MANAGEMENT SCIENCE
Vol. 30, No. 10, October 1984
Printed in U.S.A.

A QUEUEING SYSTEM WITH AUXILIARY SERVERS*

LINDA GREEN
Graduate School of Business, Columbia University, New York, New York 10027

We examine a queueing system with multiple primary servers and a fewer number of auxiliary servers. There are two classes of customers—those who require service from a primary server working alone and those who require service from a primary server who is assisted by an auxiliary server. Though the apparent Markovian state space is five-dimensional, we show that an aggregation results in an exact two-dimensional representation which is Markovian. Matrix geometric theory is used to obtain approximations for the mean delay and blocking probability of each customer type.
(MULTIPLE SERVER QUEUES; DISTINGUISHABLE CUSTOMERS)

Introduction

Many service systems have auxiliary servers that work in conjunction with the primary servers of the system on certain kinds of jobs. For example, New York City is currently considering implementation of a one-officer patrol car program in which a two-officer car would be dispatched concurrently with a one-officer car to emergency calls, while nonemergency calls would be "served" by a one-officer car alone (*New York Times* October 15, 1982). In many emergency medical systems, a parallel situation exists where a paramedic unit is dispatched simultaneously with an ambulance to provide on-the-scene medical assistance in life-threatening incidents. Another public sector example arises in the arraignment process of urban criminal court systems where some of the cases that must appear before the judge for initial disposition also require the presence of a court interpreter. This type of simultaneous multi-resource requirement is also prevalent in computer systems (see e.g. Jacobson and Lazowska 1980, Omahen 1977) where it often appears as part of a queueing network.

This paper deals with a queueing system with two types of servers and two types of customers. The "basic" servers of the system will be called *primary* servers. One primary server is assigned to each customer. For some specified proportion of arrivals, an *auxiliary* server will be necessary to assist the primary server throughout the job. In these cases, we assume that the customer cannot enter service unless both a primary and auxiliary server are available. If the number of auxiliary servers equals or exceeds the number of primary servers, there is no potential for blockage of the system due to the unavailability of an auxiliary server, and the system behaves like an ordinary multiple server system. Therefore we assume that there are fewer auxiliary servers than primary servers so that the primary servers "share" the auxiliary servers and there is a positive probability of customer delays when a primary server is free.

A more general system of this type was studied by Omahen (1977). He examined a model of a computer system with a fixed but arbitrary number of server types and customers who request simultaneous service from a random combination of them. The objective was to develop an algorithm for finding the smallest total arrival rate which would guarantee saturation of the system regardless of the service order discipline. Willemain (1974) obtained approximate results for several performance measures for a system which also has primary and secondary servers, but arranged in a hierarchical structure. In his model, customers arrive to independent primary service stations which provide initial service and, with a given probability, require additional service (a consultation) from a single secondary server after the initial service is completed. The model is similar to the one examined here in that consultations require both the primary and secondary servers. However, it differs in its assumptions of parallel arrival streams, initial service time, and a single secondary server. An approximation technique for obtaining delays in a similar system in the context of a queueing network was developed by Jacobson and Lazowska (1980). The basic approach is iteration between two simpler models, each of which represents one set of resources (primary or secondary) explicitly and the other as an infinite server.

The system under study is a member of the class of queues in which customers require a random number of servers. See Brill and Green (1984), Green (1980, 1981) and Kim (1979) for results on queues in this class with all servers identical. It shares with the other members of the class the characteristic that servers may be idle even when there are customers waiting to be served. The dynamics here are further complicated by the existence of two server types.

*Accepted by John P. Lehoczky; received November 19, 1982. This paper has been with the author 4 months for 1 revision.

0025-1909/84/3010/1207$01.25

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 §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: (1) define an aggregation of this five-dimensional state space that results in a two-dimensional Markovian state space from which the performance measures of interest can be calculated, and (2) design a method for obtaining a good approximation for the steady-state distribution of this new state space. (This approach has also been successful in the analysis of another queueing system with two server types, Green to appear.)

