Reviewed are applications of queuing models that may be economically useful in computing system analysis.

With emphasis on terminal-oriented systems with priorities, methods are given for estimating such average quantities as service time, waiting time, and response time.

Examples illustrate these methods and their ranges of efficiency, beyond which simulation techniques may be preferable.

Single-server queuing processes in computing systems

by W. Chang

Queues (waiting lines) were first studied systematically by A. K. Erlang[1] in connection with his investigations of telephone call delays at Danish telephone exchanges. Following this pioneering beginning in the early Twentieth Century, others who have made key contributions to a mathematical theory of queues are Pollaczeck,[2,3] Kolomogorov,[4] Khintchine,[5] Kendall,[6,7] Lindley,[8] and Takacs.[9-11] Queuing theory, which looks to probability theory for much of its structure, has proved useful in developing descriptive and predictive mathematical models that often lead to improvements in the services studied. As a result of this work, the theory of queues has been used in studying and improving such services as communication networks of all kinds, computing systems, production lines, transportation (harbors, tunnels, bus stations, airports, etc.) and in everyday life (e.g., banks, elevators, and supermarket checkouts).[12]

The intention of this review is to provide a practical guide for using queuing models in computing system design. The analytical methods discussed form an economical substitute for the more costly simulation techniques. Limitations of analysis are indicated—points where simulation is required for greater accuracy. System configurations involving priorities are illustrated.

Examined are mathematical models that have been found most useful for studying single-server queuing processes with Poisson

input distributions and general service times in computing systems. The *server* in a computing system can be any of a number of devices such as a channel, CPU, or communication line, that can process a *call* (an item from the server's queue). Table 1 gives a number of servers and the corresponding types of calls or *callers* that are typical of computing systems.

The first section of this paper lays a groundwork of basic definitions and basic queuing models. It is then possible to discuss imbedded Markov chain modeling, which has proved useful in studying many types of computing systems.[13-15] Priority queuing models that use imbedded Markov chaining are described and illustrated by numerical examples.

Queuing problems occur from time to time when there is sufficient caller or server irregularity in a system. For example, let us assume that a system contains a single server (i.e., only one caller can be served at a time). Also assume that either the callers arrive irregularly or randomly, or that there is appreciable variation in the time required to serve a caller (i.e., the *service time*), or both the caller-arrival and the service-time irregularity assumptions may simultaneously be true. If more than one caller is present in the system at the same time, all but one must queue up awaiting their turn for service. The rate of arrival of callers (sometimes called "traffic density" or "input rate") may be so high that large queues develop, resulting in a long waiting time per caller. On the other hand, the rate of arrival may be sufficiently low that the service facilities are not used for a proportion of time, called "idle time". Whether busy or idle-time conditions exist, a change in the system may be desirable and economically advantageous. Under certain of these conditions, we can apply queuing theory to predict what might happen under various alternatives. Such predictions are helpful to management in making corrective decisions.

Thus, queuing theory makes possible the computation of such quantities as average waiting time, average queue length, idle time, and service time, as previously mentioned. Collectively, these are some of the factors in performance evaluation. Frequently however, the queuing situations are of such complexity that mathematical models cannot be adequately descriptive. In these cases, it may be possible to obtain the desired estimates by simulating the system under a wide variety of data processing loads. The practice of simulation is a discipline in its own right and is not considered in this paper.

Basic concepts

definitions

Some of the following frequently used measures of congestion in a system have already been used in context. Definitions of these

Table 1 Typical queuing points in computer subsystems

Caller	*Caller sources*	*Server*	*Number of servers*	*Type of analysis*	*Type of queue*
Storage references	CPU; channels; storage	Bus control unit (BCU)	1	BCU	Priority; bulk service[1]
Storage references	CPU; channels; storage	Storage modules (M)	M	Main storage	Multi-server
System tasks	I/O devices; CPU; channels; storage; programs	CPU	1	Multi-programming	Priority
Messages; data; programs	Channels; CPU; communication lines	Buffers	M for constant block length; variable number for variable block length	Buffering	Multiserver
Channel requests	Channels; CPU	CPU	1	CPU; channel interference	Priority
Bytes; data records	I/O devices; CPU; communication lines	Channel	1	Channel	Contention;[2] priority; first-come first-served
I/O requests	CPU; disks; main storage	Disks and channel	$M + 1$	Disk and channel	Queues in series[3]
Terminals; programs; tasks	Terminals; I/O devices; CPU; main storage	CPU	1	Time-sharing computer	Priority with feedback
I/O requests; paging requests; tasks	I/O devices	CPU plus I/O	2	Time-sharing; paging	Queues in series with feedback[4]
Storage references; data records	CPU; main storage; data records; programs	Main storage modules (M) plus auxiliary storage units (N); high speed storage (M) plus cache (N)	$M + N$	Storage hierarchy	Queues in series with feedback
Input message	Terminals	Communication line	1	Full duplex communication input line	Multiple queues[5]
Output message	CPU	Communication line	1	Full duplex communication output line	First-come first-served; priority queues (polling messages have higher priority)
Input plus output messages	CPU plus terminals	Communication line	1	Half-duplex line	Priority queues (output messages have a higher priority)
Agent's terminal	Population	Agent's terminal	M	Industry operation (air line, bank, department store, etc.)	Multiple-server

1. *Bulk service:* During a service period, more than one caller can be simultaneously served by the BCU.
2. *Contention:* All callers can request service, and the server either randomly serves callers or serves them in a specific order.
3. *Queues in series:* The output of one queuing system is the input to the next queuing system (cascaded queues).
4. *Queues with feedback:* A caller that has been served returns to the queue for additional service. This may occur in a priority queuing system or in a cascaded queuing system (queues in series).
5. *Multiple queues:* Callers arriving at different queuing stations are to be served by the same system.

quantities are now given as they apply to computing systems and as they are used throughout this paper.

- *Queue length* is specified by the number of callers in the system waiting and being served at time t.
- *Waiting time* is the time interval between the arrival of a particular caller and the instant the caller's service begins.
- *Queuing time* is the sum of the waiting time and the service time of a caller.
- *Occupation time* is the time required to complete the service of all the callers that were in the queue at time t.
- *Idle period* is the duration of time that a particular server is not processing calls. During this period, the server is said to be in the "idle state."
- *Busy period* is the time during which a server is processing calls without entering the idle state.

Queuing theory is used to estimate the values of these quantities and their associated probability distributions. Under specified conditions, their means (averages) and variances may be used as measures of system performance.

When the probability distributions of these quantities are being studied, two kinds of solutions are of interest. The first is a *time-dependent solution*, and the second is the *stationary solution*.[16–19] In the time-dependent case, the probability distribution depends on the particular time t, and in general it also depends on the initial probability distribution (i.e., at $t = 0$). In our analysis, let $\xi(t)$ be the queue length of a system at time t. The probability distribution of $\xi(t)$ similarly depends on time t as well as on the initial probability distribution of $\xi(0)$. This is so because a queuing process is basically a *stochastic* one in which the state of the system (e.g., the number of callers present in the queue at time t) changes with a particular parameter—usually with time—in a probabilistic manner.

Since the time-dependent solution of a queuing process is often rather complicated, fruitful applications of queuing theory to computer problems usually rely on methods for finding stationary solutions. A probability distribution is stationary if it does not depend on time t. The stationary solution is important because most queuing processes have the *ergodic* property. That is, the process tends towards statistical equilibrium regardless of its initial state. Many queuing processes studied in computing systems have been observed to rapidly approach statistical equilibrium. Thus, one can usually apply the stationary approximation.[16]

In order to predict averages of one or more of the quantities—queue length, waiting time, queuing time, occupation time, idle period, busy period—we must specify a system with a sufficiently large

number of callers or a sufficiently long time period. We must also know the following:

- Arrival pattern
- Service pattern
- Service discipline

Thus, the average caller arrival rate and the statistical pattern of arrivals must be known.

arrival pattern

Suppose that callers arrive at a service facility at times $\tau_1, \tau_2, \cdots, \tau_n, \cdots$, and the times are arranged in increasing order, i.e., $\tau_n > \tau_{n-1}$. The random variables $(\tau_n - \tau_{n-1})$, where $n \geq 1$, are the *interarrival times*, and they form a sequence of independent and identically distributed random variables with a distribution

$$P\{\tau_n - \tau_{n-1} \leq x\} = A(x)$$

If the interarrival times $A(x)$ have a negative exponential distribution, the arrival pattern constitutes a Poisson input to the system.[18] In this case

$$A(x) = 1 - e^{-\lambda x} \tag{1}$$

where $1/\lambda$ is the mean interarrival time. Cox and Smith[18] show that if the arrival time is exponential, the number of arrivals of customers during time t follows a Poisson distribution of parameter λ as given by Equation 1 and by Equation C2 in Appendix C.

Another type of arrival pattern that is of interest is the regular arrival, or constant input. In this case, the interarrival time is a constant. An example of constant interarrival time is that of periodically updating a record of computation with the instant Greenwich mean time (GMT).

service pattern

In our discussion, we assume that when a server is available, the service time is specified by a statistical distribution. We shall also assume that the system has a single server (i.e., only one caller can be served at a time). The service time χ_n of caller n, where $n = 1, 2, \cdots$, is assumed to be a sequence of independent indentically distributed random variables with a general distribution of $H(x)$. That is

$$P\{\chi_n \leq x\} = H(x)$$

computer queuing

With the increasing use of time-shared and real-time systems, queuing conditions and their analyses are increasing in importance. Recalling that queuing and congestion problems frequently exist in computer systems where caller or service irregularities occur, queuing often occurs where computing facilities are shared.

Under conditions of real-time data processing—including time-sharing, guidance, and control—computer response must be fast enough to dynamically handle the processing demands. In a real-time

industrial processing application, the computing system may control an environment. In a time-sharing system, many terminal operators may simultaneously use the computer, which responds as though each user were the only one being served. This service can be achieved only through fast computer response and complete computer control over the various tasks (calls) to be performed. Consequently, the delays that occur due to congestion in such computer systems are of great concern to system designers and analysts.

The most important problem for queuing theory is the interaction of the elements of time and system capacity. When applied to the study of these elements, queuing theory can provide information such as the following: *response time* (the interval between the arrival of a call and the departure of results, including all waiting and service times), *throughput potential* (the maximum rate at which calls can be processed without the degradation of a processing facility), *queue length* (which determines the queue storage requirements), and other information that assists both in computer design and in program analysis. The following are some authors who have contributed to the analysis of queues in time-shared computer systems: E. G. Coffman,[20] L. Kleinrock,[21,22] B. Krishnamoorthi and R. C. Wood,[23] N. R. Patel,[24] L. E. Schrage,[25] and A. L. Scherr.[26]

Although many existing queuing models may be useful in analyzing modern computer systems, the systems analyst may often have to develop new models or modifications of existing models to reflect more closely the physical behavior of specific systems. In many cases, new models are needed to describe queuing problems that are growing in complexity.

service discipline

The service discipline must be specified by which a caller is selected for service out of all those awaiting service. The simplest discipline is that of first-come first-served, which consists of serving callers in order of arrival. There are other possibilities, such as a random service wherein the next caller to be served is randomly selected from the queue regardless of when that caller arrived. Another case is the priority-service discipline in which the next caller is selected from a queue on the basis of an assigned priority. We shall discuss priority disciplines in greater detail later in this paper, but a brief summary of concepts of priority queues is given here.

