Centrum voor Wiskunde en Informatica

Steady-state analysis of a queue with varying service rate

R. Nunez Queija

Probability, Networks and Algorithms (PNA)

PNA-R9712 1997

Report PNA-R9712
ISSN 1386-3711

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

Steady-State Analysis of a Queue with Varying Service Rate

R. Núñez Queija
CWI
P.O. Box 94079, 1090 GB Amsterdam, The Netherlands
email sindo@cwi.nl

ABSTRACT

In this paper we study a queueing model with a server that changes its service rate according to a finite birth and death process. The object of interest is the simultaneous distribution of the number of customers in the system and the state of the server in steady-state. The model can be applied to the performance analysis of (low priority) Available Bit Rate (ABR) traffic at an ATM switch in the presence of traffic with a higher priority such as Variable Bit Rate (VBR) traffic and Constant Bit Rate (CBR) traffic. For a specific example we illustrate by numerical experiments the influence of the latter traffic types on the ABR service.

1991 Mathematics Subject Classification: 60K25, 68M20, 90B12, 90B22.
Keywords and Phrases: Asynchronous Transfer Mode, Available Bit Rate, two-dimensional Markov process, multiserver queue, priorities, processor sharing, matrix geometric solution, spectral expansion.
Note: work carried out under project ATM in PNA2.1.

1. Introduction

In this paper we are interested in the steady-state behaviour of a system in which customers are served by a service station of which the service capacity varies over time. We model the system as a two-dimensional Markov process and study it through its steady-state distribution. The two components of the process are (i) the number of customers at the service station and (ii) the state of the service station itself. We assume that customers arrive according to a Poisson process and that the service requirements of the customers are drawn from an exponential distribution, independent of the arrival process and of each other. The storage capacity for customers at the station is assumed to be infinite. The state of the service station determines the service speed. We assume further that it changes according to some general finite birth and death process (independent of the arrival process and the service requirements of the customers). Finally, we assume that the customers present at the station are served according to the (egalitarian) *processor sharing* discipline, i.e. that all customers present equally share the available service capacity (see for instance [25] for a review on processor sharing queues). For this model we find the simultaneous steady-state distribution of the number of customers in the system and the state of the service station. We do so using arguments from the theory of *matrix geometric* solutions developed in [17] and the *spectral expansion* approach, see for instance [16] and also [14].
The presented model is a generalisation of the following *priority* model: Suppose that the service station consists of one or more (identical) servers, but that there is another type of (high priority) customers that also require service by the station and have preemptive priority over

the regular (low priority) customers. If the high priority customers also have exponentially distributed interarrival times and service requirements (independent of everything else), and the waiting space for them is finite, then the service capacity available to the low priority customers is determined by the number of high priority customers in the system (which is a finite birth and death process). This special case of the present model is the subject of [18]. In this paper we prove that all the results found in [18] are also valid for the general model. In addition we also prove some new results.
Variants of the priority model of [18] were studied by several authors. The case where both types of customers have an infinite waiting space and within each customer type the service discipline is *First Come First Served* (FCFS), was solved first by Mitrani and King in [15] and later by Gail, Hantler and Taylor in [8]. Gail et al. also studied the non-preemptive case of this model in [7]. Falin, Khalil and Stanford [3] treated the preemptive case with processor sharing among the low priority customers. A discrete-time variant modelled as an *M/G/1*-type Markov chain is considered in [6]. A more extensive treatment of the spectral analysis of *M/G/1*-type Markov chains is given in [9].

Our study was motivated by the introduction of the *Available Bit Rate* (ABR) service in *ATM* (Asynchronous Transfer Mode) networks. Typically in ATM networks there are many different types of connections (customers), of which the *Constant Bit Rate* (CBR) and the *Variable Bit Rate* (VBR) connections are the most established. More recently, the ABR service has been introduced to support data traffic connections more properly than was the case with the other services. One of the most important features of ABR connections is that (apart from some low-level guaranteed transmission rate) they can only be offered the spare capacity that is left over by other type connections (CBR, VBR). In other words, the service capacity available to ABR traffic varies (stochastically) over time. Another important characteristic of ABR traffic is that the available capacity should be shared fairly among ABR users, hence our choice of the processor sharing discipline. Further, the ABR service should guarantee very small cell loss probabilities. To achieve this, large storage buffers for ABR traffic are needed at the switches. This justifies the approximation by an infinite storage capacity. In addition, some feedback control mechanism might notify ABR users (sources) when some of the switches in the network are congested. In Section 7 we formulate an extension of the above model to implement an instantaneous (i.e. with no delays) feedback control mechanism. The assumption of Poisson arrivals appears to be reasonable for ABR traffic at the *burst* level (each 'customer' is identified with a burst of an ABR source, and the service requirement of the customer corresponds to the burst length), particularly when there are many sources connected to the communication network. For more detailed specifications of ABR we refer to [4] and [5]. Already quite some papers have appeared that address the problem of carrying the ABR service in ATM networks. Mostly, the emphasis in these papers lies on the modelling and (feedback) control aspects of ABR, see for instance [12] and [20].
The buffer dimensioning problem for ABR is addressed in [21], and the delayed feedback problem in [22] and [23]. Except for [22], all these studies do not take into account the effect of the varying service availability. In the present paper we concentrate on a specific queueing model in which the service capacity is variable over time and customers are served in (egalitarian) processor sharing fashion. In [1] Blaabjerg et al. consider a model similar to the one in [18] (which is a special case of the present model) and give various performance measures in terms of the steady-state distribution, rather than analysing this distribution in greater

detail. Our main goal is to give a detailed analysis of the steady-state distribution itself.

The same model with FIFO (First In First Out) service discipline can be used to study the performance of *non real-time* VBR traffic in the presence of CBR and *real-time* VBR traffic (which have a higher priority than the non real-time traffic). The analysis of the queue-length behaviour for that model can proceed along the same lines. We do not go into details for this application and further concentrate on the earlier model.

The paper is organised as follows. In Section 2 we present the model. In Section 3 we mention some relevant results from matrix-geometric theory for the steady-state analysis of *GI/M/1*-type Markov chains developed by M.F. Neuts in [17]. We also mention briefly the relation with the spectral expansion approach. We use the results from Section 3 to exploit the structure of the model in Section 4. In Section 5 we give the precise results for the steady-state distribution. In Section 6 we investigate the effect of fast and slowly changing service modes. We present a modification of the model in Section 7 to capture instantaneous feedback control. Numerical experiments for a special case of the model are presented in Section 8. Finally, in Section 9 we make some concluding remarks and mention some current and future research.