In §1, we give a precise definition of the model and its formulation as a bivariate Markov process. In §2, we show that approximations for the steady-state probabilities can be obtained using matrix-geometric theory introduced by Neuts (1978, 1980). Numerical results, given in §3, show that this method can be used to generate accurate queueing statistics for reasonably large systems.

1. Model Description

We consider a queueing system with r primary servers and $s < r$ auxiliary servers. We assume that arrivals occur according to a Poisson process at rate λ and all service times are exponentially distributed. Type P customers arrive at rate $\lambda_p = p\lambda$ and require a service time with mean $1/\mu_p$ from a primary server working alone. Type A customers arrive at rate $\lambda_A = q\lambda$, $q = 1 - p$, and require a service time with mean $1/\mu_A$ from a primary and auxiliary working together (i.e. the two servers begin and end service at the same moment). Therefore, while both types of customer will be blocked from entering service when all primary servers are busy, a Type A customer will also be blocked when a primary server is idle but all auxiliary servers are busy.

We assume that a customer's type is known at the time of his arrival. Therefore, it seems reasonable to assume a service order discipline which is FIFO except in the following case: when all auxiliary servers are busy, but there is a primary server available, the first Type P customer in queue, if any, will enter service, even if there are Type A customers ahead of him. This decreases the delay of the Type P customer without increasing the delay of the Type A customers in front of him. This is true because the next Type A customer in queue will still enter service when an auxiliary server next becomes free since a primary server also becomes free at that time.

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 P customers waiting in queue, the number of Type A customers waiting in queue, the number of busy primary servers, the number of busy auxiliary servers, and the number of Type A customers who have been passed into service by a Type P customer. This, of course, would lead to an intractable model.

Fortunately, there is an alternate formulation. Note that the system has two queues of waiting customers—the one consisting of both Type P and Type A customers in FIFO order, and the one that consists only of Type A customers who have been passed by a Type P customer. We will call these queues the *primary* queue and the *auxiliary* queue, respectively. We define the rules of movement for customers as follows:

○ All customers who do not find the appropriate server(s) available at their arrival epoch initially join the primary queue. Type A customers who arrive to an empty primary queue and find a primary server available, but all auxiliary servers busy, immediately move to the auxiliary queue.

○ When both a primary server and auxiliary server become free, the first customer in the auxiliary queue is taken into service. If there is no auxiliary queue, the first customer in the primary queue starts service.

○ When a primary server becomes free but all auxiliary servers are busy, the first customer in the primary queue is examined:

—if that customer is Type P, he starts service at once;

—if that customer is Type A, he is instantly moved to the auxiliary queue, the next customer in the primary queue is then examined and the process continues until either a Type P 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), $i \geqslant 0, j \geqslant 0$, where i is the number of Type P customers in service plus the number of customers (of either type) in the primary queue, and j is the number of Type A customers in service plus the number of customers in the auxiliary queue. To see this, note that: (1) for any state (i, j), the number of busy auxiliary servers is given by $N_A = \min\{j, s\}$, and the number of busy primary servers is given by $N_p = \min\{i + N_A, r\}$, and (2) the probability that any customer in the primary 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), $i + j < r$, an arrival will cause a transition to $(i + 1, j)$ with rate λ_P, or to $(i, j + 1)$ with rate λ_A. This is because when there is an idle primary server, a Type P arrival starts service immediately while a Type A arrival will either start service (if an auxiliary server is free) or join the auxiliary queue. A departure will cause a transition to: state $(i - 1, j)$ with rate $i\mu_p$ (for $i > 0$); or to $(i, j - 1)$ with rate $j\mu_A$ if $0 < j < s$, and with rate $s\mu_A$ if $j \geqslant s$.

