Journal of Applied Mathematics and Stochastic Analysis, 11:3 (1998), 339-354.

ANALYSIS OF A MULTI-SERVER QUEUEING MODEL OF ABR

R. NÚÑEZ-QUEIJA and O.J. BOXMA
CWI
P.O. Box 94079
1090 GB Amsterdam, The Netherlands

(Received October, 1997; Revised May, 1998)

In this paper we present a queueing model for the performance analysis of Available Bit Rate (ABR) traffic in Asynchronous Transfer Mode (ATM) networks. We consider a multi-channel service station with two types of customers, denoted by high priority and low priority customers. In principle, high priority customers have preemptive priority over low priority customers, except on a fixed number of channels that are reserved for low priority traffic. The arrivals occur according to two independent Poisson processes, and service times are assumed to be exponentially distributed. Each high priority customer requires a single server, whereas low priority customers are served in processor sharing fashion. We derive the joint distribution of the numbers of customers (of both types) in the system in steady state. Numerical results illustrate the effect of high priority traffic on the service performance of low priority traffic.

Key words: Asynchronous Transfer Mode, Available Bit Rate, Multi-Server Queue, Priorities, Processor Sharing, Matrix-Geometric Solution, Spectral Expansion.

AMS subject classifications: 60K25, 68M20, 90B12, 90B22.

1. Introduction

It is our pleasure to contribute this paper to the special issue in honor of Ryszard Syski. During a time span of more than forty years, Professor Syski has made many lasting contributions to Applied Probability in general and Teletraffic in particular. The second author fondly remembers the very pleasant cooperation with Ryszard Syski in editing the book *Queueing Theory and its Applications - Liber Amicorum for J.W. Cohen* (North-Holland Publ. Cy., Amsterdam, 1988).

The diverse characteristics and service requirements of the different traffic types that are carried by ATM (Asynchronous Transfer Mode) networks have led to the definition of different service categories that should be offered to users of such a network. We briefly discuss those differences, distinguishing three large categories: *Con-*

stant Bit Rate (CBR) traffic, *Variable Bit Rate* (VBR) traffic and *Available Bit Rate* (ABR) traffic.

The CBR service class guarantees a *fixed* pre-determined transmission capacity to its users. Therefore, this service is useful for traffic that requires both very small (or no) delays *and* very small (or no) losses. At the *burst-level* (where we distinguish different bursts of traffic coming from the same connection, but not the separate ATM-cells that form a burst), it is reasonable to assume that all CBR traffic requires a fixed amount of capacity over time. In all further considerations, we will leave out the CBR traffic and use the term "capacity" to indicate the total capacity minus the capacity reserved for CBR traffic.

For VBR traffic, we make a subdivision into *real-time* and *non real-time* connections. For both these subclasses, the users must specify many characterizing parameters such as minimum cell rate, mean cell rate, peak cell rate and maximum burst size. The difference lies in the requirements. The main issue for real-time connections such as voice and possibly video is the delay of the transmission; the loss of small amounts of information during the transmission is less important for these connections. This traffic lends itself very well for multiplexing. On the other hand, non real-time VBR traffic requires small losses and the delays are less important. To ensure that losses are small, large buffers are used to store non real-time VBR traffic when the communication network is heavily loaded.

The last category, ABR traffic, was introduced to cope with specific problems that arise when transmitting data. For this traffic, losses lead to retransmission of data (because of the extreme sensitivity to losses), which introduces a lot of overhead in implementations. Since transmission delays are of less importance for data traffic, the setting of non real-time VBR seems to be the appropriate one to carry data traffic. However, data traffic is very bursty and the required parameters for VBR connections are difficult to specify by the users. For ABR connections, no parameters need to be specified. Only a small amount of capacity is *reserved* for the transmission of ABR traffic. Additionally, the capacity that is not currently being required by VBR (and CBR) traffic is used for ABR traffic. When the total capacity currently available to ABR is too small, ABR traffic is stored in very large buffers, ensuring a small loss probability, until the available capacity increases again. The advantage here is that ABR traffic gets all the capacity that is left over. For the server, this means a higher utilization of the network's resources. As pointed out above, the main service guarantee for ABR traffic is a very small loss fraction or, in principle, no loss at all. No guarantee can be given on transmission delays.

A special issue of ABR is that the available capacity should be shared *fairly* among all ABR users. In queueing models, it seems reasonable to incorporate this with the queue discipline of *processor sharing*. In this discipline, all "customers" receive an equal share of the service capacity.

In addition to the large storage buffers, some feedback control mechanism can be used to keep the loss of information small. The buffers can store incoming data that can not be transmitted immediately, due to a temporarily overloaded system. Feedback control can be used to slow down data sources when the buffers are heavily loaded and an overflow may occur. We refer to the ATM Forum [1, 2] for more detailed specifications of ABR.

Since the conceptual introduction of ABR, many papers on the subject have been published. Most studies so far emphasize the modeling and (feedback) control aspects, see for instance Iliadis [12] and Ritter [20]. In [21], Ritter investigates the

problem of dimensioning the buffer for ABR traffic in order to avoid large losses. In [22] and [23], Ritter considers the case with feedback control, under the assumption that the source of ABR traffic is saturated (i.e., it sends continuously at the allowed rate).

A drawback in most studies is the assumption of a fixed available capacity for transmission of ABR traffic. As it was pointed out above, one of the essential features of ABR is that it makes use of the capacity that is *left over* by VBR traffic. Therefore, there is a need for a detailed performance analysis of ABR in the presence of other traffic. In the present paper, our goal is to devise and analyze a model that captures the influence of real-time VBR traffic on ABR traffic. We compare the performance of the ABR traffic in our model under *variable* available capacity, with the performance in an equivalent model with *fixed* available capacity.

Our model is basically a multi-server queue with two types of customers: (*i*) high priority customers (real-time VBR traffic); and (*ii*) low priority customers (ABR traffic). We assume that the high priority customers have a waiting room with a finite, typically small capacity - thus modeling the real-time requirement - and each accepted customer is served by a single server. Low priority customers have an infinite waiting room (buffer) and equally share the remaining capacity according to the processor sharing principle - this models the large storage buffers for ABR traffic and the fair sharing of the available capacity between ABR users. The servers at the service-station are divided into two groups: (*i*) there are N servers that are dedicated principally to the high priority customers (we call these the normal servers); and (*ii*) N_L servers are purely reserved for the low priority traffic (we call these the L-servers). On the normal servers, the high priority customers have *preemptive priority* over the low priority customers.

