A Queueing System with General-Use and Limited-Use Servers

LINDA GREEN
Columbia University, New York, New York
(Received April 1983; accepted January 1984)

We consider a queueing system with two types of servers and two types of customers. General-use servers can provide service to either customer type while limited-use servers can be used only for one of the two. Though the apparent Markovian state space of this system is five-dimensional, we show that an aggregation results in an exact two-dimensional representation that is also Markovian. Matrix geometric theory is used to obtain approximations for the mean delay times and other measures of interest for each customer type. We illustrate the methodology by applying it to analyze a token discount policy used by the Triborough Bridge and Tunnel Authority.

MANY SERVICE facilities have two kinds of servers—general servers who can be used for any kind of customer and specialized servers who can provide service only to a specific subset of the customer population. A common example is a toll plaza with automatic exact change lanes as well as manned booths. Another example is a repair facility in which some of the technicians have limited expertise while the others can handle jobs of any difficulty.

These systems belong to a class of models that Schwartz [1974] called lane selecting (LS) models. These LS models are characterized by multiple customer types and a server hierarchy such that the higher the level, the more types of customers the server can handle. Schwartz studied systems with one server of each type, each with its own queue, and with a set of rules that governed which "lane" a given customer type would choose upon arrival. (Roque [1980] subsequently pointed out an error in this analysis.)

The queueing system studied in this paper is an LS type model with two levels of servers, two customer types, and one queue that all customers join upon arrival if an appropriate server is not available. In developing the methodology, we will assume an arbitrary number of each server type, although numerical solution will clearly limit the size of problems that can practically be solved. The general servers, called type G, can serve either customer type while the restricted-use servers, called type R, can be used only by what we will call the type R customers. The customers who *must* use a type G server will also be called type G. A type R customer can use either kind of server, but we will assume that he

Subject classification: 802 two levels of servers.

Operations Research
Vol. 33, No. 1, January–February 1985
0030-364X/85/3301-0168 $01.25

"prefers" a type R server (so if both types are available he will choose a type R).

Examples of this kind of queueing system are found in a variety of contexts. Many police precincts in New York City, for example, use two types of patrol cars. The usual patrol cars, called radio motor patrols, have two police officers and can respond to any kind of police emergency. The other cars, referred to as "specials," are often manned by only one officer and can respond to only low danger incidents such as past burglaries. Other kinds of emergency systems also have two levels of vehicles—some ambulances carry special cardiac-care devices for coronary victims, and some firetrucks have extra-long ladders for hi-rise fires. Such a queueing situation also arises in a bank that has tellers, who can perform any ordinary banking service, and machines, that can be used only for certain kinds of transactions such as check-cashing. Another commonplace example is a restaurant with booths or tables that can seat four people and smaller tables for two.

The service order discipline will be assumed to be first-come first-served (FCFS) with the exception that a type R customer will be allowed to pass a type G customer into service if a type G server is unavailable but a type R server is free. This discipline is reasonable in many single queue situations and, as we discuss in Section 4, can also be used to approximate a multiple queue situation.

The performance measures of most interest for this system are the expected delay and probability of delay for each customer type. The standard approach for obtaining these would be to try to derive the steady-state distribution of the number of each type of customer in the system. As we describe in Section 1, even under the assumption of Poisson arrivals and exponential service times, the state space necessary to obtain this distribution would be five-dimensional. The major objectives of this paper are to (i) define an aggregation of this five-dimensional state space into a two-dimensional Markovian state space from which the performance measures of interest can be calculated, and (ii) design a method for obtaining a good approximation for the steady-state distribution for this new state space. (This approach was also successfully used in Green [1982] to analyze another queueing system with two server types.)

In Section 1, we formulate the model as a bivariate Markov process. In Section 2, we show that approximations for the steady-state probabilities can be obtained using matrix-geometric theory introduced by Neuts [1978, 1980]. Section 3 gives numerical results, and illustrates the model's use for decision making.

1. MODEL DESCRIPTION

We consider a queueing system with m restricted-use servers and n general-use servers. Arrivals occur according to a Poisson process with

rate λ and all service times are exponentially distributed. Type G customers arrive at rate $\lambda_G = q\lambda$ and must be served by a type G server. Type R customers arrive at rate $\lambda_R = p\lambda$ and "prefer" receiving service from a type R customer; i.e., if servers of both types are idle at a type R arrival epoch, the customer enters service with a type R server. Service times are exponentially distributed with rate μ_G for type G servers and rate μ_R for type R servers. Customers of either type who arrive and find all servers busy wait in queue in the order of their arrival. In addition, a type G arrival must wait in line if a type R server is idle but all type G servers are busy; while a type R arrival in this case can enter service immediately even when the system has a queue (of type G customers). The service order discipline is FCFS except that in the case when a type R server frees and the first customer in queue is of type G, the first type R customer in line, if any, will enter service next. This operational policy decreases the delay of the type R customer (and hence subsequent customers) without increasing the delay of the type G customers ahead of him.