—States (i, j), $i + j \geqslant r$, $j < s$, are those states for which all primary servers are busy, but at least one auxiliary server is free. Since all arrivals join the primary queue, transitions from (i, j) to $(i + 1, j)$ occur at rate λ. If there is no primary queue at a departure epoch, i.e., $i + j = r$, the departure will cause a transition to $(i - 1, j)$ with rate $i\mu_p$ or to $(i, j - 1)$, $j > 0$ with rate $j\mu_A$. When $i + j > r$, the transition at a departure epoch will depend on the type of customer who is first in queue as well as the type of the customer who is completing service. This is because if a Type A is first in queue and therefore next to enter service, upon starting service he leaves the primary system and joins the auxiliary system by definition of the state space. Thus, the departure transitions from (i, j), $i + j > r$ are to: $(i - 1, j)$ with rate $(r - j)\mu_P p + j\mu_A q$; $(i - 2, j + 1)$ with rate $(r - j)\mu_P q$; and $(i, j - 1)$ with rate $j\mu_A p$.

—For states (i, j), $i + j \geqslant r$, $j \geqslant s$, all servers are busy. Again, all arrivals join the primary queue and so transitions to $(i + 1, j)$ are at rate λ. Departure transitions can be considered as falling into 3 cases:

Case 1. $i = r - s$, $j = s$. No primary queue exists and a departure from state (i, j) causes a transition to $(i - 1, j)$ at rate $(r - s)\mu_P$ or to $(i, j - 1)$ at rate $s\mu_A$.

Case 2. $i > r - s$, $j = s$. There is no auxiliary queue but there is a primary queue. So when a primary-auxiliary server pair frees (i.e. a Type A customer departs), the transition is to $(i, j - 1)$ at rate $s\mu_A p$ or to $(i - 1, j)$ at rate $s\mu_A q$. If a primary server alone frees up and the first customer in queue is Type P, the transition is also to $(i - 1, j)$ with rate $(r - s)\mu_P p$. So the total transition rate to $(i - 1, j)$ is $(r - s)\mu_P p + s\mu_A q$. Finally, if a primary server alone becomes free and the first customer in queue is Type A, the transition will be to $(i - k - 1, j + k)$ where k is the number of consecutive Type A's in queue who are in front of the first Type P in queue, if any. If there are no Type P customers in queue, $k = i - r + s$ is the queue length. The transition rate from (i, j) to $(i - k - 1, j + k)$ is $(r - s)\mu_P q^k p$ for $k < i - r + s$ and $(r - s)\mu_P q^k$ for $k = i - r + s$.

Case 3. $i > r - s$, $j > s$. There is both a primary and an auxiliary queue. If a primary-auxiliary pair frees, the transition is to $(i, j-1)$ at rate $s\mu_A$. If a primary server alone becomes free, the transition is to $(i-k-1, j+k)$ where k is the number of consecutive Type A customers before the first Type P customer in the primary queue, if any. The transition rates are as in the corresponding situation in Case 2.

Note that the rules of movement guarantee that at any arrival epoch the time spent in the primary queue will be identical for both customer types. So the expected total waiting time in queue for a Type A customer is simply the sum of the expected time spent by an arbitrary customer in the primary queue plus the expected wait in the auxiliary 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 or numerical methods currently available for the efficient calculation of exact solutions. In the next section, we describe an efficient methodology for obtaining approximate steady-state probabilities. Numerical results, described in §3, indicate that the approximation can be accomplished so as to 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 examining special cases. For example, if all customers are of Type A, the system reduces to $M/M/s$ and therefore we would need to insure that $\lambda_A < s\mu_A$. Similarly, if all customers were of Type P, it would be necessary that $\lambda_p < r\mu_p$. Given customers of both types, the total departure rate when there is a primary queue can vary between $r\mu_p$ and $(r-s)\mu_p + s\mu_A$. Therefore, it is also necessary that λ not exceed the maximum of these two quantities. Numerical results indicate that if $\lambda_A < s\mu_A$ and $\lambda < \min\{(r-s)\mu_P + s\mu_A, r\mu_P\}$, 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 in which the second state variable is finite, and for which the steady-state distribution has the matrix-geometric form investigated by Neuts (1978, 1980). This allows for a simple computational procedure for obtaining the steady-state probabilities.