We point out that this is a *call-level* model: A customer represents a request of an ABR source to transmit data, and the service requirement of the customer is identified with the amount of data to be transmitted. In our analysis, we assume that arrivals occur according to two independent Poisson processes. This assumption is justified in the case where many sources are connected to the communication network.

Although we present the model in the context of (future) ABR traffic, it can just as easily be seen in the context of existing situations, where real-time VBR has priority over non real-time VBR. In this case, the processor sharing discipline for the low priority traffic should be replaced by the *First Come First Served* (FCFS) discipline. Also, the processor sharing among ABR sources is interesting in the light of *per VC* (Virtual Connection) *queueing*, where sources do not queue behind one another, but each source gets a separate access to the server (parallel to one another). The feature of processor sharing can further be generalized to *weighted fair queueing* (generalized processor sharing), where the total capacity is divided between the active sources according to some weighting factors.

Related two-dimensional Markov models have been studied in a number of papers. The case where both types of customers have an infinite waiting space, and within each customer type the service discipline is FCFS, was solved first by Mitrani and King [15], and later by Gail, Hantler, and Taylor [8]. The non-preemptive variant of that model was studied by Gail, Hantler, and Taylor [7]. Falin, Khalil, and Stanford [5] treated the preemptive case with processor sharing among the low priority customers. A discrete-time variant modeled as an M/G/1-type Markov Chain is considered in Gail, Hantler, Konheim, and Taylor [6]. A more extensive treatment of the

spectral analysis of M/G/1-type Markov Chains is given in Gail, Hantler, and Taylor [9]. A model related to the one presented in this paper, addressing the case with finite buffer capacity, is treated in Núñez-Queija [18]. In [3], Blaabjerg et al. consider a model similar to ours and give various performance measures in terms of the steady-state distribution, rather than analyzing this distribution in greater detail. Our main goal is to give a detailed analysis of the steady-state distribution itself.

In our analysis we are inspired by Gail, Hantler, and Taylor [8], but we make use of methods from other approaches. Instead of transforming the involved distributions into generating functions, the present work focuses directly on the distribution itself. It does so relying mainly on the matrix geometric approach of Neuts [17] and the spectral expansion approach (see for instance Mitrani and Chakka [14] and Mitrani and Mitra [16]).

The paper is organized as follows. We give a description of the model in Section 2. In Section 3, we mention some relevant results of the theory of matrix-geometric solutions for the steady-state analysis of GI/M/1-type Markov Chains developed by Neuts [17]. We use this in Section 4 as a starting point of our analysis. In Section 5, we give a complete characterization of the joint distribution of the numbers of customers of both types in the system at steady state. Numerical results are presented in Section 6 to illustrate the effect of high priority traffic on the service performance of low priority traffic.

2. The Model

Consider a service station consisting of $N + N_L$ identical servers (N and N_L both are positive integers) that are divided into two groups: (*i*) A number of N servers, which we call the normal servers; and (*ii*) the remaining N_L servers, henceforth called L-servers. Two types of customers - high and low priority customers - require service from the station. At the station, there is a waiting room for K high priority customer (K being a nonnegative integer) and a room of infinite capacity for low priority customers.

High priority customers arrive at the station according to a Poisson process with rate λ_H. If the N normal servers are all occupied by other high priority customers, then a newly arrived high priority customer takes his place in the finite waiting room. If there are already K other high priority customers in the waiting room, then the new customer is rejected and leaves the system without receiving service. If there are less than N other high priority customers currently being served, then a new high priority customer is immediately taken into service by one server. Also, if the service of a high priority customer is completed and the waiting room is not empty, then one of the waiting high priority customers immediately enters service. Service times of the high priority customers are assumed to be exponentially distributed with mean $1/\mu_H$ and independent of everything else.

Low priority customers arrive according to a Poisson process with rate λ_L, independent of high priority customers. Their service requirement is assumed to be exponentially distributed with mean $1/\mu_L$, independent of everything else. Furthermore, they are served according to the processor sharing discipline by the L-servers, *and* the normal servers that are not occupied by a high priority customer. Thus, if there are i high priority and $j \geq 1$ low priority customers present, then each of the low priority customers receives service at rate:

$$\frac{N_L + \max(N-i,0)}{j}\mu_L.$$

(The servers work at unit rate).

We will further use the notation $\rho_H := \lambda_H/\mu_H$ and $\rho_L := \lambda_L/\mu_L$. We are interested in the steady-state behavior of the numbers of both types of customers in the system.

Let $X_H(t)$ ($X_L(t)$) be the number of high priority (low priority) customers present in the system at time t. Then the process $(X_H(t), X_L(t))$ is an irreducible and aperiodic Markovian process. Moreover, we note that the high priority customers are not influenced by the low priority customers, and therefore follow an M/M/N/(N + K)-queue, i.e., if we define $p_i := \mathrm{P}\{X_H = i\} := \lim_{t\to\infty}\mathrm{P}\{X_H(t) = i\}$:

$$p_0 = \left(\sum_{i=0}^{N-1}\frac{(\rho_H)^i}{i!} + \frac{(\rho_H)^N}{N!}\frac{1-(\rho_H/N)^{K+1})}{1-\rho_H/N}\right)^{-1},$$

$$p_i = p_0\frac{(\rho_H)^i}{i!}, \quad i = 1,2,\ldots,N-1, \tag{1}$$

$$p_i = p_0\frac{(\rho_H)^N}{N!}\left(\frac{\rho_H}{N}\right)^{i-N}, \quad i = N, N+1,\ldots,N+K.$$

The process $(X_H(t), X_L(t))$ is ergodic if and only if the following (intuitive) condition holds:

$$\rho_L < N_L + \mathrm{E}[\max(0, N - X_H)]. \tag{2}$$

We come back to this at the end of this section.