When calls to be processed by a computer are classified according to the importance of their undelayed passage through the system, they are said to be assigned a priority. For example, in a real-time system, when the computer must be programmed so that during its normal processing of a program it can always be interrupted by communications calls, such calls have the highest priority. A computer supervises and controls a communication system and must act sufficiently fast when a message (or a message segment) is received from the multiplexor to prevent a loss of data or overflow of main storage.

This problem can be formulated as a priority queuing model, which may be outlined as follows. A single server (i.e., the computer) processes two types of calls: communications calls, which are put in a high priority queue, and the normal processing calls, which are processed on an as-available basis. Priority queuing analysis in computer systems is also used in computer channel analysis[13] and in multiprogramming system analysis.[27,28] The theory of priority queues is treated in a number of sources, such as in References 29, 30, and 31.

A normal formulation of priority queues, based on Takacs' discussion,[29] is as follows. Calls of different priorities arrive at a facility for service. Let there be N classes of priorities, 1, 2, $\cdots$, N. It is convenient to assume that an arriving call with a smaller priority number (i.e., higher priority) has preference over a call with a greater priority number. The calls are served by a single server in order of priority (and in order of arrival within each priority class). It is also often assumed that the input is a Poisson process with parameter λ_k for type-k priority calls, where $k = 1, 2, \cdots, N$. Service times for k-type calls are assumed to be mutually independent, positive, random variables with a distribution function $H_k(x)$. Two types of service disciplines are of general interest.

Two types of priority service disciplines are of general interest. One is the *preemptive-resume discipline*, wherein a server interrupts the processing of its current call and immediately beings processing a higher-priority call. When a lower-priority call—the one that was preempted—returns to service, processing continues from the point of interruption. In the *nonpreemptive discipline*, the server does not interrupt the current processing. A higher-priority call waits and obtains service immediately after the processing of the current lower-priority call.

For these types of priority queuing systems, Cobham[32,33] obtained the first moment of the waiting time distribution. Miller[34] characterized the limited distributions of the queue sizes and waiting times. Gaver,[30] Welch[35] and Jaiswal[36] studied the transient solutions of this priority system. Takacs[29] generalized the stationary solutions of priority queues. The present author[37] generalized the stationary solutions of preemptive priority queues including other service disciplines.

Other service disciplines of interest are *preemptive-repeat-identical* and *preemptive-repeat-different*.[30] In the case of the preemptive-repeat-identical discipline, when a lower-priority caller whose place has been preempted returns to processing, a service period of the same duration as the one interrupted is commenced again at the beginning. In the case of the preemptive-repeat-different discipline, when the interruption is cleared, service of the lower-priority caller begins again from the beginning but with a new independent service

period. In both cases, the service time previously allocated for the lower-priority caller before the interruption is wasted and makes no contribution to the subsequent processing time.

Imbedded Markov chains

Since queuing is generally a stochastic process that is a function of the time parameter t, a class of stochastic processes useful for systems analysis is the *Markov process*. A queue is said to exemplify a Markov process if the present state of the system, including the queue, is sufficient to predict a future state (e.g., queue length) without knowledge of the past history of the system. By definition, Markov processes are continuous in time. When the system is studied at discrete time points, the collection of state probabilities constitutes a *Markov chain*. Here we are considering only qualitative properties of Markov chains. A mathematical definition of the Markov chain is given in Appendix A.

D. G. Kendall introduced the concept of an *imbedded Markov chain* because in practical cases queuing processes are not always Markovian nature. Kendall[7] suggests that a non-Markovian process can be studied by extracting a set of points—called *regeneration points*—at which the Markov property exists. He formally defines a regeneration point as an epoch at which a knowledge of the state of the process has the characteristic Markovian consequence that a statement of the past history of the process loses its predictive value. A probabilistic definition of a regeneration point[38] is as follows: an epoch is a regeneration point for the stochastic process $\{\xi(t)\}$ if and only if, for all $t > t_0$, $P\{\xi(t) \mid \xi(t_0)\} = P\{\xi(t) \mid \xi(\tau)\}$ for all $\tau \leq t_0$. A group of epochs, t_i (for $i = 0, 1, 2, \cdots$), are regeneration points if and only if, for all $t_i < t < t_{i+1}$, $P\{\xi(t) \mid \xi(t_i)\} = P\{\xi(t) \mid \xi(\tau)\}$ for all $t_{i-1} < \tau \leq t_i$.

As an example, consider a queuing process having Poisson input (M) general service time (G), and a single server (1). (This is termed an M/G/1 queuing process by Kendall's queue-classification procedure.) It is Markovian if the present state of the process is described by the pair of random variables ξ and χ, where ξ is the instantaneous queue size, and χ is the expended service time of the customer who is currently being served. In general, however, the queue ceases to be Markovian if the state of the process is measured by the queue size alone. An exception occurs when the service time is exponentially distributed; a characteristic property of the exponential distribution ensures that a knowledge of the expended service time χ has no predictive value.

Illustrative of a set of regeneration points are the times at which callers depart from the system in the M/G/1 queueing process. Hence, the queue lengths at these departure times constitute a Markov

chain with an enumerable infinity of states that can be studied through the theory of Markov chains.[40] This technique, which is useful in many queuing systems including priority queues, is called the imbedded Markov chain technique because it involves the extracting of a discrete-time Markov chain imbedded in the continuous-time process.

Another important contribution of Kendall is a formal proof of the existence of a stationary solution for the M/G/1 queuing process if the *traffic intensity* (or system utilization) is less than unity. The traffic intensity ρ is defined as the product of the input rate λ and the mean service time α. Intuitively, when the traffic intensity is less than unity, the system should possess the characteristic "ergodic" property[39] of settling down into an equilibrium mode of behavior independent of its initial state after the elapse of a sufficiently long period of time. When $\rho > 1$ no such behavior is to be expected. Kendall's proof uses Feller's general theory of recurrent events. The stationary solutions of priority queues[29,34,37] are obtained by using the method of the imbedded Markov chain. They are basically extended solutions of Kendall's M/G/1 queuing process. In order to present some basic results, we use the mathematical notations given in the analysis of M/G/1 queues in Appendix B.

stationary distributions

We now review the technique for analyzing imbedded Markov chains, which will be useful for the later analysis of priority queues. Let the input to a single-server queuing system be a Poisson process of parameter λ. The service-time distribution is $H(x)$, and the mean service time is α. All the mathematical notations are given in Appendix B, except that the subscript k is dropped since no priorities are involved here. The callers are served in order of arrival. The stationary distributions of queue length, waiting time, and busy period are to be found.

Queue length

Let τ_n and τ_n' $(n = 1, 2, \cdots)$ be representively the arrival time and the departure time of the nth customer. Referring to the basic concepts, we define ξ_n to be the queue length immediately after the nth customer's departure. If $\xi(t)$ is the queue length at time t, then by definition $\xi_n = \xi(\tau_n)$. As previously mentioned, Kendall recognized that the queue length $\xi(t)$ is not in general a Markov process, but that regeneration points ξ_{n+1} and ξ_n form a Markov chain. Therefore, consider a sequence of such points $\xi_1, \xi_2, \cdots, \xi_n \cdots$ where $\xi_n = j$, (i.e., the nth departing customer leaves j customers in the system). The next regeneration point, ξ_{n+1}, is uniquely determined from ξ_n, and it is independent of $\xi_{n-1}, \xi_{n-2}, \cdots$. The purpose of the analysis is to determine the stationary probabilities of the queue length

$$P\{\xi_n = j\} = P_j$$

where $j = 0, 1, 2, \cdots$. To determine P_j, we use the technique given in Appendix C, according to which

$$G(z) = \sum_{j=0}^{\infty} P_j z^j$$

From Equations C8 and C9 in Appendix C

$$G(z) = \frac{(1 - \lambda\alpha)\psi[\lambda(1 - z)]}{z - \psi[\lambda(1 - z)]}$$

Here λ is the Poisson process parameter, α is the average service time, and $\psi[\lambda(1 - z)]$ is given by the Laplace transform of the service time distribution

$$\psi(s) = \int_0^{\infty} e^{-sx}\, dH(x) \qquad \text{and}$$

$$\psi[\lambda(1 - z)] = \int_0^{\infty} e^{-\lambda(1-z)x} dH(x)$$

Knowing $G(z)$, P_j can be determined by the following expression:

$$P_j = \frac{1}{j!} \frac{d^j}{dz^j} G(z)\Big|_{z=0} \tag{2}$$

From Equation 2, specific probabilities have the following forms:

$$P_0 = G(0), \qquad P_1 = G'(0), \qquad P_2 = 1/2G''(0) \cdots$$

To illustrate how the queue-size formulas can be used, consider the following example. Let the service time be exponentially distributed:

$$H(x) = 1 - e^{-x/\alpha}$$

which yields

$$\psi(s) = \frac{1}{1 + \alpha s} \tag{3}$$

The generating function $G(z)$ is obtained from Equation C8:

$$G(z) = \frac{(1 - \lambda\alpha)(z - 1)}{z[1 + \lambda\alpha(1 - z)] - 1} = \frac{1 - \lambda\alpha}{1 - \lambda\alpha z}$$

Expanding Expanding $G(z)$ in a geometric series,

$$G(z) = (1 - \lambda\alpha) \sum_{j=0}^{\infty} (\lambda\alpha z)^j$$

the queue length probability P_j is obtained as

$$P_j = (1 - \lambda\alpha)(\lambda\alpha)^j$$

for j any integer greater than or equal to zero.

Waiting time

Another problem of single-server queuing theory is that of predicting the waiting time of a caller that arrives and finds that there are j callers in the system (i.e., the queue length is j). In this case, if service is in the order of arrival, the waiting time will be $j - 1$ service times plus the time required to complete the current service (i.e., the call

being served at the time or arrival). Thus, waiting time is related directly to queue size.

