A Multiserver Queue with Narrow- and Wide-Band Customers and Wide-Band Restricted Access

YVES DE SERRES AND LORNE G. MASON

Abstract—**We consider a multiserver queueing system with two classes of customers: a type 1 (narrow-band, NB) customer requires a single server, while each type 2 (wide-band, WB) customer requests n of the m servers (n is not random). Servers allocated to a type 2 customer are seized and released simultaneously. Service times are exponentially distributed with mean $1/\mu_i$ for type i customers ($i = 1, 2$). Blocked type 1 customers are cleared while blocked type 2 customers may be delayed in an infinite waiting room. A type 1 customer enters service immediately upon arrival if at least one server is free, irrespective of the status of the type 2 queue. WB customers have restricted access to the service facility; a cutoff parameter specifies the maximum number of type 2 customers that can be in service at the same time. Two approaches, moment-generating functions and matrix-geometric techniques, are considered for the computation of the system performance; that is, the mean waiting time in queue and the probability of delay (i.e., nonzero waiting time) for type 2 customers, as well as the probability of blocking for type 1 customers.**

I. Introduction

The model considered in this paper is a member of the class of queueing systems in which some customers require service by more than one server. An important feature of such systems is that a customer cannot enter service until all required servers are available (*simultaneous seizure*); as a consequence, these systems do not qualify as batch arrival queues.

These models have wide applications in the computer and communications fields; for example, in the study, 1) of message storage systems [20], [24], 2) of demand-assigned multiple access (DAMA) circuit-switched services in communication satellite svstems [1], [2], [3], [15], and 3) of the multiplexing of multiple bit-rate data lines onto a wide-band digital trunk [7], [14], [16], [22], [23], [32], [33], as well as in multiple areas of operations research [17], [19].

The above class of queues can be divided into two subclasses distinguished by the *release mechanism* which specifies whether the servers allocated to the same customer end service independently (*independent release*) or simultaneously (*simultaneous release*).

A major contribution to the study of queues in which customers require service by more than one server is the work of Green [5], [10], [17], [18], [19]. A characteristic of the models she considers is that customers request a *random* number of servers. She first showed in [19] how the independent release of servers allocated to the same customer can be exploited to solve delay models with infinite waiting room. The key observation is that the waiting time analysis of the multiserver system can be reduced to that of a related $M/G/1$ queue with exceptional service. The technique was subsequently applied to more general systems [10], [17].

Gimpelson [16] considered a multiserver queue with two types of customers and simultaneous release. Customers of one type (narrow-band customers) require service from a single server, while each customer of the other type (wide-band customers) requests service by n of the m servers. Two methods of operation are studied. In both cases, blocked narrow-band customers are lost; blocked wide-band customers are either lost or delayed in a finite waiting room. The state equations of the birth–death process corresponding to each model are solved numerically to obtain the blocking probabilities. Gimpelson brought to light the oscillatory behavior of the blocking probability curves, a characteristic of these systems.

Kaufman [24] considered a pure loss model of a resource facility shared by several classes of customers, each class being characterized by its bandwidth and temporal requirements. Units allocated to the same customer are seized and released simultaneously. The main contributions of [24] are the generality of the residency distribution and the importance placed on efficient algorithms for the computation of the blocking probabilities. Previous studies of similar loss models are referenced in [24]; more recent contributions include [3], [7], [22], [32].

Using a system point approach, which is an extension of the system point method developed by Brill, (Brill and Green [5]) studied a multiserver queue in which customers require a random number of servers which are to begin and end service concurrently. In its general formulation, the model consists of m servers and k customer classes. Type i customers arrive according to a Poisson process at rate λ_i and require an exponentially distributed amount of service time with mean $1/\mu_i$, simultaneously from $c(i)$ servers. Customers enter service in their order of arrival. Based on system point theory, a general framework for deriving the waiting time distribution for each customer type is presented. Explicit solutions were derived for the two-server system with $\mu_i = \mu$. A numerical analysis of this model, based on the application of the block Gauss–Siedel method, is reported in [13]. The approach makes possible the analysis of systems with more than two servers, and with state-dependent or state-independent arrivals.

Kraimeche and Schwartz [27] studied a model of a broadband channel carrying a mixture of narrow- and wide-band traffics. As opposed to Gimpelson's model and to the model we propose here in which blocked wide-band customers are allowed to queue while blocked narrow-band customers are lost, Kraimeche and Schwartz consider a system in which the wide-band traffic is nonqueueable while the narrow-band traffic may queue in an infinite waiting room. The need for access control is addressed, and two access control strategies are analyzed by a moment-generating functions approach. Kraimeche and Schwartz had previously considered the bandwidth allocation problem in an all-blocked traffic model [28]. Recently, they reported an analysis of the all-queued traffic model [26].

The system studied in this paper is a multiserver queue with two types of customers: a type 1 customer (narrow-band customer) requests a single server; each type 2 customer

Paper approved by the Editor for Computer Communication Theory of the IEEE Communications Society. Manuscript received September 1, 1985; revised October 8, 1987. This work was supported by the Natural Sciences and Engineering Council of Canada under Grant G1074.

The authors are with INRS Telecommunications, Verdun, P.Q., Canada H3E 1H6.

IEEE Log Number 8820864.

0090-6778/88/0600-0675$01.00 © 1988 IEEE

(wide-band customer) requires simultaneous service from n of the m servers where n is fixed (not random). Blocked narrow-band customers are cleared while blocked wide-band customers are delayed in an infinite waiting room. A narrow-band customer enters service immediately upon arrival if there is at least one free server, irrespective of the number of wide-band customers waiting in the queue at that time. Interarrival and service times are exponential with possibly different means for each type. Finally, a cutoff parameter specifies the maximum number of wide-band customers that can be in service concurrently. The cutoff parameter protects the type 1 traffic against overload of type 2 traffic.

Our motivation for studying this queueing model is provided by the need to predict the performance of circuit-switched-based integrated services networks which can support a variety of wide-band services in addition to voice traffic. Examples include audio and video teleconferencing services [12], as well as virtual private networks. It is recognized that the assumption of Poisson arrivals for teleconference calls may be questioned, and indeed other models based on reservations have been considered in the literature. Our justification for this choice lies in its theoretical convenience in light of the absence of an accurate model for the subscriber behavior for these new services.

Our system is similar to the one studied by Gimpelson [16], however, it differs in two ways: 1) the infinite waiting room for wide-band traffic and 2) the possible cutoff imposed on the wide-band traffic.

Our model generalizes the one studied by Bhat and Fisher in [4] and by Feldman and Claybaugh in [11]; their system is a multiserver queue in which customers of two types require service from a *single* server and are served in their order of arrival. Server holding times are different for customers of different types. As in our model, customers of one type are operated with loss while customers of the other type are operated with delay; however, there is no provision for cutoff. The solution reported in [4] is based on moment-generating functions techniques while [11] presents a matrix-geometric solution of this simpler model.

Besides [4] and [11], there is abundant literature on multiserver queues with two heterogeneous classes of customers in which each customer requires service by a *single* server irrespective of its type. The interested reader is referred to [6], [21], [25], as well as the previous papers referenced therein.

In this paper, two approaches are used to compute the system performance. One solution is based on moment-generating functions techniques; the other is a matrix-geometric solution based on a result, obtained by Neuts, for multiserver queues. In each case, the mean waiting time and the probability of delay for wide-band customers, as well as the probability of blocking for narrow-band customers are computed.

