AT&T Technical Journal
Vol. 64, No. 7, September 1985
Printed in U.S.A.

# A Priority-Based Admission Scheme for a Multiclass Queueing System

By K. M. REGE and B. SENGUPTA*

(Manuscript received November 21, 1984)

We consider a queueing problem involving multiple priority classes where the station is divided into waiting and service areas. The service area has a finite number of positions where a customer of a particular class has access to only a subset of these positions. The admission into the service area is controlled by a mechanism that allows customers within a priority class to enter the service area on a first-come first-served basis. The customers of different classes are assumed to be indistinguishable once they have entered the service area. We consider service under three different disciplines: last-come first-served preemptive resume, multiple server, and processor sharing. We show that the waiting time of a customer is related to that of a customer in an equivalent M/G/1 queue. We characterize the Laplace-Stieltjes transform of the time spent in the service area. We also discuss three potential applications in the area of computer and communication systems.

## I. INTRODUCTION

This paper is concerned with investigating a queueing system in which customers from $n$ different job classes representing various priority levels receive service in a service area with a finite capacity of $m$. The capacity, $m$, of the service area refers to the maximum number of customers that can be present in the service area at any time. We describe an admission scheme that allows preferential access to the service area by the higher-priority customers. This scheme may give rise to a smaller waiting time before entry into the service area for the

* Authors are employees of AT&T Bell Laboratories.

higher-priority customers. The customer classes are assumed to be indistinguishable after they have gained access to the service area. The performance measures that we characterize are the distributions of waiting time and the mean time spent in the service area. This work has potential applications in the design of multiprogramming levels for computer systems and the design of window levels for communication systems.

More formally, let $n$ independent classes of customers arrive at a queueing station, each according to a Poisson process. Let the mean arrival rate of class $i$ be $\lambda_i$. The classes are arranged according to decreasing order of priority, that is, class 1 has the highest priority and class $n$ has the lowest priority. The queueing station is divided into $n$ waiting areas (one for each class) and a service area. The service area can hold at most $m$ customers simultaneously. Service is provided at a state-dependent rate of $\mu_i$ whenever there are $i$ customers present in the service area. We consider three service disciplines within the service area: Last-Come First-Served Preemptive Resume (LCFS-PR), $m$ Server (MS), and Processor Sharing (PS). The admission into the service area is controlled by means of a gate that allows customers from the waiting areas to enter the service area on the basis of the contents of the service area. The admission policy gives preferential treatment to higher-priority customers in gaining access to the service area. In particular, the admission policy is governed by two rules:

1. When a customer of class $i$ is admitted to the service area, the waiting areas of classes $1, \cdots, i-1$ must be empty.

2. When a class $i$ customer enters the service area, the number of customers in the service area (excluding itself) must be less than $k_i$. The sequence $\{k_i,\ i = 0, \cdots, n+1\}$ is a set of strictly decreasing, nonnegative integers with $k_0 = \infty$, $k_1 = m$, and $k_{n+1} = 0$.

This admission scheme reserves some slots in the service area for the exclusive use of the high-priority customers. In the case of the MS model, this means that at times the low-priority customers will not be allowed to enter the service area although some servers are idle. Obviously this is not the most efficient way of utilizing the servers' capacity. However, when the designer's overriding concern is to reduce the delays suffered by the high-priority customers, it is useful to reserve certain slots exclusively for the high-priority customers. The queueing station is shown schematically in Fig. 1.

In this paper, we show that this problem has an interesting structure, which can be exploited to characterize (1) the waiting-time distribution, by class, and (2) the mean time spent in the service area, by class. For the special case of the PS discipline, we show how to obtain this mean when $n$ equals two. Since writing this paper, it has been brought to our attention that Schaack and Larson[1] have independently

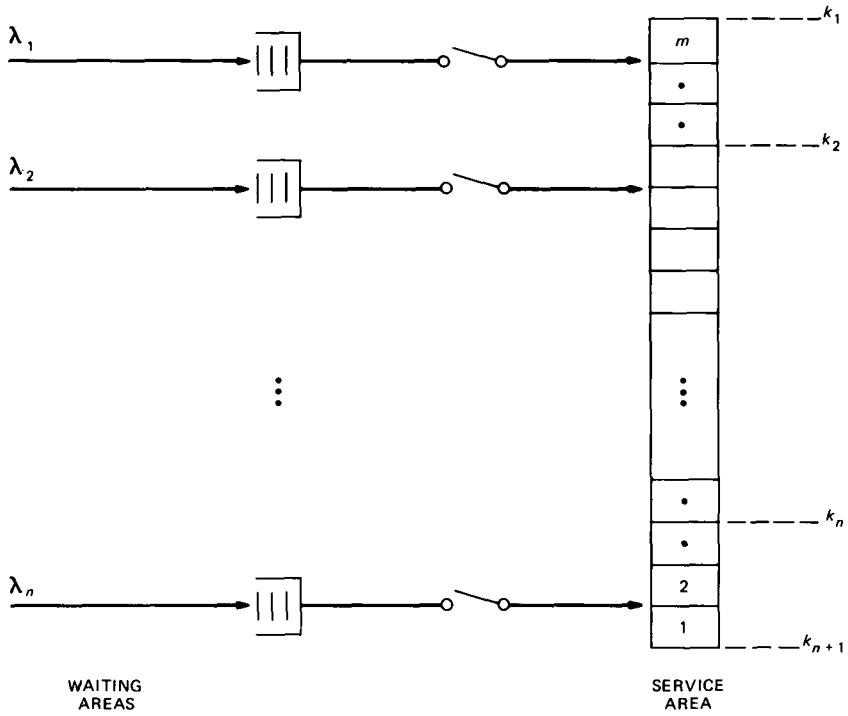


Fig. 1—The queueing station.

studied the special case of the MS discipline and reported the same results as we do here.

This paper is organized into four sections. In Section I, we discuss some potential applications for this model. In Section II, we derive the waiting-time distribution, by class. We characterize the mean time spent in the service area, by class, in Section III. Section IV summarizes our conclusions.

## II. POTENTIAL APPLICATIONS

We discuss three potential applications in this section. In the first, we propose this scheme for sharing of multiprogramming threads by several job classes in a computer system. Our second and third examples propose this admission scheme for sharing a window size by several job classes at the link level and the application level, respectively, in a communication system.

Avi-Itzhak and Heyman[2] had first proposed that a state-dependent server be used to approximate the CPU and disk subsystem of a computer. We depict a multiple CPU and disk subsystem in Fig. 2,

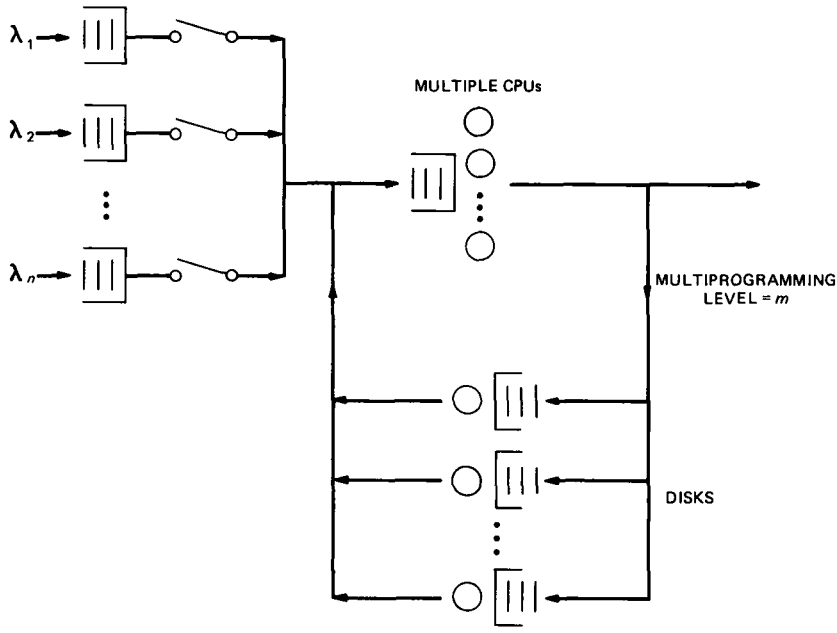


Fig. 2—A computer with shared multiprogramming threads.

which is approximated by a state-dependent server in our model. The service rate $\mu_i$ of our model is obtained by solving for the throughput in the closed queueing network of Fig. 2 with a population size of $i$. It is assumed in our model that the service requirements in terms of CPU and I/O times are approximately the same for all classes of customers. Also, the service discipline for the state-dependent server that is most appropriate for this application is PS. The closed queueing network of Fig. 2 can be solved by mean-value analysis described by Reiser and Lavenberg.[3] The circumstances under which a single state-dependent server is a good approximation of the CPU and I/O subsystem was investigated by Fredericks.[4] This approximation is usually good when each customer makes many trips to the I/O devices and when the CPU and I/O times required by a customer are not too unbalanced. In our model, $m$ is the multiprogramming level of the computer, usually determined by considerations such as available memory and the extent to which the jobs require concurrent access of the same databases. Given $m$, our model can be used to determine a way to allocate available multiprogramming threads to the various job classes so that some requirements on mean response time can be met.

