A two-level traffic shaper for an on-off source

Memorandum COSOR 99-07, 1999, Eindhoven University of Technology

A TWO-LEVEL TRAFFIC SHAPER FOR AN ON-OFF SOURCE

IVO ADAN AND JACQUES RESING*

Abstract. This paper studies a two-level traffic shaper. The cell arrival process is modeled by an on-off fluid source. The output rate of the shaper is regulated by the content of a finite token bank. Our interest focuses on the stationary joint distribution of the content of the cell buffer and the content of the token bank, from which various performance measures such as the steady-state cell delay and burst duration can be obtained. The system is analysed using an approximative model based on a stochastic discretization of the content of the token bank. We believe that this discretization technique has much wider applicability. Numerical results show the quality of the approximation and demonstrate the effect of the shaping mechanism.

Key words. Traffic shaper, Fluid model, Stochastic discretization

AMS subject classifications. 60K25, 90B22

1. Introduction.

In ATM networks traffic shaping is an important issue. Probably the best-known traffic shaping mechanism is the *leaky-bucket* or *token-bank throttle*, see, [3, 12, 13]. The token-bank throttle guarantees that the sustainable cell rate, i.e. the long-term average rate at which cells enter the network, does not exceed a specified rate. However, during periods in which the token bank is nonempty, the scheme permits a higher rate, equaling, in fact, the actual cell arrival rate. The maximum duration of such a high-rate period, also called maximum burst duration, is determined by the size of the token bank. In this paper we study a *two-level traffic shaper* which has the additional feature that during periods in which the token bank is nonempty, the rate at which cells enter the network will not exceed a second specified rate, the peak cell rate. Hence the function of the two-level traffic shaper is to shape a statistical bit rate input stream such that it conforms to three traffic parameters: sustainable cell rate, peak cell rate and maximum burst duration. In fact, this mechanism can be realized by a token-bank throttle in tandem with a cell spacer [5].

Exact analyses of various versions of the token-bank throttle have appeared in many papers, see e.g. [3, 4, 6, 12], and the references mentioned therein. Note that for the token-bank throttle at any moment in time either the cell buffer or the token bank is empty. Hence, the process describing both the number of cells waiting and the content of the token bank is essentially one-dimensional. This feature is not shared by the two-level traffic shaper and therefore its exact analysis is much more difficult.

The effect of the two-level traffic shaper on the delay and the size of the bursts of cells entering the network has been studied in [9] through simulations. Using a discrete-time model, the performance of the two-level traffic shaper is compared with that of a conventional spacer in [10]. Both studies model the arrival of cells by a renewal process. In this paper we model the cell arrivals by a Markov modulated fluid source in which the rate of fluid generation is determined by the state of a continuous-time Markov chain [2]. To keep the presentation simple we will restrict the analysis to the (insightful) case of a two-state on-off source. But the approach developed in this paper also works for multi-state fluid sources. The model with the on-off source has been analyzed independently in [8] (see also [11]), where using

* Eindhoven University of Technology, Department of Mathematics and Computer Science, P.O.Box 513, 5600 MB - Eindhoven, the Netherlands.

Laplace-transform techniques an expression for the stationary joint distribution of the content of the cell buffer and the token bank involving multi-dimensional integrals of Bessel functions is found. Although mathematically elegant, this result is not easy to use for numerically evaluating, e.g., the distribution of the cell delay or the duration of a peak-rate period. It also seems complicated to extend the analysis in [8] to models with multi-state fluid sources. Therefore, the aim of this paper is to develop an approximative model that enables us to efficiently evaluate the performance of the two-level traffic shaper in terms of cell delay and burst duration. The approximation is based on a stochastic discretization of the content of the token bank, i.e., we keep track of small stochastic quanta rather than the actual fluid content. The smaller we choose the quanta, the more accurate the approximation will be. The discretization technique has been recently used in [1] and we believe that it has much wider applicability in the area of fluid flow models.

The paper is organized as follows. After the description of the basic fluid flow model in Section 2, we introduce and analyze the approximative model in the Sections 3 and 4, resp. Next, in Section 5 we indicate how various important performance measures can be obtained. The approximative model is validated in Section 6 by comparing results for this model with simulation results for the basic model. In Section 7 we present numerical results, showing in particular the trade-off between extra delay incurred by the traffic shaper on the one hand and burstiness reduction of the regulated traffic stream on the other. Finally, in Section 8 we summarize the main results of the paper.

2. Fluid flow model. The shaper consists of two buffers, a cell buffer and a token buffer (see Fig. 1). Cells arrive as fluid from an on-off source in a cell buffer of infinite size. During an on-period the source generates fluid at constant rate v_0. The durations of the on-periods and off-periods are independent and exponentially distributed with mean $1/\mu$ and $1/\lambda$, resp. Tokens arrive (also as fluid) at constant rate v_2 in the token buffer. The size of the token buffer is finite and denoted by C. Arriving tokens which find the token buffer full are lost.

