.

Everyone is familiar with the elements of a queuing system. There is a *population* or source of potential customers, where the term "customer" means one who purchases or uses a commodity or service. In communications and computer engineering, a customer may be a message to be transmitted, an inquiry to be processed, or an I/O request. The customer desires some type of

Figure 1 Some random variables used in queuing system models



service—the transmission of a message, the processing of an inquiry, or the servicing of an I/O request—from a service facility. In the service facility, there are one or more *servers*, which are units that provide the required service for the customers. If all the servers are busy when a customer enters the system he joins a queue until a server is available.

Some random variables used in studying queuing systems are illustrated in Figure 1. Appendix 1 summarizes the queuing theory definitions used in this paper. (With a few exceptions, the notations recommended in the *Queuing Standardarization Conference Report* of May 11, 1971 issued by representatives of ORSA, AIIE, CORS, and TIMS are followed in this paper.) Certain key relationships among queuing system variables are shown in Appendix 2.

Specifications of a queue

A mathematical study of a queuing system or model requires that we discuss the following queuing specifications.

Source. The population source can be finite or infinite. A finite source system cannot have an arbitrarily long queue for service, and the number of customers in the system affects the arrival rate. In the extreme, if every customer is either waiting for or receiving service, the arrival rate drops to zero. If the source is finite but large, we assume an infinite customer population to simplify the mathematics.

Arrival process. We assume that customers enter the queuing system at times $t_0 < t_1 < t_2 \cdots t_n \cdots$. The random variables $\tau_k = t_k - t_{k-1}$ (where $k \geq 1$) are called *interarrival times.* We assume that the τ_k from a sequence of independent and identically distributed random variables, and we use the symbol τ for an arbitrary interarrival time. We specify the arrival process by giving the distribution function A of the interarrival time, $A(t) = \mathrm{P}[\tau \leq \mathrm{t}]$. The most common arrival pattern in queueing theory terminology, *random input, random arrival pattern,* or a *Poisson arrival process.* If the interarrival time distribution is exponential, that is, if $\mathrm{P}[\tau \leq t] = 1 - e^{-\lambda t}$ for each interarrival time, then the probability of n arrivals in any time interval of length t is $e^{-\lambda t}(\lambda t)^n/n!$, where $n = 0, 1, 2, \cdots$. Here λ is the average arrival rate, and the arrivals have a Poisson distribution. Other common interarrival time distributions include Erlang-k and constant distributions.[1]

Service time distribution. Let s_k be the service time required by the kth arriving customer. In this paper, the s_k are assumed to be independent, identically distributed random variables. Therefore, we can refer to an arbitrary service time as s. We also assume the common distribution function $W_s(t) = P[s \leq t]$ for service time. The most common service-time distribution in queuing theory is exponential,[1] which defines a service called *random service.* The symbol μ is reserved for average service rate, and the distribution function for random service is given by $\mathrm{W}_s(t) = 1 - e^{-\mu t}$, where $t \geq 0$. Other common service time distributions are Erlang-k and constant.[1]

A statistical parameter that is useful as a measure of the character of probability distributions for interarrival time and for service time is the *squared coefficient of variation* C_X^2, which is defined by the following equation:

$$C_X^2 = \frac{\mathrm{Var}[X]}{E[X]^2}$$

If X is a constant random variable, then $C_X^2 = 0$; if X has an exponential distribution, then $C_X^2 = 1$; and if X has an Erlang-k distribution, then $C_X^2 = 1/k$. We conclude that, for C_τ^2 nearly equal to zero, the arrival process has a regular pattern; if C_τ^2 is nearly equal to 1, the arrival process is nearly random in character; and, if C_τ^2 is greater than 1, arrivals tend to cluster. Similar statements can be made about the service time distribution, where small values of C_s^2 correspond to nearly constant services times and large values correspond to great variability in service times.

Maximum queuing system capacity. In some queuing systems, the queue capacity is assumed to be infinite. That is, every arriving customer is allowed to wait until service can be provided.

Other systems, called "loss systems," have zero waiting line capacity. That is, if a customer arrives when the service facility is fully utilized, the customer is turned away. Still other queuing systems have a positive (but not infinite) capacity.

Number of servers. The simplest queuing system is the *single-server system*, which can serve only one customer at a time. A *multiserver system* has c identical servers and can serve up to as many as c customers, simultaneously. In an *infinite-server system*, every arriving customer is immediately provided with a server.

Queue discipline. The queue discipline, sometimes called service discipline, is the rule for selecting the next customer to receive service. The most common queue discipline is "first-come, first-served," abbreviated as FCFS (or more commonly termed "first-in, first-out," and abbreviated FIFO). Another queue discipline often used is "last-come, first-served" (LCFS) or "last-in, first-out" (LIFO). "Random selection for service" (RSS) or "service in random order" (SIRO) is another queuing discipline used. Finally, we mention "priority service" (PRI).

A shorthand notation, called the Kendall notation,[2] has been developed to specify queuing systems, and has the form $A/B/c/K/m/Z$. Here A specifies the interarrival time distribution, B the service time distribution, c the number of servers, K the system capacity, m the number in the source, and Z the queue discipline. More often a shorter notation $A/B/c$ is used when there is no limit on the waiting line, the source is infinite, and the queue discipline is FIFO. The symbols used for A and B are the following:

- GI General independent interarrival time.
- G General service time, usually with the independence assumption.
- E_k Erlang-k interarrival or service time distribution.
- M Exponential interarrival or service time distribution.
- D Deterministic (constant) interarrival or service time distribution.

Thus, for example, an $M/E_4/3/20/\infty/$SIRO system has exponential interarrival time, three servers with identical Erlang-4 service time distributions, system capacity of 20 (3 in service and 17 in the queue), infinite source of customers, and service in random order (with each waiting customer having the same probability of receiving service next).

Traffic intensity. Traffic intensity u is the ratio of the mean service time $E[s]$ and the mean interarrival time $E[\tau]$. This ratio is one of the most important parameters of queuing systems and is defined by the following formula:

$$u = \frac{E[s]}{E[\tau]} = \lambda E[s] = \frac{\lambda}{\mu}$$

The traffic intensity u determines the minimum number of servers that are required to keep up with the incoming stream of customers. Thus, for example, if $E[\tau]$ is 10 seconds and $E[s]$ is 15 seconds, at least two servers are required. The unit of traffic intensity is the erlang, named after A. K. Erlang, a pioneer in queuing theory.