In order to obtain the steady-state distribution of the number of customers of each type in the system, it would be necessary to have state variables corresponding to the number of type R customers waiting in queue, the number of type G customers waiting in queue, the number of busy type R servers, the number of busy type G servers, and the number of type G customers who have been passed into service by a type P customer. This state space, of course, would lead to an intractable model.

Fortunately, there is an alternate formulation. Note that the system has two queues of waiting customers—one consisting of both type R and type G customers in FCFS order, and one consisting only of type G customers who have been passed by a type R customer. We will call these queues the *restricted* queue and the *general* queue, respectively. We define the rules of movement for customers as follows:

- All customers initially arrive to the restricted queue and enter service immediately if an appropriate server is available. Type G customers who arrive to an empty restricted queue and find a type R server available, but all type G servers busy, immediately move to the general queue.
- When a type G server becomes free, the first customer in the general queue is taken into service. If there is no general queue, the first customer in the restricted queue starts service.
- When a type R server becomes free and there is a restricted queue, (note that this alternative implies all type G servers are busy), the first customer in the restricted queue is examined:
  a. If that customer is a type R, he starts service at once;
  b. If that customer is a type G, he is instantly moved to the general

queue, the next customer in the restricted queue is then examined, and the process continues until either a type R customer is found or the queue is empty.

The system as described above can be represented as a bivariate Markov process with states (i, j), $i \geq 0, j \geq 0$, where i is the number of type R customers in service plus the number of customers (of either type) in the restricted queue, and j is the number of type G customers in service plus the number of customers in the general queue. Note, however, that since there can be no queue if a type G server is free, there are no states (i, j) where $i > m$ and $j < n$. To see that the process is Markovian, note that (i) for any state (i, j), the number of busy type R servers is given by $N_R = \min\{i, m\}$, and the number of busy type G servers is given by $N_G = \min\{j, n\}$, and (ii) the probability that any customer in the restricted queue is of a given type is just the probability that an arbitrary arrival is of this type. Thus, this two-dimensional state space contains all the information necessary to probabilistically describe the future of the system.

For any given starting state (i, j), the set of possible successor states and the associated formulae for the transition rates depend on the state of the overall system as follows:

- For states (i, j), $i < m$, $j < n$, all arrivals start service immediately with their associated server type. So transitions will be to state $(i + 1, j)$ with rate λ_R, to $(i, j + 1)$ with rate λ_G, to $(i - 1, j)$ with rate $i\mu_R$ (for $i > 0$), and to $(i, j - 1)$ with rate $j\mu_G$ (for $j > 0$).
- States $(m, j) j < n$ are those states in which all type R servers are busy, but at least 1 type G server is idle. Consequently, an arrival of either type will immediately start service with a type G server causing a transition to state $(m, j + 1)$ at rate λ. Transitions corresponding to departures take the system to state $(m - 1, j)$ and occur at rate $m\mu_R$, and transitions to state $(m, j - 1)$ occur at rate $j\mu_G$.
- For states (i, j), $i < m, j \geq n$, an arrival will cause a transition to state $(i + 1, j)$ with rate λ_R, or to state $(i, j + 1)$ with rate λ_G. These transition rates apply because a type R customer can immediately enter service with a type R server, while a type G arrival must join the general queue. A departure will cause a transition to state $(i - 1, j)$ with rate $i\mu_R$ (for $i > 0$), or to state $(i, j - 1)$ with rate $n\mu_G$.
- States (i, j), $i \geq m$, $j \geq n$, correspond to all servers busy. Since all arrivals join the restricted queue, transitions to state $(i + 1, j)$ are at rate λ. Departure transitions fall into three cases:

  Case 1. $i = m$. The system has no restricted queue and a departure from state (i, j) causes a transition to state $(i - 1, j)$ at rate $m\mu_R$ or to state $(i, j - 1)$ at rate $n\mu_G$.