The standard method of truncation would be to assume that there exists an integer $K > s$ such that at an arrival or departure epoch at which a Type A customer would otherwise cause a transition to $(\cdot, K+1)$, he is instead "lost." This 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 primary queue. Two computational difficulties would result:

(1) The number of customers simultaneously lost at a departure epoch could be as large as the number of customers in the primary queue. Since lost customers do not contribute to the workload in the system, this could lead to significant errors unless the truncation parameter is quite large.

(2) The matrix polynomial equation that must be solved would be of infinite degree. The solution could be obtained by successive substitutions. However, this would require another truncation, introducing another source of error.

Instead we assume there exists an integer $K > s$ such that at an epoch at which a Type A customer would otherwise join the auxiliary queue and cause the second state variable to increase from K to $K+1$, he changes his identity to Type P. This

$$
\begin{array}{l|cccccccccccccccccccccc|}
i=0 & B_{10} & B_{00} & 0 & \cdot \\
 & B_{21} & B_{11} & B_{00} & 0 & \cdot \\
 & 0 & B_{22} & B_{12} & B_{00} & & \cdot \\
 & \cdot & \cdot & \cdot & \cdot & & & \cdot \\
 & & & & & B_{2,r-s-1} & B_{1,r-s-1} & B_{00} & 0 \\
 & & & & & 0 & B_{2,r-s} & B_{1,r-s} & B_{0,r-s} \\
 & & & & & \cdot & B_{3,r-s+1} & B_{2,r-s+1} & B_{1,r-s+1} & B_{0,r-s+1} \\
 & & & & & \cdot & B_{4,r-s+2} & \cdot & \cdot & \cdot \\
 & & & & & \cdot & \cdot & \cdot & \cdot & \cdot \\
 & & & & & \cdot & \cdot & \cdot & \cdot & \cdot \\
i=r & & & & & 0 & B_{s+2,r} & \cdot & \cdot & B_{3,r} & B_{2,r} & A_1 & A_0 & 0 \\
 & & & & & & \cdot & \cdot & \cdot & \cdot & B_{3,r+1} & A_2 & A_1 & A_0 & 0 \\
 & & & & & & \cdot & \cdot & \cdot & \cdot & B_{4,r+2} & A_3 & A_2 & A_1 & A_0 & 0 \\
 & & & & & & \cdot & \cdot & \cdot & \cdot & B_{5,r+3} & A_4 & A_3 & A_2 & A_1 & A_0 & 0 \\
 & & & & & & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
 & & & & & & B_{K-s+2,r-2s+K} & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & A_1 & A_0 & 0 \\
i=r-2s+K+1 & & & & & & 0 & A_{K-s+2} & A_{K-s+1} & A_{K-s} & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & A_2 & A_1 & A_0 & 0 \\
 & & & & & & & 0 & A_{K-s+2} & A_{K-s+1} & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & A_3 & A_2 & A_1 & A_0 & 0 & \cdot \\
 & & & & & & & & 0 & A_{K-s+2} & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & A_4 & A_3 & A_2 & A_1 & A_0 & \cdot \\
 & & & & & & & & & & & & & & & & & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot
\end{array}
$$

truncation method overcomes both problems of the "lost" customer method:

(1) 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).

(2) The number of customers who can simultaneously move from the primary queue to the auxiliary queue will never be larger than $K - s$. 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 $\{(i,j): i \geqslant 0, 0 \leqslant j \leqslant K\}$. The generator Q of this process can be partitioned into blocks $H_i = \{(i,j), 0 \leqslant j \leqslant K\}$ and takes the following form when the states are in lexicographic order shown on p. 1211.

