W. Chang

Sequential Server Queues for Computer Communication System Analysis

Abstract: A queuing model with two sequential servers is developed to analyze performance in computer and communication systems. In one case the CPU is the first server and the terminal and its associated communications equipment are the second server. In a second case the CPU and the channel are the first server and the auxiliary storage device is the second server. We study the queuing behavior of the sequential server systems with Poisson arrivals, general service time distributions, and several service disciplines, including bulk arrivals, message priorities, and the input and output queues. The stationary distributions of the queue lengths and waiting times are determined by using an imbedded Markov chain analysis. Several examples are given to illustrate the applications of these models to practical problems.

Introduction

Networks of queues with blocking occur often in computer and communication systems because subsystems with different speeds are used to accomplish a common task. For example, in an information retrieval system data records are transferred between auxiliary storage devices and main memory. Queues may be formed when the CPU, the auxiliary storage devices or the channels are busy. Another example is in a communication system, where messages are transmitted from one station to another. If either of the stations is busy, the transmission is blocked, and a queue of messages may be formed at the station.

The simplest queuing network consists of two servers in sequence. If there is a large buffer between them (i.e., a large queue may be allowed for the second server), it is possible to analyze this system by considering the two servers as separate single-server systems. The output distribution of the first server becomes the input distribution of the second server. In particular, if the input to the first server has a Poisson distribution and the service time is exponential, the output at the first server is also Poisson [1]. In this case the second queue can be easily analyzed. If the service time at the first server is not exponential, the output distribution can be determined by using contour integration [2]. In this case, however, the analysis of the second stage is complicated because the output from the first server is not statistically independent.

The total time for a call to go through the two servers is the sum of the queuing times at these servers. More realistic models are characterized by a limitation on the queue size that is allowed between the servers. Hunt [3] has provided some useful results for this case when the service times of the servers are exponential.

A special case of the limited queue system is that in which only one call can be processed at the second server, and no waiting line is allowed between the stages. This queuing system can be reduced to a special single-server system. It has been investigated by Suzuki [4], Avi-Itzhak and Yadin [5], Prabhu [6], and Chang [7]. For an input function with Poisson distribution, Suzuki studied a Markov chain imbedded in the queuing process, Avi-Itzhak and Yadin investigated the stationary waiting time and queue length distribution, and Prabhu studied the transient behavior of this type of system. Chang studied the two servers in sequence for an input with Erlangian distribution.

In this paper we investigate the application of a queuing model to the analysis of computer and communication systems and develop some additional models for bulk arrivals and priority services. We assume a Poisson distributed input, although the model can be extended to cover an Erlangian input if the need arises. The arriving calls are served first at server 1 and then at server 2. When a call completes its service at server 1, it either goes to server 2, if it is free, or stays at server 1 and blocks further service there until the second server becomes free. Each call spends some time being served at the first server and then possibly some time waiting for the second server to become free (blocking time). The combined time that the call stays at the first server can be interpreted as its actual service time there, so that the queue at server 1 behaves like a single-server queue with a general service time distribution and an input with Poisson distribution. The analysis is complicated by the fact that a call which arrives when the queue is empty may experience a lesser degree of blocking, because the second server may have become free at the time of its

arrival. An analysis of the single-server system with two service time distributions [5] is useful in treating the sequential-server problem for this case. Numerical results are presented in a later section.

Service time distributions

Let calls n $(n = 1, 2, \cdots)$ arrive at the first server with a Poisson distribution. If the first server is busy or blocked, the arriving call joins a queue and waits there for the first server. Let the service times in both servers be identically distributed, independent, random variables with probability distributions $C(x)$ and $D(x)$, respectively. Let $x_n^{(1)}$ and $x_n^{(2)}$ be the service times of the nth call in the servers. Let x_n be the length of time that the nth call spends in the first server. This includes the actual service time at the first server and the waiting time, if any, for the second server to become free. Suppose that there is a queue when the nth call arrives. Obviously, $x_n = \max\{x_n^{(1)}, x_{n-1}^{(2)}\}$ and

$$H(x) = P\{x_n \le x\} = P\{x_n^{(1)} \le x, x_{n-1}^{(2)} \le x\} = C(x)\, D(x). \tag{1}$$

Let the Laplace-Stieltjes transform be

$$\phi(s) = \int_0^\infty e^{-sx}\, dH(x) \tag{2}$$

and the nth moment be

$$\alpha_n = \int_0^\infty x^n\, dH(x) = (-1)^n \phi^{(n)}(0). \tag{3}$$

In particular, let the first moment be α, i.e., $\alpha = \alpha_1$.

Suppose that when the $(n-1)$th call enters the second server the queue is empty. The first server is idle until the nth call arrives. Let t be the time between the instant that the $(n-1)$th call enters the second server and the instant that the nth call arrives at the first server. In this case, the length of time that the nth call spends in the first server is

$$x_n = \max\,[x_n^{(1)}, x_{n-1}^{(2)} - t]. \tag{4}$$

Define the probability distribution of x_n as

$$H(x, t) = P\{x_n^{(1)} \le x, x_{n-1}^{(2)} - t \le x\} = C(x)\, D(x + t). \tag{5}$$

Thus, the combined time of the nth call, which arrives when the queue is empty, depends on the length of the idle time t of the first server. The probability density function of t is given by $\lambda e^{-\lambda t}$; therefore we obtain the Laplace-Stieltjes transform of the combined time distribution as follows:

$$\phi(s) = \int_0^\infty \lambda e^{-\lambda t} dt \int_0^\infty e^{-sx}\, dH(x, t), \tag{6}$$

and the moments are $\beta_n = (-1)^n\, \phi^{(n)}(0)$. In particular, let the first moment be β, i.e., $\beta = \beta_1$.