  Case 2. $i > m$, $j = n$. The system has a restricted queue, but no

general queue. When a type G server becomes free, the first customer in the queue, regardless of type, will enter service and, therefore, move from the restricted system into the general system. Hence, the transition is to state $(i - 1, n)$ at rate $n\mu_G$. If a type R server becomes free and the first customer in queue is type R, the transition is again to state $(i - 1, n)$. The rate of the transition in this case will be $m\mu_R p$ and, therefore, the total rate of transition to state $(i - 1, n)$ is $m\mu_R p + n\mu_G$. Finally, if a type R server frees and the first customer in queue is type G the transition will be to state $(i - k - 1, j + k)$ where k is the number of consecutive type G's who are ahead of the first type R, if any, in queue. If there are no type R's in queue, $k = i - m$ is the queue length. The transition rate from (i, j) to $(i - k - 1, j + k)$ is $m\mu_R{}^k p$ for $k < i - m$ and $m\mu_R q^k$ for $k = i - m$.

Case 3. $i > m, j > n$. The system has both a general and a restricted queue. So when a type G server becomes free, the type G customer at the head of the general queue enters service according to the previously defined rules. The resulting transition is to state $(i, j - 1)$ at rate $n\mu_G$. Transitions that occur when a type R server becomes free are the same as in the previous case: to state $(i - 1, j)$ with rate $m\mu_R p$ and to state $(i - k - 1, j + k)$ at rate $m\mu_R q^k$ for $k < i - m$ and at rate $m\mu_R q^k$ for $k = i - m$.

Note that the rules of movement guarantee that at any arrival epoch, the time spent in the restricted queue will be identical for both customer types. So the expected total waiting time in queue for a type G customer is simply the sum of the expected time spent by an arbitrary customer in the restricted queue plus the expected wait in the general queue. Therefore, all of the usual performance measures of interest could be obtained from the steady-state distribution for this formulation of the model. However, the resulting balance equations are quite complex and there are no analytic nor numerical methods currently available for efficiently calculating exact solutions. In the next section, we describe an efficient methodology for obtaining approximate steady-state probabilities. Numerical results, described in Section 3, indicate that the approximation can yield accurate results for reasonably large systems.

Before proceeding with the development and solution of the approximate model, it is important to determine conditions for the existence of a steady-state solution for the actual model. Necessary conditions can be obtained by considering special cases. For example, if all customers are of type G, the system reduces to $M/M/n$ and, therefore, it is necessary that $\lambda_G < n\mu_G$. Similarly, if all customers are of type R, the service rate when all servers are busy is $m\mu_R + n\mu_G$, which is also the maximum departure rate when the system has both customer types. Therefore, we must have $\lambda < m\mu_R + n\mu_G$. Numerical results indicate that when these two conditions are met, a limiting distribution will exist.

2. APPROXIMATION OF THE STEADY-STATE DISTRIBUTION

The state space of the model described in the last section is infinite in both dimensions. In this section, we show how this system can be approximated by a two-dimensional Markov process with a finite second state variable, and with a steady-state distribution of the matrix-geometric form investigated by Neuts [1978]. This approximation allows us to develop a simple computational procedure for obtaining the steady-state probabilities.

The standard method of truncation would assume that there exists an integer $K > n$ such that at an arrival or departure epoch any type G customer who would cause a transition to state $(\cdot, K + 1)$ is instead "lost." This truncation would lead to a matrix-geometric model in which the maximum decrease in the first state variable is bounded only by the length of the restricted queue. Two computational difficulties would result:

(i) The number of customers simultaneously lost at a departure epoch could be as large as the number of customers in the restricted queue. Since lost customers do not contribute to the workload in the system, this loss could lead to significant errors unless the truncation parameter is quite large.
(ii) The matrix polynomial equation that must be solved would be of infinite degree. The solution could be obtained by successive substitutions. However, this procedure would require another truncation and introduce another source of error.

Instead, we assume there exists an integer $K > n$ such that at an epoch at which a type G customer would otherwise join the general queue and cause the second state variable to increase from K to $K + 1$, he changes his identity to type R, and, therefore, enters service with a type R server. This truncation method preserves the Markovian integrity of the model and overcomes both problems of the "lost" customer method:

(i) At any epoch, at most one customer will change identity. Furthermore, this customer still contributes to the workload of the system (though in a different way).
(ii) The number of customers who can simultaneously move from the restricted queue to the general queue will never be larger than $K - n$. Thus, the resulting matrix polynomial equation is finite.

