Centrum voor Wiskunde en Informatica
REPORT RAPPORT

Analysis of a multi-server queueing model of ABR

R. Nunez Queija and O.J. Boxma

Department of Operations Reasearch, Statistics, and System Theory

BS-R9613 1996

Report BS-R9613
ISSN 0924-0659

CWI
P.O. Box 94079
1090 GB Amsterdam
The Netherlands

Analysis of a Multi-Server Queueing Model of ABR

Rudesindo Núñez Queija[1] and Onno J. Boxma
CWI
P.O. Box 94079, 1090 GB Amsterdam, The Netherlands;

Abstract

In this paper we present a queueing model for the performance analysis of ABR traffic in ATM networks. We consider a multi-channel service station with two types of customers, the first having preemptive priority over the second. The arrivals occur according to two independent Poisson processes and the service times are assumed to be exponentially distributed. Each type-1 customer requires a single server, whereas type-2 customers are served in processor sharing fashion. We give a complete characterization of the joint distribution of the numbers of customers (of both types) in the system in steady state. Numerical results illustrate the effect of the high priority traffic on the service performance of the low priority traffic.

AMS Subject Classification (1991): 60K25, 68M20, 90B12, 90B22.
Keywords & Phrases: Asynchronous Transfer Mode, Available Bit Rate, multi-server queue, priorities, processor sharing.

1 Introduction

The diverse characteristics and service requirements of the different traffic types that are to be carried by ATM (Asynchronous Transfer Mode) networks have led to the definition of different categories of service that should be offered to the users of such a network. We briefly discuss these differences, distinguishing three large categories: *Constant Bit Rate* (CBR) traffic, *Variable Bit Rate* (VBR) traffic and *Available Bit Rate* (ABR) traffic.
CBR traffic requires very small (or no) delays *and* very small (or no) loss. Therefore CBR traffic is offered a *fixed* pre-determined transmission capacity. 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

[1]corresponding author: tel. +31 20 5924168, fax +31 20 5924199, email sindo@cwi.nl

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 the specific problems that arise when transmitting data. For this traffic, losses lead to retransmission of the 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. ABR traffic is stored in very large buffers and only the transmission capacity that is not currently being required by VBR (and CBR) traffic is used for ABR traffic. This has the advantage that ABR traffic gets all the capacity that is left over. For the server this is also convenient: with ABR traffic the resources can be optimally used. 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 the transmission delays.
A feature of ABR is that the available capacity should be shared *fairly* among all ABR users. In queueing models it seems reasonable to incorporate this feature 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. The feedback control can be used to slow down the data sources when the buffers are heavily loaded and an overflow may occur. We refer to [1] and [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 modelling and (feedback) control aspects, see for instance [9] and [14]. In [15] Ritter investigates the problem of dimensioning the buffer for ABR traffic in order to avoid large losses. In [16] Ritter considers the case with delayed 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 the 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 analyse 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 basically is a multi-server queue with two types of customers: The high priority customers (real-time VBR traffic) and low priority customers (ABR traffic). In a slight exaggeration of the features of real-time VBR traffic and ABR traffic, we assume that the high priority customers have no waiting room and each accepted customer is served by a single server; the low priority customers have an infinite waiting room (buffer) and equally share the remaining capacity according to the processor sharing principle. A newly arriving high priority customer is only rejected if the entire capacity is used by other high priority customers, otherwise he is accepted and one server (that is not currently serving another high priority customer) immediately starts serving this customer. Thus the high priority customers have *preemptive priority* over the low priority customers.

We point out that this is a *burst-level* model. In our analysis we will assume that the arrivals occur according to two independent Poisson processes. This assumption is justified in the case that 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. Also the processor sharing among the ABR sources is interesting in the light of *per VC* (Virtual Connection) *queueing*, where sources do not queue behind one another, but each 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 weight factors.

Variants of our model with a finite waiting room for the high priority customers and/or FCFS (First Come First Served) service discipline for the low priority customers can be analyzed in a similar manner. Also the feedback feature of ABR can be captured in a slight modification. See Section 7.

In [5], Gail et al. study a similar model. They allow an infinite waiting space for the high priority customers, and each of the two queues is served according to the FCFS queue discipline. The non-preemptive variant of that model was studied by the same authors in [4]. A discrete-time variant modelled as an $M/G/1$-type Markov Chain is considered in [3]. A more extensive treatment of the spectral analysis of $M/G/1$-type Markov Chains is given in [6].