In the second section, the model is described precisely and the notation is introduced. In Section III, the moment-generating functions approach is sketched, while Section IV is devoted to the matrix-geometric solution. In Section V, the two solutions are compared. Finally, the solutions are used in Section VI to discuss the performance of the system.

II. The Model

We consider a multiserver system (m servers) in which customers request either a single server (type 1 customers), or n servers (type 2 customers) where n is smaller or equal to m but is typically greater than one. Service requests of successive customers are independent and n is fixed (i.e., not random). Customers arrive according to a Poisson process with mean arrival rates λ_1 and λ_2 for type 1 and type 2 customers, respectively. Servers assigned to a type 2 customer begin and end service together. Service times are exponentially distributed with mean $1/\mu_i$ for type i customers ($i = 1, 2$).

Type 1 customers are operated with loss while type 2 customers are operated with delay and are allowed to queue in an infinite waiting room. Upon arrival, a type 1 customer enters service if there is at least one free server; otherwise, it is blocked and cleared. We note that narrow-band customers have access to the service facility even if there is a nonzero queue of wide-band customers, since it is possible to have free servers in front of a nonempty queue. Wide-band customers have restricted access to the service facility; the cutoff parameter r_0 specifies the maximum number of type 2 customers that can be in service at the same time. Therefore, a queue of type 2 customers forms as soon as a type 2 arrival finds r_0 type 2 customers already in service or otherwise if there are not enough free servers. Of course, a wide-band arrival joins a nonempty queue. The cutoff parameter determines, in the above manner, the sharing policy imposed to the traffic mixture.

The system is then modeled as a two-dimensional Markov process $(n_2(t), n_1(t))$ where $n_2(t)$ is the number of type 2 customers in the system (i.e., in service or in queue) and $n_1(t)$ is the number of type 1 customers in service. The state space is then the set $\{(i, j) | 0 \le i, 0 \le j \le m\}$, and the steady-state probability that the system is in state (i, j) is denoted $p(i, j)$.

If $m = rn + s$ where both r and s are nonnegative integers with $0 \le s < n$, the states (i, j) corresponding to an empty queue for the sharing policy described above are as follows:

$$0 \le j \le s + (r - r_0)n, \quad \text{and } 0 \le i \le r_0 \text{ (identified by } l = 0),$$

and for each l, $(r - r_0) + 1 \le l \le r$,

$$s + (l - 1)n + 1 \le j \le s + ln, \quad \text{and } 0 \le i \le r - l.$$

These "zero-queue" states are identified by O in Fig. 1, which is the state diagram of a system with $m = 10$, $n = 3$, and $r_0 = 2$; it follows that $r = 3$ and $s = 1$. In Fig. 1, we have indicated by an arrow only the unidirectional transitions. The segment between two neighboring states replaces (for graphical clarity) a pair of arrows in opposite directions from one state to the other. The transition rates out of state (i, j) are given in Fig. 2 below where λ_1^* is 0 when the number of busy servers is m (i.e., full service facility), and λ_1 otherwise. The states corresponding to a full service facility are easily identified as those on the top row of the state diagram, together with the states at the tip of the vertical arrows in Fig. 1.

III. A Solution Based on Moment-Generating Functions

We present in this section a solution based on moment-generating functions techniques. The method is the same as in [4], but applied to the more general model considered here. Although, in theory, this approach yields a general solution, we will see below that a restrictive condition must be imposed to the set of system parameters for the solution to be computationally tractable.

A. The Birth and Death Equations

From Figs. 1 and 2, the global balance equations are easily obtained by equating the total rate at which the process leaves a state to the total rate at which it enters that state. Here is the list of balance equations.

For the block of states identified by $l = 0$:

$\underline{0 \le j \le s + (r - r_0)n, \; 0 \le i \le r_0 - 1:}$

$$\begin{aligned}(\lambda_1 + j\mu_1 + \lambda_2 + i\mu_2)p(i, j) \\ = \lambda_1 p(i, j-1) + (j+1)\mu_1 p(i, j+1) \\ + \lambda_2 p(i-1, j) + (i+1)\mu_2 p(i+1, j) \qquad (1)\end{aligned}$$

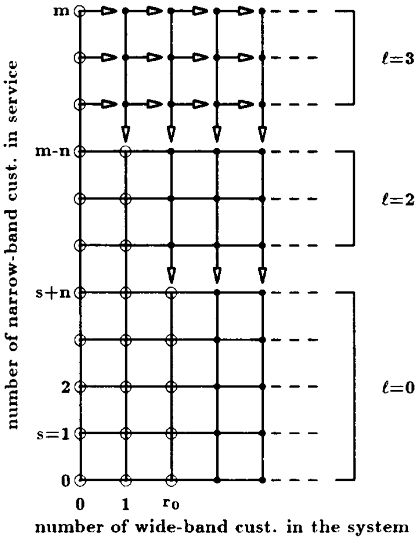

m
m-n
s+n
2
s=1
0
ℓ=3
ℓ=2
ℓ=0
0
1
r0
number of narrow-band cust. in service
number of wide-band cust. in the system

Fig. 1. State diagram.

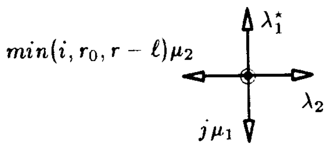


Fig. 2. Transition rates out of state (i, j).

$\underline{0 \le j \le s+(r-r_0)n-1,\ i \ge r_0}$:

$$\begin{aligned}(\lambda_1+j\mu_1+\lambda_2+r_0\mu_2)p(i,j) &= \lambda_1 p(i,j-1)+(j+1)\mu_1 p(i,j+1)\\ &\quad+\lambda_2 p(i-1,j)+r_0\mu_2 p(i+1,j)\end{aligned} \tag{2}$$

$\underline{j=s+(r-r_0)n,\ i \ge r_0}$:

$$\begin{aligned}(j\mu_1+\lambda_2+r_0\mu_2)p(i,j) &= \lambda_1 p(i,j-1)+(j+1)\mu_1 p(i,j+1)\\ &\quad+\lambda_2 p(i-1,j)+r_0\mu_2 p(i+1,j)\end{aligned} \tag{3}$$

and for each l, $(r - r_0) + 1 \le l \le r$:

$\underline{s+(l-1)n+1 \le j \le s+ln,\ 0 \le i < r-l}$:

$$\begin{aligned}(\lambda_1+j\mu_1+\lambda_2+i\mu_2)p(i,j) &= \lambda_1 p(i,j-1)+(j+1)\mu_1 p(i,j+1)\\ &\quad+\lambda_2 p(i-1,j)+(i+1)\mu_2 p(i+1,j)\end{aligned} \tag{4}$$

$\underline{j=s+(l-1)n+1,\ i=r-l}$:

$$\begin{aligned}(\lambda_1+j\mu_1+\lambda_2+(r-l)\mu_2)p(i,j) &= \lambda_1 p(i,j-1)+(j+1)\mu_1 p(i,j+1)\\ &\quad+\lambda_2 p(i-1,j)+(r-l)\mu_2 p(i+1,j)\end{aligned} \tag{5}$$

$\underline{j=s+(l-1)n+1,\ i>r-l}$:

$$\begin{aligned}(\lambda_1+j\mu_1+\lambda_2+(r-l)\mu_2)p(i,j) &= (j+1)\mu_1 p(i,j+1)+\lambda_2 p(i-1,j)\\ &\quad+(r-l)\mu_2 p(i+1,j)\end{aligned} \tag{6}$$