The queueing model under consideration is represented by a continuous-time Markov process on the state space $\{(i, j): i > 0, 0 \leq j \leq K\}$. In general, its generator Q can be partitioned into blocks $H_i = \{(i, j), 0 \leq j \leq K\}$ and takes the following form when the states are in lexicographic order:

$$
\begin{array}{l}
i = 0 \\
\\
\\
\\
\\
\\
i = m \\
\\
\\
\\
\\
\\
i = m + K - n \\
\\
\\
\\
\end{array}
\begin{bmatrix}
B_{10} & B_{00} & & & & & & & & & & & \\
B_{21} & B_{11} & B_{00} & & & & & & & & & & \\
0 & B_{22} & B_{12} & B_{00} & & & & & & & & & \\
0 & 0 & B_{23} & B_{13} & & & & & & & & & \\
\cdot & \cdot & \cdot & \cdot & & & & & & & & & \\
\cdot & \cdot & \cdot & \cdot & & & & & & & & & \\
0 & 0 & 0 & B_{2m} & B_{1m} & A_0 & & & & & & & \\
0 & 0 & 0 & B_{3,m+1} & B_{2,m+1} & A_1 & A_0 & & & & & & \\
\cdot & \cdot & \cdot & B_{4,m+2} & B_{3,m+2} & A_2 & A_1 & A_0 & & & & & \\
\cdot & \cdot & \cdot & B_{5,m+3} & B_{4,m+3} & A_3 & A_2 & A_1 & A_0 & & & & \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & & & & \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & & & & \\
0 & 0 & 0 & B_{K-n+2,m+K-n} & B_{K-n+1,m+K-n} & A_{K-n} & \cdot & \cdot & \cdot & A_1 & A_0 & 0 & \cdot \\
\cdot & \cdot & \cdot & 0 & A_{K-n+2} & A_{K-n+1} & \cdot & \cdot & \cdot & \cdot & A_1 & A_0 & 0 \\
\cdot & \cdot & \cdot & 0 & 0 & A_{K-n+2} & \cdot & \cdot & \cdot & \cdot & \cdot & A_1 & A_0 \\
\cdot & \cdot & \cdot & 0 & 0 & 0 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
\end{bmatrix}
$$

The B blocks represent transitions from boundary states while the A blocks give the transition rates for nonboundary states. Boundary states for this model are defined as $\{(i, j),\ i \leq m + K - n\}$, though some nonboundary behavior is present beginning at $i = m$. The A blocks are $(K - n + 1) \times (K - n + 1)$ and A_k gives the transition rates from $(i, \cdot)$ to $(i - k + 1, \cdot)$, $k \geq 0$. B_{ki} is the array of transition rates from $(i, \cdot)$ to $(i - k + 1, \cdot)$ and has dimensions $(K + 1) \times (K + 1)$ for $i \leq m$ and $(K - n + 1) \times (K + 1)$ for $i > m$. This difference in array size is due to the fact that this system has no states (i, j), $i > m, j < n$. The B_{0i} matrices are identical for $i = 0, \cdots, m - 1$ and so B_{00} is used to indicate them all. Let $\mu_{kj} = k\mu_R + j\mu_G$ and $\alpha_k = pq^k$. Then for the case $m = 2$, $n = 2$ and $K = 5$, the blocks are defined as follows:

$$B_{00} = \begin{bmatrix} \lambda_R & & & & & \\ & \lambda_R & & & & \\ & & \lambda_R & & & \\ & & & \lambda_R & & \\ & & & & \lambda_R & \\ & & & & & \lambda \end{bmatrix}$$

$$B_{1i} = (i = 0, 1) \begin{bmatrix} -(\lambda + i\mu_R) & \lambda_G & & & & & \\ \mu_G & -(\lambda + \mu_{i1}) & \lambda_G & & & & \\ & 2\mu_G & -(\lambda + \mu_{i2}) & \lambda_G & & & \\ & & 2\mu_G & -(\lambda + \mu_{i2}) & \lambda_G & & \\ & & & 2\mu_G & -(\lambda + \mu_{i2}) & \lambda_G & \\ & & & & 2\mu_G & -(\lambda + \mu_{i2}) \end{bmatrix}$$

$$B_{12} = \begin{bmatrix} -(\lambda + 2\mu_R) & \lambda & 0 & \cdot & \cdot & \cdot \\ \mu_G & -(\lambda + \mu_{21}) & \lambda & 0 & \cdot & \cdot \\ 0 & 2\mu_G & -(\lambda + \mu_{22}) & 0 & 0 & \cdot \\ \cdot & 0 & 2\mu_G & -(\lambda + \mu_{22}) & 0 & 0 \\ \cdot & \cdot & 0 & 2\mu_G & -(\lambda + \mu_{22}) & 0 \\ \cdot & \cdot & \cdot & 0 & 2\mu_G & -(\lambda + \mu_{22}) \end{bmatrix}$$