More specifically, let the service time distributions be exponential:

$$C(x) = 1 - e^{-\mu x}, \text{ and } D(x) = 1 - e^{-\nu x}. \tag{7}$$

In this case we have

$$\psi(s) = \frac{\mu}{s+\mu} + \frac{\nu}{s+\nu} - \frac{\mu+\nu}{s+\mu+\nu} \tag{8}$$

and

$$\phi(s) = \frac{\mu}{s+\mu} + \frac{\lambda}{\lambda+\nu}\frac{\nu}{s+\nu} - \frac{\lambda}{\lambda+\nu}\frac{\nu+\mu}{s+\nu+\mu}. \tag{9}$$

The first and second moments are

$$\alpha = 1/\mu + 1/\nu - 1/(\mu+\nu); \tag{10}$$

$$\alpha_2 = 2/\mu^2 + 2/\nu^2 - 2/(\mu+\nu)^2; \tag{11}$$

$$\beta = 1/\mu + [\lambda/(\lambda+\nu)]/\nu - [\lambda/(\lambda+\nu)]/(\mu+\nu); \tag{12}$$

$$\beta_2 = 2/\mu^2 + 2\lambda/[(\lambda+\nu)\nu^2] - 2\lambda/[(\lambda+\nu)(\mu+\nu)^2]. \tag{13}$$

Exponential service time distributions are applicable in communication system analysis; some voice and data transmissions are known to have exponential holding times. Exponential service time assumptions are also applicable in the CPU processing time analysis. Service times for other devices, however, in particular those with mechanical operation, are better approximated by an Erlang distribution. Suppose that the second server has an Erlang-2 distribution, i.e.,

$$D(x) = 1 - e^{-\nu x} - \nu x e^{-\nu s}, \tag{14}$$

and the first server has the same exponential distribution. In this case, $\psi(s)$ and $\phi(s)$ are given by

$$\psi(s) = \frac{\mu}{s+\mu} + \frac{\nu^2}{(s+\nu)^2} - \frac{\mu}{s+\mu+\nu} - \frac{\nu(\mu+\nu)}{(s+\mu+\nu)^2} \tag{15}$$

and

$$\phi(s) = \frac{\mu}{s+\mu} + \frac{\lambda}{\lambda+\nu}\frac{\nu^2}{(s+\nu)^2} + \frac{\lambda\nu}{(\lambda+\nu)^2}\frac{\nu}{s+\nu} - \frac{\lambda\mu}{(\lambda+\nu)(\mu+\nu)}\frac{\mu+\nu}{(s+\mu+\nu)} - \frac{\lambda}{\lambda+\nu}\frac{\nu}{\mu+\nu}\frac{(\mu+\nu)^2}{(s+\mu+\nu)^2} - \frac{\lambda\nu}{(\lambda+\nu)^2}\frac{\mu+\nu}{s+\mu+\nu}. \tag{16}$$

The first and second moments are

$$\alpha = \frac{1}{\mu} + \frac{2}{\nu} - \frac{\mu}{(\mu+\nu)^2} - \frac{2\nu}{(\mu+\nu)^2}, \tag{17}$$

$$\alpha_2 = \frac{2}{\mu^2} + \frac{3}{\nu^2} - \frac{2\mu}{(\mu+\nu)^3} - \frac{3\nu}{(\mu+\nu)^3}; \tag{18}$$

$$\beta = \frac{1}{\mu} + \frac{2\lambda}{\nu(\lambda+\nu)} + \frac{\lambda\nu}{\nu(\lambda+\nu)^2} - \frac{\lambda\mu}{(\lambda+\nu)(\mu+\nu)^2} - \frac{2\lambda\nu}{(\lambda+\nu)(\mu+\nu)^2} - \frac{\lambda\nu}{(\lambda+\nu)^2(\mu+\nu)}; \tag{19}$$

$$\beta_2 = \frac{2}{\mu^2} + \frac{3\lambda}{(\lambda+\nu)\nu^2} + \frac{2\lambda\nu}{(\lambda+\nu)^2\nu^2} - \frac{2\lambda\mu + 3\lambda\nu}{(\lambda+\nu)(\mu+\nu)^3} - \frac{2\lambda\nu}{(\lambda+\nu)^2(\mu+\nu)^2}. \tag{20}$$

Other types of service time distribution can be developed similarly. Once we obtain $\psi(s)$ and $\phi(s)$ and their moments, we can use the formulas developed in [4] and [5] to obtain useful information, such as the mean waiting and queuing times. For completeness of this paper, and for the development of additional models, we include some of the earlier results in the Appendix. Other useful formulas are given below. The queue-size generating function $U(z)$ is given by Eq. (A3) in the Appendix as

$$U(z) = \frac{P_0\{z\phi[\lambda(1-z)] - \psi[\lambda(1-z)]\}}{z - \psi[\lambda(1-z)]}, \tag{21}$$

where P_0 is the probability that a call arrives and finds the first server empty; P_0 is given by Eq. (A5) as

$$P_0 = \frac{1-\lambda\alpha}{1-\lambda\alpha+\lambda\beta}. \tag{22}$$

Let the queuing time in the sequential server system be defined as the time that a call spends in the first server. This time consists of the call's waiting time, service time, and blocking time. Let $\theta(s)$ be the Laplace-Stieltjes transform of the queuing time distribution. It is found from Eq. (A6) as

$$\theta(s) = U(1-s/\lambda) = \frac{P_0[\lambda\phi(s) - \lambda\psi(s) - s\phi(s)]}{\lambda - s - \lambda\psi(s)}. \tag{23}$$

The mean queuing time Y is

$$Y = -\theta'(0) = \frac{\lambda\beta_2 + \left(2 + \frac{\lambda^2\alpha_2}{1-\lambda\alpha}\right)\beta}{2(1-\lambda\alpha+\lambda\beta)}. \tag{24}$$

Let C and C_2 be the first and second moments of $C(x)$ and D and D_2 be the first and second moments of $D(x)$. The mean elapsed time T, which is the total average time that a call spends in the whole system, is simply $T = Y + D$. The mean queue length at the first server is $L = \lambda Y$.