$\underline{s+(l-1)n+1<j<s+ln,\ i \ge r-l}$:

$$\begin{aligned}(\lambda_1+j\mu_1+\lambda_2+(r-l)\mu_2)p(i,j) &= \lambda_1 p(i,j-1)+(j+1)\mu_1 p(i,j+1)\\ &\quad+\lambda_2 p(i-1,j)+(r-l)\mu_2 p(i+1,j)\end{aligned} \tag{7}$$

$\underline{j=s+ln,\ i \ge r-l}$:

$$\begin{aligned}(j\mu_1+\lambda_2+(r-l)\mu_2)p(i,j) &= \lambda_1 p(i,j-1)+(j+1)\mu_1 p(i,j+1)\\ &\quad+\lambda_2 p(i-1,j)+(r-l)\mu_2 p(i+1,j).\end{aligned} \tag{8}$$

B. Solution of Birth and Death Equations

Steady state for the process $(n_2(t),\ n_1(t))$ will exist if the stability condition $\lambda_2 < r_0\mu_2$ is satisfied. Indeed, this condition is sufficient as $r_0\mu_2$ is the maximum rate at which the system can perform work on wide-band traffic. This maximum rate might not be achieved immediately due to competing narrow-band calls, but since these are served on a loss basis enough servers (namely, $r_0 n$) are bound to become free and allocated to wide-band calls.

Under stability, the performance measures, probability of delay and mean waiting time for type 2 customers and probability of blocking for type 1 customers, are obtained in a number of steps as follows:

1) define appropriate moment-generating functions and obtain from the balance equations a set of linear equations in these moment-generating functions [Section III-B1)],

2) apply the analyticity of the moment-generating functions to compute the "zero-queue" probabilities which here play the role of "boundary" probabilities [Section III-B2)],

3) solve for the moment-generating functions at $z = 1$, and deduce the system performance [Section III-B1)].

1) Moment-Generating Functions and System Performance: We define the following moment-generating functions:

$$Q_j(z)=\begin{cases}\displaystyle\sum_{i=r_0}^{\infty} p(i,j)z^{i-r_0}, & 0 \le j \le s+(r-r_0)n\\[2ex] \displaystyle\sum_{i=r-l}^{\infty} p(i,j)z^{i-(r-l)}, & (r-r_0)+1 \le l \le r,\\ & s+(l-1)n+1 \le j \le s+ln.\end{cases} \tag{9}$$

Multiplying (2), (3), and (5) to (8) by appropriate powers of z and summing over i results in a set of linear equations involving the moment-generating functions and the "zero-queue" probabilities. These equations can be written in matrix form as

$$A(z)\begin{pmatrix} Q_0(z) \\ \vdots \\ Q_m(z) \end{pmatrix} = B(z) \tag{10}$$

where the $m + 1$ by $m + 1$ matrix $A(z)$ is a band matrix of width 3, whose upper and lower off-diagonal entries are multiples of the system parameters λ_1, λ_2, μ_1, μ_2, while the diagonal entries are of the form $\lambda_2(1 - z) + \alpha\mu_2(1 - z^{-1})$, for $\alpha \in \{0, 1, \cdots, r_0\}$. The entries of the m by 1 column vector $B(z)$ are linear combinations of the zero-queue probabilities. The matrices $A(z)$ and $B(z)$ are defined in the Appendix.

In Section III-B2) below, we show how, by using the analyticity of the moment-generating functions $Q_j(z)$, the zero-queue probabilities $p(i, j)$ can be computed up to a normalizing constant. Substituting these values into the column vector $B(z)$, the system (10) can be solved for $Q_j(1)$ and $Q_j'(1)$ as multiples of a normalizing constant. That computation is simplified by the particular Jordan-like structure of the matrix $A(z)$ (see the Appendix), and can be performed recursively. As is usual in solutions of queueing problems by moment-generating functions, the computation (actually only the last step) requires the application of Cramer's rule followed by the application of l'Hospital rule. Details will not be provided here but can be found in [9].

The normalizing constant is then obtained through the following normalizing equation:

$$1 = \sum_{j=0}^{s+(r-r_0)n} \sum_{i=0}^{r_0-1} p(i, j) + \sum_{j=s+(l-1)n+1}^{s+ln} \sum_{i=0}^{r-l-1} p(i, j) + \sum_{j=0}^{m} Q_j(1). \tag{11}$$

The model performance, namely, the mean waiting time and the probability of delay for type 2 customers, as well as the probability of loss for type 1 customers, is then easily obtained.

The mean number of type 2 customers waiting for service $\bar{Q}_2$ is given by

$$\bar{Q}_2 = \sum_{j=0}^{m} Q_j'(1). \tag{12}$$

The mean waiting time in queue $\bar{W}_2$ for type 2 customers is obtained by Little's formula

$$\bar{W}_2 = \frac{\bar{Q}_2}{\lambda_2}. \tag{13}$$

The probability P_w that a type 2 customer has to experience a delay before entering the service facility is given by

$$P_w = \sum_{j=0}^{m} Q_j(1). \tag{14}$$

On the other hand, the loss probability for type 1 customers is given by

$$P_b = \sum_{l=r-r_0}^{r} Q_{s+ln}(1). \tag{15}$$

2) Computation of the Zero-Queue Probabilities: The usual approach for obtaining the boundary probabilities is to use the analyticity of the moment-generating functions involved. In the present situation this translates into the fact that, by Cramer's rule applied to (10), the determinant of the matrix obtained from $A(z)$ by replacing any one of its columns by the column vector $B(z)$ must vanish at each zero of the determinant $|A(z)|$ of $A(z)$ on the unit disk. Since the vector $B(z)$ depends only on zero-queue probabilities, each zero of $|A(z)|$ generates a linear equation in these probabilities. If the roots of $|A(z)| = 0$ on the unit disk are distinct, the analytic property together with balance equations (1) and (4) produce as many linear equations, minus one, as there are zero-queue states. These equations can then be solved for the zero-queue probabilities $p(i, j)$ up to a normalizing constant, and the system performance can be computed as explained previously.

Unfortunately the roots of $|A(z)| = 0$ are not necessarily distinct as can be seen by setting $m = 12$, $n = 2$, $r_0 = 4$, $\mu_1 = \mu_2 = 1.0$, $\lambda_1 = 6.0$, $\lambda_2 = 2.0$, in which case 0.5 is a double root of $|A(z)| = 0$; this can be verified simply by writing down the matrix $A(0.5)$. However, it can be shown, as in [29], that the determinant of each "block" of the matrix $A(z)$ [see the Appendix for a complete description of $A(z)$] has distinct real roots.

In theory, the multiple roots case is solvable; it would require considering higher dêrivatives of $Q_j(z)$. In practice, however, this results in complex computational procedures, and we did not investigate it further. As will be seen, the matrix-geometric solution presented below does not involve the computation of the roots of $|A(z)| = 0$, and consequently circumvents the multiple roots problem.

Under the assumption of distinct roots of $|A(z)| = 0$, the adjoint matrix of $A(z)$ must be computed at each root to obtain the zero-queue probabilities. The structure of the matrix $A(z)$ makes possible an easy recursive computation of its adjoint. Details can be found in [9].