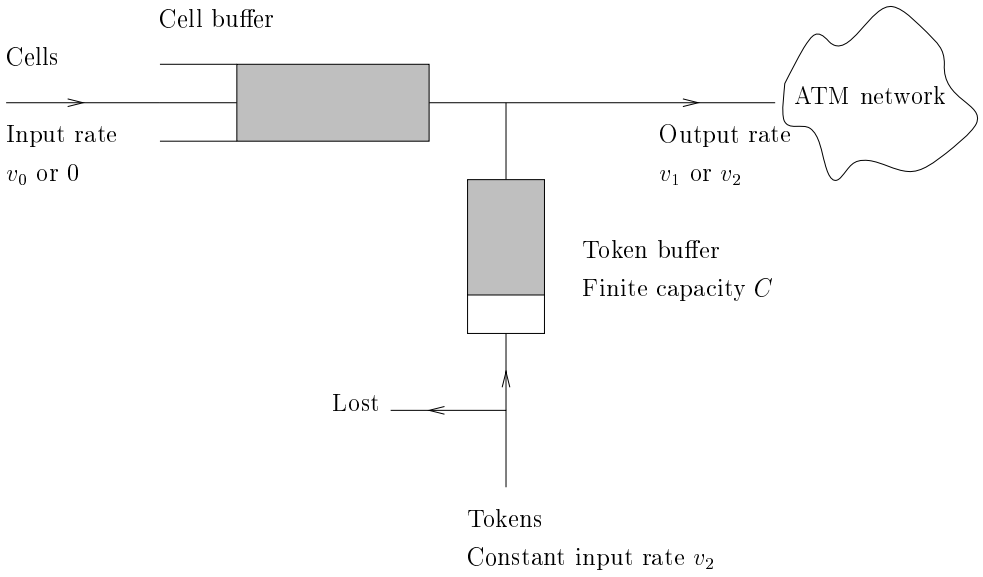

FIG. 1. *Two-level traffic shaper*

The outflow from the two buffers is regulated as follows. Only paired cells and tokens move downstream. The rate at which they depart may not exceed a peak rate v_1 and we assume that $v_2 < v_1 < v_0$. (The situation $v_1 \geq v_0$ would reduce the shaper to an ordinary token-bank throttle, whereas the uninteresting situation $v_1 \leq v_2$ would lead to a token bank which is always completely filled.) Hence, as long as the two buffers are nonempty, the outflow from each buffer is equal to the peak rate v_1. So the token buffer depletes at a constant rate $v_1 - v_2$ and the cell buffer either fills with rate $v_0 - v_1$ or depletes at rate v_1. When the token buffer becomes empty, it remains in that state and the output rate from the cell buffer drops down to v_2 (i.e. the generation rate of tokens). When also the cell buffer becomes empty, the token buffer starts to fill again with rate v_2. Sample paths of the content of the cell buffer and token buffer for the case $v_0 = 3$, $v_1 = 2$ and $v_2 = 1$ are shown in Fig. 2.

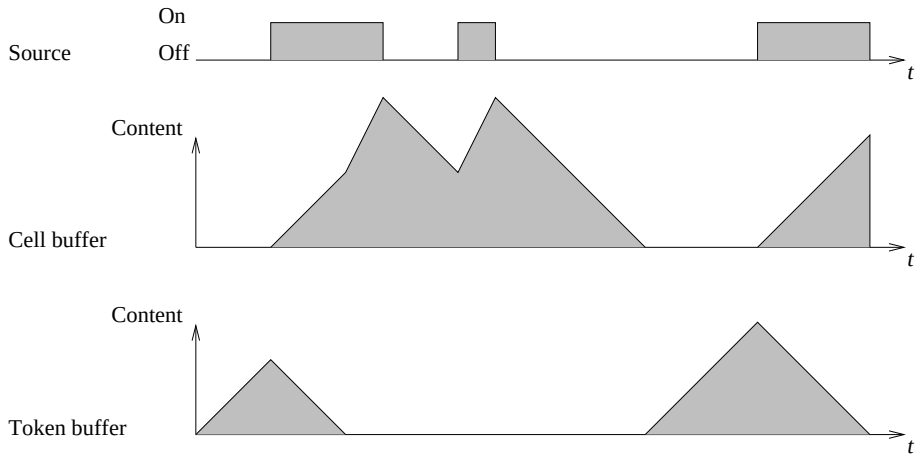

FIG. 2. *Sample paths of the content of the two buffers for* $v_0 = 3$, $v_1 = 2$ *and* $v_2 = 1$

Let U_t denote the content of the cell buffer, C_t the content of the token buffer and I_t the state of the on-off source at time t ($I_t = 1$, resp. $I_t = 0$, indicating that the source is on, resp. off). The three-dimensional process $\{(U_t, C_t, I_t),\ t \geq 0\}$ constitutes a Markov process with state space $\{(u, c, i) :\ u \geq 0,\ 0 \leq c \leq C,\ i = 0, 1\}$. It has a unique stationary distribution, provided

$$v_0 \frac{\lambda}{\lambda + \mu} < v_2. \tag{1}$$

An exact expression for the stationary distribution involving Bessel function integrals has been found in [8]. But, as mentioned before, this expression is not easy to use for numerically evaluating distributions of cell delay and burst duration. Therefore, we describe in the next section an approximative model in which we keep track of the content of the token buffer by observing small stochastic quanta rather than the actual fluid volume. It will appear that the approximative model can be evaluated efficiently and produces accurate approximations for relevant performance measures. Also, the analysis can be easily extended to multi-state fluid sources.