Server utilization. Another important parameter is the traffic intensity per server or u/c, called *server utilization* ρ when the traffic is evenly divided among the servers. Server utilization is the probability that any given server is busy, and, thus, by the Law of Large Numbers, ρ is the approximate fraction of time that every server is busy.

Probability that n customers are in the system at time t. This probability $p_n(t)$ depends not only on t, but also on the initial conditions of the queuing system, that is, the number of customers present when the service facility starts up. For most useful queuing systems, as t increases, $p_n(t)$ approaches the value p_n, which is independent on both t and the initial conditions. The system is then said to be in a steady-state condition. In this article, we consider only steady-state solutions to queuing problems because time-dependent or transient solutions are usually too complex for practical use. Also, we usually want the steady-state solution, which exists in most cases of interest.

Queuing theory provides statistical measures of queuing system performance and thus helps the systems engineer to design a minimum-cost system that provides the required level of service. These statistical measures and their variances include the following:

- Mean waiting time in the queue W_q
- Mean waiting time in the system W
- Mean number waiting for service L_q
- Mean number in the system L

These measures are not independent, and, assuming the interarrival time and service time distributions are known, the knowledge of any one of them makes it possible to calculate the other three easily from the equations of Appendix 2. Thus, if the value of W_q is computed first, then the following values are obtained:

$$L_q = \lambda W_q$$

$$W = W_q + W_s$$

$$L = \lambda W$$

Another useful performance measurement is the 90th percentile value of the response of time of the system $\pi_w(90)$, which is defined as the amount of time in the system such that 90 percent of all arriving customers spend less than this amount of time in the system. Expressed symbolically, $\pi_w(90)$ is defined by the equation $p[w \leq \pi_w(90)] = 0.9$. The 90th percentile value of time in queue $\pi_q(90)$ is similarly defined. This concept is used in Example 15 of Reference 1.

Single-server models *M/M/1*

We now consider some single-server queuing system models that are especially useful. The $M/M/1$ queuing model is widely used because the exact distributions of random variables of interest can be determined, and because they have a simple form. This model is one that has exponential interarrival and service time distributions (M/M) and a single server (1). In this respect, the $M/M/1$ system is markedly different from many queuing models for which only average or mean values and (possibly) standard deviations of the random variables of interest can be calculated. Another reason for the usefulness of the $M/M/1$ system is that it is often reasonable to assume a random arrival pattern, whereas the assumption of random service time is a conservative assumption for some queuing systems. For CPU-type service time distributions, however, the standard deviation may be much larger than the mean, and the $M/M/1$ model gives grossly optimistic predictions. The steady-state formulas for $M/M/1$ queuing system models are given in Appendix 3.

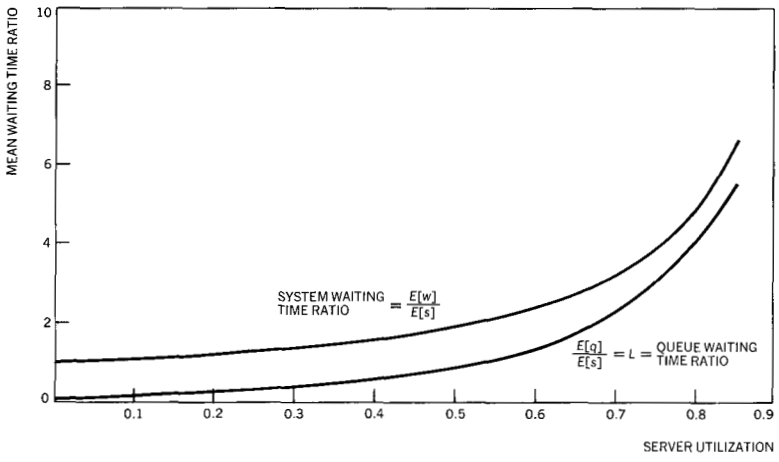
Several of the random variables for the $M/M/1$ model have a familiar form. The number of customers in the system N has a geometric distribution. The waiting time in the system or system response time w has an exponential distribution. The time that a given customer waits in the queue q has a mixed distribution that is discrete at the origin ($P[q=0] = 1-\rho$) and is continuous elsewhere. The steady-state waiting time in the queue has a distribution function that is given by the following formula:

$$W_q(t) = 1 - \rho e^{-t/E[w]}$$

valid for all $t \geq 0$.

One of the things that stands out in the formulas for the $M/M/1$ system is the highly nonlinear dependence of the random variables for the steady-state numbers of customers in the queue N_q and in the system N (plus q and w, just defined) on the server utilization ρ. Thus the mean queue waiting time ratio $E[q]/E[s]$ increases from 0.111 when $\rho = 0.1$ to 4 when $\rho = 0.8$ and to 9 when $\rho = 0.9$. The nonlinearity is illustrated in Figure 2. This figure also shows how high server utilization leads to long wait-

Figure 2 Mean system waiting time and queue waiting time ratios



ing times in the queue before receiving service, q, and to long times in the system, w, including waiting in the queue. Of course, these values increase without limit as ρ approaches 1. Figure 2 also shows that at values of ρ above 0.8 small increases in the arrival rate dramatically degrade system performance. For this reason, systems with high server utilizations are undesirable for systems without customer priorities. Properly designed priority queuing systems can function well with high server utilization. In Example 6 we examine such an example.

scaling effect

The $M/M/1$ system can be used to illustrate the "scaling effect." This effect is that given an $M/M/1$ system with mean arrival rate λ and mean service rate μ, if both λ and μ are doubled (with ρ unchanged) the effect is to halve both the mean waiting time in queue $E[q]$ and the expected or mean time spent in the system $E[w]$. The mean number waiting in queue and the mean number in the system remain unchanged. In fact, if for the new system we replace λ by $n\lambda$ and μ by $n\mu$, then we have the following scaling relationships:

$$\frac{E[q]_{\text{new}}}{E[q]_{\text{old}}} = \left(\frac{\rho/n\mu}{1-\rho}\right)\Big/\left(\frac{\rho/\mu}{1-\rho}\right) = \frac{1}{n}$$

and

$$\frac{E[w]_{\text{new}}}{E[w]_{\text{old}}} = \left(\frac{1/n\mu}{1-\rho}\right)\Big/\left(\frac{1/\mu}{1-\rho}\right) = \frac{1}{n}$$

The following argument follows an intuitively appealing line of reasoning that is illuminated by the scaling effect. If the workload of a large computer is divided equally among n smaller computers

—each with $1/n$ times the speed of the large system—then the response time does not change, and users have more conveniently located computers. The scaling effect shows, however, that the response time increases by a factor of n, on the average. Streeter[3] discusses the scaling effect more fully.