The B blocks represent transitions from boundary states which are defined for this model as $\{(i,j): i \leqslant r - 2s + K\}$. Transition rates from the nonboundary states are given by the A blocks and are identical for all $i > r - 2s + K$. Note that some nonboundary behavior is present beginning at $i = r$. A_m is the array of transition rates from $(i, \cdot)$ to $(i - m + 1, \cdot)$, $m \geqslant 0$ while B_{mn} gives the transition rates from $(n, \cdot)$ to $(n - m + 1, \cdot)$. (The B_{0i} matrices are identical for $i = 0, \ldots, r - s - 1$ and so B_{00} is used to indicate them all.) Let $\mu_{mn} = m\mu_P + n\mu_A$ and $\alpha_i = pq^i$. Then, e.g., for the case $r = 3$, $s = 1$ and $K = 4$, the blocks are defined as follows:

$$B_{00} = \begin{bmatrix} \lambda_P &&&& \\ & \lambda_P &&& \\ && \lambda_P && \\ &&& \lambda_P & \\ &&&& \lambda \end{bmatrix}, \qquad B_{02} = \begin{bmatrix} \lambda_P &&&& \\ & \lambda &&& \\ && \lambda && \\ &&& \lambda & \\ &&&& \lambda \end{bmatrix},$$

$$\underset{i=0,1,2}{B_{1i}} = \begin{bmatrix} -(\lambda + i\mu_P) & \lambda_A & 0 & 0 & \cdot \\ \mu_A & -(\lambda + \mu_{i1}) & \lambda_A & 0 & \cdot \\ 0 & \mu_A & -(\lambda + \mu_{i1}) & \lambda_A & 0 \\ \cdot & 0 & \mu_A & -(\lambda + \mu_{i1}) & \lambda_A \\ \cdot & & 0 & \mu_A & -(\lambda + \mu_{i1}) \end{bmatrix}, \qquad B_{2i} = i\mu_P I \qquad i = 1, 2,$$

$$B_{33} = \begin{bmatrix} 0 & 0 & 0 && \\ & 0 & 2\mu_P q & 0 & \\ && 0 & 2\mu_P q & 0 \\ &&& 0 & 2\mu_P q \\ &&&& 0 \end{bmatrix}, \qquad B_{34} = \begin{bmatrix} 0 & 3\mu_P q &&& \\ & 0 & 2\mu_P \alpha_1 && \\ && 0 & 2\mu_P \alpha_1 & \\ &&& 0 & 2\mu_P q \\ &&&& 0 \end{bmatrix},$$

$$B_{44} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ &&& 2\mu_P q^2 & 0 \\ &&& 0 & 2\mu_P q^2 \\ &&&& 0 \\ &&&& 0 \end{bmatrix}, \qquad B_{45} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ && 0 & 2\mu_P \alpha_2 & 0 \\ &&& 0 & 2\mu_P q^2 \\ &&&& 0 \\ &&&& 0 \end{bmatrix},$$

$$B_{55} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ &&& 0 & 2\mu_P q^3 \\ &&&& 0 \\ &&&& 0 \\ &&&& 0 \end{bmatrix}, \qquad B_{56} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ &&&& 2\mu_P q^3 \\ &&&& 0 \\ &&&& 0 \\ &&&& 0 \end{bmatrix},$$

$$A_0 = \lambda I,$$

$$A_1 = \begin{bmatrix} -(\lambda + 3\mu_P) &&&& \\ \mu_A p & -(\lambda + \mu_{21}) &&& \\ & \mu_A p & -(\lambda + \mu_{21}) && \\ && \mu_A & -(\lambda + \mu_{21}) & \\ &&& \mu_A & -(\lambda + \mu_{21}) \end{bmatrix},$$