We define the equilibrium probabilities:

$$\pi_{i,j} := \mathrm{P}\{X_H = i, X_L = j\} := \lim_{t\to\infty}\mathrm{P}\{X_H(t) = i, X_L(t) = j\}, \tag{3}$$

and partition them into vectors $\bar{\pi}_j := (\pi_{0,j}, \pi_{1,j}, \ldots, \pi_{N+K,j})$ of length $N + K + 1$. Note that $\bar{\pi}_j$ is associated with the states in which j low priority customers are present. This partition enables us to write the equilibrium vector as $\bar{\pi} = (\bar{\pi}_0, \bar{\pi}_1, \bar{\pi}_2, \ldots)$. The corresponding infinitesimal generator is given by:

$$\mathcal{Q} = \begin{bmatrix} Q_{00} & T^{(+)} & 0 & \ldots & & \\ T^{(-)} & T^{(0)} & T^{(+)} & 0 & \ldots & \\ 0 & T^{(-)} & T^{(0)} & T^{(+)} & 0 & \ldots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots \end{bmatrix}. \tag{4}$$

The matrices $T^{(+)}$, $T^{(-)}$, $T^{(0)}$ and Q_{00} are of dimension $N + K + 1$. $T^{(+)}$, $T^{(-)}$ and the off-diagonal elements of $T^{(0)}$ (and Q_{00}) are respectively associated with an arriving low priority customer, a departing low priority customer, and a change in the number of high priority customers. The diagonal entries of $T^{(0)}$ and Q_{00} are such that each row of $\mathcal{Q}$ sums up to zero. The matrices are given by:

$T^{(+)} = \lambda_L I$,

$$T^{(-)} = \text{diag}[(N_L+N)\mu_L,(N_L+N-1)\mu_L,\ldots,N_L\mu_L,N_L\mu_L,\ldots,N_L\mu_L], \tag{5}$$

$$T^{(0)} = Q_H - T^{(+)} - T^{(-)},$$

$$Q_{00} = Q_H - T^{(+)},$$

with Q_H being the infinitesimal generator of the M/M/N/(N + K) queue of the high priority traffic:

$$Q_H = \begin{pmatrix} -\lambda_H & \lambda_H & & & & & \\ \mu_H & -\mu_H-\lambda_H & \lambda_H & & & & \\ & 2\mu_H & -2\mu_H-\lambda_H & \lambda_H & & & \\ & \ddots & \ddots & \ddots & & & \\ & & N\mu_H & -N\mu_H-\lambda_H & \lambda_H & & \\ & & & \ddots & \ddots & \ddots & \\ & & & & N\mu_H & -N\mu_H-\lambda_H & \lambda_H \\ & & & & & N\mu_H & -N\mu_H \end{pmatrix}.$$

Neuts [17] has given an extensive treatment of the so-called GI/M/1 type of Markov chains, of which our present model is a special case. For our model, however, we can take the analysis further than possible for the general case. The general ergodicity condition given in Neuts [17, Theorem 1.7.1, p. 32] states that: (*i*) if the matrix R is the minimal nonnegative solution to the equation

$$T^{(+)} + RT^{(0)} + R^2T^{(-)} = 0, \tag{6}$$

then all eigenvalues of R should lie inside the complex unit disc; and (*ii*) $Q_{00} + RT^{(-)}$ should have a positive vector in its left null space. The same theorem also gives us that the first statement is equivalent with (2). As for the second statement, it can be seen that if the first statement holds, then $Q_{00} + RT^{(-)}$ is a generator, and considering Q_{00} we see that it is an irreducible generator. Therefore the second requirement is immediately satisfied. We will not go further into the details of this, but refer the interested reader to Neuts [17]. In the sequel, we assume that (2) holds.

3. Preliminaries

From the final results in Section 2, it is clear that the unique probability vector $\bar{\pi} = (\bar{\pi}_0, \bar{\pi}_1, \bar{\pi}_2, \ldots)$ satisfying $\bar{\pi}\mathcal{Q} = 0$ has the matrix-geometric form:

$$\bar{\pi}_{j+1} = \bar{\pi}_j R, \tag{7}$$

where the matrix R is the minimal nonnegative solution to (6). Equation (7) can also be argued using basic results on irreducible Markov chains.

In our analysis we shall use a different, but highly related representation based on the spectral expansion approach, see for example Mitrani and Chakka [14] and

Mitrani and Mitra [16]. The essence of this approach is that we can rewrite Relation (7) to the "spectral expansion" form:

$$\overline{\pi}_j = \sum_{k=0}^{N+K} \alpha_k (r_k)^j \overline{v}_k, \quad j \geq 0, \tag{8}$$

whenever the matrix R has $N+K+1$ different eigenvalues $r_0,\ldots,r_{N+K}$, with corresponding left eigenvectors $\overline{v}_0,\ldots,\overline{v}_{N+K}$; i.e., $\overline{v}_k R = r_k \overline{v}_k$, $k = 0,1,\ldots,N+K$. The coefficients α_k are to be chosen such that the "ground level" equations:

$$\overline{\pi}_0 Q_{00} + \overline{\pi}_1 T^{(-)} = \overline{0}\,, \tag{9}$$

are satisfied. We come back to this in Section 5.

Even if the matrix R has multiple eigenvalues, Relation (8) still holds, as long as the set of all eigenvectors spans the $(N+K+1)$-dimensional Euclidean space. When this is not the case (the matrix R is defective), the coefficients α_k become functions $\alpha_k(j)$ which are polynomials in j, and follow from the Jordan canonical form of R (see for instance Gantmacher [10]).

We now define the quadratic matrix polynomial $T(z)$ by

$$T(z) := T^{(+)} + zT^{(0)} + z^2 T^{(-)}. \tag{10}$$

Note that if $\overline{v}$ is an eigenvector of the matrix R corresponding to the eigenvalue r, then $\overline{v}$ is in the left null-space of the matrix $T(r)$, and $\det[T(r)] = 0$. It follows immediately that R is nonsingular, since $T(0) = T^{(+)} = \lambda_L I$ is nonsingular. Therefore we may write:

$$T(z) = (R - zI)(R^{-1}T^{(+)} - zT^{(-)}), \tag{11}$$

again using (6). This is a very useful factorization, since $\det[R - zI]$ is precisely the characteristic polynomial of R, and $\det[R^{-1}T^{(+)} - zT^{(-)}]$ is also a polynomial in z. Both these polynomials are of degree $N+K+1$; therefore, $\det[T(z)]$ is of degree $2(N+K+1)$. Note that if $N_L = 0$ (which we assumed not to be the case), then the degree of $\det[R^{-1}T^{(+)} - zT^{(-)}]$ is N, but the analysis can still proceed along the same lines. We come back to this in Remark 4.1.