A second application would be in the modeling of a link layer protocol such as high-level data link control. Assume that a provider of packet-switching service offers $n$ grades of service, each with its

own response-time requirement. The different grades of service may be provided by appropriately sharing a link-level window size of $m$ among the packets of $n$ service grades. The admission scheme proposed in the Introduction is one means of offering different levels of service to customers. In this application, the queueing station represents a node in the network where the customers in the service area correspond to the jobs ready for transmission. Service provided to a customer constitutes its transmission to an adjacent node and the return of the acknowledgment. Since data links are often characterized by relatively low utilizations, the value of $\mu_i$ may be approximately proportional to $i$, at least for small values of $i$. The constant of proportionality may be taken as the mean round-trip time to receive an acknowledgment on a link that has no traffic. This linearity would imply that there is hardly any wait for transmission to commence once a packet has entered the service area. For larger values of $i$, some saturation of $\mu_i$ will take place as the presence of a large number of packets in the service area starts to choke the capacity of the link. The limiting value of $\mu_i$ may be chosen as the rate at which acknowledgments can be returned in a fully utilized link. We show the closed queueing network used for calculating $\mu_i$ in Fig. 3. This is an approximation along the lines of one proposed by Schwartz.[5]

The third potential application is from the point of view of a user of a data network. This potential application is similar in spirit to the previous one except that we are concerned with the high-level protocol of host-to-host traffic using a data network. The network itself is approximated by a state-dependent server. The closed queueing network used for calculating $\mu_i$ is shown in Fig. 4. This approximation was proposed by Reiser.[6,7] The user of the network must allocate a window size of $m$ to $n$ different types of traffic. The admission scheme described earlier may enable the user to determine a way to share the

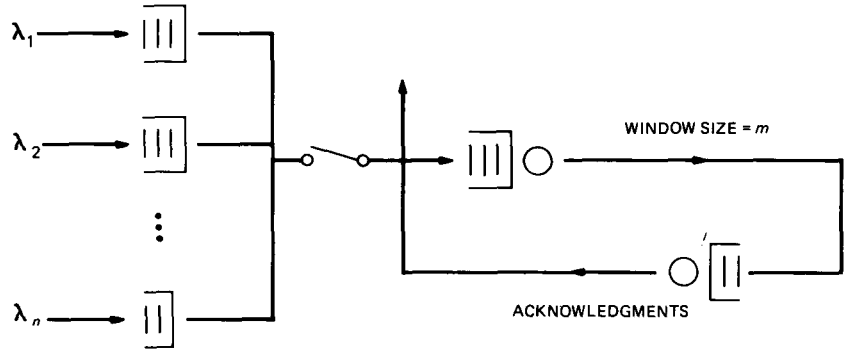


Fig. 3—Link layer protocol.

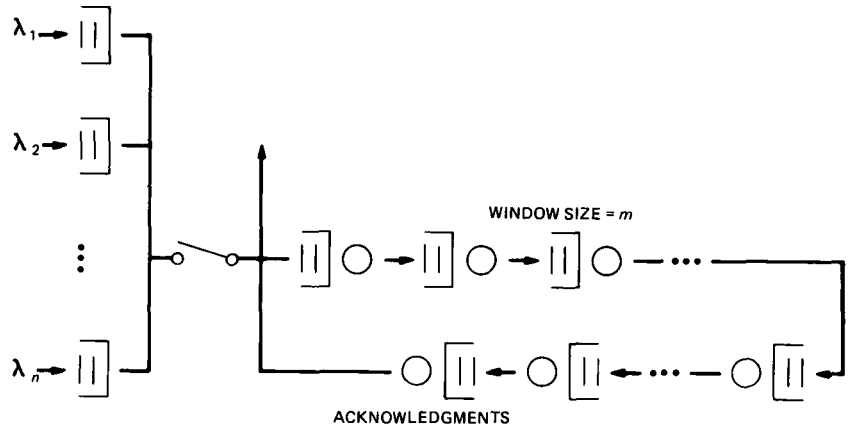


Fig. 4—High-level protocol.

window size among the types of traffic to meet certain response-time criteria.

## III. THE WAITING-TIME DISTRIBUTION

In this section, we will characterize the waiting-time distributions without assuming anything about the service discipline. The key result of this section is that, given that a customer of class $l$ is required to wait, its waiting-time distribution is related to that of a suitable M/G/1 queue.

The state of the system is completely specified by an $n$-tuple $(J_1, J_2, \cdots, J_i, \cdots, J_n)$, where $J_i (i = 2, \cdots, n)$ represents the number of customers of class $i$ waiting and $J_1$ represents the number of customers of class 1 waiting plus the number of customers present in the service area. With this description of the state space, one can, in principle, write down a system of equations for the steady-state probability vector $P(\mathbf{j}) = P(J_1 = j_1, J_2 = j_2, \cdots, J_n = j_n)$. In particular, for $l = 0, \cdots, n-1$

$$(\Lambda + \mu_{j_1})P(\mathbf{j}) = \sum_{k=l+1}^{n} \lambda_k P(\mathbf{j} - \mathbf{e}_k)\delta(j_k) + \sum_{k=1}^{l} \lambda_k P(\mathbf{j} - \mathbf{e}_1)\delta(j_1)$$

$$+ \mu_{j_1+1}P(\mathbf{j} + \mathbf{e}_1) + \mu_{j_1}P(\mathbf{j} + \mathbf{e}_{l+1})(1 - \delta(j_1 - k_{l+1}))\delta(l), \quad (1)$$

where

$$k_{l+1} \leq j_1 < k_l,$$

$$\Lambda = \sum_{k=1}^{n} \lambda_k,$$

$\delta(x) = 1$ if $x > 0$ and 0 otherwise, and $\mathbf{e}_k$ is an $n$-tuple with a 1 in the $k$th position and 0 elsewhere. For $j_1 \geq m$, $\mu_{j_1}$ is to equal to $\mu_m (= \mu)$,

since for $j_1 \geq m$, the number of customers in the service area is $m$. The solution of this equation is not easy; however, it is possible to solve it for the case where $n = 2$. Since the solution of this equation does not concern us at present, we show how to calculate this for $n = 2$ in Appendix A.

We will now concentrate our attention on the stochastic process defined by the random variable $J_1$. Let $\{u_i\}$ be the steady-state marginal probability distribution of $J_1$. Since arrivals are Poisson, a customer of class $l$ is required to wait outside with probability $\sum_{j=k_l}^{\infty} u_j$. Clearly, the waiting time is 0 whenever an arrival finds that $J_1 < k_l$. Let us now start observing the system when $J_1$ changes its value from $k_l - 1$ to $k_l$. Let $t_0$ denote this instant. At $t_0$, a customer of type $j$, with $j \leq l$, arrives to find exactly $k_l - 1$ customers in the service area and is immediately admitted for service. Let $t_f$ denote the first instant after $t_0$ when $J_1$ moves from $k_l$ to $k_l - 1$ with no type $l$ customers waiting outside. During the open interval $(t_0, t_f)$, several type $l$ customers may get admitted to the service area. If $n (n \geq 0)$ type $l$ customers are admitted to the service area during the open interval $(t_0, t_f)$, let $t_1, t_2, \cdots, t_n$ denote instants when these admissions took place (refer to Fig. 5). Note that at the instants $t_1, t_2, \cdots, t_n$, a departure occurs from the state $J_1 = k_l$ and there is at least one type $l$ customer waiting outside. Also, at these instants there cannot be any higher-priority customers in the waiting area. Now let us focus our attention on the intervals $(t_0, t_1), (t_1, t_2), \cdots, (t_{n-1}, t_n)$ and $(t_n, t_f)$. [In case $n$ equals 0, we need consider just one interval $(t_0, t_f)$.] The lengths of these intervals are governed by the customers inside the service area, which by assumption are indistinguishable, and by arrivals of