$$B_{2i} = i\mu_R I \quad (i = 1, 2)$$

$$B_{23} = \begin{bmatrix} 0 & 0 & 2\mu_R p + 2\mu_G & 0 & \cdot & \cdot \\ \cdot & \cdot & 0 & 2\mu_R p & 0 & \cdot \\ \cdot & \cdot & \cdot & 0 & 2\mu_R p & 0 \\ \cdot & \cdot & \cdot & \cdot & 0 & 2\mu_R \end{bmatrix}$$

$$B_{33} = \begin{bmatrix} \cdot & \cdot & \cdot & 2\mu_R q & 0 & 0 \\ \cdot & \cdot & \cdot & 0 & 2\mu_R q & 0 \\ \cdot & \cdot & \cdot & \cdot & 0 & 2\mu_R q \\ \cdot & \cdot & \cdot & \cdot & \cdot & 0 \end{bmatrix}$$

$$B_{34} = \begin{bmatrix} \cdot & \cdot & 0 & 2\mu_R\alpha_1 & 0 & 0 \\ \cdot & \cdot & \cdot & 0 & 2\mu_R\alpha_1 & 0 \\ \cdot & \cdot & \cdot & \cdot & 0 & 2\mu_R q \\ \cdot & \cdot & \cdot & \cdot & \cdot & 0 \end{bmatrix}$$

$$B_{44} = \begin{bmatrix} \cdot & \cdot & \cdot & 0 & 2\mu_R q^2 & 0 \\ \cdot & \cdot & \cdot & \cdot & 0 & 2\mu_R q^2 \\ \cdot & \cdot & \cdot & \cdot & \cdot & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \end{bmatrix}$$

$$B_{45} = \begin{bmatrix} \cdot & \cdot & \cdot & 0 & 2\mu_R\alpha_2 & 0 \\ \cdot & \cdot & \cdot & \cdot & 0 & 2\mu_R q^2 \\ \cdot & \cdot & \cdot & \cdot & \cdot & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \end{bmatrix}$$

$$B_{55} = \begin{bmatrix} \cdot & \cdot & \cdot & \cdot & 0 & 2\mu_R q^3 \\ \cdot & \cdot & \cdot & \cdot & \cdot & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \end{bmatrix}.$$

$$A_0 = \lambda I$$

$$A_1 = \begin{bmatrix} -(\lambda + \mu_{22}) & 0 & \cdot & \cdot \\ 2\mu_G & -(\lambda + \mu_{22}) & 0 & \cdot \\ 0 & 2\mu_G & -(\lambda + \mu_{22}) & 0 \\ \cdot & 0 & 2\mu_G & -(\lambda + \mu_{22}) \end{bmatrix}$$

$$A_2 = \begin{bmatrix} 2\mu_R p + 2\mu_G & 0 & \cdot & \cdot \\ 0 & 2\mu_R p & 0 & \cdot \\ \cdot & 0 & 2\mu_R p & 0 \\ \cdot & \cdot & 0 & 2\mu_R \end{bmatrix}$$

$$A_3 = \begin{bmatrix} 0 & 2\mu_R\alpha_1 & 0 & \cdot \\ \cdot & 0 & 2\mu_R\alpha_1 & 0 \\ \cdot & \cdot & 0 & 2\mu_R q \\ \cdot & \cdot & \cdot & 0 \end{bmatrix}$$

$$A_4 = \begin{bmatrix} \cdot & \cdot & 2\mu_R\alpha_2 & 0 \\ \cdot & \cdot & 0 & 2\mu_R q^2 \\ \cdot & \cdot & \cdot & 0 \\ \cdot & \cdot & \cdot & \cdot \end{bmatrix} \qquad A_5 = \begin{bmatrix} \cdot & \cdot & 0 & 2\mu_R q^3 \\ \cdot & \cdot & \cdot & 0 \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \end{bmatrix}$$

Let matrix $A = \sum_{i=0}^{K-n+2} A_i$. In general, it is given by

$$A = \begin{bmatrix} -m\mu_R q & m\mu_R\alpha_1 & m\mu_R\alpha_2 & \cdot & \cdot & m\mu_R q^{K-n} \\ n\mu_G & -m\mu_R q - n\mu_G & m\mu_R\alpha_1 & \cdot & \cdot & m\mu_R q^{K-n-1} \\ 0 & n\mu_G & -m\mu_R q - n\mu_G & \cdot & \cdot & \cdot \\ 0 & 0 & n\mu_G & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ & & & & 0 & m\mu_R q \\ & & & & n\mu_G & -n\mu_G \end{bmatrix}$$