In Section 4, we show that the zeros of $\det[T(z)]$ inside the unit circle coincide with the zeros of $\det[R - zI]$ (i.e., the eigenvalues of R), and that all the zeros of $\det[R^{-1}T^{(+)} - zT^{(-)}]$ lie outside the unit circle, except for the zero $z = 1$ on the unit circle.

4. Spectral Analysis

In this section we investigate the eigenvalues of R. In the ergodic case, all these eigenvalues lie *inside* the complex unit disc (see Neuts [17]). We shall show that there are $N+K+1$ of them, and that they are all real and positive.

The starting point of the analysis is (11). We investigate the zeros of $\det[T(z)]$, showing that there are $2(N+K+1)$ zeros: $N+K+1$ zeros in $(0,1)$, one at $z=1$, and $N+K$ in $(1,\infty)$. The zeros in $(0,1)$ are then identified with the eigenvalues of R.

Theorem 4.1: *For real $z \neq 0$, the matrix $T(z)$ has $N+K+1$ different real eigenvalues.*

Proof: Note that $T(z)$ is a tri-diagonal matrix with off-diagonal elements:

$$T(z)_{i-1,i} = \lambda_H z,$$
$$T(z)_{i,i-1} = \min(i,N)\mu_H z, \tag{12}$$

where $i = 1,2,\ldots,N+K$. We denote the ith diagonal element $T(z)_{i,i}$ by $t_i(z)$:

$$\begin{aligned} t_i(z) &= \lambda_L - \{\lambda_H + \min(i,N)\mu_H + \lambda_L + (N_L + \max(N-i,0))\mu_L\}z \\ &\quad + (N_L + \max(N-i,0))\mu_L z^2, \\ t_{N+K}(z) &= \lambda_L - (N\mu_H + \lambda_L + N_L\mu_L)z + N_L\mu_L z^2. \end{aligned} \tag{13}$$

Here, $i = 0,1,\ldots,N+K-1$.

We now observe that the matrix $T(z)$ is *similar* to a real symmetric matrix (i.e., there exists a nonsingular matrix D such that the matrix $S(z) := DT(z)D^{-1}$ is a real symmetric matrix). In our case we can take D to be the diagonal matrix with diagonal elements

$$D_{i,i} = \sqrt{\frac{p_i}{p_0}}, \quad i = 0,1,\ldots,N+K.$$

The p_i are given in (1). The entries of $S(z)$ are:

$$S(z)_{i,i} = t_i(z),$$
$$S(z)_{i-1,i} = S(z)_{i,i-1} = z\sqrt{\min(i,N)\mu_H\lambda_H},$$

and are zero in all other positions. For the matrix $S(z)$, it is easy to see that it has $N+K+1$ different real eigenvalues (see Parlett [19]). First, since $S(z)$ is symmetric, all its eigenvalues are real, and it is non-defective (the geometric multiplicity of each eigenvalue is equal to the algebraic multiplicity). Second, since $S(z)$ is tri-diagonal, with non-zero elements directly above and below the diagonal, each eigenvalue has a unique eigenvector (up to multiplication by a scalar), i.e., the geometric multiplicity of each eigenvalue is 1. Combining these two facts, we are done.

Since the eigenvalues of $T(z)$ and $S(z)$ coincide, the same holds for $T(z)$. □

The fact that the eigenvalues of $T(z)$ are real for real z simplifies the analysis considerably. In the sequel, we only consider the eigenvalues as real functions of the real variable z. Therefore, for real z, denote the eigenvalues of $T(z)$ by:

$$\tau_0(z) \le \tau_1(z) \le \ldots \le \tau_{N+K}(z), \tag{14}$$

the inequalities being strict if $z \neq 0$, and

$$\tau_0(0) = \tau_1(0) = \ldots = \tau_{N+K}(0) = \lambda_L. \tag{15}$$

We state that the eigenvalues $\tau_k(z)$ are continuous functions of z for real z. An indication of the proof is as follows: For $z' \neq 0$, we use the strict ordering of the eigenvalues, and the continuity of $\det[T(z)]$ as a function of z to show that for a fixed $k = 0,1,\ldots,N+K$, $\lim_{z\to z'}\tau_k(z) = \tau_k(z')$. For $z = 0$, the same can be proved using Gerŝgorin's theorem (see for instance Marcus and Minc [13]).

Theorem 4.2: *$\tau_{N+K}(1) = 0$, and for $k = 0,1,\ldots,N+K-1$, the equation $\tau_k(z) = 0$ has a solution for $z \in (0,1)$ and a solution for $z \in (1,\infty)$.*

Proof: It is clear that $\det[T(1)] = 0$, since the rows of $T(1)$ sum to 0. Furthermore, note that $T(1)$ is diagonally dominant (see Marcus and Minc [13]) with negative diagonal elements; therefore all the eigenvalues of $T(1)$ are nonpositive. This gives:

$$\tau_0(1) < \tau_1(1) < \ldots < \tau_{N+K}(1) = 0. \tag{16}$$

Since the $\tau_k(z)$ are continuous functions of z, together with (15) this immediately gives us that all the $\tau_k(z)$ for $k = 0, 1, \ldots, N+K-1$ cross the horizontal axis (at least once) somewhere in $(0,1)$.

If z increases to infinity, the matrix $T(z)$ becomes strictly diagonally dominant with positive diagonal elements (the diagonal elements are convex quadratic functions in z and the off-diagonal elements are linear in z), and so for z large enough, all the eigenvalues of $T(z)$ are positive. Therefore, all the $\tau_k(z)$ for $k = 0, 1, \ldots, N+K-1$ must cross the horizontal axis again somewhere in $(1, \infty)$. □

Theorem 4.3: *Under the ergodicity condition in* (2), $\tau_{N+K}(z) = 0$ *for some* $z \in (0,1)$.