3. The approximative model. The basic idea behind the approximative model is to discretize the state space of the token buffer by observing the number of suitably defined stochastic quanta rather than observing the actual volume of the token buffer. The reason for doing this will become clear later when we are going to solve the balance

equations for the model. In fact, by discretizing one of the state variables we have to solve a (much easier) set of ordinary differential equations instead of a set of partial differential equations. The sizes of the stochastic quanta are exponential. This implies that, once we know the number of stochastic quanta in the system we also know the distribution of the actual volume of the token buffer. This is due to the memoryless property of the exponential distribution.

The reason that our model is approximative is that the token buffer has finite capacity. In our approximative model we allow a *maximum number of exponential quanta* instead of allowing a *maximum amount of actual volume* in the token buffer. This maximum number of quanta is chosen such that, if the maximum is reached, then the *mean amount of actual volume* in the token buffer is equal to C, the maximum amount of actual volume in the original model. Notice that in the approximative model it is possible that the amount of actual volume in the token buffer is larger than C. However, by letting the mean size of a quantum tend to zero and simultaneously letting the maximum number of quanta tend to infinity we can get as close as we want to the original model. The argument here is that a deterministic quantity can be approximated arbitrarily close by an Erlang-distributed random variable. Of course, we hope that the approximative model is already accurate when the mean size of the quanta is not too small. We will come back to this issue in Section 5.

Let us now describe the approximative model in more detail. The way we collect quanta in the token buffer during an idle period of the cell buffer is the following. At the beginning of each idle period the buffer receives a quantum the size of which is initially zero but increases at rate v_2 until *either* the idle period *or* an exponentially distributed period of mean $1/\nu$ has ended, whichever happens first. In the latter case a second quantum is added the size of which again grows at rate v_2 until either the remaining idle period or the length of a new, exponentially distributed period of mean $1/\nu$, has ended. If the latter happens first, a third quantum is added, and so on, until the complete idle period has come to an end. Clearly, the total volume added during an idle period now equals v_2 times the size of that idle period, as before. Moreover, the number of quanta added to the buffer during an idle period is geometrically distributed with mean $1 + \nu/\lambda$, while the size of each quantum is exponentially distributed with mean $v_2/(\lambda + \nu)$.

While the cell buffer is not empty the quanta are drained at rate $v_1 - v_2$ in their order of arrival, and removed as soon as their sizes are reduced to zero. So a quantum allows us to send at peak rate during an exponentially distributed period with mean $1/\eta$, where

$$\frac{1}{\eta} = \frac{v_2}{\lambda + \nu} \cdot \frac{1}{v_1 - v_2}.$$

In Fig. 3 we show sample paths of the continuous content and discretized content of the token buffer for the situation in Fig. 2.

To take into account that the token buffer has a finite capacity C we assume that the number of quanta in the buffer is bounded by some N depending on the value of ν. As mentioned before, we require that when the number of quanta in the buffer has reached its maximum, i.e. N, then the *mean* buffer content is equal to C. Since the mean size of a quantum is $v_2/(\lambda + \nu)$, we have that N and ν are related by

$$N \cdot \frac{v_2}{\lambda + \nu} = C.$$

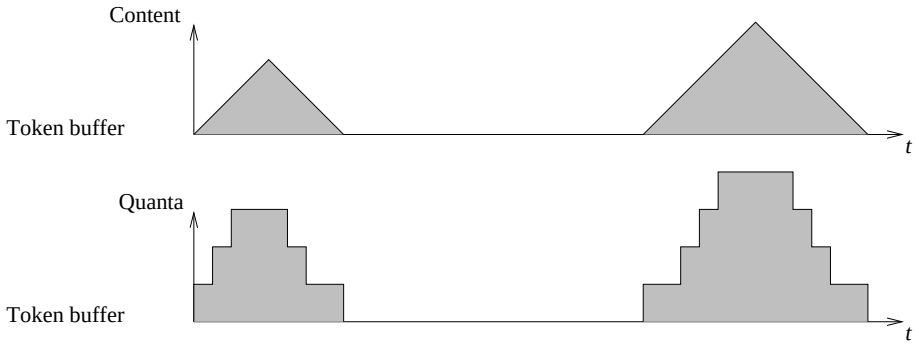

FIG. 3. *Sample paths of the continuous content and discretized content of the token buffer for the situation in Fig. 2*

Denote the number of quanta in the buffer at time t by N_t. In other words, N_t can be viewed as the state of a counter at time t, which increases by a random amount during idle periods and decreases by one each time a quantum has been removed. Our interest focuses on the three-dimensional process $\{(U_t, N_t, I_t),\ t \geq 0\}$ with state space $\{(u, n, i) : u \geq 0,\ n = 0, \cdots, N, i = 0, 1\}$. For this process we want to compute the equilibrium distribution. Once this distribution is known we can calculate the equilibrium distribution of the process $\{(U_t, C_t, I_t),\ t \geq 0\}$ since, given the number of quanta at an arbitrary point in time, their sizes are independent and identically distributed according to an exponential distribution with mean $v_2/(\lambda + \nu)$.