In our analysis we are inspired by [5], but we make use of methods from other approaches. Instead of transforming the distributions involved into generating functions, the present work focuses directly on the distribution itself. It does so relying mainly on the matrix geometric approach of M.F. Neuts (see [12]) and the spectral expansion approach (see for instance [10] and [11]).

The paper is organized as follows. In Section 2 we give a full description of the model to be analyzed. 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 M.F. Neuts in [12]. In Section 4 we use this 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 in steady state. In Section 6 numerical results are presented that illustrate the effect of the high priority traffic on the service performance of the low priority traffic. Section 7 mentions some model variants and extensions.

2 The model

Consider a service station consisting of N identical servers. High priority customers arrive to the station according to a Poisson process with rate λ_H. If all the servers are occupied by other high priority customers, then the newly arrived high priority customer is rejected and leaves the system without receiving service. If there are less then N other high priority customers currently being served, then a new high priority customer is immediately taken into service. The 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, independently of the 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 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 $\frac{N-i}{j}\mu_L$ (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 behaviour of the numbers of customers in the system of both types.

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 $M/M/N/N$-queue, i.e. for $i = 0, 1, \dots, N$:

$$\mathbf{P}\{X_H = i\} := \lim_{t\to\infty} \mathbf{P}\{X_H(t) = i\} = \frac{(\rho_H)^i/i!}{\sum_{m=0}^{N}(\rho_H)^m/m!}. \qquad (1)$$

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

$$\mathrm{E}[X_H] + \rho_L < N. \qquad (2)$$

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

We define the equilibrium probabilities

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

and partition them into vectors $\overline{\pi}_j := (\pi_{0j}, \pi_{1j}, \ldots, \pi_{Nj})$ of length $N+1$. Note that $\overline{\pi}_j$ is associated with the states in which j low priority customers are present. This partition enables us to write the equilibrium vector as $\overline{\pi} = (\overline{\pi}_0, \overline{\pi}_1, \overline{\pi}_2, \ldots)$. 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 $T^{(+)}$, $T^{(-)}$ and the off-diagonal elements of $T^{(0)}$ 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 off-diagonal elements of the matrix Q_{00} relate to a change in the number of high priority customers when there are no low priority customers. Note that Q_{00} and $T^{(0)}$ are equal in their off-diagonal entries, since the high priority customers 'do not see' the low priority customers. It is not difficult to see that $T^{(+)} = \lambda_L I$, where I is the $(N+1)$-dimensional identity matrix; $T^{(-)}$ is the diagonal matrix $T^{(-)}$ =diag$[N\mu_L, (N-1)\mu_L, \ldots, \mu_L, 0]$ and $T^{(0)}$ is a tri-diagonal matrix with the elements $T^{(0)}_{(i,i+1)} = \lambda_H$, $T^{(0)}_{(i,i)} = -\lambda_H - i\mu_H - \lambda_L - (N-i)\mu_L$, $T^{(0)}_{(i+1,i)} = (i+1)\mu_H$, for all $i = 0, \ldots, N-1$, and $T^{(0)}_{(N,N)} = -N\mu_H - \lambda_L$ (since all arriving high priority customers are rejected when there are already N high priority customers in the system). Finally, note that $Q_{00} = T^{(0)} + T^{(-)}$.

In [12] M.F. Neuts 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. The general ergodicity condition given in *Theorem 1.7.1.* on p. 32 of [12] states that (i) if the matrix R is the minimal nonnegative solution to the equation

$$T^{(+)} + RT^{(0)} + R^2T^{(-)} = 0, \tag{5}$$

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 is not difficult to see 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 [12].
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 $\overline{\pi} = (\overline{\pi}_0, \overline{\pi}_1, \overline{\pi}_2, \ldots)$ satisfying $\overline{\pi}\mathcal{Q} = 0$ has the matrix-geometric form

$$\overline{\pi}_{j+1} = \overline{\pi}_j R, \tag{6}$$

where the matrix R is the minimal nonnegative solution to (5). The relation (6) 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 e.g. [10] and [11]. The essence of this approach is that we can rewrite (6) to the 'spectral expansion' form

$$\overline{\pi}_j = \sum_{k=0}^{N} \alpha_k \left(r_k\right)^j \overline{v}_k, \tag{7}$$

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

$$\overline{\pi}_0 Q_{00} + \overline{\pi}_1 T^{(-)} = \overline{0}, \tag{8}$$

are satisfied. We come back to this in Section 5.
Even if the matrix R has multiple eigenvalues, (7) still holds, as long as the set of all eigenvectors spans the $(N+1)$-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 [7]).