Proof: Because of the continuity of $\tau_{N+K}(z)$ and the fact that $\tau_{N+K}(0) = \lambda_L > 0$, it is sufficient to show that $\tau_{N+K}(1-) < 0$. First we write:

$$\det[T(z)] = (1-z)g(z), \tag{17}$$

where $g(z)$ is the determinant of the matrix obtained by replacing the last column of $T(z)$ by the sum of all columns and then dividing that column by $1-z$:

$$g(z) = \begin{vmatrix} t_0(z) & \lambda_H z & & & & \lambda_L - (N_L+N)\mu_L z \\ \mu_H z & t_1(z) & \lambda_H z & & & \lambda_L - (N_L+N-1)\mu_L z \\ & \ddots & \ddots & \ddots & & \vdots \\ & & N\mu_H z & t_N(z) & \lambda_H z & \lambda_L - N_L \mu_L z \\ & & & \ddots & \ddots & \vdots \\ & & & & N\mu_H z \; t_{N+K}(z) & \lambda_L - N_L \mu_L z \end{vmatrix}.$$

We want to evaluate $g(1)$. Therefore, we manipulate the above matrix evaluated in $z = 1$. First divide the last column by μ_L, and all the other columns by μ_H. Then add to each column (except for the first and the last one) all columns to the left of it. We now have:

$$g(1) = \mu_L(\mu_H)^{N+K}$$

$$\times \begin{vmatrix} -\rho_H & 0 & & & & \rho_L - (N_L+N) \\ 1 & -\rho_H & 0 & & & \rho_L - (N_L+N-1) \\ & \ddots & & \ddots & \ddots & \vdots \\ & & N & -\rho_H & 0 & \rho_L - N_L \\ & & & \ddots & \ddots & \vdots \\ & & & & N & -\rho_H \quad \rho_L - N_L \\ & & & & & N \quad \rho_L - N_L \end{vmatrix}$$

$$= \mu_L(\mu_H)^{N+K}\left\{\sum_{k=0}^{N-1}(-1)^{k+N+K}[\rho_L-(N_L+N-k)](-\rho_H)^k\frac{N!}{k!}N^K\right.$$

$$\left.+\sum_{k=N}^{N+K}(-1)^{k+N+K}[\rho_L-N_L](-\rho_H)^k N^{N+K-k}\right\}.$$

The last equality follows by expanding the determinant in its last column. Rearranging some terms, we rewrite this to:

$$g(1) = \mu_L(-\mu_H)^{N+K}N!N^K\left\{(\rho_L-N_L)\sum_{k=0}^{N-1}\frac{(\rho_H)^k}{k!}-\sum_{k=0}^{N-1}(N-k)\frac{(\rho_H)^k}{k!}\right.$$

$$\left.+(\rho_L-N_L)\sum_{k=N}^{N+K}\frac{\rho_H^N}{N!}\left(\frac{\rho_H}{N}\right)^{k-N}\right\}$$

$$= \mu_L(-\mu_H)^{N+K}N!\ N^K\frac{1}{p_0}\left\{(\rho_L-N_L)\sum_{k=0}^{N+K}p_k-\sum_{k=0}^{N-1}(N-k)p_k\right\}$$

$$= \mu_L(-\mu_H)^{N+K}N!N^K\frac{1}{p_0}\{\rho_L-N_L-\mathrm{E}[\max(N-X_H,0)]\}.$$

Under the ergodicity conditions in (2), $\mathrm{sign}[g(1)] = (-1)^{N+K+1}$. Differentiating (17) gives us $\frac{d}{dz}\det[T(z)]\,|_{z=1} = -g(1)$. Together this gives us

$$\mathrm{sign}[\det[T(1-)]] = -\mathrm{sign}\left[\frac{d}{dz}\det[T(z)]\,|_{z=1}\right] = \mathrm{sign}[g(1)] = (-1)^{N+K+1}. \qquad (18)$$

On the other hand, $\det[T(1-)] = \prod_{k=0}^{N}\tau_k(1-)$, and we know that $\tau_k(1-)<0$ (because of continuity and $\tau_k(1)<0$) for $k=0,1,\ldots,N-1$. Thus, we have proved that $\tau_{N+K}(1-)<0$, and hence that $\tau_{N+K}(z)$ has a zero in $(0,1)$. □

Theorem 4.4: *$\det[T(z)]$ has $N+K+1$ roots in $(0,1)$, one at $z=1$, and $N+K$ in $(1,\infty)$. The roots inside $(0,1)$ are precisely the eigenvalues of R.*

Proof: By Theorems 4.2 and 4.3, we have found $2(N+K+1)$ roots of $\det[T(z)]$ with the required positions. Since the degree of $\det[T(z)]$ is $2(N+K+1)$ these are all the roots, proving the first assertion. Using (11), we see that the roots of the characteristic polynomial of R appear with at least the same multiplicity in $\det[T(z)]$. Since all the roots of $\det[T(z)]$ have multiplicity one, the second assertion follows. □

Remark 4.1: In Section 2, we assumed that $N_L>0$. When $N_L=0$, $\det[R^{-1}T^{(+)}-zT^{(-)}]$ becomes of degree N, and consequently the degree of $\det[T(z)]$ is $2N+K+1$. In this case, Theorems 4.1 and 4.3 remain valid, but Theorems 4.2 and 4.4 need to be modified.

It is no longer true that, for $k=N, N+1,\ldots,N+K$, $\tau_k(z)=0$ for some $z\in(1,\infty)$. Apart from this, Theorem 4.2 still holds. In Theorem 4.4, only the number of roots in $(1,\infty)$ must be changed from $N+K$ to $N-1$. The proof of the zeros in $(0,1]$ can remain unchanged, but to prove the $N-1$ zeros in $(1,\infty)$, we need an additional argument since $T(z)$ no longer becomes diagonally dominant for $z\to\infty$. However, we can show for the matrix $\hat{T}(w) := w^2T^{(+)}+wT^{(0)}+T^{(-)}$, that

$\det[\widehat{T}(w)]$ has $N-1$ roots for $w \in (0,1)$: $\widehat{T}(0)$ has N positive eigenvalues and $\widehat{T}(1)$ only has nonnegative eigenvalues. Using the fact that $\widehat{T}(w)$ has $N+K+1$ different eigenvalues for all $w \in (0,\infty)$, and that these are continuous in $w \in [0,\infty)$, it follows that (at least) $N-1$ of them must cross the horizontal axis somewhere in $w \in (0,1)$. Since, for $z \neq 0$, $T(z) = z^2\widehat{T}(\frac{1}{z})$, $\det[T(z)]$ must have $N-1$ roots in $(1,\infty)$.

5. The Equilibrium Distribution

In Section 4, we have shown that R has $N+K+1$ *different* eigenvalues in the interval $(0,1)$; therefore, the equilibrium distribution can be written as in (8). We order the eigenvalues of R as $0 < r_0 < r_1 < \ldots < r_{N+K} < 1$ (note that r_k is the root of $r_k(z)$ in the unit interval), and construct the diagonal matrix $\Lambda = \mathrm{diag}[r_0, r_1, \ldots, r_{N+K}]$. The corresponding (normalized) eigenvectors $\bar{v}_0, \bar{v}_1, \ldots, \bar{v}_{N+K}$ compose the matrix V, $\bar{v}_k$ being the $k+1$st row of V. Remember that $\bar{v}_k$ can be found as the left null-vector (unique up to multiplication by a scalar) of the matrix $T(r_k)$. We have the (obvious) Jordan decomposition $R = V^{-1}\Lambda V$.