Examples of *M/M/1* queuing system models

waiting time and server utilization

Example 1. A branch office of a large engineering firm has one on-line terminal that is connected to a central computer system during the normal eight-hour working day. Engineers, who work throughout the city, drive to the branch office to use the terminal to make routine calculations. Statistics collected over a period of time indicate that the arrival pattern of people at the branch office to use the terminal has a Poisson (random) distribution, with a mean of ten people coming to use the terminal each day. The distribution of time spent by an engineer at a terminal is exponential, with a mean of thirty minutes. The branch manager receives complaints from the staff about the terminal service. It is reported that individuals often wait over an hour to use the terminal and that it rarely takes less than an hour and a half in the office to complete a few calculations. The manager is puzzled because the statistics show that the terminal is in use only five hours out of eight, on the average. This level of utilization would not seem to justify the acquisition of another terminal. What insight can queuing theory provide?

Solution. The $M/M/1$ system is a reasonable model for this system. The arrival rate is $\lambda = 10$ customers/day $= 10$ customers/day $\times$ 1/8 day/hour $\times$ 1/60 hour/minute $=$ 1/48 customers/minute. The server utilization is computed as follows:

$$\rho = \lambda E[s] = \frac{1}{48} \times 30 = \frac{5}{8} = 0.625$$

Hence, by the formulas in Appendix 3, we have

$$P[N \geq 2] = \rho^2 = 0.391$$

Probability that there are two or more customers in the queuing system

$$L = E[N] = \frac{\rho}{1-\rho} = \frac{5/8}{1-5/8} = 1.667$$

Mean steady-state number in the queuing system

$$\sigma_N = \frac{\sqrt{\rho}}{1-\rho} = 2.108$$

Standard deviation of the number of customers in the system

$W = E[w] = \frac{E[s]}{1-\rho} = 80$ minutes	Mean time that a customer spends in the system
$\sigma_w = E[w] = 80$ minutes	Standard deviation of the time a customer spends in the system
$L_q = \frac{\rho^2}{1-\rho} = 1.04$	Mean steady-state number of customers in the queue
$E[N_q \mid N_q > 0] = \frac{1}{1-\rho} = 2.67$	Mean steady-state queue length of nonempty queues
$W_q = E[q] \frac{\rho E[s]}{1-\rho} = 50$ minutes	Mean time in the queue
$E[q \mid q > 0] = E[w] = 80$ minutes	Mean time in the queue for those who must wait
$\pi_q(90) = E[w]\log(10\rho)$ $= 80 \times 1.8326 = 146.6$ minutes	Ninetieth percentile of the time in the queue
$\pi_w(90) = 2.3\ E[w] = 184$ minutes	Ninetieth percentile of the time in the system

The overall average waiting time to use the terminal, which includes those engineers who do not wait at all, is fifty minutes. However, the average wait for those who must wait is one hour and twenty minutes, a very long wait for most people. The 90th percentile time in the office is 184 minutes, and the probability of spending ninety minutes or more in the office is shown to be nearly one-third as follows:

$$1 - W(90) = 1 - (1 - e^{-9/8}) = e^{-9/8} = 0.325$$

Thus nearly one-third of the engineers must spend more than an hour and a half in the office to achieve about half an hour of useful work, and ten percent of them require over three hours. The probability that an engineer must wait to use the terminal is $\rho = 0.625$, and the probability of his waiting more than an hour is

$$\begin{aligned} P[q > 60] &= 1 - P[q \leq 60] \\ &= 0.625e^{-1/30 \times 3/8 \times 60} = 0.625e^{-3/4} \\ &= 0.2952 \end{aligned}$$

These conclusions may seem a little startling to those who have not been exposed to queuing theory because the utilization of the

terminal seems low. Nevertheless, as Figure 2 shows, the mean waiting time in the queue $E[q]$ grows rapidly with server utilization ρ. If two terminals are provided in the branch office, the mean waiting time decreases to 3.25 minutes, with a 90th percentile value of 8.67 minutes.

single server with a limited number of customers

Example 2. Messages arrive at a switching center for a communication line in a random pattern, with a mean arrival rate of 240 messages per hour. Message lengths are distributed approximately exponentially, with a mean of 150 characters. The transmission time for a message is directly proportional to its length, and the line speed is 15 characters per second. Assuming that a very large message buffer is provided, find the following system characteristics: mean number of messages waiting L_q; mean waiting time in the queue W_q; mean waiting time for messages that are delayed $E[q|q > 0]$: mean number of messages in the system L; mean system time $E[w]$; 90th percentile waiting time in the queue $\pi_q(90)$; and the 90th percentile of total system time $\pi_w(90)$.

Solution. The formulas in Appendix 3 for an $M/M/1$ system apply to the situation in Example 2. Here we have a message arrival rate of $\lambda = 240$ messages/hour $= 1/15$ message/second; a mean service time of $E[s] = 150/15 = 10$ seconds; and a server utilization of $\rho = \lambda E[s] = \frac{10}{15} = \frac{2}{3}$. Given these specifications we can calculate the following system characteristics:

$L_q = E[N_q] = \dfrac{\rho^2}{1-\rho} = 1.33$ messages	Mean steady-state number of customers in the queue
$W_q = E[q] = \dfrac{\rho E[s]}{1-\rho} = 20$ seconds	Mean time in the queue
$E[q\|q > 0] = \dfrac{E[s]}{1-\rho} = 30$ seconds	Mean time in the queue for messages that must wait
$L = E[N] = \dfrac{\rho}{1-\rho} = 2$ messages	Mean steady-state number of messages in the system
$W = E[s] = \dfrac{E[s]}{1-\rho} = 30$ seconds	Mean steady-state time in the system
$\pi_q(90) = E[w]\log(10\rho) = 56.91$ seconds	Ninetieth percentile of time in the queue

$\pi_w(90) = 2.3\ E[w] = 69$ seconds — Ninetieth percentile of time in the system

The system of Example 2 has a large, but not unlimited, buffer. The queuing theory model that fits this system better is the $M/M/1/K$ model, that is, an $M/M/1$ system with a limit of K customers in the system. The equations for this model are shown in Appendix 4. This system is stable, moreover, even when the mean arrival rate λ exceeds the mean service rate μ because customers are turned away when the system is full.

Example 3. Consider Example 2, again. Suppose the switching center were desired to provide a sufficiently large buffer that an arriving message would be turned away less than five per cent of the time. What message buffer capacity should be provided?