The limit imposed on the computational algorithm by the distinct roots assumption prompted the research of another approach, which resulted in a matrix-geometric solution presented in next section.

IV. The Matrix-Geometric Solution

The states are ordered lexicographically. We define the row vectors p_i, for $i = 0, 1, 2, \cdots$,

$$p_i = (p(i, 0), p(i, 1), \cdots, p(i, m)). \tag{16}$$

The steady-state probability vector $p = (p_0, p_1, \cdots)$ is the solution of

$$pQ = 0$$
$$p\mathbf{1} = 1 \tag{17}$$

where $\mathbf{1}$ is the all 1 column vector of appropriate dimension, and Q is the transition rate matrix of the Markov process and is given by

$$Q = \begin{bmatrix} A_0 & B & & & & & \\ C_1 & A_1 & B & & & & \\ & \cdot & \cdot & \cdot & & & \\ & & \cdot & \cdot & \cdot & & \\ & & & C_{r_0} & A_{r_0} & B & \\ & & & & C_{r_0} & A_{r_0} & B \\ & & & & & \cdots & \\ & & & & & & \cdots \end{bmatrix}. \tag{18}$$

The matrices A_i are $m + 1$ by $m + 1$, are band matrices of width 3, have a Jordan-like structure, and are equal to

$$A_i=\begin{bmatrix} A_i^{(0)} & & & & \\ (m-in+1)\mu_1 & A_i^{(1)} & & & \\ & (m-(i-1)n+1)\mu_1 & A_i^{(2)} & & \\ & & \cdot\ \ \cdot\ \ \cdot & & \\ & & & (m-n+1)\mu_1 & A_i^{(i)} \end{bmatrix} \tag{19}$$

where the matrix $A_i^{(0)}$ is $m - in + 1$ by $m - in + 1$ and equal to (in-band matrix representation)

$$A_i^{(0)}=\begin{bmatrix} & -(\lambda+i\mu_2) & \lambda_1 \\ \mu_1 & -(\lambda+\mu_1+i\mu_2) & \lambda_1 \\ \cdot & \cdot & \cdot \\ (m-in-1)\mu_1 & -(\lambda+(m-in-1)\mu_1+i\mu_2) & \lambda_1 \\ (m-in)\mu_1 & -(\lambda_2+(m-in)\mu_1+i\mu_2) & \end{bmatrix} \tag{20}$$

while for each k, $1 \leq k \leq i$, the matrix $A_i^{(k)}$ is n by n and equal to (in-band matrix representation)

$$A_i^{(k)}=\begin{bmatrix} & -(\lambda+(j_k+1)\mu_1+(i-k)\mu_2) & \lambda_1 \\ (j_k+2)\mu_1 & -(\lambda+(j_k+2)\mu_1+(i-k)\mu_2) & \lambda_1 \\ \cdot & \cdot & \cdot \\ (j_k+n-1)\mu_1 & -(\lambda+(j_k+n-1)\mu_1+(i-k)\mu_2) & \lambda_1 \\ (j_k+n)\mu_1 & -(\lambda_2+(j_k+n)\mu_1+(i-k)\mu_2) & \end{bmatrix} \tag{21}$$

where $\lambda = \lambda_1 + \lambda_2$ and $j_k = m - (i - k + 1)n$. The matrices C_i are $m + 1$ by $m + 1$ diagonal matrices equal to

$$C_i=\begin{bmatrix} C_i^{(0)} & & & & \\ & C_i^{(1)} & & & \\ & & \cdot & & \\ & & & C_i^{(i-1)} & \\ & & & & 0I_n \end{bmatrix} \tag{22}$$

where I_n is the n by n identity matrix and

$$C_i^{(k)}=\begin{cases} i\mu_2 I_{m-in+1}, & k=0 \\ (i-k)\mu_2 I_n, & 1\leq k\leq i-1. \end{cases} \tag{23}$$

Finally, the matrix B is

$$B=\lambda_2 I_{m+1}. \tag{24}$$

As in [11], we apply theorem 2 of Neuts [31] to obtain in our case the following.

Theorem: Provided that the queue is stable, i.e., the Markov process Q is positive recurrent ($\lambda_2 < r_0\mu_2$), the steady-state probability vector $p = (p_0, p_1, \cdots)$ is given by

$$p_k=p_{r_0-1}R^{k-r_0+1}, \quad k\geq r_0. \tag{25}$$

The matrix R is the minimal nonnegative solution of the matrix equation

$$R^2C_{r_0}+RA_{r_0}+B=0. \tag{26}$$

The probability vector $\tilde{p} = (p_0, \cdots, P_{r_0-1})$ is the unique solution of

$$\tilde{p}T=0$$
$$\tilde{p}\mathbf{1}+p_{r_0-1}R(I-R)^{-1}\mathbf{1}=1 \tag{27}$$

where the matrix T is a generator ($T\mathbf{1} = 0$) and equal to

$$T=\begin{bmatrix} A_0 & B & & & \\ C_1 & A_1 & B & & \\ \cdot & \cdot & \cdot & & \\ & \cdot & \cdot & \cdot & \\ & & C_{r_0-2} & A_{r_0-2} & B \\ & & & C_{r_0-1} & A_{r_0-1}+RC_{r_0} \end{bmatrix}. \tag{28}$$

Finally,

$$RC_{r_0}\mathbf{1}=B\mathbf{1}=\lambda_2\mathbf{1}. \tag{29}$$

The solution is then obtained by a two step process:

1) First, the rate matrix R is determined by iterative substitution in (26). Efficient ways of performing that operation are examined in [30].

2) Once R is known, the r_0m "boundary" probabilities $\tilde{p}$ are determined by solving a system of r_0m linear equations. Different approaches are considered in Section IV-A.

Once the matrix R and the boundary probabilities have been computed, the model performance is easily obtained. The mean number $\bar{Q}_2$ of type 2 customers waiting for service is

$$\bar{Q}_2=\sum_{j=0}^{s+(r-r_0)n}\sum_{i=r_0}^{\infty}(i-r_0)p(i,j)+ \sum_{l=1+r-r_0}^{r}\sum_{j=s+(l-1)n+1}^{s+ln}\sum_{i=r-l}^{\infty}(i-r+l)p(i,j) \tag{30}$$

which, using (25), is easily reduced to

$$\bar{Q}_2 = p_{r_0-1} R^2 (I-R)^{-2} \mathbf{1} + p_{r_0-1} R (I-R)^{-1} v$$

$$+ \sum_{l=1+r-r_0}^{r} \sum_{j=s+(l-1)n+1}^{s+ln} \sum_{i=r-l}^{r_0-1} (i-r+l) p(i, j) \quad (31)$$

where the column vector v is $v = (0, \cdots, 0, 1, \cdots, 1, 2, \cdots, 2, \cdots, r_0, \cdots, r_0)^t$, in which there are $s + (r - r_0)n + 1$ components equal to 0, n equal to 1, n equal to 2, and so on until the n components equal to r_0. The mean waiting time for type 2 customers is obtained from Little's formula, that is

$$\bar{W}_2 = \frac{\bar{Q}_2}{\lambda_2} . \quad (32)$$

The probability of delay P_w for type 2 customers is

$$P_w = \sum_{j=0}^{s+(r-r_0)n} \sum_{i=r_0}^{\infty} p(i, j) +$$

$$\sum_{l=1+r-r_0}^{r} \sum_{j=s+(l-1)n+1}^{s+ln} \sum_{i=r-l}^{\infty} p(i, j). \quad (33)$$

Again from (25), this simplifies to

$$P_w = p_{r_0-1} R (I-R)^{-1} \mathbf{1} + \sum_{l=1+r-r_0}^{r} \sum_{j=s+(l-1)n+1}^{s+ln} \sum_{i=r-l}^{r_0-1} p(i, j). \quad (34)$$

The probability of blocking P_b for type 1 customers is

$$P_b = \sum_{l=r-r_0}^{r} \sum_{i=r-l}^{\infty} p(i, s+ln) \quad (35)$$

which becomes with the use of (25)

$$P_b = \sum_{l=r-r_0}^{r} \left\{ (p_{r_0-1} R (I-R)^{-1})_{s+ln} + \sum_{i=r-l}^{r_0-1} p(i, s+ln) \right\} \quad (36)$$

where $(\ )_j$ denotes the jth component of the vector in parentheses.