Bulk arrival

In a computer system, a service call to a device often involves several operations for that device to perform. For example, a deck of cards is to be read by a card reader that reads one card at a time. The printer is another example. When a job requires printed output, the output may consist of many lines, or pages, but the printer can print only one line at a time. In queuing theory this is known as bulk arrival. The same situation also occurs in communication systems. A message from a remote station is often segmented for transmission, and the segments are sent and received one at a time. This reduces the buffering requirement at both ends.

Single-server systems with bulk arrival and service have been investigated by Bailey [8], Downton [9], Miller [10], and Foster [11]. In what follows, we present a solution of a sequential-server system with bulk arrival.

Let m_n be the probability that an arrival consists of n calls. Let $M(z)$ be the generating function of m_n; i.e.,

$$M(z) = \sum_{n=1}^{K} m_n z^n, \tag{25}$$

where K is the maximum size of the batch. Let M_1 and M_2 be the first and second moments of the batch arrival sizes. The generating function of the queue size distribution can be obtained by following the approach of Miller [10]. The queue size at the instant of departure can be analyzed by means of an imbedded Markov chain.

The queue-size generating function satisfies the following relation:

$$U(z) = [U(z) - P_0]\psi\{\lambda[1 - M(z)]\}z^{-1} + P_0\sum_{n=1}^{N} m_n z^{n-1}\phi\{\lambda[1 - M(z)]\}, \tag{26}$$

which is obtained since an arrival consists of n calls. Each new arrival increases the queue size by n. When a new arrival finds that the system is empty, one of its n calls enters the first server for service, and the remaining $n-1$ calls join the queue. The effect of bulk arrival can be included in the formulation by replacing z with the generating function of $M(z)$ in $\phi[\lambda(1-z)]$ and $\psi[\lambda(1-z)]$ of Eq. (21) (see Miller [10]). Solving for $U(z)$, we obtain

$$U(z) = \frac{P_0(M(z)\psi\{\lambda[1-M(z)]\} - \phi\{\lambda[1-M(z)]\})}{z - \psi\{\lambda[1-M(z)]\}}. \tag{27}$$

Since by definition $U(1) = 1$, we find that

$$P_0 = \frac{1 - M_1\lambda\alpha}{M_1(1 - \lambda\alpha + \lambda\beta)}. \tag{28}$$

The queuing time distribution can be obtained by viewing the arriving batch as one composite entity. We define two new Laplace-Stieltjes transforms, $R(s)$ and $Q(s)$, as follows:

$$R(s) = \sum_{n=1}^{K} m_n[\psi(s)]^n \tag{29}$$

and

$$Q(s) = \sum_{n=1}^{K} m_n\phi(s)[\psi(s)]^{n-1}. \tag{30}$$

The Laplace-Stieltjes transform of the queuing time distribution can be obtained by substituting $R(s)$ and $Q(s)$ for $\psi(s)$ and $\phi(s)$ in Eq. (23):

$$\theta(s) = \frac{[1 - M_1\lambda\alpha)/(1 - \lambda\alpha + \lambda\beta)][\lambda Q(s) - \lambda R(s) - sQ(s)]}{\lambda - s - \lambda R(s)}. \tag{31}$$

Note that the queuing-time formula Eq. (31) is calculated on a message basis, whereas the queue-size generating function Eq. (27) is based on message segments, or individual calls. The queuing time thus computed is the time for the whole message to pass the first server, including the waiting time, the blocking time, and the service time at the first server. Since the second server and the first server can overlap somewhat, i.e., the first server handles the nth segment of a message while the second server handles the $(n-1)$th segment of the same message, the elapsed time T for the whole message is simply $T = (-1)\theta'(0) + D$.

Priority queues

To ease the congestion at a computer center, the output messages may be put on a higher priority in the use of system resources than the input messages to be polled from remote terminals. Transactions to be processed within a computer may be classified into different priority queues. A READ request to a disk unit may also be put on a higher priority queue than a WRITE request or vice versa. In this section, priority queues for sequential-server systems are described. Stationary queuing time distributions are obtained.

For a single-server system with priorities, Cobham obtained the first moment of the waiting time distribution [12]. Miller [13] and Gaver [14] characterized the stationary distributions of the queue sizes and waiting times. Welch [15] and Jaiswal [16] studied the transient solutions of priority queues. Takacs [17] generalized the stationary solutions of priority queues. The author has generalized the stationary solutions of preemptive priority queues including other service disciplines, such as the preemptive-repeat service [18].

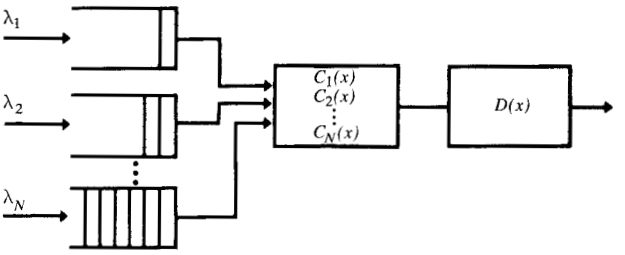

Figure 1 Sequential servers with priorities.

In the sequential-server model the second server represents a mechanical device at which the service is not interrupted until completion. Hence, we consider the non-preemptive priority queues. A higher-priority call waits and obtains service immediately after the completion of the current lower-priority service (including blocking time if any) at the first server. Calls with different priorities arrive at a facility which consists of two sequential servers. Let there be N classes of priorities, $1, 2, \cdots, N$. An arriving call with a smaller number indicates a higher priority. The calls are served in order of priority and, within each class, in order of arrival. It is also often assumed that the input is a Poisson process with parameter λ_k for k-type priority calls, where $k = 1, 2, \cdots, N$. Let the service times for k-type priority calls at the first server be mutually independent, positive, random variables with a distribution function $C_k(x)$. The service time distribution at the second server is assumed to be the same for all types of calls, $D(x)$. This assumption makes the analysis easier. If $D(x)$ is different, we have to determine the joint probability function of the queue sizes for all N priority classes. This requires the solution of a multi-dimensional generating function. The model investigated in this paper is illustrated in Fig. 1.