Solution. The traffic intensity u is 2/3. By Appendix 4, if 2 messages can be stored in the buffer ($K = 3$), then p_3 is the probability that an arriving message is turned away, or

$p_3 = \dfrac{(1-u)u^K}{1-u^{K+1}} = 0.123$ — Steady-state probability that there are three messages in the system

$p_4 = 0.076$

$p_5 = 0.048$ — Steady-state probabilities that there are 4 and 5 messages in the system

Thus the buffer must provide storage for at least four messages. The formulas of Appendix 4 also show that, for a system with sufficient buffer storage for four messages the following conditions prevail: mean number of messages waiting is 0.788; mean time a message waits for transmission is 12.417 seconds; mean waiting time for messages that are delayed is 19.57 seconds; mean time a message spends in the system is 22.417 seconds; and the probability that no messages are in the system is 0.3654.

Examples of *M/G/1* queuing system models

The most useful classical single-server queuing model is the $M/G/1$ system. For this model, however, one cannot generally obtain the distribution functions of N, N_q, w, and q as is possible for the $M/M/1$ model. For $M/G/1$ queuing system models, one usually obtains mean values. However, if the first three moments of the service time are known, then both the mean and standard deviation can be obtained for some of the random variables.

Formulas for $M/G/1$ queuing systems are given in Appendix 5. For this model, the service times must be independent, but most authors write $M/G/1$ rather than $M/GI/1$. The quantities $E[q^2]$,

σ_q^2, $E[w^2]$, σ_w^2 and σ^2 can be calculated only if the first three moments of the service time are known. The equations for the mean steady-state number in the queuing system L and W are commonly referred to as the Pollaczek-Khintchine[4,5] equations. For a given value of mean service time $E[s]$, minimum values of L, L_q, W and W_q occur when the variation of the service time is zero $\text{Var}[s] = 0$, that is, when the service time s is constant.

Example 4. A message processing center processes each incoming message in exactly 10 seconds. Messages arrive in a random pattern at a mean rate of 4 messages per minute. Calculate the descriptive statistics for the deterministic or constant service-time model $M/D/1$ as a special case of the $M/G/1$ model. Then estimate the 90th and 95th percentile values of the time in the system, w, and the time waiting in the queue before receiving service, q.

Solution. For a constant service time $E[s^2] = E[s]^2$, $E[s_s^3] = E[s_s]^3$ etc. Thus, the first three moments of service time are 10, 100 and, 1000, respectively. The formulas of Appendix 5 give the following quantities:

$\rho = \frac{4}{60} \times 10 = \frac{2}{3}$	Server utilization
$L = \frac{4}{3}$	Mean steady-state number in the system
$\sigma_N = 1.44$	Standard deviation of N
$L_q = \frac{2}{3}$	Mean steady-state number in the queue, not including those in service
$W = 20$ seconds	Mean steady-state time in the system
$\sigma_w = 12.91$ seconds	Standard deviation of w
$W_q = 10$ seconds	Mean steady-state time in the queue
$\sigma_q = 12.91$	Standard deviation of q

The queuing model for this example differs from that of Example 2 only in the service time distribution. The steady-state mean waiting time in the queue here is exactly one-half of that for Example 2, where the service time distribution is random. A general result for $M/G/1$ models with a given mean service time is that the waiting times for constant service time are one-half that for random service, and the waiting times for all Erlang service-time distributions fall somewhere in between these extremes. There are, of course, service time distributions that lead to longer waiting times in the queue and in the system than random service yields. Hyperexponential service time is one example of such a distribution.

Table 1 Percentile values for the Erlang-k distribution

	$\pi_X(r)/E[X]$							
	k							
$r/100$	1	2	3	4	5	10	20	100
0.90	2.31	1.95	1.78	1.68	1.60	1.43	1.30	1.13
0.95	3.00	2.38	2.10	1.94	1.84	1.58	1.40	1.17
0.99	4.61	3.32	2.81	2.52	2.33	1.88	1.60	1.25

Martin[6] gives a rule that the 95th percentile of response time is approximately the mean plus two standard deviations. In this paper, Martin's rule has been extended to give the 90th percentile as mean plus 1.3 standard deviations. Although Martin stated the rule for system response time, it also approximates the waiting time in queue. By Martin's rule the percentiles are as follows:

$$\pi_w(90) = 20 + 1.3 \times 12.91 = 36.8 \text{ seconds}$$

$$\pi_w(95) = 20 + 2 \times 12.91 = 45.8 \text{ seconds}$$

$$\pi_q(90) = 10 + 1.3 \times 12.91 = 26.8 \text{ seconds}$$

$$\pi_q(95) = 10 + 2 \times 12.91 = 35.8 \text{ seconds}$$

Since the squared coefficient of variation is less than one, i.e., $C_w^2 = 0.4167 < 1$, another method of estimating percentile values for total time in the queuing system w is to use an Erlang-k distribution, and obtain the percentile values from Table 1. If k is the largest integer less than or equal to $1/C_w^2$, that is, C is the floor of $(1/C_w^2) = 2.4$, which is 2, then the 90th and 95th percentiles of the total time in the queuing system are as follows:

$$\pi_w(90) = 1.95\ E[w] = 39 \text{ seconds}$$

$$\pi_w(95) = 2.38\ E[w] = 47.6 \text{ seconds}$$

These values are conservative because the reciprocal of C_w^2—which we had approximated by 2—is 2.4. Since the Erlang-k distribution is a special case of the *gamma distribution* in which the parameter k is not restricted to integer values, we can use a linear interpolation of the gamma distribution to calculate the 90th and 95th percentiles of the total time in the queuing system as follows:

$$\pi_w(90) = [1.78 + 0.6 \times (1.95 - 1.78)] \times 20$$

$$= 1.88 \times 20 = 37.6 \text{ seconds}$$

and

$$\pi_w(95) = 2.27 \times 20 = 45.4 \text{ seconds}$$

Table 2 Percentile values for the gamma distribution

	$\pi_X(r)/E[X]$												
	C_X^2												
$r/100$	1.25	1.5	1.75	2.0	2.5	3.0	4.0	5.0	6.0	8.0	10.0	20.0	100
0.90	2.44	2.54	2.63	2.71	2.83	2.91	3.01	3.03	3.01	2.87	2.67	1.53	0.0016
0.95	3.25	3.47	3.67	3.85	4.16	4.42	4.85	5.16	5.39	5.68	5.81	5.16	0.34
0.99	5.17	5.69	6.18	6.64	7.51	8.3	9.74	11.02	12.17	14.18	15.89	11.02	26.51

These values should be sufficiently accurate for design purposes because the values obtained by fitting a gamma distribution with exactly the same mean and variance as w yields the following total times in the system:

$\pi_w(90) = 37.3$ seconds

and

$\pi_w(95) = 44.8$ seconds

A gamma random variable, for which the Erlang-k random variable is a special case, is useful for estimating percentile values. A gamma random variable with parameters α and λ has the following density function:

$f(x) = \frac{\lambda(\lambda x)^{\alpha-1}}{\Gamma(\alpha)} \cdot e^{-\lambda x}$ where $x > 0$, and $\Gamma()$ is the gamma function.