In order for $\overline{\pi} = (\overline{\pi}_0, \overline{\pi}_1, \ldots)$ to be the equilibrium distribution, $\overline{x}_j = \overline{\pi}_j$ must be a solution of

$$\overline{x}_j T^{(+)} + \overline{x}_{j+1} T^{(0)} + \overline{x}_{j+2} T^{(-)} = 0, \tag{9}$$

for all $j = 0, 1, \ldots$.
It follows from (5) that for any vector $\overline{x}_0$ the sequence $\overline{x}_j = \overline{x}_0 R^j$ satisfies (9).

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 so det$[T(r)]$=0. It follows immediately that R is nonsingular, since $T(0) = T^{(+)} = \lambda_L I$ is nonsingular. Therefore we may write

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

again using (5). This is a very useful factorization, since $\det[R - zI]$ is precisely the characteristic polynomial of R (which has degree $N+1$), and $\det[R^{-1}T^{(+)} - zT^{(-)}]$ is also a polynomial in z (of degree N). In Section 4 we show that the zeros of $\det[T(z)]$ -a polynomial of degree $2N+1$- 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 [12]). We shall show that there are $N+1$ of them, and that they are all real. Starting-point of the analysis is (11). We investigate the zeros of $\det[T(z)]$, showing that there are $2N+1$ zeros: $N+1$ zeros in $(0,1)$, one at 1 and $N-1$ in $(1,\infty)$.
Note that $T(z)$ is a tri-diagonal matrix with off-diagonal elements: for $i = 1, 2, \ldots, N$,

$$\begin{aligned} T(z)_{i-1,i} &= \lambda_H z, \\ T(z)_{i,i-1} &= i\mu_H z. \end{aligned} \tag{12}$$

We denote the i^{th} diagonal element $T(z)_{i,i}$ by $t_i(z)$. For $i = 0, 1, \ldots, N-1$,

$$\begin{aligned} t_i(z) &= \lambda_L - \{\lambda_H + i\mu_H + \lambda_L + (N-i)\mu_L\} z + (N-i)\mu_L z^2, \\ t_N(z) &= \lambda_L - (N\mu_H + \lambda_L) z. \end{aligned} \tag{13}$$

As mentioned before, $\det[T(z)]$ is a polynomial in z of degree $2N+1$. We show that all $2N+1$ roots of $\det[T(z)]$ are real and positive. To see this we first note that for real z the matrix $T(z)$ is *similar* to a real symmetric matrix $S(z)$, i.e. there exists a nonsingular matrix D such that $S(z) := DT(z)D^{-1}$ is a real symmetric matrix. In our case we can take D to be the diagonal matrix $\mathrm{diag}[d_0, d_1, \ldots, d_N]$ with

$$d_i = \sqrt{\frac{(\rho_H)^i}{i!}}. \tag{14}$$

The entries of $S(z)$ are then given by $S(z)_{i,i} = t_i(z)$, $S(z)_{i-1,i} = S(z)_{i,i-1} = z\sqrt{i\mu_H\lambda_H}$ and are zero in all other positions. For real $z \neq 0$, $S(z)$ has $N+1$ different real eigenvalues (this follows from the fact that $S(z)$ is tri-diagonal with non-zero elements directly above and directly below the diagonal, see [13]), and since the eigenvalues of $S(z)$ and $T(z)$ 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 $T(z)$ (and $S(z)$) by

$$\lambda_0(z) \leq \lambda_1(z) \leq \ldots \leq \lambda_N(z), \tag{15}$$

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

$$\lambda_0(0) = \lambda_1(0) = \ldots = \lambda_N(0) = \lambda_L. \tag{16}$$

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

$$\lambda_0(1) < \lambda_1(1) < \ldots < \lambda_N(1) = 0. \tag{17}$$

It can be shown that the $\lambda_k(z)$ are continuous functions of z. This is not necessarily true for complex z, but in that case even the definition in (15) makes no sense.
Combining $\lambda_k(0) = \lambda_L > 0$, $\lambda_k(1) < 0$ for $k = 0, 1, \ldots, N-1$, and the continuity of these functions, we find that (for $k < N$) the $\lambda_k(z)$ must cross the horizontal axis (at least) once, as z increases from 0 to 1. We now show that under the ergodicity condition (2), $\lambda_N(z)$ also has a root in (0,1). To do so it is sufficient to show that $\lambda_N(1-) < 0$. First we write