The waiting time distribution can be determined from its Laplace transform, which is derived explicitly from the generating function of queue size in Appendix D with the following result:

$$\Omega(s) = \frac{1 - \lambda\alpha}{1 - \lambda\{[1 - \psi(s)]/s\}}$$

where $\Omega(s)$ is the Laplace transform of the waiting time distribution. Assuming the special case of exponentially distributed service time, we obtain $\Omega(s)$ by substituting Equation 3 into Equation D4, which is developed in Appendix D

$$\Omega(s) = \frac{(1 - \lambda\alpha)(1 + \alpha s)}{1 - \lambda\alpha + \lambda\alpha} \tag{4}$$

Inverting Equation 4 we obtain the result[40]

$$W(x) = 1 - \lambda\alpha e^{(\lambda - 1/\alpha)x}$$

Busy period

If a server is free at time zero, and if a caller is served from that instant until time b (when the caller departs and the queue is empty), the time interval $(0, b)$ is a busy period. These periods typically alternate with idle periods. Thus if the next caller arrives at time c, the time interval (b, c) is an idle period. In a Poisson-input case, the idle period can be shown to follow a negative exponential distribution that was previously discussed in connection with the arrival patterns.

Kendall[7] suggests that the busy period be found as follows. Let the caller whose arrival initiates a busy period be called the "ancestor," which constitutes the zero-order "generation." During the ancestor's service time, let n_1 more callers arrive; these callers constitute the first-order generation. During the first-generation service time, let n_2 more callers arrive, forming the second-order generation, and so on. The busy period terminates when the "family" becomes extinct, i.e., when an idle period intervenes between two such busy periods. Kandall was able to determine the busy-period using this familiar analogy. We shall, however, use the simpler method of Takacs,[41] who obtains equivalent results. The busy period distribution is treated mathematically in Appendix E; here we merely state Takacs' results.

The busy period distribution is again given as a Laplace transform from a functional equation as follows:

$$\gamma(s) = \psi\{s + \lambda[1 - \gamma(s)]\}$$

where $\gamma(s)$ is the Laplace transform of the busy-period distribution.

This is a functional equation, first given by Kendall,[7] wherein $\gamma(s)$ cannot be explicitly determined in general. In the case of exponential service time, however, $\gamma(s)$ can be explicitly determined as

$$\gamma(s) = \frac{1/\alpha + s + \lambda - [(1/\alpha + s + \lambda)^2 - 4\lambda/\alpha]^{1/2}}{2\lambda}$$

This may be inverted to give

$$D(x) = \frac{e^{-(1/\alpha+\lambda)x} I_1[2x(\lambda/\alpha)^{1/2}]}{x(\lambda\alpha)^{1/2}}$$

where $I_1(x)$ is the Bessel function of the first kind.

mean values and moments

In queuing theory, there are only a few special cases, such as the exponential service time case, in which the waiting-time and the queue-size distributions can be inverted directly from $\Omega(s)$ and $G(z)$. In other cases, we obtain the moments of the respective random variables from $\Omega(s)$ and $G(z)$ by differentiating these expressions using the technique to be presented now. Such moments as mean waiting time and the variance of the waiting time, are used as performance measures of the system. On the other hand, inversion techniques[42] have been developed to approximate the waiting-time distribution from its Laplace-Stieltjes transform. In either case, we shall emphasize the derivation of generating functions and Laplace-Stieltjes transforms in our analysis, since these consititute standard techniques in queuing theory.[16,18,38,40] The inversion techniques are of practical interest, but belong to another branch of mathematics and should be studied separately from research in queuing theory. Knowing the Laplace transforms, the moments of a distribution can be easily obtained by differentiating the transform. In general, the rth moment of the waiting-time distribution can be deduced by using the following formula:

$$W^{(r)} = \int_0^\infty x^r \, dW(x) = (-1)^r \int_0^\infty \left(\frac{d^r}{ds^r} e^{-sx}\right) dW(x)\Big|_{s=0}$$

$$= (-1)^r \frac{d^r \Omega(s)}{ds^r}\Big|_{s=0} \tag{5}$$

More specifically, the first and second moments of the waiting time are expressed as follows:

$$W^{(1)} = \frac{\lambda\alpha^{(2)}}{2(1 - \lambda\alpha)}$$

$$W^{(2)} = \frac{\lambda^2(\alpha^{(2)})^2}{2(1 - \lambda\alpha)^2} + \frac{\lambda\alpha^{(3)}}{3(1 - \lambda\alpha)}$$

The average queue length and its moments can be obtained from the queue length generating function $G(z)$ directly. Thus, from Equation C6 we obtain the class of generating functions represented by the following examples:

$$G'(z) = \sum_{j=0}^{\infty} P_j j z^{j-1} \tag{6}$$

and

$$G''(z) = \sum_{j=0}^{\infty} j(j-1)P_j z^{j-2} \tag{7}$$

The average queue length L_q and the second moment L_q^2 can be obtained from Equations 6 and 7 as follows:

$$L_q = \sum_{j=0}^{\infty} jP_j = G'(z)\Big|_{z=1} = G'(1) = \frac{\lambda^2\alpha^{(2)}}{2(1-\lambda\alpha)} + \lambda\alpha \tag{8}$$

$$\overline{L_q^2} = \sum_{j=0}^{\infty} j^2 P_j = G''(1) + G'(1)$$

More importantly, in terms of statistical analysis, the variance of the queue length $\sigma_{L_q}^2$ is given as

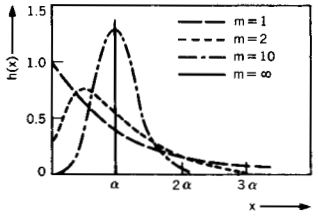
$$\sigma_{L_q}^2 = \overline{L_q^2} - L_q^2 = \frac{\lambda^3\alpha^{(3)}}{3(1-\lambda\alpha)} + \frac{\lambda^4(\alpha^{(2)})^2}{4(1-\lambda\alpha)^2} + \frac{\lambda^2(3-2\lambda\alpha)\alpha^{(2)}}{2(1-\lambda\alpha)} + \lambda\alpha(1-\lambda\alpha) \tag{9}$$

illustrative computations

We now use some of the concepts previously given for performing illustrative numerical computations. First we define a useful class of the service time distributions, the *Erlang-m distributions*, which are known to statisticians as a special class of "gamma functions." When the parameter m is an integer, the following type of gamma distribution is designated "Erlang" in honor of that pioneer's contributions to queuing theory. Here $H(x)$ is a function that represents service-time distributions

$$H(x) = 1 - e^{-mx/\alpha} \sum_{k=0}^{m-1} \frac{(mx/\alpha)^k}{k!} \tag{10}$$

Figure 1 Probability density function of the Erlang-m distribution



In Equation 10, α is the mean service time. The parameter m determines the shape of the Erlang distribution as illustrated in Figure 1. In the Erlang-1 case, $m = 1$, and $H(x)$ is an exponential service-time distribution

$$H(x) = 1 - e^{-x/\alpha}$$

Given the Erlang-∞ case, m is infinite, and $H(x)$ is a constant service-time distribution

$$H(x) = \begin{cases} 0 & \text{if } 0 \le x < \alpha \\ 1 & \text{if } x \ge \alpha \end{cases}$$

Intermediate Erlang-m values yield a family of service-time distribution curves. If the type of distribution is known, service times can be computed by means of Equation 10.

The probability density function $h(x)$, which is defined as the first derivative of $H(x)$

$$h(x) = \frac{dH(x)}{dx}$$

is plotted in Figure 1 for three values of the parameter m. Higher moments of the Erlang-m distribution is obtained as follows:

$$\alpha^{(r)} = \frac{(r + m - 1)!}{(m - 1)!} (\alpha/m)^r$$

where α is the average service time, and $\alpha^{(r)}$ is the moment.

The server utilzation ρ, or traffic intensity is given as

$$\rho = \lambda\alpha \tag{11}$$

With these concepts and refering to Equations 8 and 9 in the previous discussion of mean values and moments, we can plot the following server utilization distributions: average queue length (Figure 2), standard deviation of queue length (Figure 3), waiting time (Figure 4), and the standard deviation of the normalized mean waiting time (Figure 5). (The standard deviation is defined as the positive square root of the variance.)

As a numerical example, consider interference in a channel in which the computer is operating in the multiplex mode. Assume an average time of 0.4 milliseconds for storing the status of general-purpose registers, transferring a single byte of data, and restoring the general-purpose registers to their previous states. The channel serves a number of communications facilities, thereby having a total input rate of 0.5 kilobytes per second. The problem is to calculate the average queue length in bytes and the length of time each byte waits. Thus

$\alpha = 0.4$ milliseconds^2 and

$\lambda = 500$ bytes per second

The second moment of the service time is used in determining the average waiting time and queue length, as illustrated by Equation 8. We must know the service time distribution in order to find service time moments. In our example, however, the service time distribution is not of major concern because referring to Equation 11

$$\rho = \lambda\alpha = 0.0004 \times 500 = 0.2$$

From Figures 2 and 4, it is apparent that the variation of service time should not significantly affect the result. (In the case of high utilization, the service time distribution is important in the analysis.) If we assume that the service time is a constant, that is

$m = \infty$ then

$\alpha^{(2)} = \alpha^2 = 0.16$ milliseconds^2

The average queue length is

$$L_q = \frac{\lambda^2\alpha^{(2)}}{2(1 - \lambda\alpha)} + \lambda\alpha = 0.225 \text{ bytes}$$

Figure 2 Average queue length

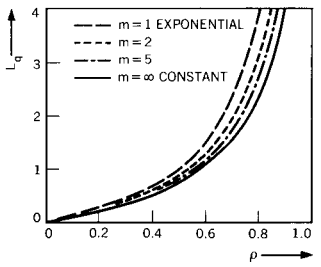


Figure 3 Standard deviation of queue length

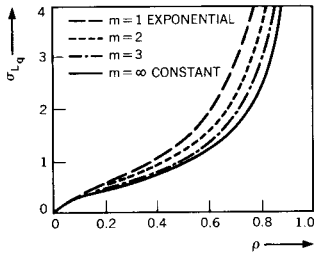


Figure 4 Normalized mean waiting time

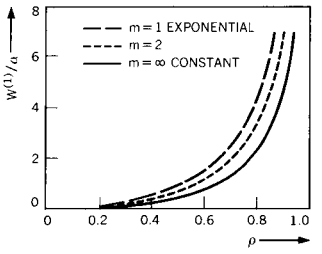
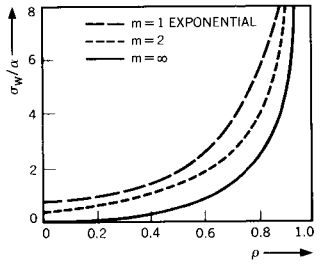


Figure 5 Standard deviation of waiting time



and the average waiting time is

$$W^{(1)} = \frac{\lambda\alpha^{(2)}}{2(1 - \lambda\alpha)} = 0.05 \text{ milliseconds}$$

Priority queues

We now use the techniques of imbedded Markov chains to analyze several classes of priority queues for the following stationary distributions: queue length, waiting time, queuing time, and busy period. These priority queues were previously defined and discussed within the context of service discipline. In order to specify these service disciplines, consider the previously discussed formulation of priority queues. There are N priority classes, each with independent Poisson processes of parameter λ_k, where $k = 1, 2, \cdots N$. These processes constitute inputs to a single-server system. Let the lower priority number indicate the higher priority, and let $H_k(x)$ be the service-time distribution for priority class k.