The mean is α/x, and the variance is α/λ^2. The random variable w can often be approximated by a gamma random variable with the same mean and variance, as can the random variable q.

Table 2 can be used to estimate percentile values for a random variable X with $C_X^2 > 1$. The use of this table is equivalent to fitting a gamma distribution to X, which has the same mean and variance as X. Since $C_q^2 = 1.67$, Table 2 (with linear interpolation) gives the 90th and 95th percentile values of waiting time in the queue as follows:

$\pi_q(90) = 2.6\ E[q] = 26$ seconds

$\pi_q(95) = 3.61\ E[q] = 36.1$ seconds

These values are close to the values obtained using Martin's rule, but they should be more precise estimates.

The one-sided inequality discussed in my article on probability for system design[1] can also be used to estimate percentile values, but that formula tends to give conservative estimates. By the one-sided inequality, for any random variable X, the estimate of

the rth percentile for $r > 50$ is expressed as follows:

$$E[X] + \sqrt{\frac{r}{100 - r}}\,\sigma_X$$

which gives the 90th and 95th percentiles of X as follows:

$$\pi_X(90) = E[X] + 3\sigma_X$$

$$\pi_X(95) = E[X] + 4.36\sigma_X$$

$M/E_k/1$ queuing system example

The $M/E_k/1$ system is an important special case of the $M/G/1$ system because many service time distributions can be approximated by an Erlang distribution. In addition, since we know all the moments of the Erlang-k distribution, we can calculate the standard deviation, as well as the mean, for the random variables, q, w, and N. Using the means and standard deviations we can estimate 90th and 95th percentile values. Reference 7 gives graphs of L, σ_N, $W/E[s]$, and $\sigma_w/E(s)$ as a function of ρ for $k = 1$, 2, 5 and ∞. (The notation used in Reference 7, however, differs from that which is used in this paper.)

The calculations for the $M/E_k/1$ system can be completed by substituting the following mean squared and cubed values of the service time:

$$E[s^2] = \frac{(k+1)(E[s])^2}{k}$$

and

$$E[s^3] = \frac{(k+1)(k+2)}{k^2}(E[s])^3$$

into equations for L_q, $E[q^2]$, and σ_q^2 of Appendix 5. Percentile estimates for q, w, and N can then be made either by Martin's rule or by calculating the squared coefficient of variation and using either Table 1 or Table 2.

Example 5. In Example 2, suppose that all the other parameters are the same, but the message length distribution has approximately an Erlang-4 distribution. Calculate the $M/G/1$ statistics and estimate the 90th percentile values for q, w, and N.

Solution.

$$E[s] = \frac{150}{15} = 10 \text{ seconds}$$

$$\lambda = \frac{1}{15} \text{ messages per second}$$

$$\rho = \lambda E[s] = 2/3$$

We calculate

$$E[s^2] = 5/4 \times 10^2 = 125$$

$$E[s^3] = \frac{5 \times 6}{4^2} \times 10^3 = 1875$$

Substituting these values into the equations of Appendix 5 gives the following statistics:

$$L_q = \frac{\lambda^2 E[s^2]}{2(1-\rho)} = 0.833 \text{ messages}$$

$$W_q = L_q/\lambda = 12.5 \text{ seconds}$$

$$E[q|q > 0] = W_q/\rho = 18.75 \text{ seconds}$$

$$E[q^2] = 437.5$$

$$\sigma_q^{\ 2} = 437.5 - 12.5^2 = 281.25$$

$$\sigma_q = 16.77$$

$$L = L_q + \rho = 1.5 \text{ messages}$$

$$W = L/\lambda = 22.5 \text{ seconds}$$

$$E[w^2] = 812.5$$

$$\sigma_w^{\ 2} = 812.5 - 22.5^2 = 306.25$$

$$\sigma_w = 17.5 \text{ seconds}$$

$$\sigma_N^{\ 2} = 2.86$$

$$\sigma_N = 1.69 \text{ seconds}$$

From these values can be obtained the following Martin's-rule estimates:

$$\pi_q(90) = 12.5 + 1.3 \times 16.77 = 34.3 \text{ seconds}$$

$$\pi_w(90) = 22.5 + 1.3 \times 17.5 = 45.25 \text{ seconds}$$

$$\pi_N(90) = 1.5 + 1.3 \times 1.69 = 3.7 \text{ messages}$$

Since

$$C_q^{\ 2} = \frac{281.25}{12.5^2} = 1.8$$

$$C_w^{\ 2} = \frac{306.25}{22.5^2} = 0.605$$

and

$$C_N^{\ 2} = \frac{2.86}{1.5^2} = 1.27$$

we can use Table 2 to estimate the 90th percentiles of the waiting time in the queue and the number of customers in the system as follows:

$\pi_q(90) = 2.65 \times 12.5 = 33.125$ seconds

$\pi_N(90) = 2.45 \times 1.5 = 3.68$ messages

Using Table 1 and linear interpolation, we estimate the 90th percentile of the total time spent in the queuing system as follows:

$\pi_w(90) = 2.08 \times 22.5 = 46.8$ seconds

since

$1/C_w^2 = 1.65$

If we fit a gamma distribution to the total time in the system w with the same mean and standard deviation, we would compute $\pi_w(90)$ to be 45.8 seconds. Thus both the Martin's-rule estimates and the estimates using the tables with linear interpolation give fairly precise percentile values.

A comparison of the statistics for Example 2 shows that the system performance is much better with Erlang-4 service than with exponential service. This is due to the smaller variance of the Erlang-4 service.

Priority queues

In many queuing systems customers are divided into priority classes, say from 1 to n, where the lower the priority class number, the higher the priority. Thus, customers of priority class i are given preference over customers in priority class j if $i < j$, and customers in priority class 1 have preference over all other customers. Customers within the same priority class are served in order of arrival (FIFO).