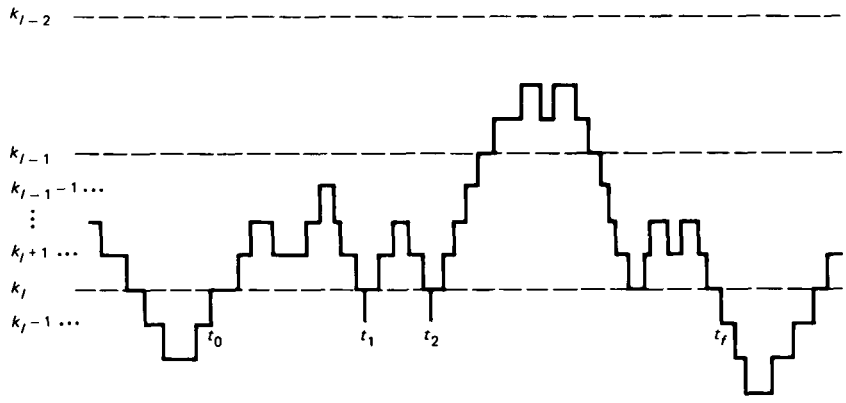


Fig. 5—A typical sample function of the process $J_1$. (The notches at $t_1$ and $t_2$ represent the event that a departure occurred from the state $J_1 = k_l$ and a waiting type $l$ customer was immediately admitted to the service area.)

customers of priority higher than that of type $l$ customers (i.e., the types of these customers are less than $l$) that are Poisson. Also, since the service requirements are memoryless and since the end points of these intervals are marked by identical states so far as customers of priority higher than that of type $l$ customers are concerned, the lengths of these intervals are independent and identically distributed (i.i.d.) random variables. Let $H_l$ denote the generic random variable corresponding to these lengths and let $H_l(s)$ denote the Laplace-Stieltjes Transform (LST) of its distribution function.

Now a type $l$ customer is required to wait if it arrives during an open interval similar to the interval $(t_0, t_f)$ described in the previous paragraph. Since the interadmission times for type $l$ customers during such an interval are i.i.d. random variables with LST of distribution function $H_l(s)$, it follows that the waiting-time distribution of a type $l$ customer, given that it is required to wait, is identical to that of a customer in an M/G/1 queue [with arrival rate $\lambda_l$ and LST of service-time distribution $H_l(s)$], which arrives to find the server busy. Let $W_l(s)$ denote the LST of the waiting-time distribution of a type $l$ customer given that it is required to wait. Then, from page 223 of Ref. 8, we have

$$W_l(s) = \frac{(1 - H_l(s))(1 - \rho_l)}{(s - \lambda_l + \lambda_l H_l(s))E(H_l)}, \tag{2}$$

where

$$\rho_l = \lambda_l E(H_l)$$

and

$$l = 2, \cdots, n.$$

The LST of the unconditional waiting-time distribution for a customer of type $l$ is given by

$$\sum_{j=0}^{k_l-1} u_j + W_l(s) \sum_{j=k_l}^{\infty} u_j. \tag{3}$$

The waiting-time distribution of a class 1 customer is easier to characterize. Given that a class 1 customer has to wait, its waiting time is the same as the sojourn time in an M/M/1 queue where the server is working at a rate of $\mu_m(=\mu$, say). So,

$$W_1(s) = \frac{\mu - \lambda_1}{\mu - \lambda_1 + s}, \tag{4}$$

and the unconditional waiting-time distribution is given by

$$\sum_{j=0}^{k_1-1} u_j + W_1(s) \sum_{j=k_1}^{\infty} u_j. \tag{5}$$

In the remainder of this section, we will show how to calculate $u_j$ and $H_l(s)$. Let us define a random variable $B_l$ to be the elapsed time from the instant $J_1$ changes from $k_l - 1$ to $k_l$ until the next instant when the value of $J_1$ drops from $k_l$ to $k_l - 1$ and $J_l = 0$. In other words, $B_l$ is the length of an interval similar to $(t_0, t_f)$ discussed earlier. From the preceding discussion it should be clear that $B_l$ constitutes a busy period of an M/G/1 queue with arrival rate $\lambda_l$ and the LST of service-time distribution $H_l(s)$. Let $B_l(s)$ be the LST of the distribution of $B_l$. Then $B_l(s)$ and $H_l(s)$ are related by (see page 212 of Ref. 8)

$$B_l(s) = H_l[s + \lambda_l - \lambda_l B_l(s)]. \tag{6}$$

Since $B_1(s)$ represents the busy period of an M/M/1 queue,

$$B_1(s) = \frac{\mu + \lambda_1 + s - [(\mu + \lambda_1 + s)^2 - 4\mu\lambda_1]^{1/2}}{2\lambda_1} \tag{7}$$

from page 215 of Ref. 8.

Next, we define the random variable $C_j$ to denote the first passage time from the state $J_1 = j$ to $J_1 = j - 1$, where $k_{l-1} > j > k_l$. Let $C_j(s)$ be the LST of the distribution of $C_j$. It should be clear that the waiting customers of class $l, \cdots, n$ play no role in determining this first passage time. Further, $J_1 = j$ implies that $J_2 = J_3 = \cdots = J_{l-1} = 0$. From these observations, now it is possible to write down the following equations for $C_j(s)$:

$$C_j(s) = \left(\frac{1}{\Lambda_l + \mu_j + s}\right)(\Lambda_l C_{j+1}(s)C_j(s) + \mu_j)$$

$$\text{for} \quad j = k_l + 1, \cdots, k_{l-1} - 2;$$

$$C_{k_{l-1}-1}(s) = \left(\frac{1}{\Lambda_l + \mu_{k_{l-1}-1} + s}\right)(\Lambda_l B_{l-1}(s)C_{k_{l-1}-1}(s) + \mu_{k_{l-1}-1});$$

$$H_l(s) = \left(\frac{1}{\Lambda_l + \mu_{k_l} + s}\right)(\Lambda_l C_{k_l+1}(s)H_l(s) + \mu_{k_l});$$

and

$$\Lambda_l = \sum_{k=1}^{l-1} \lambda_k \quad \text{for} \quad l = 2, \cdots, n. \tag{8}$$

Thus, eq. (8) defines a recursive technique for obtaining $H_l(s)$ from $B_{l-1}(s)$ via the functions $C_j(s)$ for $k_{l-1} > j > k_l$. By using eq. (6), one can obtain $B_l(s)$ from $H_l(s)$; and eq. (7) provides the value of $B_1(s)$, the boundary for eq. (8) when $l = 2$.

Finally, we show how to characterize $u_j$ for $j = 0, \cdots$, which is the steady-state marginal distribution of $J_1$. To do this, we define $n$ Semi-

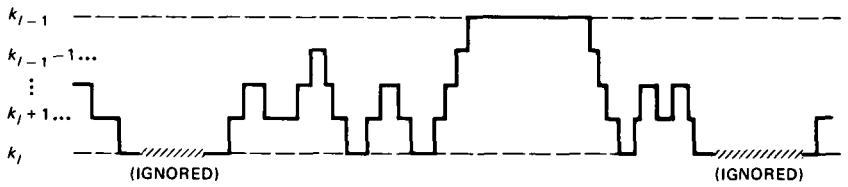


Fig. 6—Sample function of the $l$th SMP derived from the sample function of $J_1$ shown in Fig. 5.