2. The model

Consider a service station where customers arrive according to a Poisson process with rate λ. The probability distribution of the service requirement of each customer is assumed to be exponential with mean 1 and independent of everything else. All customers present at the station are served according to the *processor sharing* discipline. I.e. if the station would work at constant rate μ then the model would become the standard *M/M/1* processor sharing queue, and each customer would be served at rate μ/j, whenever there are $j > 0$ customers present. For a review on processor sharing queues we refer to [25].
However, in this paper we allow for the service speed to change according to a birth and death process. More precisely: Let $[Y(t)]_{t\geq 0}$ be a birth and death process on $\{0,1,\ldots,N\}$, N being a positive integer, with birth rate $q_i^{(+)} > 0$ $(i = 0,1,\ldots,N-1)$, and death rate $q_i^{(-)} > 0$ $(i = 1,2,\ldots,N)$, whenever $Y(t) = i$ (for notational convenience we set $q_0^{(-)} = q_N^{(+)} = 0$). We further define $q_i := q_i^{(-)} + q_i^{(+)}$. We assume $Y(t)$ to be independent of the arrival times and service requirements of the customers. The rate of service is determined by the state of the process $Y(t)$ in the following way: the station works at rate $\mu_i > 0$ when $Y(t) = i \in \{0,1,\ldots,N\}$. The restriction $\mu_i > 0$, $\forall i$, is purely for compactness of presentation, the case where some of the μ_i are equal to 0 can be treated in the same manner. We expand a little on this in Remark 4.1. A special case of the present model was treated in [18], see Section 8.

Let $X(t)$ be the number of customers present in the system at time t. Then the process $(X(t),Y(t))$ is an irreducible and aperiodic Markovian process. Moreover, by definition, $Y(t)$

is not influenced by $X(t)$, i.e. if we define $p_i := \mathbf{P}\{Y=i\} := \lim_{t\to\infty} \mathbf{P}\{Y(t)=i\}$:

$$\begin{aligned} p_0 &= \left(1+\sum_{i=1}^{N}\prod_{k=1}^{i}\frac{q_{k-1}^{(+)}}{q_k^{(-)}}\right)^{-1}, \\ p_i &= p_0\prod_{k=1}^{i}\frac{q_{k-1}^{(+)}}{q_k^{(-)}}, \qquad\qquad i=1,\ldots,N, \end{aligned} \tag{2.1}$$

see for instance Part I of [2].
By $\overline{p}$ we denote the vector of these steady-state probabilities :

$$\overline{p} = (p_0, p_1, \ldots, p_N).$$

We define the simultaneous equilibrium probabilities

$$\pi_{j,i} := \mathbf{P}\{X=j, Y=i\} := \lim_{t\to\infty} \mathbf{P}\{X(t)=j, Y(t)=i\}, \tag{2.2}$$

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

$$\mathcal{Q} := \begin{bmatrix} Q_Y-\lambda I & \lambda I & 0 & \ldots & & \\ M & Q_d & \lambda I & 0 & \ldots & \\ 0 & M & Q_d & \lambda I & 0 & \ldots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots \end{bmatrix}. \tag{2.3}$$

The matrices Q_Y, I, M and Q_d are all of dimension $(N+1)\times(N+1)$. I is the identity matrix, M is the diagonal matrix diag$[\mu_0, \mu_1, \ldots, \mu_N]$ and Q_Y is the (tri-diagonal) infinitesimal generator of the process $Y(t)$:

$$Q_Y := \begin{bmatrix} -q_0 & q_0^{(+)} & 0 & \ldots & \ldots & \ldots & 0 \\ q_1^{(-)} & -q_1 & q_1^{(+)} & 0 & \ldots & \ldots & 0 \\ 0 & q_2^{(-)} & -q_2 & q_2^{(+)} & 0 & \ldots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & \ldots & \ldots & \ldots & \ldots & q_N^{(-)} & -q_N \end{bmatrix}. \tag{2.4}$$

Finally, $Q_d = Q_Y - \lambda I - M$.

Using the theory developed by M.F. Neuts in [17] for the *GI/M/1* type of Markov Chains, we have that the process $(X(t), Y(t))$ is ergodic if and only if

$$\lambda < \overline{p}M\overline{\mathbf{e}} = \sum_{i=0}^{N} p_i\mu_i; \tag{2.5}$$

here $\overline{\mathbf{e}}$ is the $N+1$ dimensional vector consisting only of ones: $\overline{\mathbf{e}} = (1, 1, \ldots, 1)$.
In the sequel we assume that (2.5) holds, and exploiting the special structure of our model, we reduce the problem of finding the steady-state distribution to that of finding all the $N+1$ roots inside the interval (0,1) of a polynomial of degree $2(N+1)$.

3. Preliminaries

In the ergodic case, that is when (2.5) holds, the unique probability vector $\overline{\pi} = (\overline{\pi}_0, \overline{\pi}_1, \overline{\pi}_2, \dots)$ satisfying $\overline{\pi}\mathcal{Q} = 0$ has the matrix-geometric form

$$\overline{\pi}_{j+1} = \overline{\pi}_j R, \tag{3.1}$$

where the matrix R has all its eigenvalues *inside* the unit disc, and is the minimal nonnegative solution to the quadratic matrix equation

$$\lambda I + RQ_d + R^2 M = 0, \tag{3.2}$$

(see [17]). The element $[R]_{i,i'}$ is $-[Q_d]_{i,i}$ times the expected time spent in the state $(j+1, i')$ before the first return to the level j, $j \geq 0$, given that the process starts in the state (j, i) (see Section 1.7 of [17]). In particular, we immediately have that R is a strictly positive matrix. In Section 4 we show that, in our model, the matrix R has a full set of eigenvectors, i.e. the set of eigenvectors spans $\mathbb{R}^{N+1}$. When this is the case, we can rewrite (3.1) to the 'spectral expansion' form

$$\overline{\pi}_j = \sum_{k=0}^{N} \alpha_k \, (r_k)^j \, \overline{v}_k. \tag{3.3}$$

Here, $r_0, \dots, r_N$ are the (not necessarily different) eigenvalues of the matrix R and $\overline{v}_0, \dots, \overline{v}_N$ the corresponding left eigenvectors, i.e. $\overline{v}_k R = r_k \overline{v}_k$, $k = 0, 1, \dots, N$. We refer to [16] for another application of the spectral expansion approach in *quasi birth death* models and to [14] for a comparison of this method with the matrix-geometric methods based on [17].

The coefficients α_k in (3.3) are to be chosen such that the boundary equations

$$\overline{\pi}_0 \left[Q_Y - \lambda I \right] + \overline{\pi}_1 M = \overline{0}, \tag{3.4}$$

are satisfied. We come back to this in Section 5.
We remark that if R does not have a full set of eigenvectors (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 [10]).

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

$$T(z) := \lambda I + z Q_d + z^2 M. \tag{3.5}$$

Note that if $\overline{v}$ is an eigenvector of the matrix R corresponding to the eigenvalue r, then $\overline{v}$ is in the left nullspace of the matrix $T(r)$ (this can be seen by pre-multiplying (3.2) by $\overline{v}$), and so det$[T(r)]$=0. It follows immediately that R is nonsingular, since $T(0) = \lambda I$ is nonsingular. Therefore, using (3.2), we may write

$$T(z) = (R - zI)\, R^{-1} \lambda I \,(I - zG)\,, \tag{3.6}$$

where $G = \frac{1}{\lambda} RM$.
It can be shown, by probabilistic arguments, that the element $[G]_{i,i'}$ is the probability that

(for any j) starting with $j+1$ customers and the server in state i, eventually the number of customers *becomes* j at a moment that the server is in state i'. Since we assumed that the Markov chain was ergodic, the matrix G is stochastic. Further, using the above probabilistic interpretation, it can be argued that the matrix G satisfies

$$\lambda G^2 + Q_d\, G + M = 0.$$

For more on the matrix G see [17].
The factorisation (3.6) is very useful, since $\det[R - zI]$ is precisely the characteristic polynomial of R, and for $z \neq 0$: $\det[I - zG] = z^{N+1}\det[\frac{1}{z}I - G]$. Both these factors of $\det[T(z)]$ are polynomials of degree $N+1$, therefore $\det[T(z)]$ is of degree $2(N+1)$. We also have immediately that, in the ergodic case, the zeros of $\det[T(z)]$ *inside* the complex unit disk coincide with the eigenvalues of R, and that the other zeros coincide with the eigenvalues of G^{-1}. In Section 4 we show that all zeros of $\det[T(z)]$ are positive reals, and hence all eigenvalues of R and G^{-1} (and also G) are positive reals.
Note that if some of the μ_i are zero, then G is singular, and the degree of the polynomial $\det[I - zG]$ becomes smaller than $N+1$. In Remark 4.1 we come back to this.