Unfortunately, the process $\{(U_t, N_t, I_t),\ t \geq 0\}$ does *not constitute a Markov process*. Idle periods of the past determine the sizes of quanta currently present and, hence, influence future behavior of the process. However, if we disregard periods of time during which the fluid buffer is empty, then we are dealing with a Markov process, since the sizes of the quanta currently present can no longer be observed from the past of the process. Denote by $\{(\tilde{U}_t, \tilde{N}_t, \tilde{I}_t),\ t \geq 0\}$ the process $\{(U_t, N_t, I_t),\ t \geq 0\}$ restricted to periods of time during which the cell buffer is nonempty. More formally,

$$\tilde{U}_t = U_{\gamma(t)}, \quad \tilde{N}_t = N_{\gamma(t)}, \quad \tilde{I}_t = I_{\gamma(t)},$$

where $\gamma(t)$ is defined as the largest s with the property that the total time the cell buffer is nonempty during $[0, s]$ is equal to t (cf. [7]). In the next section we will determine the equilibrium distribution for the Markov process $\{(\tilde{U}_t, \tilde{N}_t, \tilde{I}_t),\ t \geq 0\}$, from which it appears to be straightforward to obtain the distribution for the original process $\{(U_t, N_t, I_t),\ t \geq 0\}$.

4. Analysis of the approximative model. Denote the unrestricted process in equilibrium by $(U_\infty, N_\infty, I_\infty)$ and define

$$F_i(u, n) = \Pr[U_\infty \leq u, N_\infty = n, I_\infty = i], \quad f_i(u, n) = \frac{\partial}{\partial u} F_i(u, n).$$

Similarly, we define $\tilde{F}_i(u, n)$ and $\tilde{f}_i(u, n)$ as distribution function and density for the restricted process in equilibrium. We first consider the restricted process and focus on the determination of $\tilde{f}_i(u, n)$. Below, we shall first formulate the balance equations for the process, one of which we shall derive afterwards (the other ones follow similarly). The main result of this section, an explicit expression for the functions $\tilde{f}_i(u, n)$, is formulated in Theorem 4.1.

The balance equations are given by

$$\begin{aligned}
(v_0 - v_2)\tilde{f}_1(u,0) + \mu\tilde{F}_1(u,0) &= \lambda\tilde{F}_0(u,0) + \eta\tilde{F}_1(u,1)\\
v_2\tilde{f}_0(0,0) + \lambda\tilde{F}_0(u,0) &= \mu\tilde{F}_1(u,0) + v_2\tilde{f}_0(u,0) + \eta\tilde{F}_0(u,1),\\
(v_0 - v_1)\tilde{f}_1(u,n) + (\mu+\eta)\tilde{F}_1(u,n) &= \lambda\tilde{F}_0(u,n) + \eta\tilde{F}_1(u,n+1)\\
&\quad + v_1\sum_{j=1}^{n-1}\tilde{f}_0(0,j)p_{n-j} + v_2\tilde{f}_0(0,0)p_n,\\
v_1\tilde{f}_0(0,n) + (\lambda+\eta)\tilde{F}_0(u,n) &= \mu\tilde{F}_1(u,n) + \eta\tilde{F}_0(u,n+1) + v_1\tilde{f}_0(u,n),\\
(v_0 - v_1)\tilde{f}_1(u,N) + (\mu+\eta)\tilde{F}_1(u,N) &= \lambda\tilde{F}_0(u,N) + v_1\sum_{j=1}^{N-1}\tilde{f}_0(0,j)\sum_{k=N-j}^{\infty}p_k\\
&\quad + v_2\tilde{f}_0(0,0)\sum_{k=N}^{\infty}p_k + v_1\tilde{f}_0(0,N),\\
v_1\tilde{f}_0(0,N) + (\lambda+\eta)\tilde{F}_0(u,N) &= \mu\tilde{F}_1(u,N) + v_1\tilde{f}_0(u,N),
\end{aligned}$$

where $n = 1,\ldots,N-1$, and

$$\begin{aligned}
p_k &= \Pr[k \text{ quanta added in a period during which the cell buffer is empty}]\\
&= \left(1 - \frac{\lambda}{\lambda+\nu}\right)^{k-1}\frac{\lambda}{\lambda+\nu}\ .
\end{aligned}$$