Markov Processes (SMP), where the $l$th SMP has states $k_l, \cdots, k_{l-1}$, where $l = 2, \cdots, n+1$. The state of the $l$th SMP is the realization of the random variable $J_1$ with two differences. When $J_1 > k_{l-1}$, we will assume that the state of the SMP is $k_{l-1}$. Further, we will simply ignore the times when $J_1 < k_l$. A typical sample function of the $l$th SMP shown in Fig. 6 may help illustrate the structure of these SMPs. The transition probability for the SMP from state $j$ to $j+1$ is $\Lambda_l/(\Lambda_l + \mu_j)$ and $j$ to $j-1$ is $\mu_j/(\Lambda_l + \mu_j)$, where $k_l < j < k_{l-1}$. The holding time in state $j (k_l < j < k_{l-1})$ is exponential with rate $\Lambda_l + \mu_j$. For state $k_{l-1}$, the holding time is $B_l$ and this state makes a transition into state $k_{l-1} - 1$ with probability 1. State $k_l$ makes a transition into $k_l + 1$ with probability 1, and the holding time is exponential with rate $\Lambda_l$. In the description of the SMP, only one statement needs clarification, that is, the holding time in state $k_l$. Since we are ignoring all times when $J_1 < k_l$, this is equivalent to ignoring the transition of $J_1$ from $k_l$ to $k_l - 1$. The result follows from observing that the transition from $k_l$ to $k_l + 1$ occurs at an exponential rate of $\Lambda_l$. It is relatively easy to solve these $n$ SMPs using methods described in Ross.[9] Let $\pi_{jl}$ be the steady-state probability of state $j$ in the $l$th SMP ($l = 2, \cdots, n+1; j = k_l, \cdots, k_{l-1}$). Then it should be clear that

$$P(J_1 = j \mid J_1 \geq k_l) = u_j \Big/ \sum_{i=k_l}^{\infty} u_i = \pi_{jl} \quad \text{for} \quad j = k_l, \cdots, k_{l-1} - 1 \quad (9)$$

and

$$P(J_1 \geq k_{l-1} \mid J_1 \geq k_l) = \sum_{i=k_{l-1}}^{\infty} u_i \Big/ \sum_{i=k_l}^{\infty} u_i = \pi_{k_{l-1}l}.$$

It is easy to use (9) to calculate $u_j (j = 0, \cdots, m-1)$ recursively, starting with the solution of the $(n+1)$st SMP and working backwards through to the second SMP. For $j \geq m$, the server always works at a rate of $\mu_m (= \mu)$, and the random variable $J_1$ behaves like the number in the system for an M/M/1 queue. Thus, we have

$$u_j = (1 - \lambda_1/\mu)(\lambda_1/\mu)^{j-m}\left(1 - \sum_{j=0}^{m-1} u_j\right)$$

for $j \geq m$. We note at this point that the probability that the system is idle is given by $u_0$.

## IV. THE TIME SPENT IN THE SERVICE AREA

In this section, we describe methods of obtaining the mean time spent in the service area by class for the LCFS-PR and MS disciplines. For the PS discipline, we describe a method for characterizing the means when $n = 2$.

### *4.1 The LCFS-PR discipline*

In this discipline, we will assume that on entry into the service area, a customer occupies the lowest-numbered service position that is empty. Further, the server renders service to the customer in the highest-numbered service position that is nonempty. Thus, a customer occupies the same service position from entry until departure. Let $T_j(s)$ be the LST of the distribution of time spent in the service area by a customer who occupies position $j$ on entry into the service area. Then,

$$T_{k_1}(s) = \mu/(\mu + s)$$

and

$$T_j(s) = \begin{cases} C_j(s) & \text{if} \quad k_{l-1} > j > k_l \\ H_l(s) & \text{if} \quad j = k_l \end{cases} \quad \text{and} \quad l = 2, \cdots, n. \tag{10}$$

It is now easy to use the results of Section II to obtain this LST or any other characterization of this distribution.

### *4.2 The multiple-server discipline*

In this discipline, let $\mu_i = i\sigma$. Then the time spent in the service area is simply exponential with parameter $\sigma$.

### *4.3 The processor sharing discipline*

The exact solution for the mean time spent in the service area can be obtained by first noting that customer classes are indistinguishable on entry into the service area. We denote the state of the system by an $n$-tuple $(J_1, J_2, \cdots, J_n)$ as seen by a customer of class $l$ after entry into the service area and let $Q_l(\mathbf{j})$ denote the probability that the state of the system is $\mathbf{j}$ when an arbitrary class $l$ customer is admitted to the service area, where $j_1$ includes the newly admitted customer. Further, let $x(\mathbf{j})$ be the mean time spent in the service area for a

customer who sees state $\mathbf{j}$ immediately on being admitted to the service area. If we let $t_l$ be the mean time spent in the service area by a class $l$ customer, then

$$t_l = \sum_{\mathbf{j} \in A_l} x(\mathbf{j}) Q_l(\mathbf{j}), \tag{11}$$

where

$$A_l = \{(j_1, j_2, \cdots, j_n) \mid j_k = 0 \text{ for } 2 \le k < l\} \quad \text{for} \quad l = 1, \cdots, n.$$

It is possible to obtain $Q_l(\mathbf{j})$ from the following observations:

1. For $j_1 < k_l$,

$$Q_l(\mathbf{j}) = P(\mathbf{j} - \mathbf{e}_1).$$

For a customer of class $l$ to see $j_1$ including itself, there must be $j_1 - 1$ ahead of it in the service area.

2. Whenever a customer of class $l$ enters the service area, $J_2 = J_3 = \cdots = J_{l-1} = 0$ and $J_1 \le k_l$.

3. For $j_1 = k_l$, one of two disjoint events must occur. Either the class $l$ customer arrived to see $k_l - 1$ customers ahead of it in the service area or it must have waited in the waiting area prior to admission. The former case is identical to the first observation above. In the latter case, we have to characterize the distribution of $(J_l, \cdots, J_n)$ at the time of entry into the service area. The random variable $J_l$ behaves like the number waiting as seen by a customer about to enter service given that it had to wait in an M/G/1 queue with an arrival rate of $\lambda_l$ and a service time of $H_l$. The distribution of $J_k (k = l + 1, \cdots, n)$ is simply the convolution of what was seen on arrival and the number of new arrivals of type $k$ during the wait of the customer of class $l$.

In principle, it is possible to write down $Q_l(\mathbf{j})$ in terms of $P(\mathbf{j})$ from the observations made above. The notation is cumbersome, so we will not go into the details here. The exact derivation when $n = 2$ is given in Appendix B.

Further, for $k_{l+1} \le j_1 < k_l$, the $x(\mathbf{j})$ satisfy

$$(\Lambda + \mu_{j_1}) x(\mathbf{j}) = 1 + \sum_{k=l+1}^{n} \lambda_k x(\mathbf{j} + \mathbf{e}_k) + \sum_{k=1}^{l} \lambda_k x(\mathbf{j} + \mathbf{e}_1)$$

$$+ \left(\frac{j_1 - 1}{j_1}\right) \delta(j_1) \mu_{j_1} \{x(\mathbf{j} - \mathbf{e}_1)\delta(j_1 - k_{l+1}) + [x(\mathbf{j} - \mathbf{e}_1)(1 - \delta(j_{l+1}))$$

$$+ x(\mathbf{j} - \mathbf{e}_{l+1})\delta(j_{l+1})](1 - \delta(j_1 - k_{l+1}))\}. \tag{12}$$

The solution of this equation is not easy; however, it is possible to solve it for the case where $n = 2$. We present this solution in Appendix C.

## V. CONCLUDING REMARKS

In the earlier sections, we have shown how to characterize distributions of the waiting time and the time spent in the service area. Of interest in many applications would also be the sojourn time (i.e., the elapsed time between arrival into and departure from the system) of customers. In principle, it is possible to characterize the sojourn time distribution for the two-class PS problem by the methods used in Ref. 10.

We note that the time spent in the service area for the first-come first-served discipline is somewhat difficult to characterize, whereas the results for the waiting time is the same as that in Section II. The reason for this difficulty can be seen by first assuming that we are about to characterize a two-class problem. Then the mean time spent in the service area has to be found conditioned on $J_1$, $J_2$ and the position of the tagged customer in the service area. The difference equations for this mean time thus will be in three variables and are hard to solve.

## REFERENCES


1. C. Schaack and R. C. Larson, "An N Server Cutoff Multi-Priority Queue," Working Paper No. OR135-85, MIT, Cambridge (February 1985).
2. B. Avi-Itzhak and D. P. Heyman, "Approximate Queueing Models for Multiprogramming Computer Systems," Oper. Res., *21*, No. 6 (November–December 1973), pp. 1212–30.
3. M. Reiser and S. S. Lavenberg, "Mean Value Analysis of Closed Multichain Queueing Networks," IBM Research Report RC7023, 1978.
4. A. A. Fredericks, "Approximations for Customer Viewed Delays in Multiprogrammed Transaction Oriented Computer Systems," B.S.T.J., *59*, No. 9 (November 1980), pp. 1559–74.
5. M. Schwartz, "Performance Analysis of the SNA Virtual Route Pacing Control," IEEE Trans. Commun., *COM-30*, No. 1 (January 1982), pp. 172–84.
6. M. Reiser, "Admission Delays on Virtual Routes With Window Flow Control," Performance of Data Communications Systems, G. Pujolle, Ed., New York: North-Holland, 1981.
7. M. Reiser, "A Queueing Network Analysis of Computer Communication Networks With Window Flow Control," IEEE Trans. Commun., *COM-27*, No. 8 (August 1979), pp. 1199–209.
8. L. Kleinrock, *Queueing Systems, Vol. 1: Theory*, New York: Wiley, 1975.
9. S. M. Ross, "Applied Probability Models With Optimization Applications," San Francisco: Holden-Day, 1969.
10. K. M. Rege and B. Sengupta, "Sojourn Time Distribution in a Multiprogrammed Computer System," AT&T Tech. J., *64*, No. 5 (May–June 1985), pp. 1077–90.