4. Spectral analysis

In this section we investigate the eigenvalues of R (and G^{-1}) through the roots of the polynomial $\det[T(z)]$. We show that all these roots are real and positive. We also show that in the ergodic case there are (indeed) $N+1$ of them in the interval (0,1), one at the point $z=1$ and N in the interval $(1,\infty)$.

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

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

$$\begin{aligned} T(z)_{i-1,i} &= q_{i-1}^{(+)} z, \\ T(z)_{i,i-1} &= q_i^{(-)} z, \end{aligned}$$

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

$$t_i(z) := \lambda - \{q_i + \lambda + \mu_i\}\, z + \mu_i z^2,$$

here $i = 0, 1, \ldots, N$. Note that for real z the matrix $T(z)$ is *similar* to a real symmetric matrix, i.e. there exists a nonsingular matrix D such that $DT(z)D^{-1}$ is a real symmetric matrix. For instance we can (and will) take D to be the diagonal matrix $\mathrm{diag}[d_0, d_1, \ldots, d_N]$ with $d_i = \sqrt{\frac{p_i}{p_0}}$. The p_i are given in (2.1). We define $S(z) := DT(z)D^{-1}$. The entries of $S(z)$ are then given by $[S(z)]_{i,i} = t_i(z)$, $[S(z)]_{i-1,i} = [S(z)]_{i,i-1} = z\sqrt{q_{i-1}^{(+)} q_i^{(-)}}$ and are zero in all other positions.
The eigenvalues of $T(z)$ and $S(z)$ coincide, and hence it remains to prove the assertions for $S(z)$. The fact that for real $z \neq 0$, $S(z)$ has $N+1$ different real eigenvalues, can be seen as follows (see also [19]): First, every eigenvalue of a real symmetric matrix is real. Second,

any real symmetric matrix has a full set of eigenvectors, therefore if $S(z)$ has an eigenvalue τ with (algebraic) multiplicity larger than 1 then there must be (at least) two independent eigenvectors corresponding to τ. But $S(z)$ is tri-diagonal with non-zero elements directly above and directly below the diagonal, and so each eigenvalue has a unique corresponding eigenvector (up to multiplication by a scalar). □

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, using Theorem 4.1, for real z, we may denote the eigenvalues of $T(z)$ by

$$\tau_0(z) < \tau_1(z) < \cdots < \tau_N(z), \qquad z \neq 0, \tag{4.1}$$

and

$$\tau_0(0) = \tau_1(0) = \cdots = \tau_N(0) = \lambda. \tag{4.2}$$

Theorem 4.2 *All eigenvalues $\tau_k(z)$, $k = 0, 1, \ldots, N$, are continuous functions of $z \in \mathbb{R}$.*

Proof. Let z_0 be an arbitrary real number, and let $(z_n^{(1)})_{n\in\mathbb{N}}$ be any row of real numbers converging to z_0:

$$\lim_{n\to\infty} z_n^{(1)} = z_0.$$

First we concentrate on one $\tau_k(z)$, $k = 0, 1, \ldots, N$. Of course, $\lim_{n\to\infty} \tau_k(z_n^{(1)})$ may not exist but the row $\left(\tau_k(z_n^{(1)})\right)_{n\in\mathbb{N}}$ does have at least one density point. This follows from Gerŝgorin's theorem: Each eigenvalue must be in at least one of the $N+1$ Gerŝgorin discs (see for instance [13]). Each Gerŝgorin disc in the complex plane corresponds to a row of the matrix: The diagonal element in the row is the center of the disc and the radius is equal to the sum of the absolute values of the off-diagonal elements in the row.
Therefore

$$|\tau_k(z) - t_k(z)| \leq q_k|z|,$$

and so $\left(\tau_k(z_n^{(1)})\right)_{n\in\mathbb{N}}$ is contained in a bounded set (for any choice of z_0).
Let $l_k \in \mathbb{R}$ be a density point of $\left(\tau_k(z_n^{(1)})\right)$ and let $(z_n^{(2)})_{n\in\mathbb{N}}$ be a subrow of $(z_n^{(1)})$ such that

$$\lim_{n\to\infty} \tau_k(z_n^{(2)}) = l_k.$$

Repeating the same procedure N more times we can find a subrow $(z_n)_{n\in\mathbb{N}}$ of $(z_n^{(2)})$ such that

$$\lim_{n\to\infty} \tau_i(z_n) = l_i,$$

for all $i = 0, 1, \ldots, N$ and certain $l_i \in \mathbb{R}$. Note that the l_i for $i \neq k$, may depend on the choice of l_k, which was an *arbitrary* density point of $\left(\tau_k(z_n^{(1)})\right)$.
Since for all $n \in \mathbb{N}$,

$$\tau_0(z_n) \leq \tau_1(z_n) \leq \cdots \leq \tau_N(z_n),$$

also

$$l_0 \leq l_1 \leq \cdots \leq l_N.$$

Further, for arbitrary τ,

$$\begin{aligned}
\det\left[T(z_0) - \tau I\right] &= \det\left[\lim_{n\to\infty} T(z_n) - \tau I\right] = \lim_{n\to\infty} \det\left[T(z_n) - \tau I\right] \\
&= \lim_{n\to\infty} \prod_{i=0}^{N} \left(\tau_i(z_n) - \tau\right) = \prod_{i=0}^{N} \left(\lim_{n\to\infty} \tau_i(z_n) - \tau\right) = \prod_{i=0}^{N} \left(l_i - \tau\right).
\end{aligned}$$

We now may conclude that $l_i = \tau_i(z_0)$. Combining this for $i = k$ with the arbitrariness of the density point l_k, we conclude that $\lim_{n\to\infty} \tau_k(z_n^{(1)})$ exists and is equal to $\tau_k(z_0)$, therefore $\tau_k(z)$ is a continuous function.
Since we took $k = 0, 1, \ldots, N$ arbitrarily, all the $\tau_k(z)$ are continuous. □

Theorem 4.3 $\tau_0(1) < \tau_1(1) < \cdots < \tau_N(1) = 0.$

Proof. It is clear that $\det[T(1)] = 0$, since the rows of $T(1)$ sum to 0. Furthermore, all the eigenvalues of $T(1)$ are nonpositive. This can be seen using again Gerŝgorin's theorem. Since (i) the diagonal elements of $T(1)$ are negative reals, (ii) the off-diagonal elements are nonnegative reals, (iii) all rows sum to 0 and (iv) the eigenvalues are real, all eigenvalues must be nonpositive. This combined with $\det[T(1)] = 0$ and (4.1) completes the proof. □