$$\det[T(z)] = (1-z)g(z), \tag{18}$$

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 $1-z$:

$$g(z) = \begin{vmatrix} t_0(z) & \lambda_H z & & & & \lambda_L - N\mu_L z \\ \mu_H z & t_1(z) & \lambda_H z & & & \lambda_L - (N-1)\mu_L z \\ & \ddots & \ddots & \ddots & & \vdots \\ & & (N-2)\mu_H z & t_{N-2}(z) & \lambda_H z & \lambda_L - 2\mu_L z \\ & & & (N-1)\mu_H z & t_{N-1}(z) & \lambda_L - \mu_L z \\ & & & & N\mu_H z & \lambda_L \end{vmatrix}.$$

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

$$\begin{aligned} g(1) &= \mu_L (\mu_H)^N \begin{vmatrix} -\rho_H & 0 & & & & \rho_L - N \\ 1 & -\rho_H & 0 & & & \rho_L - (N-1) \\ & \ddots & \ddots & \ddots & & \vdots \\ & & N-2 & -\rho_H & 0 & \rho_L - 2 \\ & & & N-1 & -\rho_H & \rho_L - 1 \\ & & & & N & \rho_L \end{vmatrix} \\ &= \mu_L (\mu_H)^N \sum_{k=0}^{N} (-1)^{k+N} (\rho_L - (N-k)) (-\rho_H)^k \frac{N!}{k!}. \end{aligned}$$

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

$$\begin{aligned} g(1) &= \mu_L \left(-\mu_H\right)^N N! \left\{ (\rho_L - N) \sum_{k=0}^{N} \frac{(\rho_H)^k}{k!} + \sum_{k=0}^{N} k \frac{(\rho_H)^k}{k!} \right\} \\ &= \mu_L \left(-\mu_H\right)^N N! \sum_{k=0}^{N} \frac{(\rho_H)^k}{k!} \left\{ \rho_L - N + \mathrm{E}[X_H] \right\}. \end{aligned} \tag{19}$$

Under the Ergodicity condition (2), $\mathrm{sign}[g(1)] = (-1)^{N+1}$. Differentiating (18) gives us $\frac{d}{dz}\det[T(z)]|_{z=1} = -g(1)$. Together this gives us

$$\mathrm{sign}\,[T(1-)] = -\mathrm{sign}\left[\frac{d}{dz}\det\,[T(z)]\,|_{z=1}\right] = \mathrm{sign}\,[g(1)] = (-1)^{N+1}. \tag{20}$$

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

Putting all together we have that $\det[T(z)]$ has (at least) $N+1$ different zeros in the interval $(0,1)$. Also $\det[T(1)] = 0$. Subsequently we show that $\det[T(z)]$ has $N-1$ different roots in the interval $(1, \infty)$. We do this by considering another quadratic matrix polynomial $\hat{T}(w)$ defined by

$$\hat{T}(w) := w^2 T^{(+)} + w T^{(0)} + T^{(-)}. \tag{21}$$

Note that for $w \neq 0$: $\hat{T}(w) = w^2 T(\frac{1}{w})$. By using the same type of argument as before we can show that for any real w (including $w = 0$), $\hat{T}(w)$ has $N+1$ different real eigenvalues

$$\hat{\lambda}_0(w) < \hat{\lambda}_1(w) < \ldots < \hat{\lambda}_N(w), \tag{22}$$

which are continuous functions in w. In fact, for all $k = 0, 1, \ldots, N$ and $w \neq 0$, $\hat{\lambda}_k(w) = w^2 \lambda_k(\frac{1}{w})$. In particular $\hat{\lambda}_k(1) < 0$ for $k = 0, 1, \ldots, N-1$, and $\hat{\lambda}_N(1) = 0$. For $w = 0$ and $k = 0, 1, \ldots, N$ we have $\hat{\lambda}_k(0) = k\mu_L$. Using the continuity, we find that each of the functions $\hat{\lambda}_k(w)$ for $k = 1, 2, \ldots, N-1$ vanishes at some point in the interval $(0,1)$. This means that there are (at least) $N-1$ different roots of $\det[\hat{T}(w)]$ lying in the interval (0,1). Or equivalently, $\det[T(z)]$ has (at least) $N-1$ roots for $z \in (1, \infty)$.