## APPENDIX A

### *The Steady-State Probabilities for a Two-Class Problem*

In this appendix, we show how to obtain the solution to eq. (1) when $n = 2$. For the sake of notational ease, we will refer to the random variables $J_1$ and $J_2$ as $I$ and $J$ in this and the other appendices. Further, $i$ and $j$ will be the realizations of the random variables $I$ and $J$, respectively. Let $P_{ij}$ be the steady-state probability that $I = i$ and $J = j$. Equation (1) reduces to the following equations when $n$ is 2:

$$(\lambda_1 + \lambda_2 + \mu_i)P_{ij} = \lambda_1 P_{i-1,j} + \lambda_2 P_{i,j-1} + \mu_{i+1}P_{i+1,j}$$
$$\text{if} \quad i > k_2, \quad j > 0, \quad (13)$$

$$(\lambda_1 + \lambda_2 + \mu_i)P_{i0} = \lambda_1 P_{i-1,0} + \lambda_2 P_{i-1,0} + \mu_{i+1}P_{i+1,0}$$
$$\text{if} \quad 0 < i < k_2, \quad (14)$$

$$(\lambda_1 + \lambda_2 + \mu_{k_2})P_{k_2,0} = (\lambda_1 + \lambda_2)P_{k_2-1,0} + \mu_{k_2}P_{k_2,1} + \mu_{k_2+1}P_{k_2+1,0} \quad (15)$$

$$(\lambda_1 + \lambda_2 + \mu_i)P_{i0} = \lambda_1 P_{i-1,0} + \mu_{i+1}P_{i+1,0} \quad \text{if} \quad i > k_2, \quad (16)$$

$$(\lambda_1 + \lambda_2 + \mu_{k_2})P_{k_2,j} = \lambda_2 P_{k_2,j-1} + \mu_{k_2}P_{k_2,j+1} + \mu_{k_2+1}, P_{k_2+1,j}$$
$$\text{if} \quad j > 0 \quad (17)$$

and

$$(\lambda_1 + \lambda_2)P_{00} = \mu_1 P_{10}. \quad (18)$$

It should be clear that $u_0$ (of Section III) is the same as $P_{00}$. So, one can obtain $P_{i0}$ for $0 \le i \le k_2$ by using (18) and (14). Even though the coefficients of $P_{ij}$ in eqs. (13) and (16) depend on $i$, it is easy to see that these are constant coefficient partial difference equations when $i \ge m$. Further, these equations are very similar to the ones solved by Rege and Sengupta.[10] Using these methods, it is easy to show that

$$P_{i0} = B_0 \sigma_2^i \quad (19)$$

and

$$P_{ij} = \frac{\rho_2}{(\sigma_1 - \sigma_2)} \left( \sum_{\nu=i+1}^{\infty} \sigma_1^{i-n} P_{\nu,j-1} + \sum_{\nu=m}^{i} \sigma_2^{i-n} P_{\nu,j-1} + B_j \sigma_2^i \right)$$
$$\text{for} \quad i \ge m \text{ and } j > 0, \quad (20)$$

where

$$\sigma_1 = (1 + \rho_1 + \rho_2 + \sqrt{(1 + \rho_1 + \rho_2)^2 - 4\rho_1})/2,$$
$$\sigma_2 = (1 + \rho_1 + \rho_2 - \sqrt{(1 + \rho_1 + \rho_2)^2 - 4\rho_1})/2,$$
$$\rho_1 = \lambda_1/\mu, \quad \rho_2 = \lambda_2/\mu,$$

and $\{B_j, j = 0, 1, \cdots\}$ constitute a sequence of unknown constants to be determined from the boundary conditions (15) and (17).

Now we will show how to determine the unknown constants in two steps. First, we will show this for $B_0$ and then for $B_j$. It is possible to determine $B_0$ by assuming two trial values and using (19) and (16) to recursively calculate two sets of $P_{i0}$ for $i = m + 1, m, \cdots, k_2$. Since each of these $P_{i0}$ is a linear function of the unknown constant $B_0$, now it is easy to use linear interpolation to obtain the correct value of $B_0$ that agrees with $P_{k_2,0}$ already obtained from (14). One can now use

(15) to obtain $P_{k_2,1}$. To calculate $B_j (j > 0)$, let us assume the $P_{k_2,j}$ has been obtained from (15) for $j = 1$ or from (17) for $j > 1$. Further assume that $P_{ik} (0 \leq k \leq j - 1$ and all $i)$ are known. As before, we start with two trial values of $B_j$ and recursively calculate two sets of $P_{ij}$ for $i = m + 1, \cdots, k_2$ by using (20) and (13). Since each of the $P_{ij}$ is a linear function of $B_j$, we can use linear interpolation to obtain the correct value of $B_j$ that agrees with $P_{k_2,j}$ already obtained from (15) for $j = 1$ or from (17) for $j > 1$. Finally, one can use (17) to obtain $P_{k_2,j+1}$.

## APPENDIX B

### *State Probability As Seen by Customers on Admission to the Service Area*

Here we describe the procedures for computing $Q_1(i, j)$ and $Q_2(i, j)$ that denote the state probabilities as seen upon admission to the service area by type 1 and type 2 customers, respectively. The procedure for computation of $Q_2(i, j)$ is described first.

It is clear that if, on arrival, a type 2 customer finds less than $k_2$ customers in the system, it does not have to wait before entering the service area. So,

$$Q_2(i, j) = P(i - 1, j) \quad \text{for} \quad i < k_2. \tag{21}$$

It is also obvious that $Q_2(i, j) = 0$ for $i > k_2$, since a type 2 customer cannot enter the service area if the number of customers in the service area other than itself is greater than or equal to $k_2$. A type 2 customer will see $I = k_2$ upon its admission to the service area in one of two ways: (1) if there are $k_2 - 1$ customers in the system just before its arrival, or (2) if $I \geq k_2$ at the time of its arrival and it has to wait until all type 2 customers ahead of it in the waiting area are admitted to the system and a departure occurs from the state $I = k_2$.

From the analysis of Section II, $W_2(s)$ is the LST of the distribution of the waiting time of a type 2 customer given that it has to wait. The generating function of the number of arrivals of type 2 during this wait is $W_2(\lambda_2(1 - z))$. Further, the probability that a type 2 customer has to wait before entering the service area is $\sum_{i=k_2}^{\infty} \mu_i$.

Thus,

$$Q_2(k_2, j) = \hat{\delta}_j P_{k_2-1,0} + \frac{(d^j/dz^j) W_2(\lambda_2(1 - z))|_{z=0}}{j!} \sum_{j=k_2}^{\infty} \mu_i, \tag{22}$$

where $\hat{\delta}_j = 1$ if $j = 0$ and 0 otherwise.

To compute $Q_1(i, j)$, we note that type 1 customers do not have to wait if there are less than $k_1$ customers in the service area at the time of their arrival. Thus,

$$Q_1(i, j) = P(i - 1, j) \qquad \text{for} \quad i < k_1. \tag{23}$$

For $i > k_1$,

$$Q_1(i,j) = \sum_{i'=k_1}^{\infty} \sum_{j'=0}^{j} P(i',j') \cdot \int_0^{\infty} \frac{\mu(\mu t)^{i'-k_1}}{(i'-k_1)!} e^{-\mu t} \frac{(\lambda_1 t)^{i-k_1}}{(i-k_1)!} e^{-\lambda_1 t} \frac{(\lambda_2 t)^{j-j'}}{(j-j')!} e^{-\lambda_2 t} dt. \quad (24)$$

For $i = k_1$, there are two possibilities: (1) if the type 1 customer finds $k_1 - 1$ customers in the service area upon arrival and does not have to wait, (2) if it has to wait but no type 1 arrivals occur during its wait. Thus,

$$Q_1(k_1,j) = P(k_1-1,j) + \sum_{i'=k_1}^{\infty} \sum_{j'=0}^{j} P(i',j') \cdot \int_0^{\infty} \frac{\mu(\mu t)^{i'-k_1}}{(i'-k_1)!} e^{-\mu t} e^{-\lambda_1 t} \frac{(\lambda_2 t)^{j-j'}}{(j-j')!} e^{-\lambda_2 t} dt. \quad (25)$$

In deriving (24) and (25) above, we have used the facts that the waiting time of a type 1 customer (given that it has to wait) has a gamma distribution and that the type 1 and type 2 arrivals that occur during this wait are independent and Poisson.