Let us derive the first balance equation. Consider the set $\{(x,0,1), 0 \le x \le u\}$. The probability that the restricted process leaves the set in a small time interval of length Δ equals $\Pr[u-(v_0-v_2)\Delta \le \tilde{U}_\infty \le u, \tilde{N}_\infty = 0, \tilde{I}_\infty = 1]\times\Pr[\text{source not switched off in }\Delta]$ plus $\Pr[\tilde{U}_\infty \le u, \tilde{N}_\infty = 0, \tilde{I}_\infty = 1] \times \Pr[\text{source switched off in }\Delta]$ which amounts to

$$\tilde{f}_1(u,0)(v_0 - v_2)\Delta + \tilde{F}_1(u,0)\mu\Delta + O(\Delta^2).$$

On the other hand, the probability that the process enters that set in a time interval of length Δ is equal to $\Pr[\tilde{U}_\infty \le u, \tilde{N}_\infty = 0, \tilde{I}_\infty = 0] \times \Pr[\text{source switched on in }\Delta]$ plus $\Pr[\tilde{U}_\infty \le u, \tilde{N}_\infty = 1, \tilde{I}_\infty = 1] \times \Pr[\text{quantum disposed in }\Delta]$ yielding

$$\tilde{F}_0(u,0)\lambda\Delta + \tilde{F}_1(u,1)\eta\Delta + O(\Delta^2).$$

Dividing the terms above by Δ and letting Δ tend to zero yields the rate out of, resp. rate into the set $\{(x,0,1), 0 \le x \le u\}$. Equating the two rates gives the desired balance equation for $\tilde{F}_1(u,0)$, i.e.,

$$(v_0 - v_2)\tilde{f}_1(u,0) + \mu\tilde{F}_1(u,0) = \lambda\tilde{F}_0(u,0) + \eta\tilde{F}_1(u,1).$$

Differentiation of the set of balance equations turns it into a system of linear differential equations for the densities $\tilde{f}_i(u,n)$. This gives

$$\begin{aligned}
(v_0 - v_2)\tilde{f}_1'(u,0) &= -\mu\tilde{f}_1(u,0) + \lambda\tilde{f}_0(u,0) + \eta\tilde{f}_1(u,1),\\
-v_2\tilde{f}_0'(u,0) &= \mu\tilde{f}_1(u,0) - \lambda\tilde{f}_0(u,0) + \eta\tilde{f}_0(u,1),\\
(v_0 - v_1)\tilde{f}_1'(u,n) &= -(\mu+\eta)\tilde{f}_1(u,n) + \lambda\tilde{f}_0(u,n) + \eta\tilde{f}_1(u,n+1),\\
-v_1\tilde{f}_0'(u,n) &= \mu\tilde{f}_1(u,n) - (\lambda+\eta)\tilde{f}_0(u,n) + \eta\tilde{f}_0(u,n+1),
\end{aligned}$$

where n runs from 1 to N and, by definition, $\tilde{f}_i(u, N+1) = 0$. It is convenient to rewrite these equations in vector-matrix notation, i.e.,

$$\begin{aligned} \tilde{f}'(u,0) &= V_2^{-1} Q_2 \tilde{f}(u,0) + \eta V_2^{-1} \tilde{f}(u,1), \\ \tilde{f}'(u,n) &= V_1^{-1} Q_1 \tilde{f}(u,n) + \eta V_1^{-1} \tilde{f}(u,n+1), \quad n = 1, \cdots, N, \end{aligned}$$

where

$$\tilde{f}(u,n) = \begin{pmatrix} \tilde{f}_1(u,n) \\ \tilde{f}_0(u,n) \end{pmatrix}, \quad V_i = \begin{pmatrix} v_0 - v_i & 0 \\ 0 & -v_i \end{pmatrix}, \quad i = 1, 2,$$

$$Q_1 = \begin{pmatrix} -(\mu+\eta) & \lambda \\ \mu & -(\lambda+\eta) \end{pmatrix}, \quad Q_2 = \begin{pmatrix} -\mu & \lambda \\ \mu & -\lambda \end{pmatrix}.$$

The initial conditions for the densities $\tilde{f}_i(u,n)$ follow by letting u tend to zero in the equations for the functions $\tilde{F}_i(u,n)$. We get

$$\begin{aligned} \tilde{f}_1(0,0) &= 0, \\ (v_0 - v_1)\tilde{f}_1(0,n) &= v_1 \sum_{j=1}^{n-1} \tilde{f}_0(0,j) p_{n-j} + v_2 \tilde{f}_0(0,0) p_n, \quad n = 1, \cdots, N-1 \\ (v_0 - v_1)\tilde{f}_1(0,N) &= v_1 \sum_{j=1}^{N-1} \tilde{f}_0(0,j) \sum_{k=N-j}^{\infty} p_k + v_2 \tilde{f}_0(0,0) \sum_{k=N}^{\infty} p_k + v_1 \tilde{f}_0(0,N). \end{aligned}$$

For the unknown densities $\tilde{f}_i(u,n)$ we now have formulated a homogeneous system of linear differential equations with constant coefficients of the first order together with initial conditions. The solution of this problem is well-known, and it can be expressed in terms of the eigenvalues and (possibly generalized) eigenvectors of the matrix of coefficients. However, the dimension of the matrix of coefficients is $2(N+1) \times 2(N+1)$. Hence, for large N, the standard solution using eigenvalues and eigenvectors of the matrix of coefficients is numerically infeasible. Below we show that this difficulty can be avoided by breaking down the problem into small pieces which can easily be solved recursively.