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:

$$\Lambda_k = \sum_{i=1}^{k} \lambda_i; \text{ in particular, let } \Lambda = \Lambda_N, \tag{32}$$

Each input is a Poisson process of parameter λ_k; therefore the sum Λ_k is also a Poisson process. Define the Laplace-Stieltjes transforms of the service time distributions as

$$\psi_k(s) = \int_0^\infty e^{-sx}\, d[C_k(x)\, D(x)] \tag{33}$$

for $k = 1, 2, \cdots, N$;

$$\phi_k(s) = \int_0^\infty \Lambda e^{-\Lambda t}\, dt \int_0^\infty e^{-sx}\, d[C_k(x)\, D(x+t)] \tag{34}$$

for $k = 1, 2, \cdots, N$;

and the nth moments as

$$a_k^{(n)} = (-1)^n \psi_k^{(n)}(0) \qquad \text{for } k = 1, 2, \cdots, N; \tag{35}$$

$$b_k^{(n)} = (-1)^n \phi_k^{(n)}(0) \qquad \text{for } k = 1, 2, \cdots, N. \tag{36}$$

Also let the Laplace-Stieltjes transforms of the weighted service-time distributions be

$$\Psi_k(s) = \sum_{i=1}^{k} (\lambda_i/\Lambda_k)\, \psi_i(s) \tag{37}$$

and

$$\Phi_k(s) = \sum_{i=1}^{k} \lambda_i\, \phi_i(s)/(\Lambda - \Lambda_k). \tag{38}$$

Define the nth moments of these weighted service time distributions as

$$h_k^{(n)} = (-1)^n \Psi_k^{(n)}(0)$$

and

$$g_k^{(n)} = (-1)^n \Phi_k^{(n)}(0), \tag{39}$$

and, in particular, let

$$a_k = a_k^{(1)},\ b_k = b_k^{(1)},\ h_k = h_k^{(1)} \text{ and } g_k = g_k^{(1)}.$$

In the previously defined non-preemptive service discipline, the 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 the completion of the lower-priority at the first server before service begins. Calls of low priority receive immediate service when no call of higher priority is waiting.

Let p_k be the probability that an arriving call is of priority class k. Clearly, $p_k = \lambda_k/\Lambda$. We now find the generating function and the queuing time distribution in terms of the Laplace-Stieltjes transform as follows.

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 transition time of the nth call that just enters into the second server. The nth call can be of any priority class. We now formulate the generating 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

$$\xi_{n+1}(k) = \begin{cases} \xi_n(k) - 1 + \eta_{n+1} & \text{if } \xi_n(k) > 0; \\ \eta_{n+1}(i) & \text{if } \xi_n(k) = 0 \text{ and } \xi_n'(k) = 0 \text{ and the next call is of priority class equal to } i; \\ \eta'_{n+1} & \text{if } \xi_n(k) = 0 \text{ but } \xi_n'(k) > 0. \end{cases} \tag{40}$$

Here, η_{n+1} is the number of new calls of priority classes less than or equal to k if the $(n+1)$th service is of priority class less than or equal to k. The parameter $\eta_{n+1}(i)$ is the number of new arrivals of priority classes less than or equal to k if the $(n+1)$th service is of priority class i. The parameter η'_{n+1} is the number of new cells of priority classes less than or equal to k if the $(n+1)$th service is of priority class greater than k. The reason we consider the cases separately is that the $(n+1)$th service time is different in each case.

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

$$U_k(z) = \sum_{n=0}^{\infty} P\{\xi_n(k) = j\}\, z^j. \tag{41}$$

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

$$P\{\xi_n(k) = 0\} = U_k(0). \tag{42}$$

Also, the probability that $\xi_n(N)$ is zero indicates that the queues are all empty, so we have

$$P\{\xi_n(N) = 0\} = U_N(0) = P_0. \tag{43}$$

There are three mutually exclusive events considered here.

1. $\xi_n(k) > 0$ and the next arrival is of priority class less than or equal to k. This event is represented by the generating function $[U_k(z) - U_k(0)]/z$.
2. $\xi_n(k) = 0$ and $\xi_n'(k) = 0$ and the next service is of priority class i. This even occurs with probability $(\lambda_i/\Lambda)P_0$, where P_0 is the probability that the first server is free.
3. $\xi_n(k) = 0$ and $\xi_n'(k) > 0$ and the next service is of priority class greater than k. This event occurs with a probability $U_k(0) - \sum_{i=1}^{N} (\lambda_i/\Lambda)P_0$.

Forming the generating functions of $\xi_{n+1}(k)$ and $\xi_n(k)$, and using the generating functions for the new arriving calls as given in the Appendix [Eq. (A2) and its equivalent], we have

$$U_k(z) = z^{-1}\,[U_k(z) - U_k(0)]\Psi_k[\Lambda_k(1-z)] + \sum_{i=1}^{N} (\lambda_i/\Lambda)P_0\phi_i[\Lambda_k(1-z)] + \left[U_k(0) - \sum_{i=1}^{N} (\lambda_i/\Lambda)P_0\right]\Phi_k[\Lambda_k(1-z)]. \quad (44)$$