There are two basic control policies for the situation wherein a customer of the ith class arrives to find a customer of the jth class in service ($i < j$), called preemptive priority and nonpreemptive priority, respectively. In a *preemptive* priority queuing system, service is interrupted and the newly arrived customer with higher priority begins service. As a further refinement, if the preemptive system is a *preemptive-resume* priority system, the lower-priority customer, whose service was interrupted, begins service at the point of interruption upon the next access to the service facility. In still another variation, a *preemptive-repeat* priority system, the lower-priority customer repeats his entire service from the beginning.

In a *nonpreemptive* priority queuing system, the newly arrived customer waits until the customer in service completes service.

Then he is allowed access to the service facility. Such a system is also called a "head-of-line" system, abbreviated HOL.

For our models we assume that each priority class has a Poisson arrival pattern with parameter λ_i and a general service time with the mean value $1/\mu_i$. Appendix 6 gives the formulas for the $M/G/1$ nonpreemptive (HOL) priority queueing system. Appendix 7 gives the formulas for the $M/G/1$ preemptive-resume priority queueing system.

Example 6. An on-line computer system processes inquiries of three basic types, each of which has an independent random arrival pattern. Type 1 inquiries arrive at the rate of 0.5 per second, and have a constant processing (service) time of 0.5 seconds. Thus, $\lambda_1 = 0.5$, $E[s_1] = 0.5$, and $E[s_1^2] = 0.25$.

Type 2 inquiries arrive at the rate of 0.1 per second and have a processing time that is exponentially distributed, with a mean of 2 seconds. Thus $\lambda_2 = 0.1$, $E[s_2] = 2$, and $E[s_2^2] = 8$.

Type 3 inquiries arrive at the rate of 0.03 per second, and have an Erlang-5 processing time, with a mean of 5 seconds. Thus $\lambda = 0.03$ $E[s_3] = 5$, and $E[s_3^2] = 30$. Here we have used the fact that, for an Erlang-k distribution of service time, $E[s^2] = ((1 + k)/k)E[s]^2$. Compare the efficiency of the system under the following conditions:

a. $M/G/1$ system with no priority classes.
b. $M/G/1$ nonpreemptive priority system with priority classes numbered in the order listed, that is, with preference in inverse order of mean processing time
c. $M/G/1$ preemptive-resume priority system, with priority classes the same as those in b.

Solution.

a. No priority classification

$\lambda = \lambda_1 + \lambda_2 + \lambda_3 = 0.63$ inquiries per second

$$E[s] = \frac{\lambda_1}{\lambda} E[s_1] + \frac{\lambda_2}{\lambda} E[s_2] + \frac{\lambda_3}{\lambda} E[s_3] = 0.95238 \text{ seconds}$$

Thus

$$\rho = \lambda E[s_3] = 0.6$$

$$E[s^2] = \sum_{i=1}^{3} \frac{\lambda_i}{\lambda} E[s_i^2] = 2.896825$$

$$E[q] = \frac{\lambda E[s^2]}{2(1-\rho)} = 2.28125 \text{ seconds}$$

$E[w] = E[q] + E[s] = 3.2336$ seconds

Using the following formula,

$E[w_j] = E[q] + E[s_j]$ where $j = 1, 2, 3$

we see that the mean times in the system for the three classes of inquiries are 2.71825 seconds, 4.28125 seconds, and 7.28125 seconds, respectively.

b. For a nonpreemptive priority system, the formulas of Appendix 6 give the following mean values of waiting time in the queue and total time in the queuing system:

$E[q] = 1.51425$ seconds

$E[w] = 2.46631$ seconds

The mean waiting times in queue $E[q_i]$ for the inquiries in the three classes are as follows:

Class 1. 1.2167 seconds
Class 2. 2.2121 seconds
Class 3. 4.1477 seconds

respectively. The corresponding mean total times in the queuing system E[w] are given as follows:

Class 1. 1.7167 seconds
Class 2. 4.2121 seconds
Class 3. 9.1477 seconds

Thus, we have seen a significant improvement in the overall system performance. Both mean waiting time in the queue and mean total time in the system are much less than they are for the nonpriority system.

c. For the preemptive-resume priority system, the equations of Appendix 7 give the following mean values of waiting time in the queue and total time in the system:

$E[q] = 0.74224$ seconds

$E[w] = 1.69462$ seconds

The average response times for the three priority classes are:

Class 1. 0.5833 seconds
Class 2. 3.7879 seconds
Class 3. 13.2386 seconds

Thus the net effect of replacing the system under **a** conditions, with no priorities, by the preemptive-resume priority system **c**

is to reduce the mean time in the queue (from 2.28125 seconds to 0.74224 seconds); to reduce the mean time in the system to about one half its previous value (from 3.2336 seconds to 1.69462); and to decrease the mean time in the system for priority *Classes 1 and 2*. The penalty is that the mean time in the queue for *Class 3* is approximately doubled. Since the messages of the third priority class comprise less than five percent of the messages, this should not cause any serious problems.

It is not difficult to show that, if the priority classes are set up to favor the customers with smallest mean service times, then the mean time in the system decreases, as we have observed in Example 6.

Many computing centers use this technique to give better overall customer service. On the other hand, if, for some reason, a substantial proportion of customers with large service times must be favored, then the overall system performance suffers. For example if the priority classes of Example 6 had been set up so that the longest messages were serviced first, then, for the preemptive-resume model, the mean queuing time would increase to 3.3245 seconds, and the mean time in the system would increase to 4.2769 seconds. (These values are only 2.28125 seconds and 3.2336 seconds, respectively, with no priority classes.)

Concluding remarks

The practical examples of the use of queuing models in computer system design and analysis have introduced basic principles and applications of queuing theory. These principles may serve as a useful introduction or review. The cited references and general references may be used to build upon this foundation by their presentations of more complex models, such as those that analyze machine interference and multiserver queues.

ACKNOWLEDGMENTS

The author thanks Gerald K. McAuliffe for helpful discussions of this paper and for providing some useful illustrative programs. Appreciation also goes to H. G. Fisher for his reading the original manuscript and suggesting improvements. The author appreciates the excellent comments and suggestions provided by the reviewers.

CITED REFERENCES

1. A. O. Allen, "Elements of probability for system design," *IBM Systems Journal* **13,** No. 4, 325–348 (1974).
2. D. G. Kendall, "Stochastic processes occurring in the theory of queues and their analysis by the method of imbedded Markov chains," *The Annals of Mathematical Statistics* **24,** 338–354 (1953).