The vector $\tilde{f}(u,N)$ satisfies

$$\tilde{f}'(u,N) = V_1^{-1} Q_1 \tilde{f}(u,N).$$

This is a homogeneous 2×2 system, the general solution of which is given by

$$\tilde{f}(u,N) = A_{0,0} y_1 e^{\sigma_1 u} + B_{0,0} y_2 e^{\sigma_2 u}, \tag{2}$$

where $A_{0,0}$ and $B_{0,0}$ are constants and σ_1 (< 0) and σ_2 (> 0) are the eigenvalues of the matrix $V_1^{-1} Q_1$ with corresponding eigenvectors y_1 and y_2, resp. Since the density $\tilde{f}(u,N)$ converges to zero as u tends to infinity, we have to set $B_{0,0} = 0$. The determination of the constant $A_{0,0}$ will be postponed.

We now proceed with the density $\tilde{f}(u,N-1)$ which satisfies

$$\tilde{f}'(u,N-1) = V_1^{-1} Q_1 \tilde{f}(u,N-1) + \eta V_1^{-1} \tilde{f}(u,N). \tag{3}$$

This is an *inhomogeneous* 2×2 system for $\tilde{f}(u, N-1)$ with $\tilde{f}(u, N)$ as inhomogeneous term. If we substitute (2) with $B_{0,0} = 0$ into (3) and express $V_1^{-1} y_1$ and $V_1^{-1} y_2$ as linear combination of the eigenvectors y_1 and y_2, i.e.,

$$V_1^{-1} y_1 = c_{1,1} y_1 + c_{1,2} y_2, \quad V_1^{-1} y_2 = c_{2,1} y_1 + c_{2,2} y_2,$$

then we get the following inhomogeneous system for $\tilde{f}(u, N-1)$,

$$\tilde{f}'(u, N-1) = V_1^{-1} Q_1 \tilde{f}(u, N-1) + \eta A_{0,0} c_{1,1} y_1 e^{\sigma_1 u} + \eta A_{0,0} c_{1,2} y_2 e^{\sigma_1 u}.$$

The general solution of this system is given by

$$\tilde{f}(u, N-1) = A_{1,0} y_1 e^{\sigma_1 u} + A_{1,1} u y_1 e^{\sigma_1 u} + B_{1,0} y_2 e^{\sigma_1 u}.$$

The first term is the solution of the homogeneous equation (notice that $y_2 e^{\sigma_2 u}$ can be excluded again, since $\tilde{f}(u, N-1)$ converges to zero as u tends to infinity), the second term is a particular solution of the equation with inhomogeneous term $\eta A_{0,0} c_{1,1} y_1 e^{\sigma_1 u}$ and, finally, the third one is a particular solution of the equation with inhomogeneous term $\eta A_{0,0} c_{1,2} y_2 e^{\sigma_1 u}$. The coefficients $A_{1,1}$ and $B_{1,0}$ satisfy

$$A_{1,1} = \eta A_{0,0} c_{1,1}, \quad B_{1,0} = \frac{1}{\sigma_1 - \sigma_2} \eta A_{0,0} c_{1,2}.$$

Again, the determination of the coefficient $A_{1,0}$ of the homogeneous term will be postponed. Repeating this procedure we can work our way down from $N-1$ to 1. This leads to the solution

$$\tilde{f}(u, N-k) = \sum_{j=0}^{k} A_{k,j} u^j y_1 e^{\sigma_1 u} + \sum_{j=0}^{k} B_{k,j} u^j y_2 e^{\sigma_1 u}, \quad k = 0, \cdots, N-1, \tag{4}$$

where the constants $A_{k,j}$ satisfy

$$A_{k,j} = \frac{1}{j} \left(\eta A_{k-1,j-1} c_{1,1} + \eta B_{k-1,j-1} c_{2,1} \right), \quad j = 1, \cdots, k.$$

and the constants $A_{k,0}$ will be determined later on. The constants $B_{k,j}$ follow recursively from

$$B_{k,k} = 0, \quad B_{k,j-1} = \frac{1}{\sigma_1 - \sigma_2} \left(\eta A_{k-1,j-1} c_{1,2} + \eta B_{k-1,j-1} c_{2,2} - j B_{k,j} \right),$$

where j runs from k downto 1. It remains to consider the differential equation for $\tilde{f}(u, 0)$ with the density $\tilde{f}(u, 1)$ given by (4) as inhomogeneous term. To formulate the general solution we need the eigenvalues σ_3 and $\sigma_4 = 0$ of the matrix $V_2^{-1} Q_2$ with corresponding eigenvectors y_3 and y_4, resp. Notice that stability condition (1) implies that $\sigma_3 < 0$. We also need to express $V_2^{-1} y_1$ and $V_2^{-1} y_2$ as linear combination of the eigenvectors y_3 and y_4, i.e.,

$$V_2^{-1} y_1 = c_{1,3} y_3 + c_{1,4} y_4, \quad V_2^{-1} y_2 = c_{2,3} y_3 + c_{2,4} y_4.$$