This generating function provides the queue lengths of priority classes less than or equal to k at every transition (departing instant from the first server), including the departure of those calls of priority classes greater than k. We formulate the queue-size generating function observed by a departing call of priority class less than or equal to k by Takacs method [17]. The $(n+1)$th call 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

$$[U_k(z) - U_k(0)]\Psi_k[\Lambda_k(1-z)]\,z^{-1} + \sum_{i=1}^{k} (\lambda_i/\Lambda)P_0\phi_i[\Lambda_k(1-z)] \quad (45)$$

represents a departing call of priority class less than or equal to k. Let $G_k^*(z)$ be this partial generating function, i.e.,

$$G_k^*(z) = z^{-1}\,[U_k(z) - U_k(\theta)]\,\Psi_k[\Lambda_k(1-z)] + \sum_{i=1}^{k} (\lambda_i/\Lambda)\,P_0\phi_i[\Lambda_k(1-z)] \quad (46)$$

and

$$G_k^*(1) = 1 - U_k(0) + \sum_{i=1}^{k} (\lambda_i/\Lambda)\,P_0. \quad (47)$$

The generating function for the queue lengths of priority classes less than or equal to k, observed by a departing caller of priority class less than or equal to k, is

$$G_k(z) = G_k^*(z)/G_k^*(1). \quad (48)$$

Combining Eqs. (44) to (48) we obtain

$$G_k(z) = \Big\{z\sum_{i=1}^{k} (\lambda_i/\Lambda)\,P_0\phi_i[\Lambda_k(1-z)] + [U_k(0) - P_0]\Phi_k[\Lambda_k(1-z)]\Psi_k[\Lambda_k(1-z)] + \sum_{i=k+1}^{N} (\lambda_i/\Lambda)\,P_0\,\phi_i[\Lambda_k(1-z)]\,\Psi_k[\Lambda_k(1-z)) - U_k(0)\,\Psi_k[\Lambda_k(1-z)]\Big\} \div G_k^*(1)\{z - \Psi_k[\Lambda_k(1-z)]\}. \quad (49)$$

For $k = N$, Eq. (44) reduces to Eq. (48) and we have

$$G_n^*(1) = 1,\ U_N(0) = P_0, \text{ and } U_N(z) = G_N(z). \quad (50)$$

Since $U_N(1) = 1$, applying L'Hospital's rule, we have

$$P_0 = U_N(0) = \frac{1 - \sum_{i=1}^{N} \lambda_i a_i}{1 - \sum_{i=1}^{N} \lambda_i a_i + \sum_{i=1}^{N} \lambda_i b_i}. \quad (51)$$

Because $G_k(1) = 1$, from Eq. (49) we have

$$U_k(0) = \Big\{1 - \sum_{i=1}^{k} \lambda_i a_i - (\Lambda_k/\Lambda)\sum_{i=1}^{N} \lambda_i b_i\,P_0 + [\Lambda_k/(\Lambda - \Lambda_k)]\sum_{i=k+1}^{N} \lambda_i a_i\,P_0\Big\} \div \Big\{1 - \sum_{i=1}^{k} \lambda_i a_i + [\Lambda_k/(\Lambda - \Lambda_k)]\sum_{i=k+1}^{N} \lambda_i a_i\Big\}. \quad (52)$$

Thus, $G_k(z)$ can be uniquely determined. Knowing $G_k(z)$, we can easily obtain the Laplace-Stieltjes transform of the queuing time distribution for callers of priority classes less than or equal to k as follows.

Let $\theta_k^*(s)$ be the Laplace-Stieltjes transform of the queuing time distribution of callers with priority classes less than or equal to k. Let $s = \Lambda_k(1-z)$ in Eq. (53); then we have

$$\theta_k^*(s) = G_k(1 - s/\Lambda_k) = \Big\{(\Lambda_k - s)\,P_0\sum_{i=1}^{k} (\lambda_i/\Lambda)\,\phi_i(s) + \Lambda_k[U_k(0) - P_0]\,\Phi_k(s)\,\Psi_k(s) + \Lambda_k\sum_{i=k+1}^{N} (\lambda_i/\Lambda)P_0\phi_i(s)\Psi_k(s) - \Lambda_k\,U_k(0)\,\Psi_k(s)\Big\} \div \Big\{[1 - U_k(0) + \sum_{i=1}^{k} (\lambda_i/\Lambda)\,P_0] \times [\Lambda_k - s - \Lambda_k\,\Psi_k(s)]\Big\}. \quad (53)$$

To find the Laplace-Stieltjes transform of the queuing time distribution for calls of priority class equal to k, $\theta_k(s)$, we use a method similar to that in [17]. Let $\gamma_k(s)$ be the smallest root within the unit circle of the equation

$$\gamma_k(s) = \psi_k\{s + \Lambda_k\,[1 - \gamma_k(s)]\}. \quad (54)$$

The first moment is

$$-\gamma_k'(0) = \frac{\sum_{i=1}^{k} \lambda_i a_i}{1 - \sum_{i=1}^{k} \lambda_i a_i}.$$

This is known as the mean busy period in a single-server queue [10].

Finally, $\theta_k(s)$ can be obtained as

$$\theta_k(s) = \theta_k^*\{s + \Lambda_{k-1}\,[1 - \gamma_{k-1}(s)]\}. \tag{55}$$

Let Y_k be the mean queuing time of a call whose priority class is k, i.e.,

$$\begin{aligned} Y_k &= -\theta_k'(0) \\ &= \Big\{G_k(1) \sum_{i=1}^{k} \lambda_i a_i^{(2)} + 2\,P_0 \Big(\sum_{i=1}^{k} \lambda_i b_i/\Lambda\Big) \\ &\quad + 2\,P_0\,(\Lambda_k/\Lambda) \sum_{i=1}^{k} \lambda_i b_i^{(2)} \\ &\quad + (\Lambda_k/(\Lambda - \Lambda_k))[U_k(0) - P_0] \sum_{i=k+1}^{N} \lambda_i a_i^{(2)} \\ &\quad + [2/(\Lambda - \Lambda_k)][U_k(0) - P_0] \sum_{i=1}^{k} \lambda_i a_i \sum_{i=k+1}^{N} \lambda_i a_i \\ &\quad + 2\,(P_0/\Lambda) \sum_{i=1}^{k} \lambda_i a_i \sum_{i=1}^{N} \lambda_i b_i \Big\} \\ &\quad \div \Big[2\,G_k(1) \Big(1 - \sum_{i=1}^{k-1} \lambda_i a_i\Big)\Big(1 - \sum_{i=1}^{k} \lambda_i a_i\Big)\Big]. \end{aligned} \tag{56}$$

The mean elapsed time T_k for callers of priority class k is $T_k = Y_k + D$.

Numerical examples.