A. Computation of the "Boundary" Probabilities

The two approaches described in this section have been considered for the solution of the linear system (30).

1) Computation by Decomposition: Due to its special structure, the solution of (27) can be obtained by decomposition as in [30]. The equation $\bar{p}T = 0$ can be written

$$p_0 A_0 + p_1 C_1 = 0$$

$$p_{i-1} B + p_i A_i + p_{i+1} C_{i+1} = 0, \quad 1 \le i \le r_0 - 2$$

$$p_{r_0-2} B + p_{r_0-1} (A_{r_0-1} + R C_{r_0}) = 0 \quad (37)$$

from which it is easily obtained that

$$p_i = p_{r_0-1} H_i, \quad 0 \le i \le r_0 - 2 \quad (38)$$

where the matrices H_i are computed recursively by

$$H_{r_0-1} = I$$

$$H_{r_0-2} = -(A_{r_0-1} + R C_{r_0}) B^{-1}$$

$$H_{r_0-i} = -(H_{r_0-i+1} A_{r_0-i+1} + H_{r_0-i+2} C_{r_0-i+2}) B^{-1},$$

$$3 \le i \le r_0. \quad (39)$$

From the first equation in (37) and the normalizing equation in (27), we have that p_{r_0-1} is the solution of

$$p_{r_0-1} (H_0 A_0 + H_1 C_1) = 0$$

$$p_{r_0-1} \left\{ \sum_{i=0}^{r_0-2} H_i \mathbf{1} + (I-R)^{-1} \mathbf{1} \right\} = 1. \quad (40)$$

This linear system of equations was solved by the most accurate, double precision, IMSL subroutine (international mathematical and statistical libraries). Experience with our Fortran program on a DEC/VENUS 8600 has revealed that computational inaccuracies appear when the number of recursions in the above scheme (r_0 iterations) exceeds 7 (see Table I below). As an alternative to the above solution of (37), we have examined iterative methods of solution for linear systems of equations.

2) Computation by Iterative Methods: Gauss-Siedel method with successive overrelaxation [34] was applied to the linear system (27). The number of iterations required can be reduced by appropriately selecting the value of the relaxation factor. Since our interest was mainly an analysis of the range of the algorithm, we did not consider computing an exact or approximate value of the optimum relaxation factor [34]; the relaxation factor was prespecified; the value 1.5 appeared to be secure and efficient for this particular problem. Iteration is continued until the maximum point-wise relative difference between the last two iterates is below some prespecified threshold, for example, 10^{-4}.

The linear system (27) could be solved that way for slightly larger systems; for example, it took 456 iterations to solve (27) in the special case $m = 48$, $n = 4$, $r_0 = 12$, $\lambda_1/m\mu_1 = 0.7$, $\lambda_2/r_0\mu_2 = 0.1$, $\mu_1/\mu_2 = 5.0$, and a stopping threshold equal to 10^{-4}.

V. Comparison of the Two Solutions

As the system size increases, both algorithms suffer a loss of precision which manifests itself by the appearance of negative "boundary" probabilities. These probabilities are solutions of a linear system of equations and are computed by the most accurate double-precision IMSL subroutine. This routine performs iterative improvement of the solution, and indicates upon exit, the degree of accuracy obtained (the approximate number d of digits unchanged after improvement). Table I shows the decrease of accuracy for a given set of model parameters as the number of servers is increased; the parameters are determined by $\lambda_1/m\mu_1 = 0.7$, $\lambda_2/r_0\mu_2 = 0.1$, and $\mu_1 = 1.0$. We note that 16 digits is machine precision for double precision arithmetic.

The matrix-geometric solution is more attractive than the moment-generating functions solution, both theoretically and algorithmically.

On a theoretical basis, the main advantage of the matrix-geometric solution is that it covers all combinations of system parameters resulting in a stable model. We recall that the solution based on moment-generating functions requires the computation of the roots of the determinant of a complex matrix; the solution is easily obtained only under the assumption that these roots are distinct, a condition that is not always satisfied.

On a computational basis, the solution based on moment-generating functions and the matrix-geometric solution with boundary probabilities computed by decomposition have more or less the same range. For moderatively large systems ($m = 48$), it appears that the matrix-geometric solution in which the boundary probabilities are computed by iterative methods can be used to determine the system performance. However, convergence may be slow and more sophisticated ways of accelerating it should be considered if that procedure is to be

TABLE I
ACCURACY AND RANGE OF THE TWO SOLUTIONS; d_1 REFERS TO THE GENERATING FUNCTIONS SOLUTION, d_2 TO THE MATRIX-GEOMETRIC SOLUTION. THE HYPHEN INDICATES THAT THE ALGORITHM FAILED TO COMPUTE THE SOLUTION

m	n	r_0	μ_1/μ_2	d_1	d_2
4	2	2	5.0	16	16
12	6	2	5.0	15	15
24	12	2	5.0	13	14
36	18	2	5.0	–	13
8	2	4	5.0	12	14
16	4	4	5.0	12	13
24	6	4	5.0	10	11
36	9	4	5.0	–	10
12	2	6	5.0	9	13
24	4	6	5.0	8	9
36	6	6	5.0	–	7
16	2	8	5.0	6	9
24	3	8	5.0	4	7
32	4	8	5.0	–	–
24	4	6	0.5	9	12
24	4	6	1.0	9	11
24	4	6	5.0	8	11
24	4	6	10.0	7	9
24	4	6	20.0	6	8
24	2	12	5.0	–	–
24	3	8	5.0	4	7
24	4	6	5.0	8	9
24	6	4	5.0	10	11
24	8	3	5.0	12	12
24	12	2	5.0	13	14

efficient enough to justify the cost of an exact solution in such cases. Unfortunately, the computational range of these solutions is far from covering practical communications systems. Their usefulness is then limited to the qualitative investigation of the behavior of such systems. Large systems must be studied through numerical solutions or through approximations.

VI. Discussion

The nonmonotonic variation of the blocking probability as a function of the percentage of one traffic type was discovered by Gimpelson [16]. This oscillatory variation is due to the fact that in competing for capacity, a single narrow-band customer can keep a wide-band customer out of service. This narrow-band customer effectively "reserves" a number of channels for the narrow-band traffic, resulting in an improved performance for that traffic type. The effect depends mainly on n, the number of servers required by a wide-band customer, called the bandwidth factor by Gimpelson. Gimpelson considered a loss model; the oscillatory variation appeared in the probability of blocking. In our delay/loss model, the phenomenon affects both the mean waiting time and the probability of blocking, as is shown in Figs. 3 and 4. This behavior, characteristic of systems carrying mixtures of traffics with different bandwidth requirements, can have dramatic consequences when designing a system in which the relative percentages of traffics are not known exactly.