$$A_2 = \begin{bmatrix} 3\mu_P p & & & & \\ & 2\mu_P p + \mu_A q & & & \\ & & 2\mu_P p & & \\ & & & 2\mu_P p & \\ & & & & 2\mu_P \end{bmatrix}, \quad A_3 = \begin{bmatrix} 0 & 3\mu_P q & & & \\ & & 2\mu_P \alpha_1 & & \\ & & & 2\mu_P \alpha_1 & \\ & & & & 2\mu_P \alpha_1 \\ & & & & 0 \end{bmatrix},$$

$$A_4 = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ & & 0 & 2\mu_P \alpha_2 & 0 \\ & & & 0 & 2\mu_P q^2 \\ & & & & 0 \\ & & & & 0 \end{bmatrix}, \quad A_5 = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ & & & 0 & 2\mu_P q^3 \\ & & & & 0 \\ & & & & 0 \\ & & & & 0 \end{bmatrix}.$$

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

$$A = \begin{bmatrix} -r\mu_P q & r\mu_P q & 0 & & & & & \\ \mu_A p & a_{r-1,1} & (r-1)\mu_P q & 0 & & & & \\ 0 & 2\mu_A p & a_{r-2,2} & (r-2)\mu_P q & & & & \\ \vdots & \vdots & \vdots & \vdots & & & & \\ & & & s\mu_A & b_{r-s,s} & (r-s)\mu_P \alpha_1 & (r-s)\mu_P \alpha_2 \cdots & (r-s)\mu_P q^{K-s} \end{bmatrix}$$

where $a_{ij} = -i\mu_P q - j\mu_A p$, $b_{ij} = -i\mu_P q - j\mu_A$.

From Neuts (1980), Q is positive recurrent if

$$\boldsymbol{\pi} A_0 \mathbf{e} \leqslant \sum_{i=2}^{K-s+2} (i-1) \boldsymbol{\pi} A_i \mathbf{e} \tag{1}$$

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

$$\boldsymbol{\pi} A = 0, \qquad \boldsymbol{\pi} \mathbf{e} = 1. \tag{2}$$

In this case, the stationary probability vector $\mathbf{x} = [\mathbf{x}_0, \mathbf{x}_1, \mathbf{x}_2, \ldots]$ of Q exists and satisfies the matrix-geometric form

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

where R is the minimal solution to

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

The vector $[\mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_r]$ is obtained by solving the equations

$$\begin{bmatrix} \mathbf{x}_0, & \mathbf{x}_1, & \vdots & \mathbf{x}_r \end{bmatrix}^T$$

$$\begin{bmatrix} B_{10} & B_{00} & 0 & \cdot & & & & & & \cdot & \\ B_{21} & B_{11} & B_{00} & 0 & \cdot & & & & & \cdot & \\ 0 & B_{22} & B_{12} & B_{00} & \cdot & & & & & \cdot & \\ & & & \cdot & & & & & & & \\ & & & & \sum_{i=0}^{K'} R^i B_{i'+2,j'} & \sum_{i=0}^{K'} R^i B_{i'+1,j'} + R^{K'+1} A_{K-s+2} & \cdots & \sum_{i=0}^{K'} R^i B_{2,j'} + & \sum_{i=K'+1}^{K-s-r} R^i A_{j'+2} & \sum_{i=0}^{K-s+2} R^i A_{i+1} \end{bmatrix}$$

$$= [\mathbf{0} \quad \mathbf{0} \quad \cdot \quad \cdot \quad \cdot \quad \mathbf{0}]^T, \tag{5}$$

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

where $i' = s + i$, $j' = r + i$, $K' = K - 2s$. R can be solved by iterative substitution.