3. D. N. Streeter, "Centralization or dispersion of computing facilities," *IBM Systems Journal* **12,** No. 3, 283–301 (1973).
4. F. Pollaczek, "Uber das Warteproblem," *Mathematische Zeitschrift* **38,** 429–537 (1934).
5. A. Y. Khintchine, "Mathematical theory of a stationary queue," *Mathematicheskii Schornik* **39,** No. 4, 73–84 (1932).
6. James Martin, *Systems Analysis for Data Transmission*, Prentice Hall, Englewood Cliffs, New Jersey (1972).
7. *Analysis of Some Queuing Models in Real-Time Systems*, Form GF20–0007, IBM Corporation, Data Processing Division, White Plains, New York 10604.
8. J. D. C. Little, "A proof for the queuing formula $L = \lambda W$," *Operations Research* **9,** 383–387 (1961).

GENERAL REFERENCES

W. Chang, "Single-server queuing processes in computer systems," *IBM Systems Journal* **9,** No. 1, 36–71 (1970).

W. Chang, "Computer interference analysis," *IBM Journal of Research and Development* **17,** No. 1, 13–26 (1973).

P. M. Morse, *Queues, Inventories and Maintenance*, John Wiley and Sons, Inc., New York, New York (1958).

D. R. Cox, W. L. Smith, *Queues*, Methuen, London (1961).

A. M. Lee, *Applied Queuing Theory*, The Macmillan Company, New York, New York (1966).

R. B. Cooper, *Introduction to Queuing Theory*, The Macmillan Company, New York, New York (1972).

D. Gross and C. M. Harris, *Fundamentals of Queuing Theory*, John Wiley and Sons, Inc., New York, New York (1974).

L. Kleinrock, *Queuing Systems Volume 1: Theory*, John Wiley and Sons, Inc., New York, New York (1975).

Appendix 1: Queuing notation and definitions

$A(t)$	Distribution function of interarrival time, $A(t) = P[\tau \leq t]$.
c	Number of identical servers.
D	Deterministic (constant) interarrival- or service-time distribution.
E_k	Erlang-k distribution of interarrival or service time.
$E[N_q \mid N_q > 0]$	Mean (expected or average) steady-state queue length of nonempty queues.
$E[q \mid q > 0]$	Mean steady-state time in queue for nonempty queues.
FCFS	First come, first served queuing discipline.
FIFO	First in, first out queuing discipline (identical with FCFS).
G	General probability distribution of service time, with independence usually assumed.

GI	General independent interarrival time distribution, sometimes used to describe the service time distribution.
K	Maximum number allowed in the queuing system, including both those waiting for service and those receiving service.
L	$E[N]$, the mean steady-state number in the queuing system.
L_q	$E[N_q]$, mean steady-state number in the queue, not including those in service.
LCFS	Last come, first served queuing discipline.
LIFO	Last in, first out queuing discipline (identical to LCFS).
λ	Mean arrival rate to the queuing system.
M	Exponential interarrival- or service-time distribution.
μ	Mean service rate per server.
$N(t)$	Random variable describing the number in the queuing system at time t.
N	Random variable describing the steady-state number in the queuing system.
$N_q(t)$	Random variable describing number in the queue (excluding those in service) at time t.
N_q	Random variable describing the steady-state number in the queue.
$N_s(t)$	Random variable describing number receiving service at time t.
N_s	Random variable describing the steady-state number receiving service.
$p_n(t)$	Probability that there are n customers in the queuing system at time t.
p_n	Steady-state probability that there are n customers in queuing system.
PRI	Priority queuing discipline.
q	Random variable describing the time a customer spends in the queue (waiting line) before receiving service.
RSS	Random selection for service queuing discipline.

ρ	Server utilization $\rho = \lambda/c\mu$.
s	Random variable describing the service time.
SIRO	Service in random order queuing discipline (identical with RSS). Each waiting customer has the same probability of being served next.
τ	Random variable describing interarrival time.
u	Traffic intensity $u = \dfrac{E[s]}{E[\tau]} = \lambda E[s] = \dfrac{\lambda}{\mu}$.
w	Random variable describing the total time a customer spends in the queueing system, including both the time spent in the queue waiting for service and the service time.
$W(t)$	Distribution function for w, $W(t) = P[w \leq t]$.
W	$E[w]$, mean steady-state time in the system, including both time in the queue and service time.
$W_q(t)$	Distribution function for the time in the queue, $Wq(t) = P[q \leq t]$.
W_q	E[q], mean steady-state waiting time in the queue excluding service time.
$W_s(t)$	Distribution function for service time, $W_s(t) = P[s \leq t]$.
W_s	$E[s]$, mean service time, $W_s = 1/\mu$.

Appendix 2: Relationships used in queuing theory models

u	$= \dfrac{E[s]}{E[\tau]} = \lambda E[s] = \dfrac{\lambda}{\mu}$	
ρ	$= u/c$	
w	$= q + s$	
W	$= E[w] = E[q] + E[s] = W_q + W_s$	
$N(t)$	$= N_q(t) + N_s(t)$	
N	$= N_q + N_s$	
L	$= E[N] = \lambda W = E[N_q] + E[N_s]$	Little's formula.[8]
L_q	$= E[N_q] = \lambda W_q$	Little's formula.[8]

Appendix 3: Steady-state formulas for M/M/1 queuing system models

$$P_n = P[N = n] = (1 - \rho)\rho^n \qquad n = 0, 1, 2, \cdots$$

$$P[N \geq n] = \sum_{k=n}^{\infty} p_k = \rho^n \qquad n = 1, 2, 3, \cdots$$

$$L = E[N] = \frac{\rho}{1-\rho}$$

$$\sigma_N^2 = \frac{\rho}{(1-\rho)^2}$$

$$W(t) = P[w \leq t] = 1 - e^{-\mu(1-\rho)t} = 1 - e^{-t/E[w]}$$

$$W = E[w] = \frac{E[s]}{1-\rho} = \frac{1}{\mu(1-\rho)}$$

$$\sigma_w^2 = E[w]^2$$

$$L_q = E[N_q] = \frac{\rho^2}{1-\rho}$$

$$\sigma_{N_q}^2 = \frac{\rho^2(1+\rho-\rho^2)}{(1-\rho)^2}$$

$$E[N_q | N_q > 0] = \frac{1}{1-\rho}$$

$$\text{Var}\,[N_q | N_q > 0] = \frac{\rho}{(1-\rho)^2}$$

$$W_q(t) = P[q \leq t] = 1 - \rho e^{-\mu(1-\rho)t} = 1 - \rho e^{-t/E[w]}$$

$$W_q = E[q] = \frac{\rho E[s]}{1-\rho}$$

$$\sigma_q^2 = \frac{(2-\rho)\rho(E[s])^2}{(1-\rho)^2}$$

$$E[q | q > 0] = \frac{E[s]}{1-\rho}$$

$$\text{Var}\,[q | q > 0] = \left(\frac{E[s]}{1-\rho}\right)^2$$

$$\pi_w(r) = \frac{E[s]}{1-\rho} \log\left(\frac{100}{100-r}\right) = E[w] \log\left(\frac{100}{100-r}\right)$$