Corollary 4.1 *For $k = 0, 1, \ldots, N-1$ the equation $\tau_k(z) = 0$ has (at least) one solution for $z \in (0, 1)$ and (at least) one solution for $z \in (1, \infty)$.*

Proof. The roots in (0,1) follow immediately from (4.2), Theorem 4.3 and the continuity of the $\tau_k(z)$: Each of the $\tau_k(z)$, for $k = 0, 1, \ldots, N-1$, must cross the horizontal axis (at least once) somewhere in (0,1).
If z increases to infinity the matrix $T(z)$ becomes strictly diagonally dominant (for each row the absolute value of the diagonal element exceeds the sum of the absolute values of the other entries in the row) with positive diagonal elements (the diagonal elements are convex quadratic functions of z and the off-diagonal elements are linear in z), and so (again by Gerŝgorin's theorem) for z large enough, all the eigenvalues of $T(z)$ are positive. Therefore all the $\tau_k(z)$ for $k = 0, 1, \ldots, N-1$ must cross the horizontal axis again somewhere in $(1, \infty)$.
□

Theorem 4.4 *Under the ergodicity condition (2.5), $\tau_N(z) = 0$ for some $z \in (0, 1)$.*

Proof. It is sufficient to show that $\tau_N(1-) < 0$ or equivalently, that $\mathrm{sign}[\tau_N(1-)] = -1$. Then, by $\tau_N(0) = \lambda > 0$ and the continuity of $\tau_N(z)$, $\tau_N(z)$ must cross the horizontal axis somewhere in the interval (0,1).

Since $\det[T(1-)] = \prod_{k=0}^{N} \tau_k(1-)$, also $\mathrm{sign}[\det[T(1-)]] = \prod_{k=0}^{N} \mathrm{sign}\,[\tau_k(1-)]$. By Theorem 4.3 and the continuity of the $\tau_k(z)$, for $k = 0, 1, \dots, N-1$ we have: $\mathrm{sign}\,[\tau_k(1-)] = \mathrm{sign}\,[\tau_k(1)] = -1$. We now show that $\mathrm{sign}[\det[T(1-)]] = (-1)^{N+1}$. First we write

$$\det[T(z)] = (1-z)g(z), \tag{4.3}$$

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 $1-z$:

$$g(z) = \begin{vmatrix} t_0(z) & q_0^{(+)}z & & & & \lambda-\mu_0 z \\ q_1^{(-)}z & t_1(z) & q_1^{(+)}z & & & \lambda-\mu_1 z \\ & \ddots & \ddots & \ddots & & \vdots \\ & & \ddots & \ddots & \ddots & \vdots \\ & & & q_{N-1}^{(-)}z & t_{N-1}(z) & \lambda-\mu_{N-1}z \\ & & & & q_N^{(-)}z & \lambda-\mu_N z \end{vmatrix},$$

(all non-specified entries are zero).
We want to show that $\mathrm{sign}\,[g(1-)] = (-1)^{N+1}$. Therefore we evaluate $g(1)$ by manipulating the above matrix evaluated in $z=1$. First add to each column (except for the first and the last one) all columns to the left of it. We now have

$$\begin{aligned} g(1) &= \begin{vmatrix} -q_0^{(+)} & 0 & & & & \lambda-\mu_0 \\ q_1^{(-)} & -q_1^{(+)} & 0 & & & \lambda-\mu_1 \\ & \ddots & \ddots & \ddots & & \vdots \\ & & q_{N-2}^{(-)} & -q_{N-2}^{(+)} & 0 & \vdots \\ & & & q_{N-1}^{(-)} & -q_{N-1}^{(+)} & \lambda-\mu_{N-1} \\ & & & & q_N^{(-)} & \lambda-\mu_N z \end{vmatrix} \\ &= \sum_{i=0}^{N} (-1)^{N+i}(\lambda-\mu_i) \prod_{k=0}^{i-1}(-q_k^{(+)}) \prod_{k=i+1}^{N} q_k^{(-)}. \end{aligned}$$

The last equality follows by expanding the determinant in its last column. Using (2.1) we rewrite this to

$$\begin{aligned} g(1) &= (-1)^N \sum_{i=0}^{N} (\lambda-\mu_i)\frac{p_i}{p_0} \prod_{k=0}^{N} q_k^{(-)} \\ &= (-1)^N \frac{\prod_{k=0}^{N} q_k^{(-)}}{p_0} \left(\lambda - \overline{p}M\overline{\mathbf{e}}\right). \end{aligned}$$

Under the Ergodicity condition (2.5), $\mathrm{sign}[g(1)] = (-1)^{N+1}$. This completes the proof. □

Theorem 4.5 *All roots of* $\det[T(z)]$ *are different.* $N+1$ *of them lie in (0,1), one at* $z=1$ *and* N *in* $(1,\infty)$.

Proof. By Theorems 4.3 and 4.4 and Corollary 4.1 we have found $2(N+1)$ roots of $\det[T(z)]$ with the required positions. Since the degree of $\det[T(z)]$ is $2(N+1)$, these are all the roots. □

As we remarked before, the roots of $\det[T(z)]$ inside the interval (0,1) are precisely the eigenvalues of R (and the roots in $[1,\infty)$ are precisely the eigenvalues of G^{-1}). Since the eigenvalues of R (resp. G) are all different, we also have immediately that the set of eigenvectors of R (resp. G) is a basis for $\mathbb{R}^{N+1}$.

Remark 4.1 When one or several of the μ_i are equal to zero, $\det[T(z)]$ is no longer of degree $2(N+1)$. It becomes of degree $2(N+1)-n_0$, where n_0 is the number of states i of the process $Y(t)$ for which $\mu_i = 0$. In that case $\det[T(z)]$ still has $N+1$ roots in (0,1) and one at the point $z=1$, but the number of zeros in $(1,\infty)$ decreases to $N-n_0$. This can be proved using the same arguments as in the analysis with strictly positive service rates.
The $N+1$ roots in (0,1) assure that the spectral expansion form (3.3) still applies.

Remark 4.2 The case $N=1$ results in an $M/M/1$ queue with server breakdown and repair (or vacation), which is a known model. Generalisations were analysed by Neuts in [17] and Takagi in [24]. 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.