To produce these graphs, the system performance was computed with the parameter μ_1 set to 1. It follows that the unit of time for the mean waiting time curves is the mean service time of a narrow-band customer.

The nonmonotonic variation of the mean waiting time and the probability of blocking also depends on the ratio of mean service times for both traffic types. Figs. 5 and 6 show the effect of the ratio μ_1/μ_2 on the model performance.

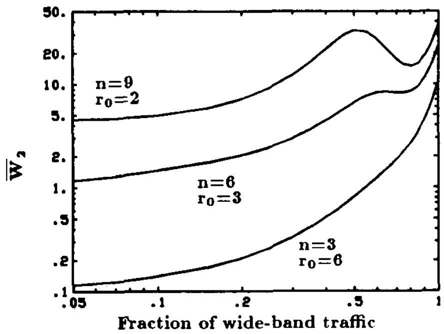


Fig. 3. Effect of the bandwidth factor n/m on the mean waiting time for $m = 18$, total load = 16.0, $\mu_1/\mu_2 = 10.0$.

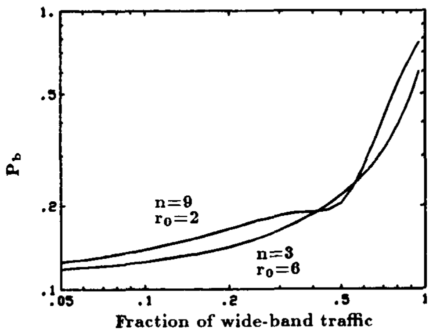


Fig. 4. Effect of the bandwidth factor n/m on the probability of blocking for $m = 18$, total load = 16.0, $\mu_1/\mu_2 = 10.0$.

As mentioned in the Introduction, the cutoff parameter protects the narrow-band traffic against overload of the wide-band traffic. Reducing the cutoff parameter results in smaller narrow-band blocking and larger wide-band delay (Figs. 7 and 8). As observed, the cutoff imposed on the wide-band traffic results in an improved performance of the narrow-band traffic, especially when the percentage of wide-band traffic is large. The point at which the cutoff effect manifests itself (i.e., the point at which the curves depart from one another) depends on the total load. For the total load selected here, the system becomes unstable as r_0 is reduced and the percentage of wide-band traffic increases; that explains why the curve corresponding to $r_0 = 3$ does not cover the full range of traffic divisions. Finally, we note that when access restriction is exercised, part of the service capacity is reserved to narrow-band customers so that as the fraction of wide-band traffic approaches 1 (i.e., the fraction of narrow-band approaches 0) the probability of blocking approaches 0 unless instability has occurred. On the other hand, if wide-band customers have access to the full service capacity, the limit of the probability of blocking, as the percentage of wide-band traffic approaches 1, depends on the total load.

The combined effect of blocking and delay is measured by the power factor

$$\text{PR} = \frac{1 - P_b}{1 + \mu_2 \overline{W}_2}. \tag{41}$$

The larger the power, the better is the system. As shown in Fig. 9, reducing the cutoff results in overall smaller power. To protect the narrow-band traffic, a nontrivial cutoff parameter (i.e., $r_0 \neq m/n$) is needed to avoid NB traffic to be blocked for long periods.

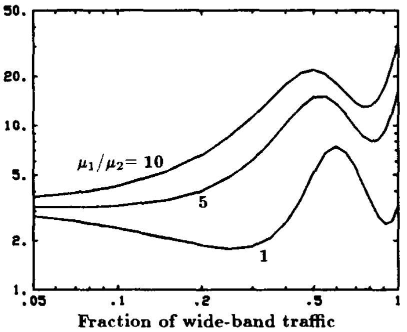


Fig. 5. Effect of the ratio μ_1/μ_2 on the mean waiting time, for $m = 16$, $n = 8$, $r_0 = 2$, total load = 14.0.

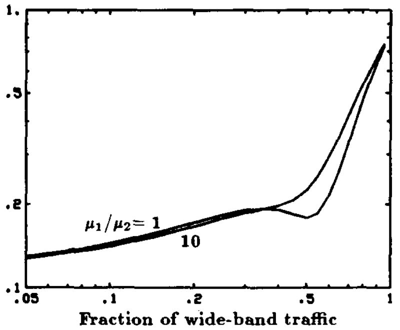


Fig. 6. Effect of the ratio μ_1/μ_2 on the probability of blocking, for $m = 16$, $n = 8$, $r_0 = 2$, total load = 14.0.

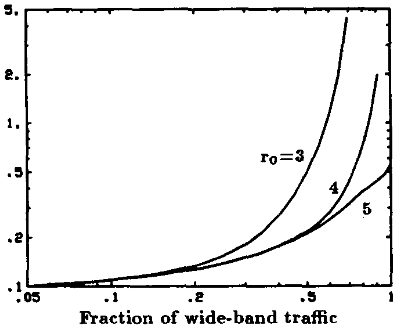


Fig. 7. Effect of the cutoff parameter on the mean waiting time, for $m = 15$, $n = 3$, $\mu_1/\mu_2 = 1.0$, total load = 12.0.

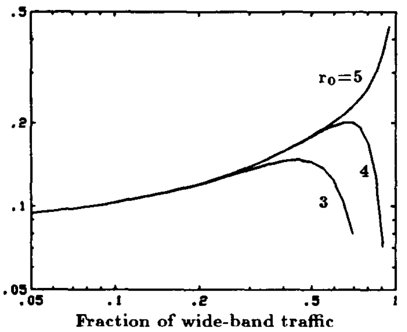


Fig. 8. Effect of the cutoff parameter on the probability of blocking, for $m = 15$, $n = 3$, $\mu_1/\mu_2 = 1.0$, total load = 12.0.

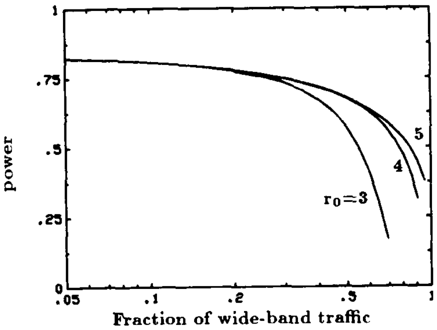


Fig. 9. Effect of the cutoff parameter on the power for $m = 15$, $n = 3$, $\mu_1/\mu_2 = 1.0$, total load = 12.0.

VII. Conclusion

Two basic techniques have been used to analyze a multiserver, delay/loss queue, with narrow- and wide-band traffics and WB restricted access. This queue is a member of an important class of multiserver models that are found in the analysis of many practical systems.

Much effort has been devoted to the computational aspects of both solutions so as to determine and possibly extend the range of parameters over which the performance can be computed. Partial success has been obtained but only dimensions much smaller than those of practical systems remained computationally tractable. There is ongoing research on approximations for large systems of this kind; the exact solutions described here can be used for validation of these approximations.

The solutions were used to examine the model performance. The oscillatory variation of the probability of blocking versus the percentage of wide-band traffic, first pointed out by Gimpelson for a loss model, was shown to be present also in our delay/loss model where both the probability of blocking and the mean waiting time exhibit oscillations. A cutoff parameter is introduced as a protection mechanism for narrowband traffic against overload of wide-band traffic; its effect on system performance has been described.