• *Sequential server queues*

Consider a heavily used data collection terminal. Assume that it takes 0.50 s for the terminal to prepare a message for transmission. The prepared message is then placed in a terminal buffer which can hold one message at a time. The buffer is polled by a central computer through a communication line. Suppose that it takes about 0.67 s to transmit a message to the central computer. Assume that both of these service times are exponentially distributed, i.e., $\mu = 1/0.5 = 2$ and $\nu = 1/0.67 = 1.5$, and that the arrival rate is one message per second, i.e., $\lambda = 1$. Based on these data, we calculate that $\alpha = 0.88$, $\alpha_2 = 1.277$, $\beta = 0.652$, and $\beta_2 = 0.791$. The mean queuing time is $Y = 5.6$ s and the mean elapsed time is $T = 6.27$ s.

• *Bulk arrival queues*

Consider an automatic retail store. Several cash registers are connected to a small in-house computer; each of these registers has a keyboard, a display, and a small printer. Customers with merchandise arrive at the counters for service. Consider each individual counter. Assume that in the peak traffic period, customers arrive at the counter at the rate 1/100 s. Also assume that each customer may bring either 10 or 16 items of merchandise with equal probability. The cashier enters each merchandise identification number on the keyboard, which takes an average of 5 s each with an exponential service time distribution; the operator can enter the nth item while the system is processing the $(n-1)$th item from the same terminal. When the processing is complete, the computer sends the resultant $(n-1)$th message to the printer. If the work on the $(n-1)$th transaction (processing and printing) has not been completed while the nth entry has been completed and is waiting at the terminal buffer, the keyboard is locked so that the $(n+1)$th entry cannot be made.

Suppose that, from other analysis, it is known that it takes about 1 s for the small computer to complete the processing. (This analysis includes all other loads from the other terminals.) Assume that the printer takes another second to print the output message. Combine the computer and the printer as the second server. The mean service time is 2 s. For simplicity, assume that the second server has an exponential distribution. (Note that the sequential-server model can handle other types of distributions; we use exponential service time distributions here for illustration.) Our problem is to determine the elapsed time of a customer and the mean queue size at each of the counters during the peak traffic situation. For the example, $\lambda = 0.01$, $\mu = 0.2$, and $\nu = 0.5$. From these data, we calculate $\alpha = 5.57$, $\alpha_2 = 53.92$, $\beta = 5.01$, and $\beta_2 = 50.07$. The generating function of the arrival batch is $M(z) = 0.5\,z^{10} + 0.5\,z^{16}$, with first and second moments $M_1 = 13$ and $M_2 = 178$.

The composite service time moments can be calculated as follows:

$$\alpha^* = -R'(0) = M_1\alpha, \tag{57}$$

$$\alpha_2^* = R''(0) = M_1\alpha_2 + (M_2 - M_1)\alpha^2, \tag{58}$$

$$\beta^* = -Q'(0) = (M_1 - 1)\alpha + \beta, \tag{59}$$

and

$$\beta_2^* = Q''(0) = \beta_2 + 2(M_1 - 1)\alpha\beta + (M_2 - 3M_1 + 2)\alpha^2 + (M_1 - 1)\alpha_2. \tag{60}$$

Thus, we obtain $\alpha^* = 72.4$, $\alpha_2^* = 5820$, $\beta^* = 71.85$ and $\beta_2^* = 5715$. Finally, the mean queuing time is

$$Y = -\theta'(0) = \frac{\lambda\beta_2^* + \beta^*[2 + \lambda^2\alpha_2^*/(1 - \lambda\alpha^*)]}{2(1 - \lambda\alpha^* + \lambda\beta^*)} = 190\text{s}.$$

The mean queue size is $L = \lambda Y = 1.90$ customers and the mean elapsed time (i.e., total time spent by a customer at a counter) is $T = Y + D = 190 + 2 = 192$ s.

As another example, consider a computer communication system. A remote station generates an input traffic of 0.1 message/s. The input message is to be transmitted to a host computer for processing. Assume that each input message consists of ten segments and that each

segment has an average of 128 characters, with an exponential distribution. Assume that the speed of communication line is 300 characters/s.

To send a segment of a message, assume that a line service time, which includes the transmission time and the time for some line overhead functions (e.g., the polling of the input station), is 0.6 s per message segment. The host computer handles other input messages from other lines and places the input message segments in a queue. Assume that the host computer has a mean response time of 435 ms, with an exponential distribution. The response time includes all the queuing times and service times that a segment requires at the host. Assume that the communication protocol is synchronized and that the message segments are sequenced. (Each segment is given a sequence number.) One segment may be prepared at the remote station and sent over the communication line, and a previous segment may be processed at the host computer. This provides some overlapping operation and an error recovery capability if something goes wrong.

This problem may be formulated as a sequential server problem. The remote station and the communication line are treated as the first server and the host computer as the second server. The queue to be studied is the input message queue at the remote station for which $\lambda = 0.1$, $\mu = 1/0.6$, $\gamma = 1/0.435$, $M_1 = 10$, and $M_2 = 100$. Using these parameters we obtain the moments

$\alpha = 0.782$, $\alpha_2 = 0.972$, $\beta = 0.6076$, and $\beta_2 = 0.7305$;

$\alpha^* = 7.82$, $\alpha_2{}^* = 80.22$, $\beta^* = 7.64$ and $\beta_2{}^* = 63.50$.

The mean queuing time and the mean elapsed time are determined to be

$Y = 25.4$ s and $T = 25.8$ s.

Note that only the last segment's host time is needed in computing T because the other host times are overlapped with the line times.