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 (3.3). We order the eigenvalues of R as $0 < r_0 < r_1 < \cdots < r_N < 1$, and construct the diagonal matrix $\Lambda = \mathrm{diag}[r_0, r_1, \ldots, r_N]$. 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 (3.4), or using (3.1) for $j=0$,

$$\overline{\pi}_0\,[Q_Y - \lambda I + RM] = \overline{0}. \tag{5.1}$$

Next we observe that $[Q_Y - \lambda I + RM]\,\overline{\mathbf{e}} = \overline{0}$. To see this, we first post-multiply (3.2) by $\overline{\mathbf{e}}$ and use $Q_Y\overline{\mathbf{e}} = \overline{0}$ to conclude that also $(I-R)(\lambda I - RM)\overline{\mathbf{e}} = \overline{0}$. Since $I-R$ is nonsingular, also $(\lambda I - RM)\overline{\mathbf{e}} = \overline{0}$.
Since Q_Y is the generator of a Markov chain and R is a nonnegative matrix, also the *off-diagonal* entries of $[Q_Y - \lambda I + RM]$ must be nonnegative. Therefore $[Q_Y - \lambda I + RM]$ is (also) the generator of a Markov chain, and because of the structure of Q_Y it is easy to see that it is an *irreducible* generator. So (5.1) has a positive solution, which is unique up to multiplication by a scalar. Obviously, it must be that

$$\overline{\pi}_0\,(I-R)^{-1}\,\overline{\mathbf{e}} = \overline{\pi}_0 \sum_{j=0}^{\infty} R^j \overline{\mathbf{e}} = \sum_{j=0}^{\infty} \overline{\pi}_j \overline{\mathbf{e}} = 1. \tag{5.2}$$

Together (5.1) and (5.2) completely determine $\overline{\pi}_0$ and therefore $\overline{\pi}$. Since we want to have the $\overline{\pi}_k$ as in (3.3), or equivalently in matrix form:

$$\overline{\pi}_j = \overline{\alpha}\Lambda^j V, \tag{5.3}$$

we rewrite (5.1) and (5.2) to

$$\begin{aligned} \overline{\alpha}\left[V(Q_Y - \lambda I) + \Lambda V M\right] &= \overline{0}, \\ \overline{\alpha}\left(I - \Lambda\right)^{-1} V\overline{\mathbf{e}} &= 1; \end{aligned} \tag{5.4}$$

which uniquely 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 (2.1). If we sum the $\overline{\pi}_j$ over all $j \geq 0$ we get the marginal distribution of the state of the server:

$$\overline{\alpha}\left(I - \Lambda\right)^{-1} V = \sum_{j=0}^{\infty} \overline{\pi}_j = \overline{p}. \tag{5.5}$$

In particular, the marginal queue length distribution is given by

$$\mathbf{P}\{X = j\} = \overline{\alpha}\Lambda^j V\overline{\mathbf{e}} = \sum_{k=0}^{N} \alpha_k \left(r_k\right)^j \overline{v}_k\overline{\mathbf{e}}. \tag{5.6}$$

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

$$\mathbf{P}\{X = j\} = \sum_{k=0}^{N} \alpha_k \left(r_k\right)^j. \tag{5.7}$$

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 (5.7) would vanish.

Remark 5.1 From (5.7) the moments of the number of customers in the system are easily determined, in particular the mean $\mathbb{E}[X]$ and the variance $\mathrm{var}[X]$. Using Little's formula we immediately obtain the mean processing time (or sojourn time).

6. Fast- and slowly-changing service modes

If the service capacity changes very fast compared to the rate at which customers arrive (and are served), we might expect the 'queue'-length process to behave as the 'queue'-length process in the regular *M/M/1* (processor-sharing) model with *constant* service capacity (we use the word 'queue' between quotation marks because in processor-sharing models there is no real queue: all customers are served at the same time). Also we might expect the 'queue'-length process to become highly unstable when the capacity changes very slowly.
Next we formalise these two statements in two theorems, but first we introduce some new notation. Let

$$Q_Y(s) := sQ_Y, \qquad s \in (0, \infty).$$

Note that – for any *time-scale parameter* $s \in (0,\infty)$ – $Q_Y(s)$ is the generator of a Markov chain having the same structure as the process $Y(t)$, only the time-scale has changed: transitions occur s times faster. We will simply take $Q_Y(s)$ as the new generator for the process $Y(t)$ (instead of introducing a new process $Y_s(t)$. In particular, the steady-state probability distribution (2.1) of the process $Y(t)$ is independent of s, and so is the ergodicity condition (2.5).
Generalising the matrix $T(z)$, we define the matrix $T_s(z)$, but we write it slightly different (by substituting $Q_d = Q_Y(s) - \lambda I - M$):

$$T_s(z) := (1-z)\lambda I + zsQ_Y + z(z-1)M.$$

All the results we have proved in Section 4 for $T(z)$ and its eigenvalues remain true for $T_s(z)$ as a function of z, when $s \in (0,\infty)$ (we emphasise the fact that the case $s=0$ is *not* contained). For $s \in (0,\infty)$ we also define the eigenvalues of $T_s(z)$

$$\tau_{s,0}(z) \le \tau_{s,1}(z) \le \cdots \le \tau_{s,N}(z), \qquad z \in \mathbb{R},$$

and the roots of $\det[T_s(z)]$ in (0,1)

$$r_{s,0} < r_{s,1} < \cdots < r_{s,N}.$$

In the theorems we will also need the following permutation of the μ_i in *decreasing* order:

$$\mu_{[0]} \ge \mu_{[1]} \ge \cdots \ge \mu_{[N]} \ge 0.$$

Lemma 6.1 *The eigenvalues $\tau_{s,k}(z)$, $k = 0,1,\ldots,N$, are continuous functions of $s \in (0,\infty)$. Moreover,*

$$\lim_{s\to 0} \tau_{s,k}(z) = \begin{cases} \lambda - (\lambda + \mu_{[k]})z + \mu_{[k]}z^2 & \text{if } z \in [0,1), \\ \lambda - (\lambda + \mu_{[N-k]})z + \mu_{[N-k]}z^2 & \text{if } z \in [1,\infty), \end{cases}$$

and, for $z \in [0,\infty)$,

$$\lim_{s\to\infty} \frac{\tau_{s,k}(z)}{s} = z\tau_{1,k}(1).$$

Proof. To prove that the $\tau_{s,k}(z)$ are continuous functions of s and that the limits exist, we can mimic the arguments from Theorem 4.2. Because of the lengthy notation we will leave these details to the reader.
To find $\lim_{s\to 0} \tau_{s,k}(z)$, we can then plug in $s=0$ in $T_s(z)$ (because of the continuity), which becomes a diagonal matrix. Taking into account the ordering of the eigenvalues, we then have the required limits immediately.
In the same way we can use

$$\lim_{s\to\infty} \frac{T_s(z)}{s} = zQ_Y,$$

and $Q_Y = T_1(1)$, to find $\lim_{s\to\infty} \frac{\tau_{s,k}(z)}{s} = z\tau_{1,k}(1)$. □

The next theorem says that when the service modes change very slowly ($s \to 0$), the 'queue'-length process becomes arbitrarily unstable when in at least one of the states of the process $Y(t)$ the 'queue'-length process has a nonnegative drift. When the 'queue'-length process has a negative drift for all states of the process $Y(t)$, then the *smallest* μ_i provides a bound on the stability of the 'queue'-length process.

Theorem 6.1 *For* $k \in \{0, 1, \ldots, N\}$*:*

$$\lim_{s \to 0} r_{s,k} = \frac{\lambda + \mu_{[k]} - \sqrt{\left(\lambda + \mu_{[k]}\right)^2 - 4\lambda\mu_{[k]}}}{2\mu_{[k]}} = \begin{cases} \dfrac{\lambda}{\mu_{[k]}}, & \textit{if } \mu_{[k]} > \lambda, \\ 1, & \textit{if } \mu_{[k]} \leq \lambda. \end{cases}$$

Proof. Using Lemma 6.1 and the ordering of the eigenvalues $\tau_{s,k}(z)$ it can be shown that each $\tau_{s,k}(z)$ 'takes its roots with it' as $s \to 0$. This gives the required limits for this theorem. Again we omit the lengthy notation needed for an exact proof and refer the interested reader to the proof of Theorem 4.2 to fill in the details. □

The following theorem says that if the service mode changes very rapidly, then the 'queue'-length process behaves as the regular *M/M/1* (processor-sharing) model with a fixed service rate $1/\sum_{i=0}^{N} p_i\mu_i$. In fact this means that the model starts to behave as if the service rate is fixed at the average service rate.