An analysis of the blocking period has been carried out for both the exact model reported herein, and for various approximate models. These results will be reported in a forthcoming paper.

Appendix

In this Appendix, the matrices $A(z)$ and $B(z)$ of (10) are defined. As mentioned in Section III-B1), (10) is the matrix formulation of the set of linear equations [in the moment-generating functions $Q_j(z)$] obtained by applying (9) to the balance equations, and grouping terms. The operations involved are elementary.

Before describing $A(z)$, we comment on its form. $A(z)$ is a band matrix of width 3 with a Jordan-like structure. The nonzero entries of the matrix $A(z)$ corresponding to the example in Fig. 1 are shown below:

$$\begin{bmatrix}
\bullet & \bullet & & & & & & & & & \\
\bullet & \bullet & \bullet & & & & & & & & \\
 & \bullet & \bullet & \bullet & & & & & & & \\
 & & \bullet & \bullet & \bullet & & & & & & \\
 & & & \bullet & \bullet & * & & & & & \\
 & & & & \bullet & \bullet & & & & & \\
 & & & & & \bullet & \bullet & \bullet & & & \\
 & & & & & & \bullet & \bullet & * & & \\
 & & & & & & & & \bullet & \bullet & \\
 & & & & & & & & \bullet & \bullet & \bullet \\
 & & & & & & & & & \bullet & \bullet
\end{bmatrix}. \tag{42}$$

This structure of $A(z)$ is pivotal to the computation of the roots of $|A(z)| = 0$, as well as the recursive computation of its adjoint matrix and the moment-generating functions $Q_j(1)$ and $Q_j'(1)$.

In general, defining $c_i = s + (r - r_0 + i)n + 1$, matrix $A(z)$ is of the form

$$A(z)=\begin{bmatrix} A^{(0)}(z) & -c_0\mu_1 z^{-1} & & & \\ & A^{(r-r_0+1)}(z) & -c_1\mu_1 z^{-1} & & \\ & & \cdot & & \\ & & & A^{(r-1)}(z) & -c_{r_0-1}\mu_1 z^{-1} \\ & & & & A^{(r)}(z) \end{bmatrix} \tag{43}$$

where the submatrices $A^{(i)}(z)$ (defined below) correspond to the • entries in (42), while the entries $-c_i\mu_1 z^{-1}$ correspond to the * in (42). The submatrix $A^{(0)}(z)$ is the $s + (r - r_0)n + 1$ by $s + (r - r_0)n + 1$ band matrix of width 3

$$A^{(0)}(z)=f_{r_0}(z)\cdot I_{s+(r-r_0)n+1}+\begin{bmatrix} & \lambda_1 & -\mu_1 \\ -\lambda_1 & \lambda_1+\mu_1 & -2\mu_1 \\ \cdot & \cdot & \cdot \\ -\lambda_1 & \lambda_1+(s+(r-r_0)n-1)\mu_1 & -(s+(r-r_0)n)\mu_1 \\ -\lambda_1 & (s+(r-r_0)n)\mu_1 & \end{bmatrix} \tag{44}$$

where I_k is the identity matrix of order k, $f_\alpha(z) = \lambda_2(1 - z) + \alpha\mu_2(1 - z^{-1})$, and the second term of (44) is shown in-band matrix representation.

For each l, $(r - r_0) + 1 \le l \le r$ and $j_l = s + (l - 1)n$, $A^{(l)}(z)$ is the n by n band matrix of width 3

$$A^{(l)}(z)=f_{r-l}(z)\cdot I_n+\begin{bmatrix} & \lambda_1+(j_l+1)\mu_1 & -(j_l+2)\mu_1 \\ -\lambda_1 & \lambda_1+(j_l+2)\mu_1 & -(j_l+3)\mu_1 \\ \cdot & \cdot & \cdot \\ -\lambda_1 & \lambda_1+(j_l+n-1)\mu_1 & -(j_l+n)\mu_1 \\ -\lambda_1 & (j_l+n)\mu_1 & \end{bmatrix} \tag{45}$$

where again the second term of (45) is shown in-band matrix representation.

The column vector $B(z)$ has the form $(B^{(0)}(z), B^{(r-r_0+1)}(z), \cdots, B^{(r)}(z))^t$ where

$$B^{(0)}(z)=\begin{bmatrix} \lambda_2 p(r_0-1,0)-r_0\mu_2 z^{-1}p(r_0,0) \\ \lambda_2 p(r_0-1,1)-r_0\mu_2 z^{-1}p(r_0,1) \\ \vdots \\ \lambda_2 p(r_0-1,s+(r-r_0)n-1)-r_0\mu_2 z^{-1}p(r_0,s+(r-r_0)n-1) \\ \lambda_2 p(r_0-1,s+(r-r_0)n)-r_0\mu_2 z^{-1}p(r_0,s+(r-r_0)n)-(s+(r-r_0)n+1)\mu_1 z^{-1}p(r_0-1,s+(r-r_0)n+1) \end{bmatrix} \tag{46}$$

and for each l, $(r - r_0) + 1 \le l \le r$ and $j_l = s + (l - 1)n$

$$B^{(l)}(z)=\begin{bmatrix} \lambda_2 p(r-l-1,j_l+1)-(r-l)\mu_2 z^{-1}p(r-l,j_l+1)+\lambda_1 p(r-l,j_l) \\ \lambda_2 p(r-l-1,j_l+2)-(r-l)\mu_2 z^{-1}p(r-l,j_l+2) \\ \lambda_2 p(r-l-1,j_l+3)-(r-l)\mu_2 z^{-1}p(r-l,j_l+3) \\ \vdots \\ \lambda_2 p(r-l-1,j_l+n-1)-(r-l)\mu_2 z^{-1}p(r-l,j_l+n-1) \\ \lambda_2 p(r-l-1,j_l+n)-(r-l)\mu_2 z^{-1}p(r-l,j_l+n)-(j_l+n+1)\mu_1 z^{-1}p(r-l-1,j_l+n+1) \end{bmatrix}. \tag{47}$$

References

[1] J. M. Aein, "A multi-user-class, blocked-calls-cleared, demand access model," *IEEE Trans. Commun.*, vol. COM-26, pp. 378-385, Mar. 1978.

[2] J. M. Aein and O. S. Kosovych, "Satellite capacity allocation," *Proc. IEEE*, vol. 65, pp. 332-342, Mar. 1977.

[3] G. Barberis and R. Brignolo, "Capacity allocation in a DAMA satellite system," *IEEE Trans. Commun.*, vol. COM-30, pp. 1750-1757, July 1982.

[4] U. N. Bhat and M. J. Fisher, "Multichannel queueing systems with heterogeneous classes of arrivals," *Naval Res. Logist. Quart.*, vol. 23, no. 2, pp. 271-282, 1976.

[5] P. H. Brill and L. Green, "Queues in which customers receive simultaneous service from a random number of servers: A system point approach," *Management Sci.*, vol. 30, no. 1, pp. 51-68, Jan. 1984.

[6] G. Brune, "On delay and loss in a switching system for voice and data with internal overflow," in *Proc. 11th Int. Teletraffic Congress,* Kyoto, Japan, Sept. 1985, pp. 2.2A2.1-2.2A2.7.