Since $\det[T(z)]$ has exactly $2N+1$ roots we have localized them all: $N+1$ of them lie in (0,1), $N-1$ of them lie in $(1, \infty)$ and the point 1 is a root. Moreover, all these roots are different.

It follows from (11) that the roots of $\det[T(z)]$ in (0,1) are exactly the eigenvalues of R.

5 The stable distribution

In Section 4 we have shown that R has $N+1$ *different* eigenvalues in the interval (0,1); therefore the equilibrium distribution can be written as in (7). We order the eigenvalues of R as $0 < r_0 < r_1 < \ldots < r_N < 1$, and construct the diagonal matrix $\Lambda = \text{diag}[r_0, r_1, \ldots, r_N]$. The corresponding (normalised) eigenvectors $\overline{v}_0, \overline{v}_1, \ldots, \overline{v}_N$ compose the matrix V, $\overline{v}_k$ being the $k+1^{\text{st}}$ row of V. We have the (obvious) Jordan decomposition $R = V^{-1}\Lambda V$.

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

$$\overline{\pi}_0 \left[Q_{00} + RT^{(-)} \right] = \overline{0}. \tag{23}$$

We already mentioned at the end of Section 2 that $Q_{00}+RT^{(-)}$ is an irreducible generator, and therefore (23) has a positive solution, which is unique up to multiplication by a scalar. Obviously, if we let $\mathbf{e}$ be the $(N+1)$-dimensional vector with all elements equal to 1, it must be that

$$\overline{\pi}_0 \left(I - R\right)^{-1} \mathbf{e} = \overline{\pi}_0 \sum_{j=0}^{\infty} R^j \mathbf{e} = \sum_{j=0}^{\infty} \overline{\pi}_j \mathbf{e} = 1. \tag{24}$$

Together (23) and (24) completely determine $\overline{\pi}_0$ and therefore $\overline{\pi}$. Since we want to have the $\overline{\pi}_k$ as in (7), or equivalently in matrix form:

$$\overline{\pi}_j = \overline{\alpha}\Lambda^j V, \tag{25}$$

we rewrite (23) and (24) to

$$\begin{aligned} \overline{\alpha} \left[VQ_{00} + \Lambda VT^{(-)} \right] &= \overline{0}, \\ \overline{\alpha} \left(I - \Lambda\right)^{-1} V\mathbf{e} &= 1; \end{aligned} \tag{26}$$

this determines $\overline{\alpha} = (\alpha_0, \alpha_1, \ldots, \alpha_N)$.

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

$$\overline{\alpha} \left(I - \Lambda\right)^{-1} V = \sum_{j=0}^{\infty} \overline{\pi}_j = \overline{p}. \tag{27}$$

In particular the low priority queue length distribution is given by

$$\mathbf{P}\{X_L = j\} = \overline{\alpha}\Lambda^j V\mathbf{e} = \sum_{k=0}^{N} \alpha_k \left(r_k\right)^j \overline{v}_k \mathbf{e}. \tag{28}$$

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

$$\mathbf{P}\{X_L = j\} = \sum_{k=0}^{N} \alpha_k \left(r_k\right)^j. \tag{29}$$

However, note that it remains to be verified whether the elements of some $\overline{v}_k$ sum up to 0. If that would be the case, the corresponding term in (29) would vanish.

Remark 5.1 From (29) 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 immediately obtain the mean processing time (or sojourn time) of the low priority customers.

Remark 5.2 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 in [12] and Takagi in [17]. In the present setting the stable distribution of this model can be analytically determined: $\det[T(z)]$ is then a polynomial of degree 3, and we know that $z = 1$ is a root, which leaves us with a quadratic function. We omit the details.

6 Numerical results

In this section we present some numerical results to illustrate the influence of the varying server availability on the performance of the 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).
If for fixed λ_L, μ_L and ρ_H we let μ_H (or equivalently λ_H) go to infinity, then all low priority customers are (with respect to the high priority customers) so long in the system, that they experience the server availability 'in its steady-state behaviour'. In other words, during the sojourn time of any low priority customer, the mean number of servers available will be $N - \mathrm{E}[X_H]$. 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 $M/M/c$ ($c = N - \mathrm{E}[X_H]$) queue with processor sharing (note that its queue length distribution coincides with that of an $M/M/1$ 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 server availability for the low priority traffic will have a large variance. Whenever there are many servers occupied by high priority customers, it will take a long time (with respect to the low priority traffic) before they become available to the low priority customers.