• *Priority queues*

Consider an application of the priority queues with bulk arrivals in a computer system. Consider a computer partition with two disk file units as shown in Fig. 2. Two types of transactions form two queues with type 1 having a higher processing priority than type 2. Let the arrival rates be $\lambda_1 = 0.6$ and $\lambda_2 = 0.5$, respectively. Each type 1 transaction consists of five units of work. Each unit of work requires the processing of one data record from disk unit 1 and one data record from disk unit 2. Each type 2 operation consists of six units of work. Each unit of work involves the processing of two data records from disk unit 1 and one data record from disk unit 2. Suppose that the average processing time for a data record in the

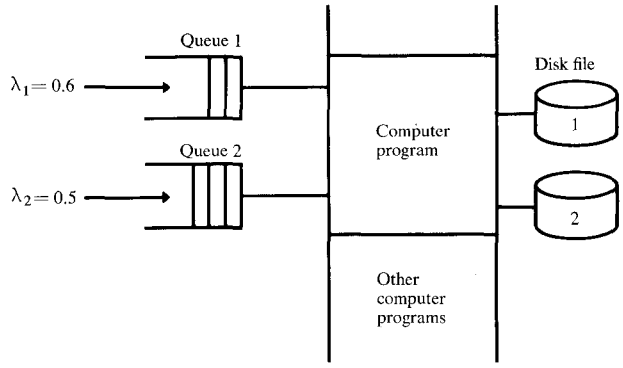

Figure 2 Partition with two queues.

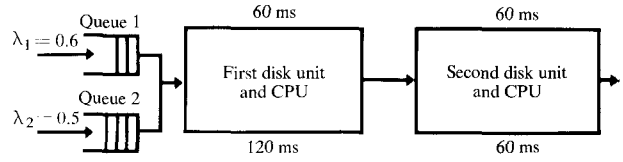

Figure 3 Sequential server system with priorities.

CPU is 10 ms and the access and data transfer time in a disk unit is 50 ms for both types of transaction and for both types of file. Thus, for each data record, the mean CPU time and the disk file service time are 10 ms + 50 ms = 60 ms. We can treat this problem as two servers in sequence, disk unit 1 and some CPU processing as the first server, and disk unit 2 and some CPU processing as the second server. This is shown in Fig. 3.

Let $M_r(k)$ be the rth moment of the batch arrival sizes in priority class k. We have the following data: $M_1(1) = 5$, $M_2(1) = 25$, and $M_1(2) = 6$, $M_2(2) = 36$. Assume exponential service time distributions; the service rates are $\mu_1 = 1/0.06 = 16.66$, $\mu_2 = 1/(0.06 + 0.06) = 8.33$ and $\nu = 1/0.06 = 16.66$. From Eqs. (10)-(13), we obtain

$a_1 = 0.09$, $a_1{}^{(2)} = 0.0126$, $a_2 = 0.14$, and $a_2{}^{(2)} = 0.0184$;

$b_1 = 0.061$, $b_1{}^{(2)} = 0.0074$, $b_2 = 0.1206$, and $b_2{}^{(2)} = 0.1441$.

Since we have bulk arrivals, the composite service time moments can be obtained from Eqs. (57)-(60):

$a_1{}^* = 0.45$, $a_1{}^{(2)*} = 0.225$, $a_2{}^* = 0.84$, and $a_2{}^{(2)*} = 0.6984$;

$b_1{}^* = 0.3661$, $b_1{}^{(2)*} = 0.1594$, $b_2{}^* = 0.8206$, and $b_2{}^{(2)*} = 0.797$.

Substituting these values into Eq. (51), we have $P_0 = 0.33$. From Eq. (52), $U_1(0) = 0.636$. From Eq. (47),

$G_1^*(1) = 0.544$ and, finally, from Eq. (56), we obtain $Y_1 = 0.77$ and $Y_2 = 3.2$. The mean elapsed time to process a transaction of type 1 is $T_1 = Y_1 + D = 0.83$ s and the mean elapsed time to process a transaction of type 2 is $T_2 = Y_2 + D = 3.26$ s. The mean queue lengths are $L_1 = \lambda_1 Y_1 = 0.462$ and $L_2 = \lambda_2 Y_2 = 1.6$.

Concluding remarks

In this paper we are treating simple environments with practical applications, and we have developed some sequential-server queuing models for the analysis of computer/communication system problems. (For more complex environments using sequential-server queuing models with finite queues between stages, see Neuts [19].) Several numerical examples were given to illustrate the use of the models. In the examples, mean queue lengths and mean queuing times are used as performance measures. It is also possible to use the second moments and variances that can be obtained from the second derivatives of the generating function and the Laplace-Stieltjes transform studied in this paper. The variances are useful in predicting the percentiles of a probability distribution, which are often needed in designing real-time computer systems.

Appendix: Sequential server queue

• *Queue length distribution*

Let ξ_n be the queue length immediately after the departure of the nth call from the first server and the entrance into the second server. Let η_n be the number of new calls which arrive during the aggregate service time of the nth call under the condition that the nth call began its service when the queue was not empty. Let η_n' be the number of new calls that arrive during the aggregate service time of the nth call if the nth call began its service when no queue was present. The queue lengths ξ_{n+1} and ξ_n and the number of newly arriving calls are related by the following equation:

$$\xi_{n+1} = \begin{cases} \xi_n - 1 + \eta_n & \text{if } \xi_{n+1} > 0, \\ \eta'_{n+1} & \text{if } \xi_n = 0. \end{cases} \tag{A1}$$

Assume that the stationary distribution of queue length exists, then ξ_{n+1} and ξ_n must have the same marginal distribution. The ξ_n calls form an imbedded Markov chain which we study by using the generating function technique.

The generating function for η_n can be written as

$$\sum_{i=0}^{\infty} P\{\eta_n = i\}\, z^i = \int_0^{\infty} \sum_{i=0}^{\infty} [(\lambda x)^i / i!]\, z^i\, e^{-\lambda x}\, dH(x)$$

$$= \int_0^{\infty} e^{-\lambda(1-z)x}\, dH(x) = \psi[\lambda(1-z)]. \tag{A2}$$

Equation (A2) is obtained because the number of new arrivals follows a Poisson process. If each of the new arrivals generates a batch of calls, and the batch size has a generating function $M(z)$, then Eq. (A2) can be written as

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

which is useful for the bulk arrival model studied in this paper.