The modeling of a priority queueing system is a two-step procedure. We first treat the system as though no priorities are involved. Every caller is served on a first-come first-served basis. A queue-length generating function for an imbedded Markov chain previously discussed, is derived as shown in Appendix C for a queuing model having no priorities. We then take into consideration the effect of the priorities by modifying the results obtained in the first step to reflect the influence of caller priorities on the waiting-time distribution.

Preparatory to discussing the priority queuing systems, some additional concepts must be defined. Let the sum of the Poisson-process parameters be expressed as follows:

$$\Lambda_k = \sum_{i=1}^{k} \lambda_i$$

Each input is a Poisson process of parameter λ_k, therefore the sum Λ_k is also a Poisson process.[34] Also, let the Laplace-Stieltjes transforms of the weighted service-time distributions be

$$\Psi_k(s) = \sum_{i=0}^{k} \frac{\lambda_i}{\Lambda_k} \psi_i(s) \tag{12}$$

and

$$\Phi_k(s) = \sum_{i=k+1}^{N} \frac{\lambda_i}{\Lambda_n - \Lambda_k} \psi_i(s)$$

Let a_k be defined as the first moment of the weighted service time, i.e.

$$a_k = \sum_{i=0}^{k} \frac{\lambda_i}{\Lambda_k} a_i$$

In a priority queuing system, two priority-service disciplines are of interest—the preemptive-resume and the non-preemptive disciplines. Their stationary solutions are now presented.

preemptive-resume service

In a preemptive-resume priority system, the presence of callers of priority numbers greater than k does not influence the stochastic law of the waiting time of callers with priority numbers less than or equal to k.) Thus, callers with priority numbers greater than k are considered as not present in the system when studying the waiting time of callers with priorities less than or equal to k. To determine the stationary distribution of the waiting time for priority-class k, we first consider a modified queuing process with the following characteristics: a Poisson input of Λ_k, a service time distribution that is weighted by the input rates and characterized by $\psi_k(s)$ as defined in Equation 12, and a single server. The generating function can be obtained by using the technique of the imbedded Markov chain previously discussed. The Laplace-Stieltjes transform of the waiting time in this modified queuing process can be obtained from Equation D4 of Appendix D

$$\Omega_k^*(s) = \frac{1 - \Lambda_k a_k}{1 - \dfrac{\Lambda_k[1 - \Psi_k(s)]}{s}} \tag{13}$$

If $\Lambda_k a_k$ is less than one, a stationary solution exists. If $\Lambda_k a_k$ is greater than or equal to one, there is no stationary solution[31] for priority classes less than or equal to k.

The waiting-time distribution $W_k(x)$ for the priority-k caller is obtained from Equation 13 in the following way. During the waiting time of the modified queuing process, assume that there are j ($j = 0, 1, \cdots$) arrivals of priority number less than k. The additional delay experienced by the caller with priority number k is then identical to the total of j independent busy periods in a single-server queuing process with Poisson input Λ_{k-1} and a service time distribution characterized by $\Psi_{k-1}(s)$. Let $D_{k-1}(x)$ be the distribution function of the length of a busy period in this process, and $\gamma_{k-1}(s)$ be the Laplace-Stieltjes transform of $D_{k-1}(x)$. From Equation E2 we can complete the formulation of the transform as follows:

$$\gamma_{k-1}(s) = \Psi_{k-1}\{s + \Lambda_{k-1}[1 - \gamma_{k-1}(s)]\} \tag{14}$$

Using similar reasoning to that given for Equation E1, we obtain from Equation 13 the following Laplace-Stieltjes transform of the waiting-time distribution for priority class k:

$$\Omega_k(s) = \Omega_k^*\{(s + \Lambda_{k-1}[1 - \gamma_{k-1}(s)]\}$$

Using the same line of thought, the Laplace-Stieltjes transform $\theta_k(s)$ of the queuing-time distribution is found to be

$$\theta_k(s) = \Omega_k(s)\psi_k\{s + \Lambda_{k-1}[1 - \gamma_{k-1}(s)]\}$$

This equation shows that during the service of priority k callers under the preemptive-resume service discipline, other callers with priority numbers less than k may arrive and preempt the service.

nonpreemptive service

In the previously defined nonpreemptive service discipline, service time of a caller of any priority class is not interruptable. Consequently the presence of a low-priority call can affect the waiting time of a high-priority call. For example, if a low-priority call is being served when a call of a high-priority class arrives, the high-priority call must wait for completion of the lower-priority service before service begins. Calls of low priority receive immediate service when no calls of higher priority are waiting.

To find the waiting-time stationary distribution for priority class k, we use an approach similar to the one presented in the preemptive-resume priority queues. The generating function for this case, however, is slightly more complicated than the one for preemptive resume service. Consider a queuing process in which callers are classified into two queues. Let $\xi_n(k)$ be the queue length of calls having priority classes less than or equal to k, and let $\xi'_n(k)$ be the queue length of priority classes greater than k at the nth departure. The nth caller can be of any priority class. We now formulate the qenerating function for $\xi_n(k)$.

For a stationary process, $\xi_{n+1}(k)$ and $\xi_n(k)$ have the same probability distribution, and are related by

$$\xi_{n+1}(k) = \begin{cases} \xi_n(k) - 1 + \nu_{n+1} & \text{if} \quad \xi_n(k) > 0 \\ \nu_{n+1} & \text{if} \quad \xi_n(k) = 0 \text{ and } \xi'_n(k) = 0 \text{ (i.e., both queues are empty, and the next call is of priority class less than or equal to } k) \\ \nu'_{n+1} & \text{if} \quad \xi_n(k) = 0 \text{ and } \xi'_n(k) > 0 \text{ (or the next call is of priority class greater than } k \text{ if } \xi_n(k) = 0 \text{ and } \xi'_n(k) = 0) \end{cases} \tag{15}$$

Here, ν_{n+1} is the number of new calls of priority classes less than or equal to k, if the $n+1$st service is of priority class less than or equal to k. The parameter ν'_{n+1} is the number of new calls of priority classes less than or equal to k if the $n+1$st service is of priority class greater than k.

Let $U_k(z)$ be the generating function of $\xi_n(k)$ so that

$$U_k(z) = \sum_{j=0}^{\infty} P\{\xi_n(k) = j\} z^j \tag{16}$$

Thus the probability that $\xi_n(k)$ is zero is expressed as follows:

$$P\{\xi_n(k) = 0\} = U_k(0)$$

Notice that Equation 15 expresses the following three mutually exclusive events:

- $\xi_n(k) > 0$, and the next arrival is of priority class less than or equal to k. The first event, $\xi_n(k) > 0$, is represented by the generating function

$$\frac{U_k(z) - U_k(0)}{z}$$

- $\xi_n(k) = 0$ and $\xi_n'(k) = 0$, and the next service is of priority class less than or equal to k. This event occurs with probability $(\Lambda_k/\Lambda_n)P_0$, where P_0 is the probability that the system is empty.

- $\xi_n(k) = 0$ and $\xi_n'(k) \geq 0$, and the next service is of priority class greater than k. The third event occurs with a probability of $U_k(0) - (\Lambda_k/\Lambda_n)P_0$, where P_0 is the probability that the system is empty, and

$$P_0 = 1 - \sum_{i=1}^{n} \lambda_i \alpha_i$$

Forming the generating functions on both sides of Equation 15 and using the technique of the imbedded Markov chain as given in the Appendices, we obtain Takacs' expression[29] for the generating function of $\xi_n(k)$

$$U_k(z) = [U_k(z) - U_k(0)]\Psi_k[\Lambda_k(1 - z)] + \frac{\Lambda_k}{\Lambda_n} P_0 \Psi_k[\Lambda_k(1 - z)] + \left[U_k(0) - \frac{\Lambda_k}{\Lambda_n} P_0\right]\Phi_k[\Lambda_k(1 - z)] \tag{17}$$

Equation 17 generates the queue lengths of priority classes less than or equal to k at every departing instant, including the departure of those callers of priority classes greater than k. We now obtain a relation similar to the one in Equation D3 in Appendix D, which allows us to obtain the Laplace-Stieltjes transform of the waiting-time distribution from the queue-size generating function. We formulate the queue-size generating function observed by a departing caller of priority class less than or equal to k by Takacs' method.

In Equation 17, the $n + 1$st customer is of priority class less than or equal to k, if the service-time distribution is of priority class less than or equal to k. Hence, the partial generating function of the right-hand side of Equation 17

$$\left[\frac{U_k(z) - U_k(0)}{z} + \left(\frac{\Lambda_k}{\Lambda_n}\right)P_0\right]\Psi[\Lambda_k(1 - z)]$$

represents a departing caller of priority class less than or equal to k.

Takacs called this generating function $G^*(z)$. (Note that Takacs actually uses $G_k(z)$ in Reference 29; we add the asterisk to denote a modified queuing process.) The generating function of the queue size, $G_k(z)$, observed by a departure of priority class less than or equal to k, can be obtained from $G_k^*(z)$ by the following normalization:

$$G_k(z) = \frac{G_k^*(z)}{G_k^*(1)}$$

The Laplace-Stieltjes transform of the waiting time distribution of the modified queuing process is given by

$$G_k(z) = \Omega_k^*[\Lambda_k(1 - z)]\Psi_k[\Lambda_k(1 - z)]$$

If we replace $\Lambda_k(1 - z)$ by s, we obtain

$$\Omega_k^*(s) = \frac{1 - \sum_{i=1}^{k} \lambda_i \alpha_i + (\Lambda_N - \Lambda_k)\Phi_k(s)}{1 - \Lambda_k[1 - \Psi_k(s)]/s} \tag{18}$$

Finally, the Laplace-Stieltjes transform of the waiting time distribution for customers of priority class k is obtained in a way similar to the case of the preemptive-resume discipline as follows:

$$\Omega_k(s) = \Omega_k^*\{s + \Lambda_{k-1}[1 - \gamma_{k-1}(s)]\}$$

The Laplace-Stieltjes transform of the queuing time distribution is represented as

$$\theta_k(s) = \Omega_k(s)\Psi_k(s)$$

The transform takes this form because the service time of priority k cannot be interruptable during its service by any new arrivals of priority classes less than k.

waiting-time moment calculations

When a stationary solution exists (i.e., the Laplace transform of the waiting time exists), the moments such as the mean waiting time can be obtained by differentiating the Laplace transform as given in Equation 5. Let $a_k^{(r)}$ be the rth moment of the weighted service time distribution

$$a_k^{(r)} = (-1)^r \frac{d^r}{ds^r} \Psi(s)\bigg|_{s=0} = \sum_{i=1}^{k} \frac{\lambda_i}{\Lambda_k} \alpha_i^{(r)}$$

where $\Psi_k(s)$ is given in Equation 12.