## APPENDIX C

### *Characterization of the Time Spent in the Service Area by a Tagged Customer in a Two-Class Processor Sharing System*

Let $X(s)$ denote the LST of the distribution of the time spent in the service area by an arbitrary "tagged customer." Similarly, let $X_{i,j}(s)$ denote the LST of the conditional distribution of this random variable given that the state $\mathbf{J}$ of the system at the time the tagged customer was admitted to the service area was $(i,j)$. (Here $i$ is assumed to include the tagged customer.) Let $\mathbf{X}_i(s)$ denote the row vector with entries

$$\{\mathbf{X}_i(s)\}_j = X_{i,j}(s) \quad \text{for} \quad i \geq k_2, \quad j \geq 0. \quad (26)$$

When $i < k_2$, there can be no type 2 customers waiting outside, that is, $j = 0$. So we let $X_i(s)$ denote the quantity $\mathbf{X}_i(s)$. We are interested in determining the quantities $\mathbf{X}_i(s)$ for $i \geq k_2$ and $X_i(s)$ for $1 \leq i < k_2$.

Assume that the tagged customer was admitted to the service area at time 0 and let $\tilde{T}$ denote the time at which the tagged customer finishes service and quits the system. Then,

$$\{\mathbf{X}_i(s)\}_j = E[e^{-s\tilde{T}} \mid \mathbf{J}_{0^+} = (i,j)] \quad \text{for} \quad i \geq k_2, \quad j = 0, 1, \cdots,$$

and

$$X_i(s) = E[e^{-s\tilde{T}} \mid \mathbf{J}_{0^+} = (i,0)] \quad \text{for} \quad i \leq i \leq k_2 - 1. \quad (27)$$

Let $T_i$ denote the first passage time (after time 0) into the state $(i, .)$, that is,

$$T_i = \text{Min}\{t: t \geq 0, \mathbf{J}_t = (i, .)\} \quad \text{for} \quad i \geq 1. \tag{28}$$

Also, for $i \geq k_2$ and $j$, $k = 0, 1, 2, \cdots$, let ${}_iR_{j,k}(s)$ denote the quantity

$$E[e^{-sT_{i-1}} 1\{\tilde{T} > T_{i-1}\} 1\{\mathbf{J}_{T_{i-1}^+} = (i-1, j)\} \mid \mathbf{J}_{0^+} = (i, k)],$$

where $1\{A\}$ denotes the indicator function of the event $A$. Observe that for $i \geq m$ the server continues to work at the maximum multiprogramming level until the first passage time into the state $(i-1, \cdot)$ so that, sample path by sample path, the above expectation is independent of $i$. Thus,

$${}_iR_{j,k}(s) = R_{j,k}(s) \quad \text{for} \quad i \geq m. \tag{29}$$

Also, during this time no type 2 jobs are admitted to the service area so that $j$ cannot be less than $k$, that is,

$$R_{j,k}(s) = 0 \quad \text{for} \quad j < k. \tag{30}$$

Let $R(s)$ denote the matrix with entries

$$\{R(s)\}_{j,k} = R_{j,k}(s) \quad \text{for} \quad j, k = 0, 1, 2, \cdots. \tag{31}$$

Then $R(s)$, clearly, is lower triangular. Moreover, by a sample path argument it can be shown that $R_{j,k}(s)$ depends upon the difference, $j - k$, which represents the number of type 2 arrivals during $(0, T_{i-1})$, and not on $k$, which refers to the number of type 2 customers waiting outside at time $0^+$. Thus $R(s)$ has the form

$$R(s) = \begin{bmatrix} r_0(s) & & & 0 \\ r_1(s) & r_0(s) & & \\ r_2(s) & r_1(s) & r_0(s) & \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix}. \tag{32}$$

### *C.1 A matrix equation for R(s)*

By conditioning on the first event to occur, we write $R_{j,k}(s)$ as

$$R_{j,k}(s) = \frac{\mu(m-1)/m}{\Lambda + \mu + s}\,\delta_{j,k} + \frac{\Lambda_1}{\Lambda + \mu + s} E[e^{-sT_{i-1}} 1\{\tilde{T} > T_{i-1}\}$$

$$\cdot 1\{\mathbf{J}_{T_{i-1}^+} = (i-1, j)\} \mid \mathbf{J}_{0^+} = (i+1, k)] + \frac{\lambda_2}{\Lambda + \mu + s} E[e^{-sT_{i-1}}$$

$$\cdot 1\{\tilde{T} > T_{i-1}\} 1\{\mathbf{J}_{T_{i-1}^+} = (i-1, j)\} \mid \mathbf{J}_{0^+} = (i, k+1)], \tag{33}$$

where the first term corresponds to the event "departure of a nontagged customer," the second to the event "arrival of a type 1 customer," and

the third to the event "arrival of a type 2 customer." Now the expectation in the second term can be written as

$$E[e^{-sT_{i-1}}1\{\tilde{T} > T_{i-1}\}1\{\mathbf{J}_{T_{i-1}^+} = (i-1, j)\} \mid \mathbf{J}_{0^+} = (i+1, k)]$$

$$= \sum_{j'=0}^{\infty} E[e^{-s(T'+T'')}1\{\tilde{T} > T' + T''\}1\{\mathbf{J}_{T'+T''} = (i-1, j)\}$$

$$\cdot 1\{\mathbf{J}_{T'} = (i, j')\} \mid \mathbf{J}_{0^+} = (i+1, k)], \quad (34)$$

where $T'$ is the first passage time into the state $(i, \cdot)$ and $T''$ is the time elapsed since the first passage into $(i, \cdot)$ until the first passage into the state $(i-1, \cdot)$. Clearly, $T_{i-1} = T' + T''$ since the Markov process $\mathbf{J}$ is skip-free. From the Markov property of $\mathbf{J}_{T'}$ it follows that the right-hand size of (34) equals

$$\sum_{j'=0}^{\infty} E[e^{-sT''}1\{\tilde{T} - T' > T''\}1\{\mathbf{J}_{(T'+T'')^+} = (i-1, j)\} \mid \mathbf{J}_{T'^+}$$

$$= (i, j'), \tilde{T} > T'] \times E[e^{-sT'}1\{\tilde{T} > T'\}1\{\mathbf{J}_{T'^+} = (i, j')\} \mid \mathbf{J}_{0^+}$$

$$= (i+1, k)],$$

which is nothing but

$$\sum_{j'=0}^{\infty} R_{j,j'}(s)R_{j',k}(s) = \{R^2(s)\}_{j,k}.$$

Thus, (34) can be written as

$$R_{j,k}(s) = \frac{\mu(m-1)/m}{\Lambda + \mu + s}\,\delta_{j,k} + \frac{\lambda_1}{\Lambda + \mu + s}\,\{R^2(s)\}_{j,k}$$

$$+ \frac{\lambda_2}{\Lambda + \mu + s}\,R_{j,k+1}(s) \quad (35)$$

or in a matrix form

$$R(s) = \frac{\mu((m-1)/m)}{\Lambda + \mu + s}\,I + \frac{\lambda_1}{\Lambda + \mu + s}\,R^2(s) + \frac{\lambda_2}{\Lambda + \mu + s}\,R(s)\Delta. \quad (36)$$

In (35) and (36), $\delta_{j,k}$ is the Kronecker $\delta$, that is, $\delta_{j,k}$ equals 1 if $j = k$ and 0 otherwise; and the matrix $\Delta$ is given by

$$\Delta = \begin{bmatrix} 0 & & & & \\ 1 & 0 & & 0 & \\ & 1 & 0 & & \\ 0 & & 1 & 0 & \\ & & \cdot & \cdot & \cdot \\ & & \cdot & & \cdot \end{bmatrix}. \quad (37)$$

The structure of $R(s)$ as given in (32) makes it possible to compute the entries $r_i(s)$ recursively. $r_0(s)$ satisfies the quadratic equation

$$r_0(s) = \frac{\mu(m-1)/m}{\Lambda + \mu + s} + \frac{\lambda_1}{\Lambda + \mu + s} r_0^2(s) \tag{38}$$

so that $r_0(s)$ is given by

$$r_0(s) = \frac{(\Lambda + \mu + s) \pm \sqrt{(\Lambda + \mu + s)^2 - 4\lambda_1\mu(m-1)/m}}{2\lambda_1}. \tag{39}$$

Since, for $s \geq 0$, $|r_0(s)|$ must be less than or equal to 1, the larger of the two roots is unacceptable. Thus,

$$r_0(s) = \frac{(\Lambda + \mu + s) - \sqrt{(\Lambda + \mu + s)^2 - 4\lambda_1\mu(m-1)/m}}{2\lambda_1}. \tag{40}$$

Once $r_0(s)$ is known, $r_i(s)$, for $i \geq 1$, can be computed recursively since $r_i(s)$ is expressible in terms of $r_0(s), \cdots, r_{i-1}(s)$. We note here that $r_0(s)$ has a form similar to that of $\sigma_2(s)$ in Ref. 10. In fact, if $\lambda_2 = 0$, then $r_0(s)$ and $\sigma_2(s)$ are identical and represent the same quantity.