[7] L. E. N. Delbrouck, "On the steady-state distribution in a service facility carrying mixtures of traffic with different peakedness factors and capacity requirements," *IEEE Trans. Commun.*, vol. COM-31, pp. 1209-1211, Nov. 1983.

[8] Y. De Serres, "A multi-server, non-preemptive, cutoff-priority queue in which some customers require more than one server; a matrix-geometric solution," INRS-Telecommun., Tech. Rep. 85-41, Dec. 1985.

[9] ——, "A multi-server, non-preemptive, cutoff-priority queue in which some customers require more than one server," INRS-Telecommun., Tech. Reps. 85-09 and 85-34, Mar. and Oct. 1985.

[10] A. Federgruen and L. Green, "An *M/G/c* queue in which the number of servers required is random," *J. Appl. Prob.*, vol. 21, no. 3, pp. 583-601, Sept. 1984.

[11] R. M. Feldman and C. A. Claybaugh, "A note on a computational model for a data/voice communication queueing system," *Naval Res. Logist. Quart.*, vol. 29, no. 3, pp. 529-534. Sept. 1982.

[12] M. J. Ferguson and L. Mason, "Network design for a large class of teleconferencing systems," *IEEE Trans. Commun.*, vol. COM-32, pp. 789-796, July 1984.

[13] G. Y. Fletcher, H. G. Perros, and W. J. Stewart, "A queueing network model of a circuit switching access scheme in an integrated services environment," *IEEE Trans. Commun.*, vol. COM-34, pp. 25-30, Jan. 1986.

[14] G. F. W. Fredrikson, "Analysis of channel utilization in traffic concentrators," *IEEE Trans. Commun.*, vol. COM-22, pp. 1122-1129, Aug. 1974.

[15] G. Frenkel, "The grade of service in multiple-access satellite communications systems with demand assignments," *IEEE Trans. Commun.*, vol. COM-22, pp. 1681-1685, Oct. 1974.

[16] L. A. Gimpelson, "Analysis of mixtures of wide- and narrow-band traffic," *IEEE Trans. Commun. Technol.*, vol. COM-13, pp. 258-266, Sept. 1965.

[17] L. Green, "A multiple dispatch queueing model of police patrol operations," *Management Sci.*, vol. 30, no. 6, pp. 653-664, June 1984.

[18] ——, "Comparing operating characteristics of queues in which customers require a random number of servers," *Management Sci.*, vol. 27, no. 1, pp. 65-74, Jan. 1981.

[19] ——, "A queueing system in which customers require a random number of servers," *Oper. Res.*, vol. 28, no. 6, pp. 1335-1346, Nov.-Dec. 1980.

[20] S. Hattori, S. Morita, Y. Fujii, and M. W. Kim, "A design model for real-time voice storage system," *IEEE Trans. Commun.*, vol. COM-30, pp. 53-57, Jan. 1982.

[21] H. D. Ide and R. G. Schehrer, "On application and performance of cut-off priority queues in switching systems with overload protection," in *Proc. 11th Int. Teletraffic Congress,* Kyoto, Japan, Sept. 1985, pp. 2.1B2.1-2.1B2.7.

[22] S. Iisaku and Y. Urano, "Performance analysis of integrated communication systems with heterogeneous traffic," in *Proc. 11th Int. Teletraffic Congress,* Kyoto, Japan, Sept. 1985, pp. 2.1A2.1-2.1A2.6.

[23] L. Katzschnor and R. Scheller, "Probability of loss of data traffics with different bit rates hunting one common PCM-channel," in *Proc. 8th Int. Teletraffic Congress,* Melbourne, Australia, Nov. 1976, pp. 525/1-525/8.

[24] J. S. Kaufman, "Blocking in a shared resource environment," *IEEE Trans. Commun.*, vol. COM-29, pp. 1474-1481, Oct. 1981.

[25] K. Kawashima, "Efficient numerical solutions for a unified reservation system with two classes," *Rev. Elec. Commun. Lab.*, vol. 31, no. 3, pp. 419-429, 1983.

[26] B. Kraimeche and M. Schwartz, "Bandwidth allocation strategies in wide-band integrated networks," *IEEE J. Select. Areas Commun.*, vol. SAC-4, pp. 869-878, Sept. 1986.

[27] ——, "Analysis of traffic access control strategies in integrated service networks," *IEEE Trans. Commun.*, vol. COM-33, pp. 1085-1093, Oct. 1985.

[28] ——, "Traffic access control strategies in integrated digital networks," in *Proc. IEEE INFOCOM '84*, San Diego CA, Apr. 1984, pp. 230-235.

[29] I. L. Mitrani and B. Avi-Itzhak, "A many-server queue with service interruptions," *Oper. Res.*, vol. 16, no. 3, pp. 628-638, 1968.

[30] M. F. Neuts and D. M. Lucantoni, "A Markovian queue with *N* servers subject to breakdowns and repairs," *Management Sci.*, vol. 25, no. 9, pp. 849-861, Sept. 1979.

[31] M. F. Neuts, "Further results in the *M/M/*1 queue with randomly varying rates," *Opsearch*, vol. 15, no. 4, pp. 158-168, 1978.

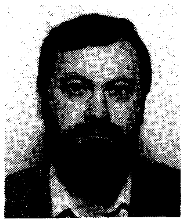
[32] V. Ramaswami and Rao K. Aswath, "Flexible time slot assignment: A performance study for the integrated services digital network," in *Proc. 11th Int. Teletraffic Congress,* Kyoto, Japan, Sept. 1985, pp. 2.1A3.1-2.1A3.7.

[33] M. Yokoyama, H. Yamamoto, and M. Kajiwara, "A multi-bit-rate substitute-communication system with high call-carrying capacity," *IEEE Trans. Commun.*, vol. COM-31, pp. 799-803, June 1983.

[34] D. M. Young, *Iterative Solution of Large Linear Systems.* New York: Academic, 1971.

Yves De Serres received the B.S. and M.S. degrees in mathematics from the University of Montreal, Canada, in 1972 and 1974, respectively. He obtained the B.Eng. degree in electrical engineering from McGill University in 1980, and an M.S. degree in telecommunications from the University of Quebec at INRS-Telecommunications in 1983.

Since 1980 he has been associated with INRS-Telecommunications as a Student and as a Research Assistant. His current research interest is the dynamic control of queueing systems.



Lorne G. Mason received the B.Sc. and Ph.D. degrees in mechanical engineering, control division, from the University of Saskatchewan, Saskatoon, Canada, in 1963 and 1973, respectively.

From 1963 to 1965 he was a Rocket Design Engineer with Bristol Aerospace, Winnipeg, Man., Canada. Since that time he has been involved in the areas of telecommunications and automatic control in both industrial and academic environments. He was employed with the B.C. Telephone Company, Vancouver, B.C., Canada, in the traffic studies and network planning areas. Following the completion of his postgraduate studies, he served as a consultant to Yale University, New Haven, CT, concerning the application of learning techniques to network management and control. From 1974 to 1977, he was employed by Bell-Northern Research, Ottawa, Ont., and Montreal, P.Q., Canada, where he was instrumental in developing a planning system for the introduction of digital transmission and switching equipment into telephone networks. In 1977 he joined INRS-Telecommunications, Montreal, where he pioneered a course on network planning which has subsequently been offered also at McGill University, Montreal. He currently holds the position of Full Professor of Telecommunications. His research interests involve the application of control and optimization techniques to network design and management.