The equilibrium distribution is fully determined as soon as we have $\bar{\pi}_0$, which must satisfy:

$$\bar{\pi}_0[Q_{00} + RT^{(-)}] = \bar{0}. \tag{19}$$

We already mentioned at the end of Section 2 that $Q_{00} + RT^{(-)}$ is an irreducible generator, and therefore (19) has a positive solution, which is unique up to multiplication by a scalar. Obviously, if we let $\mathbf{e}$ be the $(N+K+1)$-dimensional vector with all elements equal to 1, it must be that:

$$\bar{\pi}_0(I-R)^{-1}\mathbf{e} = \bar{\pi}_0 \sum_{j=0}^{\infty} R^j \mathbf{e} = \sum_{j=0}^{\infty} \bar{\pi}_j \mathbf{e} = 1. \tag{20}$$

Together, (19) and (20) completely determine $\bar{\pi}_0$, and therefore $\bar{\pi}$. Since we want to have the $\bar{\pi}_k$ as in (8), or equivalently in matrix form:

$$\bar{\pi}_j = \bar{\alpha}\Lambda^j V, \tag{21}$$

we rewrite (19) and (20) to:

$$\begin{aligned} \bar{\alpha}[VQ_{00} + \Lambda V T^{(-)}] &= \bar{0}, \\ \bar{\alpha}(I-\Lambda)^{-1} V \mathbf{e} &= 1. \end{aligned} \tag{22}$$

This determines $\bar{\alpha} = (\alpha_0, \alpha_1, \ldots, \alpha_{N+K})$.

An alternative way of finding the coefficients α_k in the present model is by using (1). Denoting by $\bar{p}$ the vector $(p_0, p_1, \ldots, p_{N+K})$, with $p_i = \mathbf{P}\{X_H = i\}$ (which are known quantities, see (1)), it must hold that:

$$\bar{\alpha}(I-\Lambda)^{-1} V = \sum_{j=0}^{\infty} \bar{\pi}_j = \bar{p}. \tag{23}$$

In particular, the low priority queue length distribution is given by:

$$\mathbf{P}\{X_L = j\} = \bar{\alpha}\Lambda^j V \mathbf{e} = \sum_{k=0}^{N+K} \alpha_k (r_k)^j \bar{v}_k \mathbf{e}. \tag{24}$$

If we had used the normalization $\bar{v}_k \mathbf{e} = 1$ for the eigenvectors, this would have become:

$$\mathrm{P}\{X_L = j\} = \sum_{k=0}^{N+K} \alpha_k (r_k)^j. \tag{25}$$

However, note that it remains to be verified whether the elements of some $\bar{v}_k$ sum up to 0. If that is the case, the corresponding term in (25) vanishes.

Remark 5.1: From (25), the mean length $\mathrm{E}[X_L]$ and variance $\mathrm{var}[X_L]$ of the low priority queue are easily determined. Using Little's formula we also obtain the mean processing time (or sojourn time) of the low priority customers.

Remark 5.2: When $N_L = 0$, the case $N = 1$ results in an M/M/1 queue with server breakdown and repair (or vacation), which is a known model. Generalizations were analyzed by Neuts [17] and Takagi [24].

6. Numerical Results

In this section, we present some numerical results to illustrate the influence of varying server availability on the performance of low priority traffic. For normalization purposes we choose $\mu_L = 1$, and in all cases we take $N = 17$ (in accordance with data supplied by KPN Research for The Netherlands). Further, we choose the extreme cases where there is no waiting room for high priority customers ($K = 0$), and no reserved capacity for low priority customers ($N_L = 0$).

Before discussing the numerical experiments, we first make some intuitive remarks about the cases when λ_H and μ_H are very large - or very small - compared to λ_L and μ_L. These intuitions can be proved formally. First, we fix λ_L, μ_L, and ρ_H and let μ_H (or equivalently λ_H) go to infinity. Note that with fixed ρ_H, the mean number of servers available to the low priority customers, $N - \mathrm{E}[X_H]$, is also fixed. As $\mu_H \to \infty$, low priority customers are (with respect to the service times of high priority customers) in the system so long, that the mean number of available servers during the sojourn time of a low priority customer will be close to the mean number of available servers in steady state (i.e., close to $N - \mathrm{E}[X_H]$). Therefore, it is to be expected that the low priority traffic in the limit (as $\mu_H \to \infty$) experiences the system as if it were an M/M/1 processor-sharing queue with server capacity $c = N - \mathrm{E}[X_H]$. For the queue length distribution, this model coincides with that of the regular M/M/1 queue with traffic load $\frac{\rho_L}{c}$.

On the other hand, if we let $\mu_H \to 0$ (again for fixed λ_L, μ_L, and ρ_H), the opposite happens: the number of servers available to low priority customers changes very slowly compared to their sojourn times. An arriving low priority customer finding no available server (there are N high priority customers present) must wait until one becomes available before receiving any service. The mean of this waiting time is $\frac{1}{N\mu_H}$, and tends to infinity as $\mu_H \to 0$. Since the probability of finding all servers occupied is positive (and completely determined by ρ_H), the expected sojourn time of the low priority customers also goes to infinity, and by Little's law, so does $\mathrm{E}[X_L]$.

In our experiments, we are interested in the behavior of the mean and variance of the number of low priority customers in the system, at some *fixed system load* $\rho := \rho_L + \mathrm{E}[X_H]$. Therefore, for different values of μ_H and with μ_L normalized to 1, we vary λ_L and λ_H, keeping ρ constant.