Similarly, the generating function for η_n' can be obtained as $\phi[\lambda(1-z)]$. Define the probability that there are j calls in a queue of length ξ_n as

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

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

$$U(z) = \sum_{j=0}^{\infty} P_j\, z^j.$$

If the stationary solution for queue length exists, the generating functions for ξ_{n+1} and ξ_n must be the same. From Eq. (A1) and from the fact that the generating function of the sum of two independent variables is the product of the two generating functions, it follows that

$$U(z) = P_0\, \phi(\lambda(1-z)) + [U(z) - P_0]\, z^{-1}\, \psi[\lambda(1-z],$$

Solving for $U(z)$, we obtain

$$U(z) = \frac{P_0\, [z\, \phi(\lambda(1-z)) - \psi(\lambda(1-z))]}{z - \psi[\lambda(1-z)]}, \tag{A3}$$

where P_0 remains to be determined. Because

$$\sum_{j=0}^{\infty} P_j = 1,$$

we find from Eq. (A3) that

$$U(1) = 1. \tag{A4}$$

Using L'Hospital's rule and Eq. (A4), we obtain

$$P_0 = \frac{1 - \lambda\alpha}{1 - \lambda\alpha + \lambda\beta}. \tag{A5}$$

Thus, Eq. (A3) is uniquely determined.

• *Queuing time distribution*

Let y_n be the nth call's queuing time at the first server, including the three segments of the call's waiting time, service time, and blocking time. Let $Y(x)$ be its probability distribution and $\theta(s)$ be its Laplace-Stieltjes transform. Because the number of new calls which arrive during the queuing time y_n must equal the queue size at the nth call's departure instant from the first server, we have $\theta[\lambda(1-z)] = U(z)$. Let $z = 1 - s/\lambda$; then we find that

$$\theta(s) = \left(\frac{1-\lambda\alpha}{1-\lambda\alpha+\lambda\beta}\right)\left[\frac{(\lambda-s)\phi(s)-\lambda\psi(s)}{\lambda-s-\lambda\psi(s)}\right]. \quad \text{(A6)}$$

The nth moment of the queuing time distribution is given by

$$Y_n = \int_0^\infty x^n \, dY(x) = (-1)^n \, \theta^{(n)}(0). \quad \text{(A7)}$$

Knowing $\theta(s)$, one can determine the waiting time distribution $W(x)$ at the first server as follows. Let

$$\Omega(s) = \int_0^\infty e^{-sx} \, dW(x);$$

then,

$$\theta(s) = [\Omega(s) - P_0] \, \psi(s) + P_0 \, \phi(s).$$

Solving for $\Omega(s)$, we obtain

$$\Omega(s) = \left(\frac{1-\lambda\alpha}{1-\lambda\alpha+\lambda\beta}\right)\left[\frac{\lambda\phi(s)-s-\lambda\psi(s)}{\lambda-s-\lambda\psi(s)}\right]. \quad \text{(A8)}$$

The total elapsed time is defined as the duration between the instant that a call arrives and the instant that it departs from the second server. Let $T(x)$ be the elapsed time distribution. Its Laplace-Stieltjes transform has the relation

$$\int_0^\infty e^{-sx} \, dT(x) = \theta(s) \int_0^\infty e^{-sx} \, dD(x). \quad \text{(A9)}$$

References

1. P. J. Burke, "The Output of a Queuing System," *Oper. Res.* **4,** 699 (1956).
2. W. Chang, "Output Distribution of a Single-Channel Queue," *Oper. Res.* **11,** 620 (1963).
3. G. C. Hunt, "Sequential Arrays of Waiting Lines," *Oper. Res.* **4,** 674 (1956).
4. T. Suzuki, "On a Tandem Queue with Blocking," *J. Oper. Res. Japan* **6,** 137 (1964).
5. B. Avi-Itzhak and M. Yadin, "A Sequence of Two Servers with No Intermediate Queue," *Management Science* **11,** 553 (1965).
2. N. U. Prabhu, "Transient Behavior of a Tandem Queue," *Management Science* **13,** 631 (1967).
7. W. Chang, "Two Servers in Sequence with Erlang Input," Proceedings of the Symposium on Computer-Communications Networks and Teletraffic, Polytechnic Press of the Polytechnic Institute of Brooklyn, New York, 1972, p. 409.
8. N. T. J. Bailey, "On Queuing Processes with Bulk Service," *J. Roy. Statist. Soc. B* **16,** 80 (1954).
9. F. Downton, "Waiting Time in Bulk Service Queues," *J. Roy. Statist. Soc. B* **21,** 256 (1955).
10. R. G. Miller, "A Contribution to the Theory of Bulk Queues," *J. Roy. Statist. Soc. B* **21,** 320 (1959).
11. F. G. Foster, "Batched Queuing Processes," *Oper. Res.* **12,** 441 (1964).
12. A. Cobham, "Priority Assignment in Waiting Line Problems," *Oper. Res.* **2,** 70 (1954).
13. R. G. Miller, "Priority Queues," *Ann. Math. Statist.* **31,** 86 (1960).
14. D. P. Gaver, "A Waiting Line with Interrupted Service, Including Priorities," *J. Roy. Statist. Soc. B* **13,** 73 (1962).
15. P. D. Welch, "On Preemptive Resume Priority Queues," *Ann. Math. Statist.* **35,** 600 (1964).
16. N. K. Jaiswal, "Time-dependent Solution of the Head-of-the-Line Priority Queue," *J. Roy. Statist. Soc. B* **24,** 91 (1962).
17. L. Takacs, "Priority Queues," *Oper. Res.* **12,** 63 (1964).
18. W. Chang, "Preemptive Priority Queues," *Oper. Res.* **13,** 620 (1965).
19. M. F. Neuts, "Two Queues in Series with a Finite Intermediate Waiting Room," *J. Appl. Prob.* **5,** 123, (1968).

Received March 8, 1974; revised April 9, 1975

The author is located at the Data Processing Division Headquarters, White Plains, New York 10601.