To find the waiting-time moments, the first step is to find the busy period moments by differentiating Equation 14. For example, the first two moments of the busy period distribution are the following:

$$d_{k-1}^{(1)} = \frac{a_{k-1}^{(1)}}{1 - \Lambda_{k-1} a_{k-1}^{(1)}}$$

$$d_{k-1}^{(2)} = \frac{a_{k-1}^{(2)}}{[1 - \Lambda_{k-1} a_{k-1}^{(1)}]^3}$$

The next step is to find an expression for the moment of the modified waiting time distribution, $W_k^{*(r)}$, by differentiating $\Omega_k^*(s)$ in Equation 18. Since we have two expressions for $\Omega_k^*(s)$, preemptive and nonpreemptive, we have the following two cases:

Case 1 Preemptive-resume discipline

$$W_k^{*(1)} = \frac{\Lambda_k a_k^{(2)}}{2[1 - \Lambda_k a_k^{(1)}]}$$

$$W_k^{*(2)} = \frac{\Lambda_k a_k^{(3)}}{3[1 - \Lambda_k a_k^{(1)}]} + \frac{\Lambda_k^2 [a_k^{(2)}]^2}{2[1 - \Lambda_k a_k^{(1)}]^2}$$

Case 2 Nonpreemptive discipline

$$W_k^{*(1)} = \frac{\Lambda_N a_N^{(2)}}{2[1 - \Lambda_k a_k^{(1)}]}$$

$$W_k^{*(2)} = \frac{\Lambda_N a_N^{(3)}}{3[1 - \Lambda_k a_k^{(1)}]} + \frac{\Lambda_N a_N^{(2)} \Lambda_k a_k^{(2)}}{2[1 - \Lambda_k a_k^{(1)}]^2}$$

Finally, we compute the waiting-time moments from $\Omega_k(s)$ by differentiating as follows:

$$W_k^{(1)} = W_k^{*(1)}(1 + \Lambda_{k-1} d_{k-1}^{(1)}) = \frac{W_k^{*(1)}}{1 - \Lambda_{k-1} a_{k-1}^{(1)}}$$

and

$$W_k^{(2)} = W_k^{*(2)}[1 + \Lambda_{k-1} d_{k-1}^{(1)}]^2 + W_k^{*(1)} \Lambda_{k-1} d_{k-1}^{(2)}$$

$$= \frac{W_k^{*(2)}}{[1 - \Lambda_{k-1} a_{k-1}^{(1)}]^2} + \frac{W_k^{*(1)} \Lambda_{k-1} a_{k-1}^{(2)}}{[1 - \Lambda_{k-1} a_{k-1}^{(1)}]^3}$$

Since both $W_k^{*(1)}$ and $W_k^{*(2)}$ apply to Case 1 as well as Case 2, the following numerical examples are similarly separated into two cases.

numerical examples

Let the inputs to a priority queuing system be

$\lambda_1 = \lambda_2 = \lambda_3 = 0.3$ calls/second

The average times are all the same

$\alpha_1 = \alpha_2 = \alpha_3 = 1$ second

and identically distributed with the same exponential distribution

$H_1(x) = H_2(x) = H_3(x) = 1 - e^{-x}$

To calculate the mean waiting times for both types of priority disciplines, we first construct the Laplace transforms of the service-time distributions

$$\psi_1(s) = \psi_2(s) = \psi_3(s) = \frac{1}{s + 1}$$

The weighted service-time distributions have the following Laplace transforms

$$\Psi_1(s) = \psi_1(s) = \frac{1}{s+1}$$

$$\Psi_2(s) = \frac{\lambda_1\psi_1(s) + \lambda_2\psi_2(s)}{\lambda_1 + \lambda_2} = \frac{1}{s+1}$$

$$\Psi_3(s) = \frac{\lambda_1\psi_1(s) + \lambda_2\psi_2(s) + \lambda_3\psi_3(s)}{\lambda_1 + \lambda_2 + \lambda_3} = \frac{1}{s+1}$$

The weighted service-time moments (from Equation 12) are the following:

$$a_1^{(1)} = a_2^{(1)} = a_3^{(1)} = 1 \text{ second}$$

$$a_1^{(2)} = a_2^{(2)} = a_3^{(2)} = 2 \text{ seconds}^2$$

Let the busy periods for the priority system be defined as follows:

- $d_1^{(1)}$ is the busy period of the server for the first priority class.
- $d_2^{(1)}$ is the busy period of the server for the first and second priorities.
- $d_3^{(1)}$ is the busy period of the server for all three classes.

In this example, the mean busy periods for the three priority classes are as follows:

$$d_1^{(1)} = \frac{1}{1 - 0.3} = 1.429 \text{ seconds}$$

$$d_2^{(1)} = \frac{1}{1 - 0.3 - 0.3} = 2.5 \text{ seconds}$$

$$d_3^{(1)} = \frac{1}{1 - 0.3 - 0.3 - 0.3} = 10 \text{ seconds}$$

The mean waiting times for the two cases can now be calculated.

Case 1 Preemptive-resume discipline

Mean waiting time for the first priority:

$$W_1^{*(1)} = \frac{0.3 \times 2}{1 - 0.3} = 0.855 \text{ seconds}$$

$$W_1^{(1)} = W_1^{*(1)} = 0.855 \text{ seconds}$$

Meaning waiting time for the second priority:

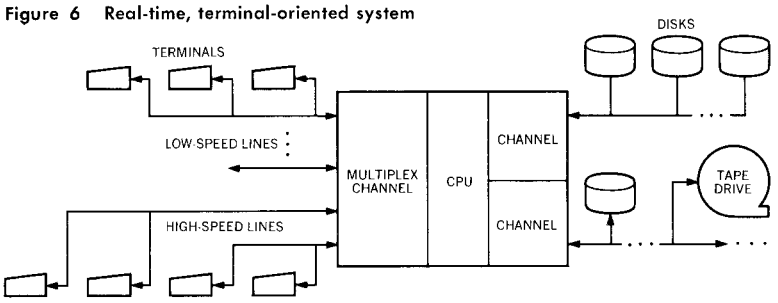
$$W_2^{*(1)} = \frac{0.6 \times 2}{1 - 0.3 - 0.3} = 3 \text{ seconds}$$

$$W_2^{(1)} = W_2^{*(1)}(1 + \lambda_1 d_1^{(1)}) = 3 \times 1.436 = 4.31 \text{ seconds}$$

Mean waiting time for the third priority:

$$W_3^{*(1)} = \frac{0.9 \times 2}{1 - 0.3 - 0.3 - 0.3} = 18 \text{ seconds}$$

Figure 6 Real-time, terminal-oriented system



$$W_3^{(1)} = \frac{W_3^{*(1)}}{1 - \lambda_1 a_1 - \lambda_2 a_2} = \frac{18}{0.4} = 45 \text{ seconds}$$

Case 2 Nonpreemptive discipline

Mean waiting time for the first priority:

$$W_1^{*(1)} = \frac{0.9 \times 2}{1 - 0.3} = 2.57 \text{ seconds}$$

$$W_1^{(1)} = 2.57 \text{ seconds}$$

Mean waiting time for the second priority:

$$W_2^{*(1)} = \frac{0.9 \times 2}{1 - 0.3 - 0.3} = 4.5 \text{ seconds}$$

$$W_2^{(1)} = \frac{4.5}{1 - 0.3} = 6.4 \text{ seconds}$$

Mean waiting time for the third priority:

$$W_3^{*(1)} = \frac{0.9 \times 2}{1 - 0.3 - 0.3 - 0.3} = 18 \text{ seconds}$$

$$W_3^{(1)} = \frac{18}{1 - 0.3 - 0.3} = 45 \text{ seconds}$$

The mean waiting times of the first and second priorities in Case 2 are higher than those for Case 1 because the server in Case 2 completes the execution of a lower-priority call in service before interrupted to serve a higher-priority call.

Example terminal-oriented system

We now analyze a hypothetical real-time airline reservation system to illustrate the use of the queuing models presented in this paper for computing average system response time. In the overall system, shown in Figure 6, messages are sent and received between remote terminals and the data center through high-speed and low-speed teletype lines.

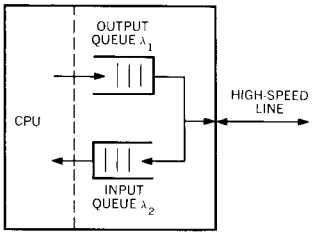
In the hypothetical network, several high-speed lines are rated at 600 characters per second, and low-speed lines are rated at 14 characters per second. Input messages have an average length of 27 characters with an assumed Erlang-2 distribution as shown in Figure 1. Output messages from the data center have an average length of 90 characters with an assumed Erlang-3 distribution.

Both high-speed and low-speed lines are half-duplex data links, wherein input and output messages are sent through the same line. We also assume a one-to-one ratio of input to output messages, i.e., for every input message there is an output message.

high-speed line analysis

We begin by breaking down the system and analyzing its parts, starting with an analysis of the high-speed lines. During peak-traffic periods, high-speed lines are assumed to have an input traffic rate of 3 messages per second per line, which are acquired by the computer through the process of polling. Polling messages are 3 characters in length. To further simplify the analysis, we assume that an input message has an average length of 30 characters (i.e., 27 data characters plus 3 polling characters). Output messages have a higher priority than input messages. Referring to Figure 7, the high-speed communications problem can be formulated as a priority queuing system as follows:

Figure 7 Queuing model of a communication line



$\lambda_1 = 3$ messages/second

$\lambda_2 = 3$ messages/second

Average service times for output messages (Erlang-3 case) are the following:

$$\alpha_1^{(1)} = \frac{90}{600} = 0.15 \text{ seconds}$$

$$\alpha_1^{(2)} = \frac{(3 + 2 - 1)!}{(3 - 1)!}\left(\frac{\alpha_1}{3}\right)^2 = \frac{4 \times 3 \times 2}{2 \times 1} \cdot \frac{(0.15)^2}{9}$$

$$= 0.03 \text{ seconds}^2$$

Average service times for input messages (Erlang-2 case) are the following:

$$\alpha_2^{(1)} = \frac{30}{600} = 0.05 \text{ seconds}$$

$$\alpha_2^{(2)} = \frac{(2 + 2 - 1)!}{(2 - 1)!}\left(\frac{0.05}{2}\right)^2 = 0.00375 \text{ seconds}^2$$

Using the nonpreemptive-priority-queuing formula, we find the following average waiting times:

Output message

$$W_1^{(1)} = \frac{\lambda_1\alpha_1^{(2)} + \lambda_2\alpha_2^{(2)}}{2[1 - \lambda_1\alpha_1^{(1)}]} = \frac{3 \times 0.03 + 3 \times 0.00375}{2(1 - 3 \times 0.15)}$$

$$= 0.092 \text{ seconds}$$

Input message

$$W_2^{(1)} = \frac{\lambda_1\alpha_1^{(2)} + \lambda_2\alpha_2^{(2)}}{2[1 - \lambda_1\alpha_1^{(1)} - \lambda_1\alpha_2^{(2)}][1 - \lambda_1\alpha_1^{(1)}]}$$

$$= 0.154 \text{ seconds}$$

Thus, the average response time for the input message plus the output message in a high-speed communication line is as follows:

$$T_c = W_1^{(1)} + \alpha_1^{(1)} + W_2^{(1)} + \alpha_2^{(1)}$$

$$= 0.092 + 0.15 + 0.154 + 0.05 = 0.446 \text{ seconds}$$

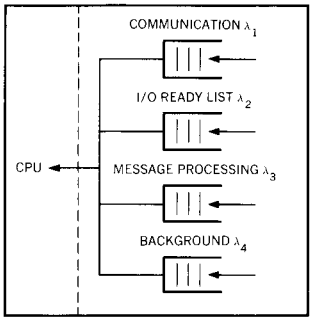
low-speed line analysis

To analyze the low-speed lines, we single out the line with the highest peak traffic (worst-case line). Certain facts must be provided by the customer such as the following. The average service time for an input message is 2 seconds, and the average service time for an output message is 6 seconds. Thus, we can compute the delay (waiting time) and response time for the given line traffic using a method similar to that of high-speed line analysis.

CPU analysis

We now discuss the effect of the CPU. A real-time teleprocessing system of the type illustrated by Figure 6 is often organized as a priority queuing system, wherein important tasks are processed by the computer immediately upon arrival. In the following analysis, the CPU has 4 processing queues (or queue lists) as shown in Figure 8. Input messages arriving from the communication network and output messages ready for transmission are handled in the communication queue λ_1 (highest priority).

Figure 8 CPU queuing model



Since the data base (customer's records, etc.) may be too large for main storage, it is often stored on disk files. Thus, an incoming message must be processed against this data base, requiring a number of I/O accesses to the data base. (For efficiency, the CPU is often programmed in a multiprogramming mode, i.e., after an I/O access is made, the computer may process other tasks.) After the I/O data is in main storage, the computer is called to process messages. The I/O-ready records are the "callers," and they may be placed in a second-priority queue, the I/O Ready List λ_2.

Message processing itself is handled by the third-priority queue λ_3. Again, background programs can be run on an as-available basis. These have the fourth priority and are stacked in the fourth queue λ_4.

We use a preemptive priority queuing model to analyze this system. Let the total input plus output message rate of the first queue during a peak-traffic period be the following:

λ_1 = 10 input plus output messages/second

Assuming an Erlang-2 service-time distribution with a mean service time of 15 milliseconds, we have the following:

$\alpha_1^{(1)}$ = 15 milliseconds = 0.015 seconds

$$\alpha_1^{(2)} = \frac{3 \times 2}{1}\left(\frac{0.015}{2}\right)^2 = 0.00034 \text{ seconds}^2$$

Assuming an average of 6 I/O accesses per message, then

$\lambda_2 = 60$ accesses/second

Also assuming that for each access a constant CPU processing time of 3 milliseconds is required, then

$\alpha_2^{(1)} = 0.003$ seconds

$\alpha_2^{(2)} = 0.000009$ seconds^2

For the message processing queue, we assume that the processing time (service time) has an Erlang-3 distribution with an average of 30 milliseconds. Thus

$\lambda_3 = \lambda_1 = 10$ message/second

$\alpha_3^{(1)} = 0.03$ seconds

$$\alpha_3^{(2)} = \frac{4 \times 3 \times 2}{2 \times 1} \cdot \frac{0.03^2}{9} = 0.0012 \text{ seconds}^2$$

The average waiting time for the communication queue λ_1 is

$$W_1^{(1)} = \frac{10 \times 0.00034}{2(1 - 10 \times 0.015)} = 0.002 \text{ seconds}$$

The average waiting time for the I/O ready list λ_2 is

$$W_2^{(1)} = \frac{10 \times 0.00034 + 60 \times 0.000009}{2(1 - 10 \times 0.015)(1 - 10 \times 0.015 - 60 \times 0.003)}$$

$$= \frac{0.00394}{2 \times 0.85 \times 0.77} = 0.0031 \text{ seconds}$$

The average waiting time in the message-processing queue λ_3 is

$$W_3^{(1)} = \frac{10 \times 0.00034 + 60 \times 0.000009 + 10 \times 0.0012}{2(1 - 10 \times 0.015 - 60 \times 0.003)(1 - 10 \times 0.015 - 60 \times 0.003 - 10 \times 0.003)}$$

$$= \frac{0.00514}{2 \times 0.77 \times 0.47} = 0.0071 \text{ seconds}$$

Omitting background jobs, which would have been similarly computed, the average response time for a message in our hypothetical system is computed as follows:

$$T_p = W_1^{(1)} + \alpha_1^{(1)} + 6[W_2^{(1)} + \alpha_2^{(1)}] + W_3^{(1)} + a_3^{(1)}$$

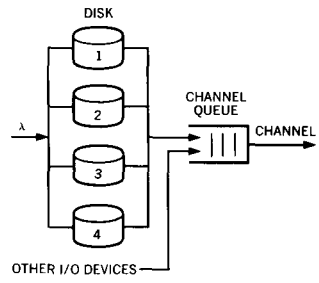
$$= 0.0917 \text{ seconds}$$

disk-file analysis

Assume an auxiliary storage of 8 disk files, which are connected to 2 channels, i.e., 4 files to each channel. In addition to disks, which are used for real-time applications, an unspecified number of tape drives are also attached to the channels. The configuration for one of the two channels is shown in Figure 9. Assume that the traffic

(I/O accesses) is evenly divided among the disks. Also, the time required for a channel to serve a desk request is calculated as a sum of rotational delay time plus data transfer time, which is a variable quantity that depends on record length. Thus, the average rotational delay equals 17 milliseconds, and the data transfer time (variable) equals 10 milliseconds.

Figure 9 Model of a disk system



Since the channel may also serve other devices, the average channel service time additionally depends on service parameters of those devices. Assuming a mean channel service time of 30 milliseconds and a variance of 1000 milliseconds, we use a single-server queuing formula to compute the channel waiting time and its variance. For example, assume that for a certain traffic rate on the channel, we obtain the following data: The average channel waiting time is

$W_c = 10$ milliseconds

and the variance is

$\sigma_c^2 = 100$ milliseconds^2

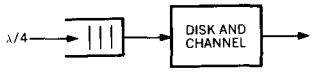
With the disk system cascaded, we can modify the service time of a disk to include the service time and the waiting time of a channel. Although this additive model is an approximation, we know that the channel time is small compared to the service time (disk arm motion) and, therefore, expect a reasonably accurate result. Assume that the arm-motion time in a disk file has the following statistics:

$\alpha_d^{(1)} = 100$ milliseconds average arm motion time

$\alpha_d^{(2)} = 15000$ milliseconds^2 second moment of the arm motion

Service time computations for the simplified disk-channel model (shown in Figure 10) are obtained as follows:

Figure 10 Simplified model



$\alpha^{(1)} = 100 + 30 + 10 = 140$ milliseconds average service time

The variance of the service time is

$\sigma^2 = 15000 - 100^2 + 1000 + 100 = 6100$ milliseconds^2

The second moment of the service time is

$\alpha^2 = (140)^2 + 6100 = 24700$ milliseconds^2

(Note that the second moment is equal to the square of the mean plus the variance.) The traffic rate is

$\lambda = 60/2 = 30$ accesses per second per channel (i.e. $30/4$ accesses per second per disk)

The average waiting time on a disk

$$W^{(1)} = \frac{(30/4)(24700/10^6)}{2[1 - (30/4)(0.14)]} = 0.46 \text{ seconds}$$

The average response time for a disk is

$$T_d = 0.46 + 0.140 = 0.60 \text{ seconds}$$

total response time

The total average response time can now be calculated for a terminal attached to a high-speed line as follows:

$$T = T_c + 6T_d + T_p = 0.446 + 6 \times 0.60 + 0.0917$$

$$= 4.138 \text{ seconds}$$

Thus, disk files form a complex queuing system in which we repeatedly apply the single-server queuing analysis. First the channel is considered as a single-server queue, and its waiting time and moments are computed. The disk-file system is then considered, and channel delays are added to disk service times. A single disk is treated as a single server with a traffic rate equal to that of the other disks. Effects of other approximations are discussed in Reference 19.

Although one may calculate second moments in each model and thereby calculate the total response-time variance as a sum of the individual variances, the procedure may not yield sufficiently accurate results because the variance is sensitive to the assumptions made in the simplified model. Such a procedure should only be used to give the analyst an idea of the degree of variation of the total response time. For a more detailed analysis, simulation techniques should be used.

Concluding remarks

Queuing theory has proved useful for analyzing service and congestion in many computer subsystems. Emphasis here has been on the basic principles and logical steps required to solve queuing problems because many variations of computer congestion problems do not fit standard models. Thus single-server queuing processes with random input, general service times, and priorities are reviewed. Waiting-time, response-time, and busy-period distributions are found by using their Laplace transforms. Queue-size distribution is obtained from the queue-size generating function. Methods of determining the means and second moments are also given. To further aid the analyst in creating his own model, the technique of imbedded Markov chains is presented.

Several examples are presented to illustrate these various techniques. Both the utility and limits of queuing analysis are illustrated by a detailed analysis of a practical (but hypothetical) teleprocessing system and communication network. This example implies that the present state of the art of queuing theory does not permit the detailed analysis of a complete system, but it is useful for subsystem analysis. Advanced research in computer queuing analysis currently includes the study of time-sharing algorithms[22,43,44] and multiserver systems.

Appendix A: Markov chain

A sequence of random variables $\xi_0, \xi_1, \cdots \xi_n \cdots$, where each random variable may assume any integer value or zero, forms a Markov chain if for all n (where $n = 1, 2, 3, \cdots$) and for all possible values of the random variables ξ_n

$$P\{\xi_n = j \mid \xi_0 = i_0, \xi_1 = i_1, \cdots, \xi_{n-1} = i_{n-1}\}$$
$$= P\{\xi_n = j \mid \xi_{n-1} = i_{n-1}\}$$

If $\xi_n = j$, the system is in state j at the nth step. The probability distribution of the random variable ξ_0

$$P\{\xi_0 = j\}$$

for j any integer value greater than or equal to zero, is called the "initial distribution." The conditional probabilities

$$P\{\xi_n = j \mid \xi_{n-1} = i\}$$

are called the "transition probabilities" and are often given in matrix form.