In Figures 1 and 2, we have chosen $\rho = \frac{7}{10}N$. We consider three values for μ_H: $\mu_H = \frac{1}{5}$, 1, and 5. Further, we also plot the results for the M/M/1 queue with server capacity $c = N - \mathrm{E}[X_H]$. We have already argued that this model corresponds to the

case $\mu_H = \infty$. Therefore, in this case:

$$\mathbf{P}\{X_L = j\} = \left(1 - \frac{\rho_L}{c}\right)\left(\frac{\rho_L}{c}\right)^j, \; j = 0, 1, 2, \ldots. \tag{26}$$

Since $\mu_L = 1$, in the experiments we vary λ_L from 0 to $\frac{7}{10}N$. At the same time, λ_H decreases such that at all times $\rho = \rho_L + \mathrm{E}[X_H] = \frac{7}{10}N$. In Figures 1 and 2, the mean $\mathrm{E}[X_L]$ and variance $\mathrm{var}[X_L]$ of the number of low priority customers are plotted, respectively. On the horizontal axis λ_L is normalized to ρ_L/ρ.

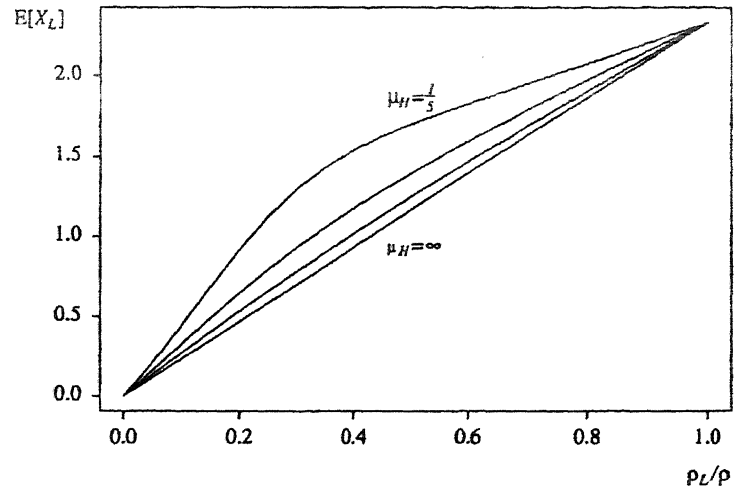

Figure 1

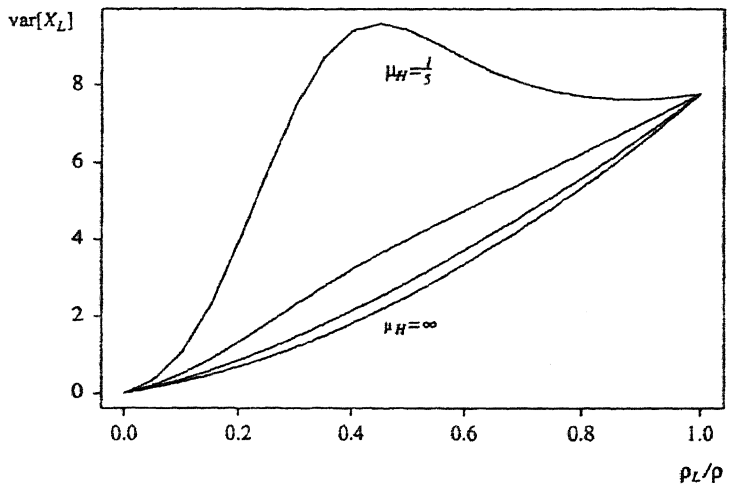

Figure 2

In both figures, the top curve belongs to the case $\mu_H = \frac{1}{5}$, the second to $\mu_H = 1$, the third to $\mu_H = 5$, and the bottom curve to $\mu_H = \infty$. We see that $\mathrm{E}[X_L]$ and $\mathrm{var}[X_L]$ are particularly sensitive to μ_H when ρ_L and $\mathrm{E}[X_H]$ are of the same order.

In Figures 3 and 4, the same procedure is repeated for a system load of $\rho = \frac{9}{10}N$. We see this in this case $\mathrm{E}[X_L]$ and $\mathrm{var}[X_L]$ are more sensitive to μ_H.

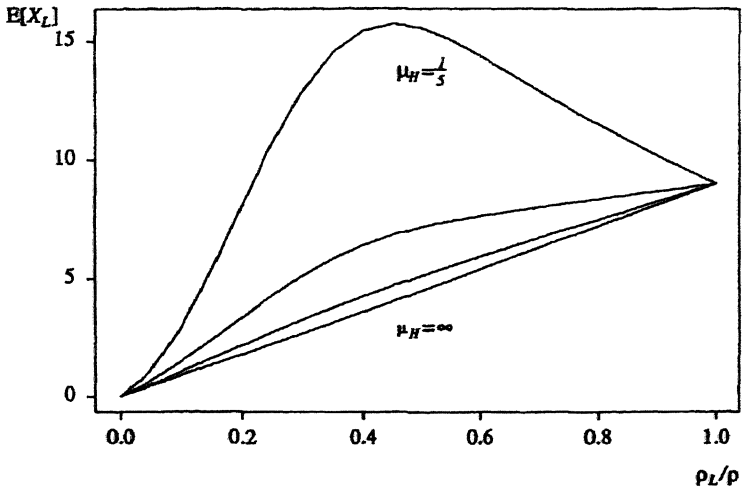

Figure 3

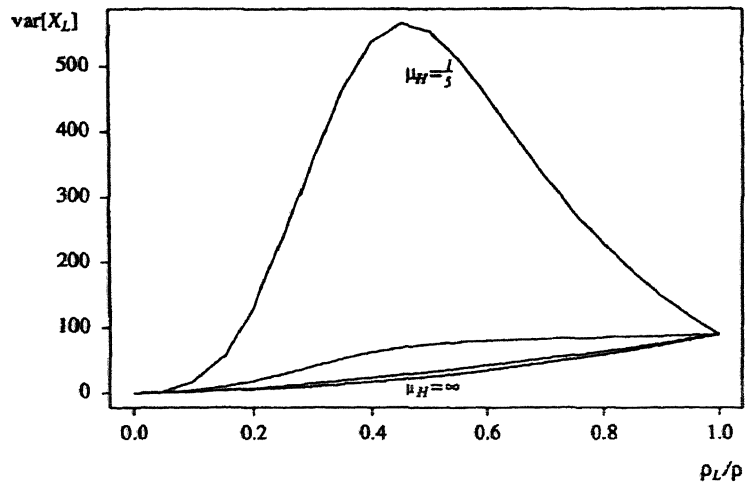

Figure 4

Note that from (26), it follows that in the case $\mu_H = \infty$,

$$\mathrm{E}[X_L] = \frac{\rho_L/c}{1-\rho_L/c} = \frac{\rho_L}{N-\mathrm{E}[X_H]-\rho_L} = \frac{\rho_L}{N-\rho},$$

$$\mathrm{var}[X_L] = \frac{\rho_L/c}{1-\rho_L/c}\left(1+\frac{\rho_L/c}{1-\rho_L/c}\right) = \frac{\rho_L}{N-\rho}\left(1+\frac{\rho_L}{N-\rho}\right).$$

Therefore in Figures 1 and 3, the bottom curve is a straight line; in Figures 2 and 4 the bottom curve is a quadratic curve.