Neuts [1980] shows that Q is positive recurrent if

$$\underline{\pi} A_0 \underline{e} < \sum_{i=2}^{K-n+2} (i-1)\pi A_i \underline{e} \tag{1}$$

where $\underline{\pi}$ is the unique solution to

$$\underline{\pi} A = 0 \qquad \underline{\pi} e = 1. \tag{2}$$

In this case, the stationary probability vector $\underline{x} = [\underline{x}_0, \underline{x}_1, \cdots]$ of Q satisfies the matrix-geometric form

$$\underline{x}_i = \underline{x}_{i-1} R, \qquad i > m \tag{3}$$

where R is the minimal solution to

$$\sum_{i=0}^{K-n+2} R^i A_i = 0. \tag{4}$$

The vector $[\underline{x}_0, \underline{x}_1, \cdots, \underline{x}_m]$ is obtained by solving:

$$\begin{bmatrix} \underline{x}_0 \\ \underline{x}_1 \\ \cdot \\ \cdot \\ \cdot \\ \\ \\ \underline{x}_m \end{bmatrix}^T \begin{bmatrix} B_{10} & B_{00} & 0 & \cdot & \cdot & \cdot & \cdot \\ B_{21} & B_{11} & B_{00} & 0 & \cdot & \cdot & \cdot \\ 0 & B_{22} & B_{12} & B_{00} & \cdot & \cdot & \cdot \\ \cdot & & & \cdot & \cdot & \cdot & \cdot \\ \cdot & & & \cdot & \cdot & \cdot & \cdot \\ \cdot & & & \cdot & \cdot & \cdot & \cdot \\ & & & & \cdot & \cdot & \cdot \\ 0 & \cdot & \cdot & 0 & x & y & z \end{bmatrix} = \begin{bmatrix} \underline{0} \\ \underline{0} \\ \cdot \\ \cdot \\ \cdot \\ \\ \\ \underline{0} \end{bmatrix}^T \tag{5}$$

$$\underline{x}_0 e + \underline{x}_1 e + \cdots + \underline{x}_m (I - R)^{-1} e = 1,$$

where

$$x = \sum_{i=1}^{K-n} R^i B_{i+2,m+i},$$

$$y = \sum_{i=1}^{K-n} R^i B_{i+1,m+i} + \bar{R}^{K-n+1} \bar{A}_{K-n+2},$$

$$z = \sum_{i=1}^{K-n+2} R^{i-1} A_i;$$

($\bar{A}_i$ and $\bar{R}$ denote the A and R matrices "padded" with n columns of zeros to make the resulting dimension $(K - n + 1) \times (K + 1)$). R can be solved by iterative substitution.

3. NUMERICAL RESULTS AND EXAMPLES

Computer runs were performed on an IBM 4341, for varying proportions of restricted to general servers and restricted to general customers under several levels of overall system congestion. (We define system congestion here to be $\rho = \lambda/(m\mu_R + n\mu_G)$.) The total number of servers ranged from 6 to 10. One of the primary issues examined was the effect of the truncation parameter K on solution accuracy. This issue was studied by finding the minimum K, denoted by $\bar{K}$, necessary to obtain a specified level of numerical stability in the mean number of customers in each part of the system. In particular, let $L^{(R)}(K)$ be the mean number of customers in the restricted system as computed when the truncation parameter is K, and let $L^{(G)}(K)$ be the analogous measure for the general system, i.e.,

$$L^{(R)}(K) = \sum_{i=0}^{m} \sum_{j=0}^{K} ip_{ij} + \sum_{i=m+1}^{\infty} \sum_{j=n}^{K} ip_{ij}$$

$$L^{(G)}(K) = \sum_{i=0}^{m} \sum_{j=0}^{K} jp_{ij} + \sum_{i=m+1}^{\infty} \sum_{j=n}^{K} jp_{ij}$$

and define

$$K^{(R)} = \min_{K>m+1}\{K: |L^{(R)}(K) - L^{(R)}(K-1)|/L^{(R)}(K) < 0.02\}$$

$$K^{(G)} = \min_{K>m+1}\{K: |L^{(G)}(K) - L^{(G)}(K-1)|/L^{(G)}(K) < 0.02\}.$$

Then $\bar{K} =^{\text{def}} \max\{K^{(R)}, K^{(G)}\}$. (The calculation of the probability vector $\underline{x}_i$ is carried out until each element of $\underline{x}_i \leq 10^{-5}$.)