If we know the initial probability distribution

$$P\{\xi_0 = j\}$$

where $j = 0, 1, 2, \cdots$

and transition probabilities in a Markov chain, then we can uniquely determine the probability distribution of each random variable

$$\xi_n (n = 1, 2, \cdots)$$

by the following formula:

$$P\{\xi_n = j\} = \sum_{i=0}^{\infty} P\{\xi_n = j \mid \xi_{n-1} = i\} P\{\xi_{n-1} = i\}$$

where $n = 1, 2, \cdots$

A Markov chain is called "homogeneous" if the transition probabilities are independent of n. Let P_{ij} be defined as

$$P_{ij} = P\{\xi_n = j \mid \xi_{n-1} = i\}$$

A probability distribution is stationary if it is independent of n

$$P\{\xi_n = j\} = P_j$$

The stationary probability distribution of a homogeneous Markov chain can be found by solving the following equation:

$$P_j = \sum_{i=0}^{\infty} P_{ij} P_i$$

where $j = 0, 1, \cdots$

In the case of a finite Markov chain, i.e., the number of states in the system is N, the probability P_j (where $j = 0, 1, \cdots, N$) can be found by solving the $N + 1$ linear equations

$$P_i = \sum_{i=0}^{N} P_{ij} P_i$$

and

$$\sum_{j=0}^{N} P_j = 1$$

When there is an infinite number of states, i.e., $N = \infty$, P_j can be determined by the generating function technique discussed in Appendix C.

Appendix B: Notations in an M/G/1 queuing process

Let the service times for priority k customers be mutually independent, positive random variables with a distribution function $H_k(x)$. When there is only one kind of call present in the system (i.e., there are no priorities), the subscript k is dropped.

It is convenient to use the Laplace-Stieltjes transforms of the distritions H_k, W_k, and T_k as used in Equations B1, B3, and B5.
Let

$$\psi_k(s) = \int_0^\infty e^{-sx}\, dH_k(x) \tag{B1}$$

and the rth moment is given as

$$\alpha_k^{(r)} = \int_0^\infty x^r\, dH_k(x) \tag{B2}$$

The first moment of Equation B2

$$\alpha_k = \alpha_k^{(1)}$$

is the mean service time for the priority k caller. Referring to Equation B4, where $W_k(x)$ is the stationary waiting-time distribution for callers of priority-class k, we define the second Laplace-Stieltjes transform by

$$\Omega_k(s) = \int_0^\infty e^{-sx}\, dW_k(x) \tag{B3}$$

where the rth moment is given by

$$W_k^{(r)} = \int_0^\infty x^r\, dW_k(x) \tag{B4}$$

Further in Equation B6, where $T_k(x)$ is the stationary response time distribution (i.e., waiting time plus service time) for calls of priority class k, the third transform is defined as follows:

$$\theta_k(s) = \int_0^\infty e^{-sx}\, dT_k(x) \tag{B5}$$

and the rth moment is given by

$$T_k^{(r)} = \int_0^\infty x^r \, dT_k(x) \tag{B6}$$

Let $P_j(k)$ be the probability that the queue length is j for priority classes less than or equal to k, when the system is observed at the end of service of a priority class less than or equal to k. Define the generating function

$$G_k(z) = \sum_{j=0}^{\infty} P_j(k) z^j$$

In a priority queuing system, *busy period* for priority class less than or equal to k is defined as beginning when a call of priority class less than or equal to k finds the server free of calls of priority classes less than or equal to k and continuing until the instant at which the server is again free of calls of priority classes less than or equal to k. Let $D_k(x)$ be the busy period distribution for priority classes less than or equal to k, and define the Laplace transform

$$\gamma_k(s) = \int_0^\infty e^{-sx} \, dD_k(x)$$

and the moments of the busy-period distribution as

$$d_k^{(r)} = \int_0^\infty x^r \, dD_k(x)$$

Appendix C: Queue length distribution

Let ν_n be the number of calls that arrive during the service of the nth caller, where ν_n is a conditional random variable that depends on the service time. The probability distribution of ν_n can be determined as follows. Let χ_n be the service time of the nth customer. Then

$$P\{\nu_n = i\} = \int_0^\infty P\{\nu_n = i \mid \chi_n = x\} \, dH(x) \tag{C1}$$

Using as input a Poisson process of parameter λ, we have from Reference 16

$$P\{\nu_n = i \mid \chi_n = x\} = \frac{(\lambda x)^i e^{-\lambda x}}{i!} \tag{C2}$$

Substituting Equation C2 into Equation C1 yields the probability distribution

$$P\{\nu_n = i\} = \int_0^\infty \frac{(\lambda x)^i e^{-\lambda x}}{i!} \, dH(x)$$

The queue lengths ξ_{n+1} and ξ_n and the number of calls received during the service of the nth caller are related by the following equation

$$\xi_{n+1} = \begin{cases} \xi_n - 1 + \nu_n & \text{if} \quad \xi_{n+1} > 0 \\ \nu_{n+1} & \text{if} \quad \xi_n = 0 \end{cases} \tag{C3}$$

Equation C3 can be explained as follows. When $\xi_n > 0$, the head of the queue of ξ_n becomes the $n + 1$st caller, entering the service at time $t = \tau'_n$ and leaving at time $t = \tau'_{n+1}$. If $\nu_{n+1} = 0$, the queue size at $t = \tau'_{n+1}$ in the system is reduced by one because the $n + 1$st caller has just left the system. Since, in general, ν_{n+1} callers arrive during the service time of the $n + 1$st caller, we have

$$\xi_{n+1} = \xi_n - 1 + \nu_{n+1}$$

However, if $\xi_n = 0$, the system is empty immediately after τ'_n. In this case, the $n + 1$st caller arrives at τ_{n+1}, where $\tau_{n+1} > \tau'_n$ and leaves at τ'_{n+1}. Thus $\tau'_{n+1} - \tau_{n+1}$ is the service time of the $n + 1$st caller. If ν_{n+1} new callers arrive during the service time of the $n + 1$st caller, then when the $n + 1$st departs, ν_{n+1} callers are present in the system. Hence

$$\xi_{n+1} = \nu_{n+1}$$

If we introduce a notation

$$a^+ = \begin{cases} a & \text{if } a > 0 \\ 0 & \text{if } a \leq 0 \end{cases}$$

then the queue lengths in Equation C3 can be written as

$$\xi_{n+1} = (\xi_n - 1)^+ + \nu_{n+1} \tag{C4}$$

Assuming that the stationary distribution of the queue length exists, then ξ_{n+1} and ξ_n must have the same marginal distribution. (See Reference 7 for aid in proving the existence of a stationary distribution under the condition $\lambda\alpha < 1$, where α is the average service time.) For our purpose here, we shall particularize the suggested proof by showing that if $\lambda\alpha < 1$, a stationary distribution of ξ_n exists. The ξ_n callers form an imbedded Markov chain, which we study by using the generating functions discussed in Reference 39. The generating function for ν_n can be written as

$$\sum_{i=0}^{\infty} P\{\nu_n = i\} z^i = \int_0^{\infty} \sum_{i=0}^{\infty} \frac{(\lambda x)^i}{i!} z^i e^{-\lambda x} \, dH(x)$$

$$= \int_0^{\infty} e^{-\lambda(1-z)x} \, dH(x) = \psi[\lambda(1 - z)] \tag{C5}$$

Equation C5 implies that if we replace s in $\psi(s)$ by $\lambda(1 - z)$, we obtain the generating function of ν_n.

Define the probability that there are j callers in queue of length ξ_n as

$$P\{\xi_n = j\} = P_j$$

and define a new generating function $G(z)$ for P_j as

$$G(z) = \sum_{j=0}^{\infty} P_j z^j \tag{C6}$$

If the stationary solution in queue length exists, ξ_{n+1} and ξ_n must have the same marginal distribution. Their generating functions

must also be the same. If we let $\Gamma(z)$ be the generating function of $(\xi_n - 1)^+$, then

$$\Gamma(z) = E[z^{(\xi_n - 1)^+}] = P_0 + P_1 + P_2 z + P_3 z^2 + \cdots$$

$$= P_0 + \frac{G(z) - P_0}{z}$$

From Equation C4, ξ_{n+1} is the sum of two random variables, ν_{n+1} and $(\xi_{nn} - 1)_i^+$. Using the theorem that the generating function of the sum of two independent variables is the product of the two generating functions,[39] it follows that

$$G(z) = \left[P_0 + \frac{G(z) - P_0}{z}\right]\psi[\lambda(1 - z)] \tag{C7}$$

It should be noted that in a stationary process ξ_{n+1} and ξ_n must have the same generating function as given in Equation C5. Solving for $G(z)$ in Equation C7, we obtain

$$G(z) = \frac{P_0(z - 1)\psi[\lambda(1 - z)]}{z - \psi[\lambda(1 - z)]} \tag{C8}$$

where P_0 remains to be determined. From Equation C8, since

$$\sum P_i = 1$$

then from Equation C6

$$G(1) = 1$$

By Equation C5

$$\psi[\lambda(1 - z)]|_{z=1} = \psi(0) = 1$$

and

$$1 = P_0 \lim_{z \to 1} \frac{(z - 1)\psi[\lambda(1 - z)]}{z - \psi[\lambda(1 - z)}.$$

Using the L'Hospital's rule

$$1 = P_0 \frac{1}{1 + \lambda\psi'(0)}$$

However

$$\psi'(0) = -\int_0^\infty x \, dH(x) = -\alpha$$

the negative average service time. Thus

$$P_0 = 1 - \lambda\alpha \tag{C9}$$

If $\lambda\alpha$ in Equation C9 is greater than 1, this leads to a contradiction of positive probabilities, and therefore a stationary distribution of the queue size cannot exist. If $\lambda\alpha < 1$, then

$$P_0 = 1 - \lambda\alpha$$

so that the generating function of the queue size exists as given in Equation C7. Therefore, the stationary solution of the system exists.

Appendix D: Waiting-time distribution

The waiting-time distribution of the M/G/1 queuing process is discussed here using concepts of the queue-size generating function given in Appendix C.

Let η_n be the waiting time, χ_n the service time, and β_n the queuing time of the nth caller. Then

$$\beta_n = \eta_n + \chi_n \tag{D1}$$

Suppose that the nth caller arrives at time τ_n and departs at time τ_n'. Then the total time the nth caller spends in the system (i.e., queuing time) is

$$\beta_n = \tau_n' - \tau_n$$

If there are no new arrivals during the time period β_n, the queue length ξ_n must be zero when the nth customer leaves. However if there are five new arrivals during β_n, the queue length ξ_n must then be five. Thus, in general, the number of new arrivals during the queuing time of the nth customer must be equal to the queue length ξ_n at the nth departure.