Theorem 6.2 *For all* $k = 0, 1, \ldots, N-1$

$$\lim_{s \to \infty} r_{s,k} = 0,$$

and

$$\lim_{s \to \infty} r_{s,N} = \frac{\lambda}{\sum_{i=0}^{N} p_i\mu_i}.$$

Proof. Remember that in Theorem 4.3 we proved that

$$\tau_{1,0}(1) < \tau_{1,1}(1) < \cdots < \tau_{1,N}(1) = 0. \tag{6.1}$$

Combining this with the result found in Lemma 6.1 for $s \to \infty$ we can conclude that for $k = 0, 1, \ldots, N-1$ the zero $r_{s,k}$ of $\tau_{s,k}(z)/s$ in (0,1) goes to zero as s goes to infinity (in fact the zero of $\tau_{s,k}(z)/s$ in $(1, \infty)$ goes to infinity). Here we use again (like in the proof of Theorem 6.1) that the functions $\tau_{s,k}(z)/s$ 'take their zeros with them' as $s \to \infty$.
Noting that the zeros of $\tau_{s,k}(z)/s$ as a function of z coincide with the zeros of $\tau_{s,k}(z)$ we have proved the first part of the theorem. To prove the result for $r_{s,N}$ we need to introduce $\alpha_{s,k}$, for $k = 0, 1, \ldots, N$, and $\overline{\pi}_{s,j}$, for $j = 0, 1, 2, \ldots$, as the analogues of α_k and $\overline{\pi}_j$ in the model with time-scale parameter s. We also use the matrix V_s with rows $\overline{v}_{s,k}$, the diagonal matrix Λ_s with diagonal entries $r_{s,k}$, $k = 0, 1, \ldots, N$ and the rate-matrix R_s. The first part of equation (5.4) becomes

$$\overline{\alpha}_s \left[V_s(sQ_Y - \lambda I) + \Lambda_s V_s M\right] = \overline{0},$$

and determines $\overline{\alpha}_s = (\alpha_{s,0}, \alpha_{s,1}, \dots, \alpha_{s,N})$ up to multiplication by a scalar (which can be found using the second part of equation (5.4)).
Now we use the following observation: In steady-state the frequency of transitions from level j to level $j+1$, $j = 0, 1, 2, \dots$, must equal the frequency of transitions from level $j+1$ to level j, i.e.

$$\overline{\pi}_{s,j} \lambda I \overline{\mathbf{e}} = \overline{\pi}_{s,j+1} M \overline{\mathbf{e}}.$$

We can write

$$\begin{aligned} 1 &= \lim_{s\to\infty} \lim_{j\to\infty} \frac{\overline{\pi}_{s,j+1} M \overline{\mathbf{e}}}{\overline{\pi}_{s,j} \lambda I \overline{\mathbf{e}}} \\ &= \lim_{s\to\infty} \lim_{j\to\infty} \frac{\sum_{k=0}^{N} \alpha_{s,k} (r_{s,k})^{j+1} \overline{v}_{s,k} M \overline{\mathbf{e}}}{\lambda \sum_{k=0}^{N} \alpha_{s,k} (r_{s,k})^{j} \overline{v}_{s,k} \overline{\mathbf{e}}} \\ &= \lim_{s\to\infty} \frac{\alpha_{s,N}\ r_{s,N}\ \overline{v}_{s,N} M \overline{\mathbf{e}}}{\lambda \alpha_{s,N}\ \overline{v}_{s,N} \overline{\mathbf{e}}} \\ &= \lim_{s\to\infty} \frac{r_{s,N}\ \overline{v}_{s,N} M \overline{\mathbf{e}}}{\lambda \overline{v}_{s,N} \overline{\mathbf{e}}}. \end{aligned}$$

Remember that the vector $\overline{v}_{s,N}$ consists of strictly positive elements (it is the eigenvector corresponding to the largest eigenvalue of the strictly positive rate-matrix R_s), so we can normalise the vector $\overline{v}_{s,N}$ such that $\overline{v}_{s,N} \overline{\mathbf{e}} = 1$.
Using $\overline{v}_{s,N} T_s(r_{s,N}) = \overline{0}$, we have (after dividing by s)

$$\lim_{s\to\infty} r_{s,N} \overline{v}_{s,N} Q_Y = \lim_{s\to\infty} \overline{v}_{s,N}\ \frac{1}{s} \left[\lambda I + r_{s,N}\ s Q_Y + (r_{s,N})^2 M\right] = \overline{0}.$$

Since $r_{s,N} \overline{v}_{s,N}$ can not go to the nullvecor as $s \to \infty$ (because $\lim_{s\to\infty} r_{s,N} \overline{v}_{s,N} M \overline{\mathbf{e}} = \lambda$) it must be that

$$\lim_{s\to\infty} \overline{v}_{s,N} = \overline{p}. \qquad \square$$

7. Instantaneous feedback control

An important issue in the ABR-service specification is that the network can force ABR sources to (temporarily) reduce their sending rate when one or more links are heavily loaded and a buffer overflow in the network might occur. In practice, the idea is that the network detects congestion at the links and then sends back a signal to the ABR sources forcing them to reduce their sending rates. In this section we show how the model presented in Section 2 can easily be adapted to incorporate such a feedback control mechanism.

Suppose that the allowed customer arrival rate depends on the number of customers in the system as long as the number of customers is below some threshold-level $J \in \{0, 1, 2, \dots\}$. Beyond that threshold the customers are allowed to arrive at some *guaranteed* rate, which will be typically small.
More precise, let the arrival rate of customers be equal to λ_j when the number of customers

in the system is j, for $j = 0, 1, \ldots, J-1$, and equal to λ when the number of customers is at least J. In this context it is natural to have

$$\lambda_0 \geq \lambda_1 \geq \cdots \geq \lambda_{J-1} \geq \lambda > 0,$$

but this ordering is not important for the analysis.
The transition matrix (2.3) now becomes

$$\mathcal{Q} := \begin{bmatrix} Q_d^{(0)} & \lambda_0 I & & & & & & \\ M & Q_d^{(1)} & \lambda_1 I & & & & & \\ & \ddots & \ddots & \ddots & & & & \\ & & M & Q_d^{(J-1)} & \lambda_{J-1} I & & & \\ & & & M & Q_d & \lambda I & & \\ & & & & M & Q_d & \lambda I & \\ & & & & & \ddots & \ddots & \ddots \end{bmatrix},$$

where $Q_d^{(0)} = Q_Y - \lambda_0 I$ and $Q_d^{(j)} = Q_Y - \lambda_j I - M$, for $j = 1, 2, \ldots, J-1$. $Q_d = Q_Y - \lambda I - M$ as before.
Since the modification of the rates – with respect to the original model of Section 2 – only concerns a finite number of states, the ergodicity condition (2.5) remains the same. Moreover, since the equilibrium equations for the levels $j \geq J$ remain exactly the same as in the original model, in the ergodic case we still have relation (3.1) for $j \geq J$, and with *exactly the same matrix R as before.* Therefore, the analysis with respect to the *homogeneous* part of the state space (all levels $j \geq J$) remains unchanged. However, once the matrix R has been determined we need to solve the following finite set of equations to find the $\overline{\pi}_0, \overline{\pi}_1, \ldots, \overline{\pi}_J$ (up to multiplication by a scalar):

$$\begin{aligned} \overline{\pi}_0 Q_d^{(0)} + \overline{\pi}_1 M &= \overline{0}, \\ \overline{\pi}_{j-1}\lambda_{j-1} I + \overline{\pi}_j Q_d^{(j)} + \overline{\pi}_{j+1} M &= \overline{0}, \qquad j = 1, \ldots, J-1, \\ \overline{\pi}_{J-1}\lambda_{J-1} I + \overline{\pi}_J Q_d^{(J)} + \overline{\pi}_J R M &= \overline{0}. \end{aligned} \tag{7.1}$$