Table I illustrates how $\bar{K}$ fluctuates under varying system parameters. In the examples given, $\mu_R = \mu_G$ although the results are almost identical for $\mu_R = 2\mu_G$. The most significant observation is that although $\bar{K}$ tends to increase somewhat as the overall system congestion increases, it is most sensitive to the general customer traffic intensity $\rho_G = \lambda_G/n\mu_G$. For the cases examined, a $\bar{K}$ of 14 was found to be the maximum necessary truncation parameter whenever $\rho_G < 0.75$. This truncation value remained valid even when the overall system congestion ρ was 0.9. The CPU time involved for this size problem is approximately 10 seconds. This CPU time increases dramatically as K becomes larger—for $K = 18$, it is approximately 47 seconds and for $K = 22$, about 100 seconds. Table I also shows p_K, the probability that the waiting room is full, for each $\bar{K}$. Since the approximation affects only transitions from this state, the small magnitude of these numbers confirms that the $\bar{K}$'s are large enough to keep the approximation errors small.

As an illustration of how the model could be used for decision making,

we studied a situation involving toll booths on the section of the Triborough Bridge which connects the boroughs of Queens and the Bronx in New York City. Although a toll plaza is a multiple queue situation, the model presented in this paper can be used as an approximation since the service order discipline does not permit the blocking of restricted customers (in this case, cars with exact change or tokens) by general customers (cars which need change) when a restricted server (automatic booth) is available. The use of a single queue model to approximate a multiple queue system is not uncommon and was, in fact, the method used by Edie [1954] in his classical study of toll booth delays at the Lincoln Tunnel. The resulting delays will, of course, underestimate the

TABLE I

$\bar{K}$ FOR VARIOUS SYSTEM CONFIGURATIONS ($n = 5$ TYPE G SERVERS)

p	$m = 1$			$m = 3$			$m = 5$		
	ρ_G	$\bar{K}$	p_K	ρ_G	$\bar{K}$	p_K	ρ_G	$\bar{K}$	p_K
$\rho = 0.4$									
0.2	0.38	8	0.0016	0.51	9	0.0047	0.64	11	0.0072
0.4	0.29	8	0.0004	0.38	8	0.0021	0.48	8	0.0067
0.6	0.19	8	0.0001	0.26	8	0.0003	0.32	8	0.0008
0.8	0.10	8	0.0000	0.13	7	0.0000	0.16	8	0.0000
$\rho = 0.6$									
0.2	0.58	11	0.0027	0.77	14	0.0110	0.96	>25	
0.4	0.43	8	0.0049	0.57	9	0.0120	0.72	13	0.0088
0.6	0.29	8	0.0013	0.38	8	0.0051	0.48	8	0.0115
0.8	0.14	8	0.0001	0.19	8	0.0006	0.24	8	0.0011
$\rho = 0.8$									
0.2	0.77	13	0.0110	Unstable			Unstable		
0.4	0.57	8	0.0191	0.77	14	0.0129	0.96	>25	
0.6	0.38	8	0.0058	0.51	8	0.9246	0.64	11	0.0129
0.8	0.19	8	0.0007	0.26	8	0.0033	0.32	8	0.0075

delays of an equivalent multiple queue system where customers may be waiting in one lane while a server in another lane is available. (Edie's study found, however, that the error due to this underestimation compensated for the overestimation due to the assumption of exponential service times for toll booth transactions which are, in reality, more nearly deterministic.)

This toll plaza of the Triborough Bridge has 8 toll booths in each direction, 4 of which are automatic exact change lanes that can be used as manual lanes, if necessary. Since automatic booths are cheaper to operate, and the use of exact change reduces transaction times, the Triborough Bridge and Tunnel Authority (TBTA) sells tokens to customers at a discount to encourage their use. Thus, the TBTA is interested in determining the tradeoffs between the size of the discount and the

number of automatic booths that can be effectively used. That is, when traffic is heavy, they would like to be able to operate all 4 of the exact change lanes automatically and have p, the proportion of cars using tokens, large enough so that disproportionately long queues do not build up in the manual lanes.

To determine what this discount should be, we first must determine the effect of the resulting p on system performance. We used the model to obtain queueing statistics for several situations, using an arrival rate to reflect a moderate rush hour. The mean transaction time for manual lanes ($1/\mu_G$) is approximately 10 seconds, and for automatic lanes ($1/\mu_R$) is approximately 8 seconds. λ was chosen so that the traffic intensity of the system operating with all manual lanes would be $\hat{\rho} = \lambda/8\mu_G = 0.95$.

Using this level of system congestion, we looked at the effects of operating 3 of the 4 exact change booths automatically and 1 manually versus all 4 automatically, assuming a p value of either 0.52 or 0.75. (For the system with 4 exact change lanes, the system is unstable for $p < 0.48$.) For each set of system parameters, we computed the expected number of cars in queue, the overall expected delay, and the expected delays for each customer type, as well as the blocking probabilities for each customer type and the probability α that a car needing change is delayed while an automatic booth is available. This probability, given by $\alpha = \sum_{i<m} \sum_{j>n} p_{ij}$, is a good measure for assessing whether there is an adequate number of manual booths for the proportion of customers who require them.