### *C.2 The structure of $\mathbf{X}_i(s)$*

Now that we have an explicit expression for $R(s)$, we shall attempt to express the quantities $\mathbf{X}_i(s)$ in terms of $R(s)$.

The $k$th entry of $\mathbf{X}_i(s)$ can be written as

$$\{\mathbf{X}_i(s)\}_k = E[e^{-s\tilde{T}}1\{\tilde{T} > T_{i-1}\} \mid \mathbf{J}_{0^+} = (i, k)] + E[e^{-s\tilde{T}}1\{\tilde{T} \leq T_{i-1}\} \mid \mathbf{J}_{0^+} = (i, k)]. \tag{41}$$

The first term on the right-hand side of (41) can be expanded as

$$E[e^{-s\tilde{T}}1\{\tilde{T} > T_{i-1}\} \mid \mathbf{J}_{0^+} = (i, k)] = \sum_{j=0}^{\infty} E[e^{-s[(\tilde{T}-T_{i-1})+T_{i-1}]} \cdot 1\{\tilde{T} > T_{i-1}\}1\{\mathbf{J}_{T_{i-1}^+} = (i-1, j)\} \mid \mathbf{J}_{0^+} = (i, k)], \tag{42}$$

which, because of the Markov property of $\mathbf{J}_t$, reduces to

$$\sum_{j=0}^{\infty} E[e^{-s(\tilde{T}-T_{i-1})} \mid \mathbf{J}_{T_{i-1}^+} = (i-1, j),\ \tilde{T} > T_{i-1}]E[e^{-sT_{i-1}} \cdot 1\{\tilde{T} > T_{i-1}\}1\{\mathbf{J}_{T_{i-1}^+} = (i-1, j)\} \mid \mathbf{J}_{0^+} = (i, k)].$$

It follows from the memoryless property of the tagged customer's service-time distribution that $E[e^{-s(\tilde{T}-T_{i-1})} \mid \mathbf{J}_{T_{i-1}^+} = (i-1, j),\ \tilde{T} > T_{i-1}]$ equals $\{\mathbf{X}_{i-1}(s)\}_j$; and, for $i \geq m$, $E[e^{-sT_{i-1}}1\{\tilde{T} > T_{i-1}\}1\{\mathbf{J}_{T_{i-1}^+} = (i-1, j)\} \mid \mathbf{J}_{0^+} = (i, k)]$ equals $\{R(s)\}_{j,k}$. Thus (42) reduces to

$$E[e^{-s\tilde{T}}1\{\tilde{T} > T_{i-1}\} \mid \mathbf{J}_{0^+} = (i, k)] = \{\mathbf{X}_{i-1}(s)R(s)\}_k \quad \text{for} \quad i \geq m. \tag{43}$$

To derive an expression for the second term in (41), we note that, for $i \geq m$, no sample path that figures in the expectation

$E[e^{-s\hat{T}}1\{\hat{T} \le T_{i-1}\} \mid \mathbf{J}_{0^+} = (i, k)]$ allows the multiprogramming level to fall below $m$ before the departure of the tagged customer. Thus the server continues to operate at rate $\mu$ until the tagged customer's departure. Also, for any two initial states $(i_1, k_1)$ and $(i_2, k_2)$, as long as $i_1, i_2 \ge m$, there is a one-to-one correspondence between sample paths describing the trajectory of $\mathbf{J}_t$ between 0 and $\hat{T}$, which make equal contributions to the expectation. Thus,

$$E[e^{-s\hat{T}}1\{\hat{T} \le T_{i-1}\} \mid \mathbf{J}_{0^+} = (i, k)] = a(s) \quad \text{for} \quad i \ge m, \tag{44}$$

where $a(s)$ is independent of $i$ and $k$.

To derive an expression for $a(s)$, it will be convenient to introduce a quantity $\sigma(s)$ defined by

$$\sigma(s) = E[e^{-sT_{i-1}}1\{\hat{T} > T_{i-1}\} \mid \mathbf{J}_{0^+} = (i, k)] \quad \text{for} \quad i \ge m, \quad k = 0, 1, \cdots. \tag{45}$$

Note that although the definition of $\sigma(s)$ involves the initial state $(i, k)$, $\sigma(s)$ is independent of the latter as long as $i \ge m$. Also note that

$$\sigma(s) = \sum_{k=0}^{\infty} r_k(s). \tag{46}$$

By conditioning on the first relevant event to occur, $\sigma(s)$ can be written as

$$\sigma(s) = \frac{\mu(m-1)/m}{\lambda_1 + \mu + s} + \frac{\lambda_1}{\lambda_1 + \mu + s} \cdot E[e^{-s(T_{i-1} - T_{i+1})}1\{\hat{T} > T_{i-1}\} \mid \mathbf{J}_{T_{i+1}} = (i+1, \cdot),\ T_{i-1} > T_{i+1}]. \tag{47}$$

In (47) it can be seen that arrivals of type 2 customers are completely ignored since they do not affect the mechanics involved. Since the first passage time from the state $(i+1, \cdot)$ to $(i-1, \cdot)$ involves the sum of the first passage times from $(i+1, \cdot)$ to $(i, \cdot)$ and from $(i, \cdot)$ to $(i-1, \cdot)$, which are i.i.d., (47) reduces to

$$\sigma(s) = \frac{\mu(m-1)/m}{\lambda_1 + \mu + s} + \frac{\lambda_1}{\lambda_1 + \mu + s}\sigma^2(s). \tag{48}$$

Equation (48) is identical to the one describing $\sigma_2(s)$ in Ref. 10, with $\lambda$ replaced by $\lambda_1$. The desired root of (48) is the smaller of the two roots so that

$$\sigma(s) = \frac{(\lambda_1 + \mu + s) - \sqrt{(\lambda_1 + \mu + s)^2 - 4\mu\lambda_1(m-1)/m}}{2\lambda_1}. \tag{49}$$

Now $a(s)$ can be obtained in a straightforward manner from $\sigma(s)$. By conditioning on the first relevant event to occur, we write

$$a(s) = \frac{\mu/m}{\mu + \lambda_1 + s} + \frac{\lambda_1}{\mu + \lambda_1 + s} E[e^{-s(\tilde{T}-T_{i+1})}1\{\tilde{T} \le T_{i-1}\} \mid \mathbf{J}_{T_{i+1}^+}$$

$$= (i + 1, \cdot), \tilde{T} > T_{i+1}] = \frac{\mu/m}{\mu + \lambda_1 + s} + \frac{\lambda_1}{\mu + \lambda_1 + s}$$

$$\cdot [E[e^{-s(\tilde{T}-T_{i+1})}1\{\tilde{T} \le T_{i-1}\}1\{\tilde{T} \le T'\} \mid \mathbf{J}_{T_{i+1}^+} = (i + 1, \cdot), \tilde{T} > T_{i+1}]$$

$$+ E[e^{-s(\tilde{T}-T_{i+1})}1\{\tilde{T} \le T_{i-1}\}1\{\tilde{T} > T'\} \mid \mathbf{J}_{T_{i+1}^+} = (i + 1, \cdot), \tilde{T} > T_{i+1}]], \quad (50)$$

where $T'$ denotes the first passage time into the state $(i, \cdot)$ after the state has reached $(i + 1, \cdot)$ at time $T_{i+1}$. Following arguments similar to the ones used earlier, it can be shown that (50) can be written as

$$a(s) = \frac{\mu/m}{\mu + \lambda_1 + s} + \frac{\lambda_1}{\mu + \lambda_1 + s} [a(s) + \sigma(s)a(s)],$$

that is,

$$a(s) = \mu/m[\mu + s - \lambda_1\sigma(s)]^{-1}. \quad (51)$$

The desired vector $\mathbf{X}_i(s)$ can now be written as

$$\mathbf{X}_i(s) = \mathbf{a}(s) + \mathbf{X}_{i-1}(s)R(s) \quad \text{for} \quad i \ge m, \quad (52)$$

where $\mathbf{a}(s)$ is the row vector $(a(s), a(s), a(s), \cdots)$.