Here, $\overline{\pi}_{J+1}$ in the last equation has been replaced with $\overline{\pi}_J R$ according to (3.1).
Having solved these equations, we can find the coefficients α_k of (3.3), which are now only valid for $j \geq J$, by solving

$$\overline{\alpha} \Lambda^J V = \overline{\pi}_J.$$

From this equation we can determine $\overline{\alpha}$ up to multiplication by the same scalar as $\overline{\pi}_0, \overline{\pi}_1, \ldots, \overline{\pi}_J$. This multiplicative scalar can then be found by requiring the resulting distribution $\overline{\pi}$ to sum up to 1.

Remark 7.1 An appealing special case of this feedback model is when $\lambda_0 = \lambda_1 = \cdots = \lambda_{J-1}$, i.e. there are only two possible arrival rates. The equations (7.1) are then the truncation of a homogeneous set of equations. For such a truncation the matrix-geometric relation (3.1) does not hold in general, but some generalised form of (3.3) does:

$$\overline{\pi}_j = \sum_{k=0}^{2N+1} \gamma_k (r_k)^j \overline{v}_k,$$

where the r_k for $k = 0, 1, \ldots, N$ are (as before) the roots of $\det[T(z)]$ inside the interval (0,1), and for $k = N+1, N+2, \ldots, 2N+1$, they are the roots of $\det[T(z)]$ in the interval $[1, \infty)$. The $\overline{v}_k$ are the corresponding left nullvectors of $T(r_k)$.

Remark 7.2 In the feedback-control mechanism we presented in this section we assume that the arrival rate is changed immediately as soon as the triggering event (a customers leaving or arriving leading to a different allowed arrival rate) takes place. In practice the feedback-information signal sent from the system to the sources has to go through the network itself, and suffers some delay. Nevertheless, the presented model could give useful insight into issues such as the effect of using several arrival rates compared to the two-valued model mentioned in Remark 7.1 (or no feedback at all).

8. Numerical example

In this section we show some numerical experiments for a special case of our model, which was treated in [18]. To avoid notational confusion within this paper we use a somewhat different notation than in [18].

In this special case the varying service availability has the following interpretation: It is caused by the presence of another type of traffic that also requires service from the same service station, and has (preemptive) priority over the regular traffic. Thus, the regular traffic we have been studying so far will be called low-priority traffic and the new traffic type will be called high-priority traffic. This interpretation of the varying service availability is a very natural one: it arises in the performance analysis of ABR (low-priority) traffic in the presence of CBR and VBR (high-priority) traffic.

Before we get to the numerical results, we first give a more precise specification of this model. Suppose that the service station consists of N identical service units and that there are two types of customers requiring service from the station: low-priority and high-priority customers. The high-priority customers arrive according to a Poisson process with intensity ν and each of them requires service from a *single* service unit for an exponentially distributed period with mean $\frac{1}{\eta}$. We assume that the service requirements are independent of the arrival process and of each other. If a high-priority customer arrives and finds less than N other high-priority customers being served at the station, then a server which is not currently serving a high-priority customer is assigned to serve the new customer for the total length of his required service. If a new high-priority customer finds all N servers occupied by other high-priority customers, he is rejected by the station and leaves without having received any service.

The low-priority customers are assumed to arrive according to a Poisson process (independent of everything else) with intensity λ, each having an exponentially distributed service requirement with mean $\frac{1}{\mu}$ (independent of everything else and of each other). Further, there is an infinite storage capacity to hold low-priority customers and all present low-priority customers equally share the service units that are not currently required by high-priority customers (processor sharing). Note that an arriving high-priority customer can take away (immediately) a service unit that is currently serving the low-priority customers and thus reduce the service capacity available to the low-priority traffic.

We now specify how this model fits into the general one. In this context we can take the process $Y(t)$ to be the number of high-priority customers in the system at time t, and its transition rates become:

$$\begin{aligned} q_i^{(+)} &= \nu, \quad i = 0, 1, \ldots, N-1, \\ q_i^{(-)} &= i\eta, \quad i = 1, 2, \ldots, N, \end{aligned}$$

and as usual, $q_N^{(+)} = q_0^{(-)} = 0$. The process $Y(t)$ in this case evolves as the number of customers in the regular $M/M/N/N$ loss-model.

From (2.1) we immediately obtain the following expression for the steady-state distribution of $Y(t)$:

$$p_i = \frac{(\rho_Y)^i/i!}{\sum_{m=0}^{N}(\rho_Y)^m/m!}, \qquad i = 0, 1, \ldots, N,$$

where $\rho_Y := \nu/\eta$ is the *offered* load by the high-priority customers (this is *not* equal to the accepted load $\mathbb{E}[Y]$ because of rejections of high-priority customers). Similarly we will use $\rho_X := \lambda/\mu$ to denote the (actual) load on the system caused by the low-priority customers.

Further, it is clear that

$$\mu_i = (N-i)\mu,$$

for all $i = 0, 1, \ldots, N$.

It is easy to check that the ergodicity condition (2.5) becomes

$$\rho_X < N - \mathbb{E}[Y].$$

For more details on this special model we refer to [18]. For this model we have done numerical experiments and we show some results to illustrate the effect of the high-priority traffic (CBR, VBR) on the low-priority traffic (ABR). Throughout these experiments we have chosen $\mu = 1$ (time-normalisation) and $N = 17$ (in accordance with data supplied by KPN Research for The Netherlands). Further, we have also fixed $\rho_Y = \frac{\nu}{\eta} = 10$. This value for ρ_Y is not of any particular interest, the resulting graphs are of the same form for other values of ρ_Y. Note however, that for fixed ρ_Y the steady-state probabilities p_i of the process $Y(t)$, and in particular the average number of service units available to the low-priority traffic ($N - \mathbb{E}[Y]$), are also fixed.

The two remaining free parameters, λ and ν (or $\eta = \nu/\rho_Y$), will now be varied in the graphs.

In Figure 1 we have plotted – for several values of ν (and η) – the expected number of low-priority customers in the system against an increasing load of the low-priority customers. Instead of λ itself, we have chosen $\rho_X/(N - \mathbb{E}[Y])$ so that the horizontal axis ranges from 0 to 1 (remember that $\mathbb{E}[Y]$ is fixed with $\rho_Y = 10$). The lowest curve, denoted by 'limit', corresponds to the regular *M/M/1* model with fixed service capacity $N - \mathbb{E}[Y]$. As stated in Theorem 6.2, the curves converge to this lowest curve as $\nu \to \infty$.