The steady-state number of customers waiting in queue is given by

$$L_q = L_q^{(R)} + L_q^{(G)}$$

where

$$L_q^{(R)} = \sum_{i=m+1}^{\infty} \sum_{j=n}^{K} (i - m)p_{ij}$$

is the steady-state number of customers in the "restricted queue" and

$$L_q^{(G)} = \sum_{i=0}^{\infty} \sum_{j=n+1}^{K} (j - n)p_{ij}$$

is the number in the "general queue."

By the rules of movement, we see that the amount of time spent waiting in the restricted queue is identical for both customer types and so the arrival rate to this queue is λ. Therefore, the mean delay for the cars with exact change is given by $W_q^{(R)} = L_q^{(R)}/\lambda$ and for cars requiring change is given by $W_q^{(G)} = L_q^{(R)}/\lambda + L_q^{(G)}/\lambda_G$.

The probability that a type R customer has a positive delay is the probability that all servers are busy and is given by $p_B^{(R)} = \sum_{i \geq m} \sum_{j \geq n} p_{ij}$. The corresponding probability for a type G customer is the probability that all type G servers are busy and is given by $p_B^{(G)} = \sum_{i=0}^{\infty} \sum_{j=n}^{K} p_{ij}$.

The results are shown in Table II. We observe that when only 52% of

cars have exact change, using all 4 automatic lanes results in significantly worse performance than using one of them as a manual booth. This result changes, however, when p increases to 0.75. With this proportion of exact change customers, operating all 4 exact change booths automatically results in improved system performance. Thus, for all 4 exact change booths to be used effectively, the discount would have to be large enough to result in close to 75% of cars having exact change.

Another interesting observation is that when only 3 lanes are being operated automatically, and the proportion of cars with exact change increases from 0.52 to 0.75, the overall system performance does not

TABLE II

COMPARISON OF SYSTEM PERFORMANCE FOR 8 LANES USING 3 OR 4 AUTOMATIC BOOTHS

	3 Automatic Lanes	4 Automatic Lanes
$p = 0.52$		
E (No. cars in queue)	5.13	11.09
E (overall delay)	6.75 sec	14.59 sec
$W_q^{(G)}$	8.62	26.53
$W_q^{(R)}$	4.97	3.52
$p_B^{(G)}$	0.77	0.83
$p_B^{(R)}$	0.58	0.48
α	0.14	0.32
$p = 0.75$		
E (No. cars in queue)	4.21	3.31
E (overall delay)	5.53	4.36
$W_q^{(G)}$	6.35	5.94
$W_q^{(R)}$	5.25	3.83
$p_B^{(G)}$	0.68	0.68
$p_B^{(R)}$	0.61	0.54
α	0.03	0.07

improve very much. This result, at first glance, is counterintuitive. However, it becomes more understandable if one considers that the exact change customers have smaller delays when there is a larger percentage of cars without exact change since these other customers do not block them. Thus, increasing the percentage of exact change cars when there are not "enough" automatic lanes causes slightly greater delays for the exact change customers, somewhat smaller delays for the other cars, with the total effect resulting in only a small overall improvement.

This example, as well as others we have looked at, demonstrates the complexity of the dynamics in this type of queueing system and the resulting difficulty in predicting the effects of proposed policies. Thus, it appears to be particularly important to use a model that captures the

essential characteristics of such a system. For this reason, the model presented in this paper should be a valuable tool for the analysis of these systems.

ACKNOWLEDGMENT

I am very grateful to Debashis Guha for his programming assistance. I also thank Peter Kolesar for his helpful suggestions.

REFERENCES

EDIE, L. 1954. Traffic Delays at Toll Booths. *Opns. Res.* **2,** 107–138.

GREEN, L. 1984. A Queueing System with Auxiliary Servers. *Mgmt. Sci.* **30,** 1207–1216.

NEUTS, M. F. 1978. Markov Chains with Applications in Queueing Theory which Have a Matrix-Geometric Invariant Probability Vector. *Adv. Appl. Prob.* **10,** 185–212.

NEUTS, M. F. 1980. *Matrix-Geometric Solutions in Stochastic Models.* Johns Hopkins University Press, Baltimore.

ROQUE, D. R. 1980. A Note on "Queueing Models with Lane Selection." *Opns. Res.* **28,** 419–420.

SCHWARTZ, B. 1974. Queueing Models with Lane Selection: A New Class of Problems. *Opns. Res.* **22,** 331–339.