Then the general solution of the differential equation for $\tilde{f}(u, 0)$ is given by

$$\tilde{f}(u, 0) = D y_3 e^{\sigma_3 u} + \sum_{j=0}^{N-1} A_{N,j} u^j y_3 e^{\sigma_1 u} + \sum_{j=0}^{N-1} B_{N,j} u^j y_4 e^{\sigma_1 u}. \tag{5}$$

The first term is the solution of the homogeneous equation. The coefficient D has to be determined yet. The other terms constitute a particular solution of the inhomogeneous equation. The coefficients $A_{N,j}$ and $B_{N,j}$ recursively follow from

$$A_{N,N} = 0, \quad A_{N,j-1} = \frac{1}{\sigma_1 - \sigma_3}\left(\eta A_{N-1,j-1}c_{1,3} + \eta B_{N-1,j-1}c_{2,3} - jA_{N,j}\right),$$
$$B_{N,N} = 0, \quad B_{N,j-1} = \frac{1}{\sigma_1 - \sigma_4}\left(\eta A_{N-1,j-1}c_{1,4} + \eta B_{N-1,j-1}c_{2,4} - jB_{N,j}\right),$$

where j runs from N downto 1. This completes the general solution of the system of differential equations for the densities $\tilde{f}(u,n)$. It is given by the expressions (4)–(5) in which the coefficients $A_{0,0},\ldots,A_{N-1,0}$ and D have to be determined yet. In fact, these coefficients follow from the initial conditions and the normalization equation

$$\sum_{n=0}^{N}\int_{u=0}^{\infty}(\tilde{f}_0(u,n) + \tilde{f}_1(u,n))du = 1.$$

These findings are summarized in the theorem below.

THEOREM 4.1. *The densities $\tilde{f}(u,n)$ of the restricted process can be expressed as*

$$\begin{aligned}\tilde{f}(u,N-k) &= \sum_{j=0}^{k} A_{k,j}u^j y_1 e^{\sigma_1 u} + \sum_{j=0}^{k} B_{k,j}u^j y_2 e^{\sigma_1 u}, \quad k = 0,\cdots,N-1,\\ \tilde{f}(u,0) &= Dy_3 e^{\sigma_3 u} + \sum_{j=0}^{N-1} A_{N,j}u^j y_3 e^{\sigma_1 u} + \sum_{j=0}^{N-1} B_{N,j}u^j y_4 e^{\sigma_1 u}.\end{aligned}$$

Here, σ_1 and σ_3 are the negative eigenvalues of the matrices $V_1^{-1}Q_1$ and $V_2^{-1}Q_2$ respectively, y_1 and y_2 are eigenvectors of $V_1^{-1}Q_1$, y_3 and y_4 are eigenvectors of $V_2^{-1}Q_2$, and the coefficients $A_{k,j}$, $B_{k,j}$ and D follow from recursion relations, initial conditions and a normalization equation given in this section.

This concludes the determination of the densities $\tilde{f}_i(u,n)$ (and thus also of the distribution functions $\tilde{F}_i(u,n)$) for the restricted Markov process. We will now connect these results to those for the original (unrestricted) process. Notice that

$$(6)\quad f_i(u,n) = \tilde{f}_i(u,n)(1 - \Pr[\text{cell buffer is empty}]) = \tilde{f}_i(u,n)(1 - \sum_{n=1}^{N} F_0(0,n)).$$

Hence it remains to find the probabilities $F_0(0,n)$ (the probabilities $F_0(0,0)$ and $F_1(0,n)$ are all zero). These probabilities satisfy the following balance equations,

$$\begin{aligned}F_0(0,1)(\lambda+\nu) &= f_0(0,0)v_2,\\ F_0(0,n)(\lambda+\nu) &= f_0(0,n-1)v_1 + F_0(0,n-1)\nu, \quad n = 2,\cdots,N-1,\\ F_0(0,N)\lambda &= f_0(0,N)v_1 + f_0(0,N-1)v_1 + F_0(0,N-1)\nu.\end{aligned}$$

Substitution of (6) into these equations yields a set of N linear equations for the unknowns $F_0(0,n)$, which may be readily solved.

REMARK 4.2 (FINITE-CAPACITY CELL BUFFER). The case of a cell buffer with finite capacity can be treated similarly. To solve for the densities we also have to

use the solutions $y_2 e^{\sigma_2 u}$ and $y_4 e^{\sigma_4 u}$. This implies that we get extra terms in the solution formulated in Theorem 4.1. The larger set of unknown coefficients can be determined from the initial conditions and the extra boundary conditions in the states $(K,0),\ldots,(K,N)$, where K denotes the finite capacity of the cell buffer.