3. Numerical Results

Computer runs were performed for varying levels of system congestion and for varying proportions of auxiliary to primary servers. For simplicity, we let $\mu_P = \mu_A$. The first issue that we examined was the effect of the truncation parameter K on solution accuracy. This was done by finding the minimum K necessary to obtain a specified level of numerical stability in the mean number of customers in each part of the system. In particular, if we let $L^{(P)}(N)$ be the mean number of customers in the primary part of the system as computed when the truncation parameter is N and $L^{(A)}(N)$ be the analogous measure for the auxiliary queue, i.e.

$$L^{(P)}(N) = \sum_{i=0}^{\infty} \sum_{j=0}^{N} i x_{ij}, \qquad L^{(A)}(N) = \sum_{i=0}^{\infty} \sum_{j=0}^{N} j x_{ij} \quad \text{then} \tag{6}$$

$$\begin{aligned} K^{(P)} &= \min_{N>s} \left\{ N: \frac{|L^{(P)}(N+1) - L^{(P)}(N)|}{L(N)} \right\} < 0.02, \\ K^{(A)} &= \min_{N>s} \left\{ N: \frac{|L^{(A)}(N+1) - L^{(A)}(N)|}{L^{(A)}(N)} \right\} < 0.02, \quad \text{and} \quad K = \max\{K^{(P)}, K^{(A)}\}. \end{aligned} \tag{7}$$

The results show that the single most important factor that determines K is $\rho_A = \lambda_A / s\mu$, the traffic intensity of Type A customers. As expected, K increases as ρ_A increases. Figure 1 illustrates this relationship for $r = 5$, $s = 2$. In addition, for fixed ρ_A, K generally remains constant for various choices of r, s and overall traffic intensity ($\rho = \lambda / r\mu$). (Since by definition, $K > s$, K must eventually increase as s increases.) Another significant finding is that K does not get unreasonably large as long as there are a sufficient number of Type A servers to assure that the probability of a Type A customer waiting while a primary server is free is not large. This probability, given by $\alpha = \sum_{i<r-s} \sum_{j>s} x_{ij}$, is an important measure of performance for this system. In particular, we examined systems with up to 5 primary servers and $K = 13$ and found that this choice of K was adequate whenever $\alpha \leqslant 0.5$ if $\rho \leqslant 0.5$ and whenever $\alpha \leqslant 0.15$ for $0.5 < \rho < 0.8$. Thus, if such a system is designed with an objective of efficient utilization of servers, the imposition of the truncation parameter K will not be a practical constraint in obtaining accurate numerical solutions by this method.

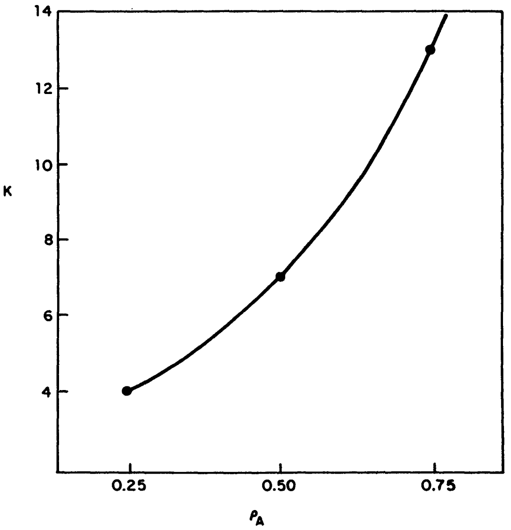

FIGURE 1. Required K as a Function of Type A Traffic Intensity $r = 5$, $s = 2$.