In Figure 2 we have done the same for the variance of the number of low-priority customers in the system. Again we observe the indicated convergence as ν increases.

Remark 8.1 The general model presented in Section 2 allows to incorporate more realistic features of ABR-traffic performance within the framework of this section. For example, the

Figure 1: $\mu = 1$, $\rho_Y = \frac{\nu}{\eta} = 10$ and $N = 17$.

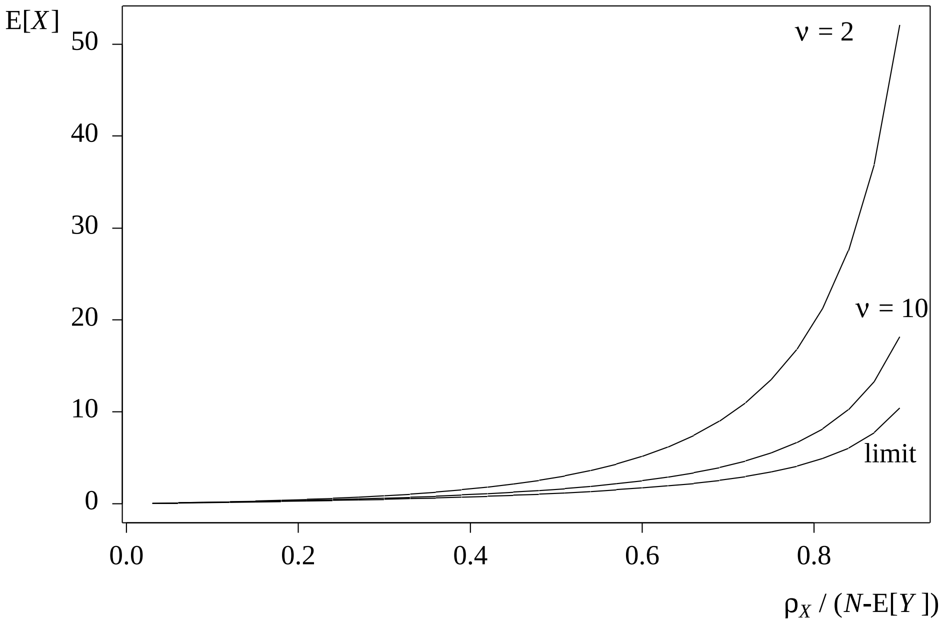
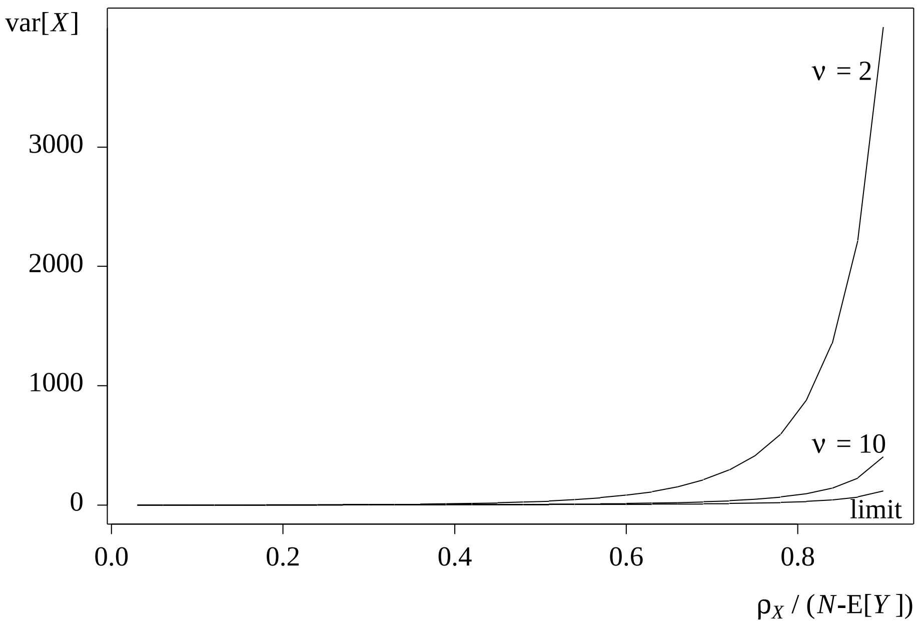


Figure 2: $\mu = 1$, $\rho_Y = \frac{\nu}{\eta} = 10$ and $N = 17$.



model of Section 2 captures the following variants. (i) There is an additional number of service units at the service station which are *always* available to the low-priority customers; this corresponds to the allocation of a minimum cell rate to ABR-traffic. (ii) There is a finite waiting-room for the high-priority traffic (typically small, because these traffic types do not allow for significant delays); this would lead to the $M/M/c/c+K$ loss-model for the process $Y(t)$ (in this case $N = c + K$). Also a combination of these two variants are captured in the general model.

9. Concluding remarks

In this paper we have studied a queueing model with a server that changes its service rate according to a finite birth and death process. This model captures many features of the behaviour at the burst-level of ABR traffic at an ATM link in the presence of other traffic types with a higher priority (CBR, VBR). Because of the simplifying assumptions regarding the distributions of the service requirements (exponential) and the process regulating the available service capacity (finite birth-death process) we were able to give a detailed analysis of the distribution of the number of customers in the system. Through Little's formula we also obtained the mean processing time. Of course, in practice one would also like to know the variance of the processing time and possibly higher moments, or even the complete distribution. These issues are the subject of our current research, and some preliminary results have been obtained.

In Section 7 we have modified the model to cope with a simple feedback mechanism. The issue of feedback is highly important in the context of the ABR service. An important drawback of our feedback model is that it assumes no delays for the feedback signals. In the future we plan to study the possibility of allowing (stochastic) delays for feedback information, i.e. although the system has detected a congestion of traffic, it might take some time before this information can be used to lower the arrival rate.

Acknowledgement

The author is 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. The author wants to thank in particular Professor O.J. Boxma for reading previous versions of this paper and providing many helpful comments on them.

Bibliography

[1] S. Blaabjerg, G. Fodor, M. Telek, A.T. Andersen. *A partially blocking-queueing system with CBR/VBR and ABR/UBR arrival streams.* Institute of Telecommunications, Technical University of Denmark (internal report).

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

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

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

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

[6] 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.

[7] 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.

[8] 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.

[9] 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.

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

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

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

[13] M. Marcus, H. Minc. *A survey of matrix theory and matrix inequalities.* Allyn and Bacon, Inc., Boston, 1964.

[14] 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.

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

[16] 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.

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

[18] R. Núñez Queija, O.J. Boxma. *Analysis of a multi-server queueing model of ABR.* CWI Report BS-R9613 (1996).

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

[20] 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).

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

[22] 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).

[23] M. Ritter. *Analysis of a queueing model with delayed feedback and its application to the ABR flow control.* University of Würzburg, Institute of Computer Science, Research Report Series 164 (1997).

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

[25] S.F. Yashkov. *Processor sharing queues: Some progress in analysis.* Queueing Systems 2 (1987), 1-17.