In the remainder we denote the system load by $\rho := \rho_L + \mathrm{E}[X_H]$. In each of the experiments we keep ρ fixed and for $\mu_H = \frac{1}{5}, 1, 5$ and ∞, we vary ρ_L from 0 to ρ (at the same time ρ_H decreases such that $\mathrm{E}[X_H]$ goes from ρ to 0).

In Figure 1 the average number of low priority customers in the system, $\mathrm{E}[X_L]$, is shown for fixed $\rho = \frac{7}{10}N$ and $\frac{\rho_L}{\rho}$ increasing from 0 to 1. 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 $\mu_H = \infty$,which is the case where there is a fixed server availability $c = N - \mathrm{E}[X_H]$ for the low priority traffic. Note that the bottom curve is a straight line:

$$\mathrm{E}[X_L] = \frac{\rho_L/c}{1 - \rho_L/c} = \frac{\rho_L}{N - \mathrm{E}[X_H] - \rho_L} = \frac{\rho_L}{N - \rho}, \tag{30}$$

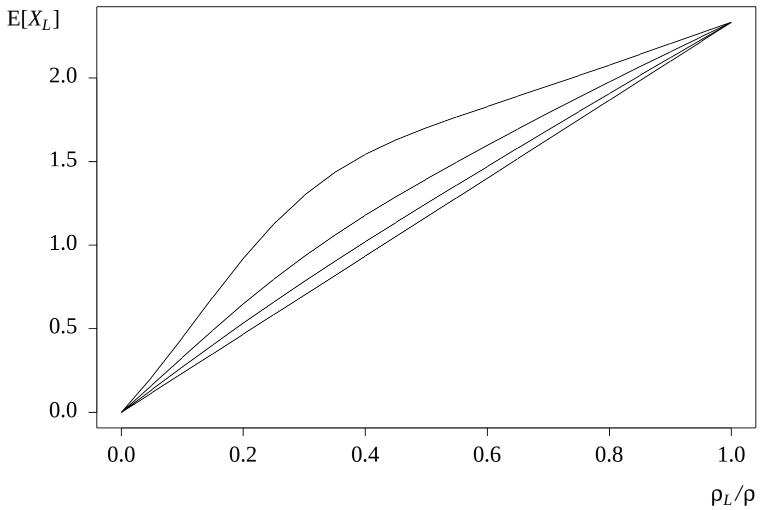

Figure 1: $\rho = \frac{7}{10}N$, $\mu_H = \frac{1}{5}, 1, 5, \infty$.

when μ_H goes to infinity (and $\lambda_H = \rho_H\mu_H$).

In Figure 2 the same is done for var$[X_L]$ with again the property that the larger μ_H, the

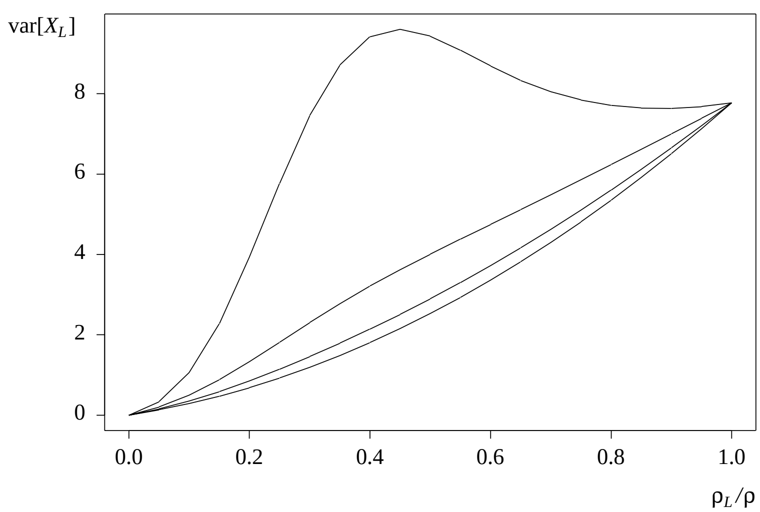

Figure 2: $\rho = \frac{7}{10}N$, $\mu_H = \frac{1}{5}, 1, 5, \infty$.

lower the corresponding curve and the bottom curve again being the limiting situation.