If we introduce two more vectors $\alpha(s)$ and $\beta(s)$, where

$$\beta(s) = \frac{\mu/m}{\mu/m + s} [1, 1, 1, \cdots], \quad (53)$$

and

$$\alpha(s) = \mathbf{a}(s) - \beta(s) + \mathbf{X}_{m-1}(s)R(s), \quad (54)$$

the vectors $\mathbf{X}_i(s)$ for $i \ge m$ can be expressed as

$$\mathbf{X}_i(s) = \beta(s) + \alpha(s)[R(s)]^{i-m}. \quad (55)$$

(The above equation can be proved by mathematical induction.)

It can be seen that as $i \to \infty$, the term $\alpha(s)[R(s)]^{i-m}$ vanishes, so that $\mathbf{X}_i(s)$ approaches its limiting value $\beta(s)$. In other words, when $i$ is large, the tagged customer receives its entire service at the rate $\mu/m$ as expected.

### *C.3 Boundary conditions*

For $i \ge m$, eq. (55) characterizes the vectors $\mathbf{X}_i(s)$ in terms of known quantities $\beta(s)$, $a(s)$, $R(s)$, and the unknown $\mathbf{X}_{m-1}(s)$. In other words, if $\mathbf{X}_{m-1}(s)$ is known, $\mathbf{X}_i(s)$ can be obtained, for $i \ge m$, directly from (55). To completely characterize the time spent in the service area by a tagged customer, it remains to derive the boundary conditions, that

is, a system of equations from which the quantities $X_1(s)$, $X_2(s)$, $\cdots$, $X_{k_1-1}(s)$, $\mathbf{X}_{k_1}(s)$, $\cdots$, $\mathbf{X}_{m-1}(s)$ can be obtained.

Without going into the details of derivation—it involves arguments similar to the ones used in the earlier analysis—we state the boundary conditions, which are as follows:

$$\mathbf{X}_i(s) = \frac{\mu_i/i}{\Lambda + \mu_i + s}\,\mathbf{1} + \frac{(i-1)\mu_i/i}{\Lambda + \mu_i + s}\,\mathbf{X}_{i-1}(s) + \frac{\lambda_1}{\Lambda + \mu_i + s}\,\mathbf{X}_{i+1}(s) + \frac{\lambda_2}{\Lambda + \mu_i + s}\,\mathbf{X}_i(s)\Delta \quad \text{for} \quad k_2 + 1 \le i \le m-1, \tag{56}$$

$$\mathbf{X}_{k_2}(s) = \frac{\mu_{k_2}/k_2}{\Lambda + \mu_{k_2} + s}\,\mathbf{1} + \frac{(k_2-1)\mu_{k_2}/k_2}{\Lambda + \mu_{k_2} + s}\,\{\mathbf{X}_{k_2}(s)\Delta^T + X_{k_2-1}(s)\mathbf{e}_0\} + \frac{\lambda_1}{\Lambda + \mu_{k_2} + s}\,\mathbf{X}_{k_2+1}(s) + \frac{\lambda_2}{\Lambda + \mu_{k_2} + s}\,\mathbf{X}_{k_2}(s)\Delta, \tag{57}$$

$$X_{k_2-1}(s) = \frac{\mu_{k_2-1}/(k_2-1)}{\Lambda + \mu_{k_2-1} + s} + \frac{(k_2-2)\mu_{k_2-1}/(k_2-1)}{\Lambda + \mu_{k_2-1} + s}\,X_{k_2-2}(s) + \frac{\Lambda}{\Lambda + \mu_{k_2-1}}\,\{\mathbf{X}_{k_2}(s)\}_0, \tag{58}$$

$$X_i(s) = \frac{\mu_i/i}{\Lambda + \mu_i + s} + \frac{(i-1)\mu_i/i}{\Lambda + \mu_i + s}\,X_{i-1}(s) + \frac{\Lambda}{\Lambda + \mu_i + s}\,X_{i+1}(s) \quad \text{for} \quad 2 \le i < k_2 - 1, \tag{59}$$

and

$$X_1(s) = \frac{\mu_1}{\Lambda + \mu_1 + s} + \frac{\Lambda}{\Lambda + \mu_1 + s}\,X_2(s), \tag{60}$$

where $\mathbf{e}_0 = [1\ 0\ 0\ 0\ \cdots]$ and $\mathbf{1} = [1\ 1\ 1\ \cdots]$.

Equations (55) through (60) give a characterization of the LST of the distribution of the time spent in the service area by the tagged customer given the state of the system at the time it was admitted to the service area. To derive the mean time spent in the service area, we need to differentiate these quantities at $s = 0$. Noting that $\alpha(0) = \mathbf{0}$, we have

$$\mathbf{X}_i'(0) = \beta'(0) + \alpha'(0)[R(0)]^{i-m} \quad \text{for} \quad i \ge m. \tag{61}$$

The boundary conditions also are obtained by differentiating the corresponding equations at $s = 0$:

$$\mathbf{X}_i'(0) = -\frac{1}{(\Lambda + \mu_i)}\,\mathbf{1} + \frac{(i-1)\mu_i/i}{(\Lambda + \mu_i)}\,\mathbf{X}_{i-1}'(0) + \frac{\lambda_1}{\Lambda + \mu_i}\,\mathbf{X}_{i+1}'(0) + \frac{\lambda_2}{\Lambda + \mu_i}\,\mathbf{X}_i'(0)\Delta \quad \text{for} \quad k_2 + 1 \le i < m - 1 \tag{62}$$

$$\mathbf{X}_{k_2}'(0) = -\frac{1}{\Lambda + \mu_{k_2}} + \frac{(k_2 - 1)\mu_{k_2}/k_2}{\Lambda + \mu_{k_2}}\,[\mathbf{X}_{k_2}'(0)\Delta^T + X_{k_2-1}'(0)\mathbf{e}_0] + \frac{\lambda_1}{\Lambda + \mu_{k_2}}\,\mathbf{X}_{k_2+1}'(0) + \frac{\lambda_2}{\Lambda + \mu_{k_2}}\,\mathbf{X}_{k_2}'(0)\Delta, \tag{63}$$

$$X_{k_2-1}'(0) = -\frac{1}{\Lambda + \mu_{k_2-1}} + \frac{(k_2 - 2)\mu_{k_2-1}/(k_2 - 1)}{\Lambda + \mu_{k_2-1}}\,X_{k_2-2}'(0) + \frac{\Lambda}{\Lambda + \mu_{k_2-1}}\,\{\mathbf{X}_{k_2}'(0)\}_0, \tag{64}$$

$$X_i'(0) = -\frac{1}{\Lambda + \mu_i} + \frac{(i-1)\mu_i/i}{\Lambda + \mu_i}\,X_{i-1}'(0) + \frac{\Lambda}{\Lambda + \mu_i}\,X_{i+1}'(0) \quad \text{for} \quad 2 \le i < k_2 - 1, \tag{65}$$

and

$$X_1'(0) = -\frac{1}{\Lambda + \mu_1} + \frac{\Lambda}{\Lambda + \mu_1}\,X_2'(0). \tag{66}$$

## AUTHORS

**Kiran M. Rege**, B. Tech. (Electrical Engineering), 1977, I.I.T., Bombay; Ph.D. (Electrical Engineering), 1981, University of Hawaii; AT&T Bell Laboratories, 1982—. Mr. Rege spent 1984 on leave of absence, teaching at I.I.T. Bombay in the Department of Electrical Engineering. His technical work at AT&T Bell Laboratories includes modeling and analysis of switching, computer, and communication systems. His research interests include communication theory, queueing theory, and performance analysis of computer and communication systems.

**Bhaskar Sengupta**, B. Tech. (Electrical Engineering), 1965, I.I.T. Kharagpur, Eng. Sc.D. (Operations Research), 1976, Columbia University; AT&T Bell Laboratories, 1981—. Mr. Sengupta has worked in IBM and Service Bureau Company and was an Assistant Professor at the State University of New York at Stony Brook. He was also a consultant to Turner Construction Company in New York. At AT&T Bell Laboratories he works on performance problems for communication, computer, and manufacturing systems.