$$\pi_w(90) = 2.3\ E[w]$$

$$\pi_w(95) = 3E[w]$$

$$\pi_q(r) = E[w] \log\left(\frac{100\rho}{100-r}\right) = \frac{E[q]}{\rho} \log\left(\frac{100\rho}{100-r}\right)$$

$$\pi_q(90) = E[w] \log\,(10\rho),\ \pi_q(95) = E[w] \log\,(20\rho)$$

Appendix 4: Formulas for *M/M/1* queuing system models with a limited waiting line (*M/M/1/K* models where $K \geq 1$ and $N \leq K$)

$$P_n = \begin{cases} \dfrac{(1-u)u^n}{1-u^{K+1}} \text{ if } \lambda \neq \mu \text{ and } n = 0, 1, \cdots, K \\ \text{or} \\ \dfrac{1}{K+1} \text{ if } \lambda = \mu \text{ and } n = 0, 1, \cdots, K \end{cases}$$

P_K = Probability that an arriving customer is turned away.

$\lambda_a = (1 - p_K)\lambda$, where λ_a is the actual arrival rate at which customers enter the system.

$$L = E[N] = \begin{cases} \dfrac{u[1-(K+1)u^K + Ku^{K+1}]}{(1-u)(1-u^{K+1})} \text{ if } \lambda \neq \mu \\ \text{or} \\ \dfrac{K}{2} \text{ if } \lambda = \mu \end{cases}$$

$W = E[w] = L/\lambda_a$

$L_q = E[N_q] = L - (1 - p_0)$

$W_q = E[q] = L_q/\lambda_a$

$E[q|q > 0] = W_q/(1 - p_0)$

$\rho = (1 - p_K)u$ True server utilization, that is, the traffic intensity experienced by the server.

Appendix 5: Formulas for M/G/1 queuing system models

$$L_q = E[N_q] = \frac{\lambda^2 E[s^2]}{2(1-\rho)} = \frac{\lambda^2 \sigma_s^2 + \rho^2}{2(1-\rho)}$$

$W_q = E[q] = L_q/\lambda$

$E[q|q > 0] = W_q/\rho$

$$E[q^2] = \frac{\lambda E[s^3]}{3(1-\rho)} + \frac{1}{2}\left(\frac{\lambda E[s^2]}{(1-\rho)}\right)^2$$

$\sigma_q^2 = E[q^2] - W_q^2$

$L = E[N] = L_q + \rho$

$W = E[w] = L/\lambda$

$$E[w^2] = E[q^2] + \frac{E[s^2]}{1-\rho}$$

$\sigma_w^2 = E[w^2] - W^2$

$$\sigma_N^2 = \frac{\lambda^2 E[s^3]}{3(1-\rho)} + \left(\frac{\lambda^2 E[s^2]}{2(1-\rho)}\right)^2 + \frac{\lambda^2(3-2\rho E[s^2])}{2(1-\rho)} + \rho(1-\rho)$$

Appendix 6: Formulas for an *M/G/1* nonpreemptive priority queuing system

$$\lambda = \lambda_1 + \lambda_2 + \cdots + \lambda_n$$

$$E[s] = \frac{\lambda_1}{\lambda} E[s_1] + \frac{\lambda_2}{\lambda} E[s_2] + \cdots + \frac{\lambda_n}{\lambda} E[s_n]$$

$$E[s^2] = \frac{\lambda_1}{\lambda} E[s_1^2] + \frac{\lambda_2}{\lambda} E[s_2^2] + \cdots + \frac{\lambda_n}{\lambda} E[s_n^2]$$

$$u_j = \lambda_1 E[s_1] + \cdots + \lambda_j E[s_j] \qquad j = 1, 2, \cdots, n$$

$$u_n = u = \lambda E[s]$$

$$E[q_j] = \frac{\lambda E[s^2]}{2(1-u_{j-1})(1-u_j)} \qquad j = 1, 2, \cdots, n; \; u_0 = 0$$

$$E[q] = \frac{\lambda_1}{\lambda} E[q_1] + \frac{\lambda_2}{\lambda} E[q_2] + \cdots + \frac{\lambda_n}{\lambda} E[q_n]$$

$$E[w_j] = E[q_j] + E[s_j] \qquad j = 1, 2, \cdots, n$$

$$E[w] = E[q] + E[s]$$

$$L_q = E[N_q] = \lambda E[q]$$

$$L = E[N] = \lambda E[w]$$

Appendix 7: Formulas for an *M/G/1* preemptive-resume priority queuing system

$$\lambda = \lambda_1 + \lambda_2 + \cdots + \lambda_n$$

$$E[s] = \frac{\lambda_1}{\lambda} E[s_1] + \frac{\lambda_2}{\lambda} E[s_2] + \cdots + \frac{\lambda_n}{\lambda} E[s_n]$$

$$E[s^2] = \frac{\lambda_1}{\lambda} E[s_1^2] + \frac{\lambda_2}{\lambda} E[s_2^2] + \cdots + \frac{\lambda_n}{\lambda} E[s_n^2]$$

$$u_j = \lambda_1 E[s_1] + \cdots + \lambda_j E[s_j] \qquad j = 1, 2, \cdots, n$$

$$u = u_n = \lambda E[s]$$

$$E[w_j] = \frac{1}{1-u_{j-1}}\left[E[s_j] + \frac{\sum_{i=1}^{j} \lambda_i E[s_i^2]}{2(1-u_j)}\right]; \; u_0 = 0, j = 1, 2, \cdots, n$$

$$E[q_j] = E[w_j] - E[s_j]$$

$$W_q = E[q] = \frac{\lambda_1}{\lambda} E[q_1] + \frac{\lambda_2}{\lambda} E[q_2] + \cdots + \frac{\lambda_n}{\lambda} E[q_n]$$

$$L_{q_j} = E[N_{q_j}] = \lambda_j E[q_j] \qquad j = 1, 2, \cdots, n$$

$$L_q = E[N_q] = \lambda E[q] = \lambda W_q$$

$$W = E[w] = E[q] + E[s]$$

$$L = E[N] = \lambda E[w] = \lambda W$$

Reprint form No. G321-5009