Since the number of new arrivals has a Poisson distribution, we have the following queue-length queuing-time probability

$$P\{\xi_n = j \mid \beta_n = x\} = \frac{e^{-\lambda x}(\lambda x)^j}{j!}$$

If we let $T(x)$ be the queuing-time distribution, we obtain the queue-length distribution by the following integration:

$$P = \{\xi_n = j\} = \int_0^\infty \frac{e^{-\lambda x}(\lambda x)^j}{j!}\, dT(x)$$

Forming the generating function, and recalling that

$$P\{\xi_n = j\} = P_j$$

from Equation C6, we obtain

$$\sum_{j=0}^{\infty} P\{\xi_n = j\}z^j = G(z) = \int_0^\infty \sum_{j=0}^{\infty} \frac{e^{-\lambda x}(\lambda x z)^j}{j!}\, dT(x)$$

$$= \theta[\lambda(1 - z)] \tag{D2}$$

where $\theta(s)$ is the Laplace-Stieltjes transform of the queuing-time distribution. The queue-length generating function (Equation D2) can be obtained by replacing s by $\lambda(1 - z)$ in $\theta(s)$. However, from Equation D1, we have from the convolution of the two functions

$$\theta(s) = \Omega(s)\psi(s)$$

where $\Omega(s)$ is the Laplace-Stieltjes transform of the waiting-time distribution. Hence

$$\Omega[\lambda(1 - z)]\psi[\lambda(1 - z)] = \frac{P_0(z - 1)\psi[\lambda(1 - z)]}{z - \psi[\lambda(1 - z)]} \tag{D3}$$

If we replace $\lambda(1 - z)$ by s, then

$$\Omega(s) = \frac{1 - \lambda\alpha}{1 - \lambda\{[1 - \psi(s)/s]\}} \tag{D4}$$

This is the Laplace-Stieltjes transform of the waiting-time distribution for the M/G/1 queuing process as developed by Pollaczek and Khintchine.[3,5,7]

Appendix E: Busy-period distribution

Let $D(x)$ be the busy-period distribution. Then define

$$D^n(x) = \int_0^\infty D^{(n-1)}(x - y)\, dD(y)$$

where $n > 1$

as the nth folded convolution of $D(x)$ with itself. Also define

$D_0(x) = 1$ and

$D_1(x) = D(x)$

Suppose that n arrivals appear during the interval of the first service time y in a busy period; the probability of this event is

$e^{-\lambda y}(\lambda y)^n/n!$

If $n = 0$, the busy period consists of serving the first caller during a service time y; if $n = 1$, then the busy period is the sum of y and an additional busy period that is initiated by the new arrival. Since the duration of the busy period is independent of the order of service, for example when $n = 3$, it may be assumed that the busy period is the sum of four random variables with distributions $H(x)$, $D(x)$, $D(x)$, and $D(x)$, or simply assume two random variables with distribution functions $H(x)$ and $D^{(3)}(x)$.

Following this line of thought, one can obtain the busy-period distribution by solving

$$D(x) = \int_0^x \sum_{n=0}^{\infty} \frac{e^{-\lambda y}(\lambda y)^n}{n!}\, D^n(x - y)\, dH(y) \tag{E1}$$

Designating $\gamma(s)$ as the Laplace-Stieltjes transform of $D(x)$, then

$$[\gamma(s)]^n = \int_0^\infty e^{-sx}\, dD^n(x)$$

It follows from Equation E1 that

$$\begin{aligned}\gamma(s) &= \int_0^\infty e^{-(s+\lambda)y} \sum_{n=0}^{\infty} \frac{(\lambda y)^n[\gamma(s)]^n}{n!}\, dH(y) \\ &= \int_0^\infty e^{-[s+\lambda-\lambda\gamma(s)]y}\, dH(y) \\ &= \psi\{s + \lambda[1 - \gamma(s)]\}\end{aligned} \tag{E2}$$

CITED REFERENCES AND FOOTNOTES

1. E. Brockmeyer, H. L. Halstrom, and A. Jensen, "The Life and Works of A. K. Erlang," *Translations of the Danish Academy of Technical Sciences,* No. 2, 1–277 (1948).
2. F. Pollaczek, "Ueber eine Aufgabe der Wahrscheinlichkeitstheorie I," *Mathematische Zeitschrift* **32,** 64–100 (1930); *ibid* II, **32,** 729–750 (1930).
3. F. Pollaczek, "Ueber das Warteproblem," *Mathematische Zeitschrift* **38,** 429–537 (1934).
4. A. N. Kolmogorov, Sur le problème d'attente," *Mathematicheskii Sbornik* **38,** 1–2, 101–106 (1931).
5. A. Y. Khintchine, "Mathematical theory of a stationary queue," *Mathematicheskii Sbornik* **39,** 4, 73–84 (1932).
6. D. B. Kendall, "Some problems in the theory of queues," *Journal of the Royal Statistical Society, Series B* **XIII,** No. 2, 151–185 (1951).
7. 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)
8. D. V. Lindley, "The theory of queues with a single server," *Proceedings of the Cambridge Philosophical Society* **48,** 277–289 (1952).
9. L. Takacs, "Investigation of waiting time problems by reduction to Markov processes," *Acta Mathematica Academiae Scientiarum Hungaricae* **6,** 101–129 (1955).
10. L. Takacs, "Transient behavior of single-server queuing processes with recurrent input and exponentially distributed service times," *Operations Research* **8,** 231–245 (1960).
11. L. Takacs, "The transient behavior of a single server queuing process with Poisson input," *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability* **II,** 535–567, University of California Press, Berkeley, California (1961).
12. T. L. Saaty, *Elements of Queuing theory with Applications,* McGraw-Hill Book Co., New York, New York (1961).
13. W. Chang and D. J. Wong, "Computer channel interference analysis," *IBM Systems Journal* **4,** No. 2, 162–170 (1965).
14. W. Chang, "Congestion analysis of a computer core storage system," *Naval Research Logistics Quarterly* **14,** No. 3, 367–379 (1967).
15. W. Chang, "Queues with feedback for time-sharing computer system analysis," *Operations Research* **16,** No. 3, 613–627 (May–June 1968).
16. L. Takacs, *Introduction to the theory of queues,* Oxford University Press, New York, New York (1962).
17. N. V. Prabhu, *Queues and Inventories: A Study of their Basic Stochastic Processes,* John Wiley and Sons, New York, New York (1964).
18. D. R. Cox and W. L. Smith, *Queues,* John Wiley and Sons, New York, New York (1962)
19. *IBM Data Processing Techniques Manual—Analysis of Some Queuing Models in Real-Time* may be obtained through IBM Branch Offices by ordering F20-0007-1.
20. E. G. Coffman, *Stochastic Models of Multiple and Time-Shared Computer Operations,* Report No. 66-38, Department of Engineering, University of California at Los Angeles (June 1966). This report may also be obtained through the Clearinghouse for Federal Scientific and Technical Information by ordering AD636 976).
21. L. Kleinrock, "Analysis of a time-shared processor," *Naval Research Logistics Quarterly* **11,** No. 10, 59–73 (1964).
22. L. Kleinrock, "Time-shared systems: a theoretical treatment," *Journal of the Association for Computing Machinery* **14,** No. 2, 242–261 (April 1967).
23. B. Krishnamoorthi and R. C. Wood, "Time-shared computer operations with both interarrival and service times exponential," *Journal of the*

Association for Computing Machinery **13,** No. 3, 317–338 (July 1966).
24. N. R. Patel, "A mathematical analysis of computer time-sharing systems," Interim Technical Report No. 20, Army Research Office (Durham), Grant No. DA-ARO(D)-31-124-G158, Operations Research Center, MIT, Cambridge, Massachusetts (1964).
25. L. E. Schrage, *Some Queuing Models for a Time-Shared Facility,* Ph.D. Dissertation, Department of Industrial Engineering, Cornell University, Ithaca, New York (1966). (Also see *Dissertation Abstracts* **26,** Order No. 66-3686, p. 7186.)
26. A. L. Scherr, *An Analysis of Time-Shared Computer Systems,* Research Monograph No. 36, The MIT Press, Cambridge, Massachusetts (1967).
27. W. Chang, "Queuing with nonpreemptive and preemptive-resume priorities," *Operations Research* **13,** No. 6, 1020–1022 (November–December 1965).
28. W. Chang and D. J. Wong, "Analysis of real-time multiprogramming," *Journal of the Association for Computing Machinery* **12,** No. 4, 581–588 (October 1965).
29. L. Takacs, "Priority queues," *Operations Research* **12,** No. 1, 63–74 (January–February 1964).
30. D. P. Gaver, Jr., "A waiting line with interrupted service, including priorities," *Journal of the Royal Statistical Society,* Series **13,** No. 24, 73–90 (1962).
31. W. Chang, "Priority queues with recurrent input," *Fifth International Teletraffic Congress,* 85–88 (Bell Telephone System and United States Independent Telephone Association 1967).
32. A. Cobham, "Priority assignment in waiting line problems," *Operations Research* **2,** No. 1, 70–76 (February 1954).
33. A. Cobham, "Priority assignment—a correction," *Operations Research* **3,** No. 4, 547 (November 1955).
34. R. G. Miller, "Priority Queues," *The Annals of Mathematical Statistics* **31,** 86–103 (1960).
35. P. D. Welch, "On preemptive resume priority queues," *The Annals of Mathematical Statistics* **35,** 600–612 (1964).
36. N. K. Jaiswal, "Time-dependent solution of the head-of-the-line priority queue," *Journal of the Royal Statistical Society, Series B* **24,** 91–101 (1962).
37. W. Chang, "Preemptive priority queues," *Operations Research* **13,** No. 5, 620–623 (September–October 1965).
38. N. K. Jaiswal, *Priority Queues,* Academic Press, New York, New York (1968).
39. W. Feller, *An Introduction to Probability Theory and its Applications, Vol. I* and *Vol. II,* John Wiley and Sons, New York, New York (1968 and 1966).
40. J. Riordan, *Stochastic Service Systems,* John Wiley and Sons, New York, New York (1962).
41. L. Takacs, "The time dependence of a single server queue with Poisson input and general service times," *The Annals of Mathematical Statistics* **33,** No. 4, 1340–1348 (1962).
42. D. P. Gaver, Jr., "Observing stochastic processes, and approximate transform inversion," *Operations Research* **14,** No. 3, 444–459 (May–June 1966).
43. D. Chazan, A. G. Konheim, and B. Weiss, "A note on time-sharing," *Journal of Combinatorial Theory* **5,** 4, 344–369 (December 1968).
44. A. G. Konheim, "A note on time sharing with preferred customers," *Zeitschrift fuer Wahrscheinlichkeitstheorie und verwandte Gebiete* **9,** 112–130 (1968).