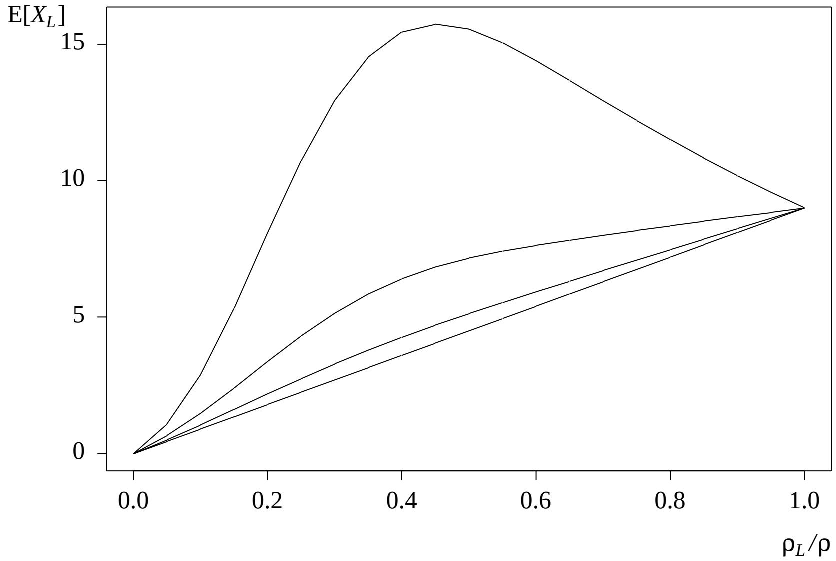

Figure 3: $\rho = \frac{9}{10}N$, $\mu_H = \frac{1}{5}, 1, 5, \infty$.

In Figures 3 and 4 the same procedure is repeated for a system load of $\rho = \frac{9}{10}N$. We see that in this case the experienced effects are stronger.

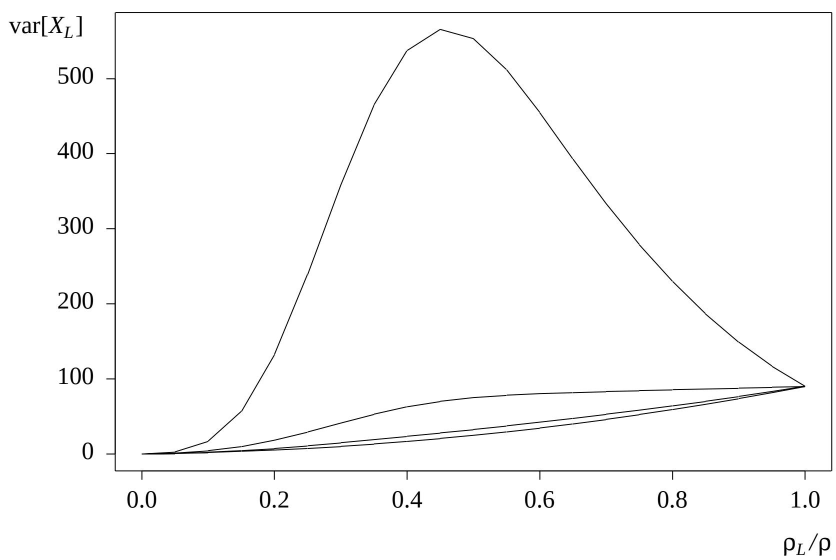

Figure 4: $\rho = \frac{9}{10}N$, $\mu_H = \frac{1}{5}, 1, 5, \infty$.

Our main conclusions, based on the experiments, are that (i) it is easy to numerically evaluate the system under consideration, and that (ii) a relatively small value of μ_H leads to a large variation of the server availability for low priority traffic and hence large values of $\mathrm{E}[X_L]$ and $\mathrm{var}[X_L]$, particularly if ρ_L and $\mathrm{E}[X_H]$ are of the same order.

7 Variants and extensions

The presented model can be modified or extended to capture more realistic features, which we are currently investigating. We briefly discuss some of them in this section.