The expected waiting time in queue can be found for each customer type by Little's formula. The steady-state average number of customers waiting in the primary queue is given by

$$L_q^{(P)} = \sum_{j=0}^{s-1} \sum_{i=r-j+1}^{\infty} (i-r+j)x_{ij} + \sum_{j=s}^{K} \sum_{i=r-s+1}^{\infty} (i-r+s)x_{ij} \tag{8}$$

and the steady-state number of customers in the auxiliary queue is given by

$$L_q^{(A)} = \sum_{i=0}^{\infty} \sum_{j=s+1}^{K} (j-s)x_{ij}. \tag{9}$$

Since the dynamics of the system are defined such that the amount of time spent in the primary part of the system is identical for both customer types, the mean delay for Type P customers is

$$W_q^{(P)} = L_q^{(P)}/\lambda \tag{10}$$

and for Type A customers is

$$W_q^{(A)} = L_q^{(P)}/\lambda + L_q^{(A)}/\lambda_A. \tag{11}$$

Table 1 shows some expected waiting times and values of α for $\rho = 0.5$ with varying proportions of Type A customers and servers. (Cases denoted as "unstable" violate one of the necessary conditions given in §1 for the existence of a limiting distribution in the actual system.) We observe that the expected delay for Type A customers is quite sensitive to increases in their proportion, and particularly when the ratio of auxiliary to primary servers is less than 0.5. The lower this proportion of auxiliary to primary servers, the sooner the system will become unstable as the proportion of Type A customers increases.[1]

TABLE 1

Expected Delays and α for Moderate Traffic Intensity

$\lambda/r\mu = 0.5$						
$r = 3$		$s = 1$			$s = 2$	
q	E (Type P delay)	E (Type A delay)	α	E (Type P delay)	E (Type A delay)	α
0.2	1.35	4.86	0.06	1.41	1.51	0.00
0.4	1.13	13.51	0.27	1.36	1.90	0.02
0.5	0.98	25.08	0.48	1.31	2.32	0.03
0.6	0.83	43.04	0.65	1.23	2.81	0.06
0.8				0.87	4.88	0.19
	unstable					
$r = 5$		$s = 2$			$s = 3$	
q	E (Type P delay)	E (Type A delay)	α	E (Type P delay)	E (Type A delay)	α
0.2	0.75	1.62	0.02	0.77	0.83	0.00
0.4	0.63	5.07	0.13	0.74	1.22	0.02
0.6	0.39	17.27	0.43	0.61	2.45	0.08
0.8				0.31	5.92	0.26
	unstable					

[1] This work was supported by the Faculty Research Fund of the Graduate School of Business, Columbia University. I am very grateful to Marcel F. Neuts, David Lucantoni and Awi Federgruen for helpful discussions. I also thank Debashis Guha for carrying out the extensive programming.

References

Brill, P. H. and L. Green, "Queues in which Customers Receive Simultaneous Service From a Random Number of Servers," *Management Sci.*, 30 (1984), 51–68.

Green, L., "A Queueing System in Which Customers Require a Random Number of Servers," *Oper. Res.*, 28 (1980), 1335–1346.

———, "Comparing Operating Characteristics of Queues in Which Customers Require a Random Number of Servers," *Management Sci.*, 27 (1981), 65–74.

———, "A Queueing System with General-Use and Limited-Use Servers," *Oper. Res.* (to appear).

Jacobson, P. A. and E. D. Lazowska, "The Method of Surrogate Delays: Simultaneous Resource Possession in Analytic Models of Computer Systems," Technical Report No. 80-04-05, Department of Computer Science, University of Washington, 1980.

Kim, S., "$M/M/s$ Queueing System Where Customers Demand Multiple Server Use," Ph.D. Dissertation, Southern Methodist University, 1979.

Neuts, M. F., "Markov Chains with Applications in Queueing Theory, Which Have a Matrix-Geometric Invariant Probability Vector," *Adv. in Appl. Probab.*, 10 (1978), 185–212.

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

New York Times, "City Restudying One-Officer Cars for Police Patrol," October, 15, 1982.

Omahen, K. J., "Capacity Bounds for Multiresource Queues," *J. Assoc. Comput. Mach.*, 24 (1977), 646–663.

Willemain, T. R., "Approximate Analysis of a Hierarchical Queueing Network," *Oper. Res.*, 22 (1974), 522–544.