Remark 6.1: We have presented only a small portion of our numerical experiments. The numerical evaluation of this model proved to be fast and stable in many experiments over a wide range of parameter values. No serious problems were encountered in finding the roots of $\det[T(z)]$ inside $(0,1)$. Even when the traffic load is close to the system capacity (say 95%), this problem turned out of be numerically stable. All the roots but the one closest to 1 can be found using a grid method; the last one can then be determined using a bisection method between the second largest root and 1. Also, finding the vectors $\bar{v}_k$ and $\bar{\alpha}$ gives no rise to serious problems, even when the dimension of $T(z)$ is of the order of several hundreds. This numerical stability is partly due to the tridiagonal structure of $T(z)$.

Remark 6.2: Based on the experiments presented in this section, we conclude that approximating the variable server availability by its mean leads to a serious underestimation of: (*i*) $\mathrm{E}[X_L]$ and the mean processing time (by Little's law); and (*ii*) $\mathrm{var}[X_L]$, when μ_H and λ_H are relatively small compared to μ_L and λ_L. This sensitivity is larger when ρ_L and $\mathrm{E}[X_H]$ are of the same order, and when the total system load is larger.

Acknowledgment

The authors are indebted to Dr. J.L. van den Berg (KPN Research) and Dr. I. Norros (VTT) for interesting discussions about the modeling aspects of ABR, and to Professor J.W. Cohen for several discussions and comments.

References

[1] ATM Forum, ATM user-network interface specification 3.1, *ATM Forum Contribution* (September 1994).

[2] ATM Forum, ATM traffic management specification 4.0, *ATM Forum Contribution* **95-0013R7.1** (August 1995).

[3] Blaabjerg, S., Fodor, G., Telek, M. and Andersen, A.T., A partially blocking-queueing system with CBR/VBR and ABR/UBR arrival streams, *Inst. of Telecomm.*, Techn. Univ. of Denmark (internal report).

[4] Cohen, J.W., *The Single Server Queue*, North-Holland Publishing Company, Amsterdam, 2nd edition 1982.

[5] Falin, G., Khalil, Z. and Stanford, D.A., Performance analysis of a hybrid switching system where voice messages can be queued, *Queueing Systems* **16** (1994), 51-65.

[6] Gail, H.R., Hantler, S.L., Konheim, A.G. and Taylor, B.A., An analysis of a class of telecommunications models, *Perf. Evaluation* 21 (1994), 151-161.

[7] Gail, H.R., Hantler, S.L. and Taylor, B.A., Analysis of a non-preemptive priority multi-server queue, *Adv. in Applied Probability* **20** (1988), 852-879.

[8] Gail, H.R., Hantler, S.L. and Taylor, B.A., On a preemptive Markovian queue with multiple servers and two priority classes, *Math. of Operations Res.* **17** (1992), 365-391.

[9] Gail, H.R., Hantler, S.L. and Taylor, B.A., Spectral analysis of M/G/1 and G/M/1 type Markov chains, *Adv. in Appl. Probab.* **28** (1996), 114-165.

[10] Gantmacher, F.R., *The Theory of Matrices*, Chelsea Publishing Company, New York 1977.

[11] Gohberg, I., Lancaster, P. and Rodman, L., *Matrix Polynomials*, Academic Press, New York 1982.

[12] Iliadis, I., A new feedback congestion control policy for long propagation delays, *IEEE J. on Selected Areas in Commun.* **13** (1995), 1284-1295.

[13] Marcus, M. and Minc, H., *A Survey of Matrix Theory and Matrix Inequalities*, Allyn and Bacon, Inc., Boston 1964.

[14] Mitrani, I. and Chakka, R., Spectral expansion solution for a class of Markov models: Application and comparison with the matrix-geometric method, *Perf. Evalu.* **23** (1995), 241-260.

[15] Mitrani, I. and King, P.J.B., Multiprocessor systems with preemptive priorities, *Perf. Evalu.* 1 (1981), 118-125.

[16] Mitrani, I. and Mitra, D., A spectral expansion method for random walks on semi-infinite strips, In: *Iterative Methods in Linear Algebra* (ed. by R. Beauwens and P. de Groen), Proc. of the IMACS International Symp., Brussels, Belgium, Elsevier Science Publishers, Amsterdam (1991).

[17] Neuts, M.F., *Matrix-Geometric Solutions in Stochastic Models - An Algorithmic Approach*, The Johns Hopkins University Press, Baltimore, MD 1981.

[18] Núñez-Queija, R., A queueing model with varying service rate for ABR, In: *Lect. Notes in Computer Sci.*, Proc. of the 10th Inter. Confer. on Modeling Tech. and Tools for Comput. Perf. Eval., Palma de Mallorca, Spain (1998), to appear.

[19] Parlett, B.N., *The Symmetric Eigenvalue Problem*, Prentice-Hall, Englewood Cliffs, NJ 1980.

[20] Ritter, M., Steady-state analysis of the rate-based congestion control mechanism for ABR services in ATM networks, *Univ. of Würzburg, Inst. of Compu. Sci., Res. Rep. Series* **114** (1995).

[21] Ritter, M., Network buffer requirements of the rate-based control mechanism for ABR services, *Proc. of IEEE INFOCOM '96*, San Francisco (1996), 1090-1097.

[22] Ritter, M., Analysis of a rate-based control policy with delayed feedback and variable bandwidth availability, *Univ. of Würzburg, Inst. of Compu. Sci., Res. Rep. Series* **133** (1996).

[23] Ritter, M., Analysis of a queueing model with delayed feedback and its application to the ABR flow control, *Univ. of Würzburg, Inst. of Compu. Sci., Res. Rep. Series* **164** (1997).

[24] Takagi, H., *Queueing Analysis - A Foundation of Performance Evaluation. Volume* 1: *Vacation and Priority Systems*, Elsevier Science Publishers, Amsterdam 1991.