Variant 7.1 The analysis does not change essentially if we introduce a finite waiting room of size K for the high priority customers. This results in an $M/M/N/N+K$ queue for the high priority customers. The submatrices $T^{(+)}$, $T^{(0)}$ and $T^{(-)}$ become of dimension $N+1+K$, and their elements are such that whenever $i \in \{N, N+1, \ldots, N+K\}$, where i is again the number of high priority customers present, then the low priority customers receive no service and the high priority customers are served at rate $N\mu_H$. Further, a newly arriving high priority customer is admitted iff $i < N+K$.

Variant 7.2 If the low priority customers are served according to the FCFS queue discipline instead of processor sharing, the analysis remains almost unchanged. Using the terminology of level j whenever there are j low priority customers present, we have that beyond the $N-1^{\text{st}}$ level the equilibrium equations do not change, and for the levels $j \in \{0, 1, \ldots, N-1\}$ the departure rate of low priority customers becomes $\mu_L * \min\{N-i, j\}$, where i is, as usual, the number of high priority customers. The distribution remains of the form (25) for $j \geq N$, with exactly the same values for the r_k as in the processor sharing case. The coefficients α_k and the $\overline{\pi}_j$ for $j = 0, 1, \ldots, N-1$ are to be found by solving the remaining finite Markov process.

Variant 7.3 The presented model can also easily be adapted to incorporate a feedback control mechanism, which is important in the context of ABR. Suppose the low priority traffic is permitted to maintain a given arrival rate λ_L^* as long as the number of low priority customers in the system is below a given threshold $J \in \{0, 1, 2, \ldots\}$. But as soon as the level J is reached, the permitted arrival rate drops to λ_L until there are again less than J low priority customers in the system. The relation $\lambda_L < \lambda_L^*$ is of no importance to the analysis, but given the above interpretation, $\lambda_L \geq \lambda_L^*$ makes no sense.
For the homogeneous part of the state space, that is for all $j \geq J$, relation (6) still holds for exactly the same matrix R as before. So the problem is reduced to finding the $\overline{\pi}_0, \overline{\pi}_1, \ldots, \overline{\pi}_J$, which satisfy

$$\begin{aligned} \overline{0} &= \overline{\pi}_0 Q_{00}^* + \overline{\pi}_1 T^{(-)}, \\ \overline{0} &= \overline{\pi}_{j-1} T^{(+)*} + \overline{\pi}_j T^{(0)*} + \overline{\pi}_{j+1} T^{(-)}, j = 1, 2, \ldots, J-1, \\ \overline{0} &= \overline{\pi}_{J-1} T^{(+)*} + \overline{\pi}_J \left[T^{(0)} + R T^{(-)} \right]. \end{aligned} \tag{31}$$

Like in Variant 7.2, these equations define a finite Markov process (with $(J+1)(N+1)$ states).
In fact our analysis can handle a further gradation of arrival rate levels: Instead of taking arrival rate λ_L^* for all $j \in \{0, 1, \ldots, J-1\}$, the more general case of arrival rate $\lambda_L^{(j)}$ at level $j \in \{0, 1, \ldots, J-1\}$ may be considered.

Acknowledgement

The authors are indebted to dr. J.L. van den Berg (KPN Research) and dr. I. Norros (VTT) for interesting discussions about the modelling 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] H.R. Gail, S.L. Hantler, A.G. Konheim, B.A. Taylor. *An analysis of a class of telecommunications models.* Performance Evaluation 21 (1994), 151-161.

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

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

[6] H.R. Gail, S.L. Hantler, B.A. Taylor. *Spectral analysis of $M/G/1$ and $G/M/1$ type Markov chains.* Advances in Applied Probability 28 (1996), 114-165.

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

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

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

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

[11] I. Mitrani, D. Mitra. *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, Proceedings of the IMACS international symposium, Brussels, Belgium (1991). Elsevier Science Publishers B.V., Amsterdam.

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

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

[14] M. Ritter. *Steady-state analysis of the rate-based congestion control mechanism for ABR services in ATM networks.* University of Würzburg, Institute of Computer Science, Research Report Series 114 (1995).

[15] M. Ritter. *Network buffer requirements of the rate-based control mechanism for ABR services.* To appear in IEEE INFOCOM '96, San Francisco.

[16] M. Ritter. *Analysis of a rate-based control policy with delayed feedback and variable bandwidth availability.* University of Würzburg, Institute of Computer Science, Research Report Series 133 (1996).

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