REMARK 4.3 (MULTI-STATE FLUID SOURCES). The extension from 2-state to M-state fluid sources is straightforward. The density vectors $f(u,n)$ of, in this case, dimension M can then be expressed in terms of the eigenvectors and eigenvalues of the matrices V_i and Q_i of dimensions $M \times M$. The resulting expressions are similar to the ones in Theorem 4.1. The numerical evaluation, of course, will be more involving.

5. Performance characteristics. In this section we will indicate how various important performance characteristics may be obtained from the steady-state distribution of the process $(U_\infty, N_\infty, I_\infty)$.

Mean content of the cell buffer: The mean content $E[U]$ of the cell buffer is given by

$$E[U] = \sum_{n=0}^{N} \int_{u=0}^{\infty} u(f_0(u,n) + f_1(u,n))du\,.$$

Substitution of (6) and the expressions in Theorem 4.1 for the densities $\tilde{f}_i(u,n)$ yields a finite sum expression for $E[U]$.

Mean cell delay: By Little's law we have the following relation between the mean cell delay $E[S_c]$ and the mean content of the cell buffer $E[U]$.

$$E[U] = \delta E[S_c]\,, \tag{7}$$

where δ denotes the average input rate of cells, i.e., $\delta = (v_0\lambda)/(\lambda+\mu)$. Hence, once $E[U]$ is known, the mean cell delay directly follows from (7).

Distribution of the cell delay: Let $a(u,c)$ denote the density that an arriving cell finds the shaper in state (u,c). The density that an arriving cell finds an amount of u in the cell buffer and n exponential quanta in the token buffer is equal to $f_1(u,n)\cdot(\lambda+\mu)/\lambda$. Given that there are n exponential quanta, the content of the token buffer is Erlang-n distributed with mean $nv_2/(\lambda+\nu)$. So it follows that

$$a(u,0) = f_1(u,0)\cdot\frac{\lambda+\mu}{\lambda}$$

and for $c>0$ that

$$a(u,c) = \frac{\lambda+\mu}{\lambda}\sum_{n=1}^{N} f_1(u,n)\left(\frac{\lambda+\nu}{v_2}\right)^n \frac{c^{n-1}}{(n-1)!}\,e^{-c(\lambda+\nu)/v_2}\,.$$

Next, let $S_c|(u,c)$ denote the cell delay given that an arriving cell finds the shaper in state (u,c). Since the cell buffer depletes at rate $v_1 - v_2$, the shaper is allowed to send at peak rate for at most $c/(v_1-v_2)$ units of time. Hence, if $u/v_1 < c/(v_1-v_2)$, then

$$S_c|(u,c) = \frac{u}{v_1} \tag{8}$$

and otherwise,

$$S_c|(u,c) = \frac{c}{v_1 - v_2} + \frac{1}{v_2}\left(u - v_1\frac{c}{v_1 - v_2}\right) = \frac{u-c}{v_2}\,. \tag{9}$$

By conditioning on the state seen on arrival we find that

$$\Pr[S_c > t] = \int_{(u,c):S_c|(u,c)>t} a(u,c)dudc + \int_{u=v_2 t}^{\infty} a(u,0)du\,. \tag{10}$$

Substitution of the expressions for $a(u,c)$ and $f_1(u,n)$ into (10) yields after some tedious algebra a finite sum expression for $\Pr[S_c > t]$.

Mean frame delay: The burst of cells arriving at the shaper during an on-period may be considered as a *frame* (or packet, or message) offered by a user of the network. A user will not be primarily interested in the delay of a single cell at the shaper, he is more interested in the *frame delay*. This is defined as the time elapsing from the arrival of the first cell of a frame till the departure of the last cell of the frame. The mean frame delay, $E[S_f]$, equals the expected time elapsing from the arrival of the first cell of a frame till the arrival of the last cell of the frame plus the mean delay of the last cell in the frame. Clearly, the first term is equal to $1/\mu$, the expected length of an on-period. The second term is equal to $E[S_c]$. Here we use the fact that the last cell in a frame sees at arrival exactly the same situation as an arbitrary cell does. This property is a consequence of the exponentiality of the frame sizes. The number of cells in a frame arriving before an arbitrary cell has the same distribution as the number of cells arriving before the last cell in a frame. We conclude that

$$E[S_f] = E[S_c] + 1/\mu\,.$$

Distribution of the frame delay: This distribution can be obtained similarly as the distribution of the cell delay. Let $b(u,c)$ denote the density that an arriving frame finds the shaper in state (u,c). The density that an arriving frame finds an amount of u in the cell buffer and n exponential quanta in the token buffer, is equal to $f_0(u,n)\cdot(\lambda+\mu)/\mu$. Using again that, if there are n exponential quanta, the content of the token buffer is Erlang-n distributed with mean $nv_2/(\lambda+\nu)$, it follows that

$$b(u,0) = f_0(u,0)\cdot\frac{\lambda+\mu}{\mu}$$

and for $c > 0$ that

$$b(u,c) = \frac{\lambda+\mu}{\mu}\sum_{n=1}^{N} f_0(u,n)\left(\frac{\lambda+\nu}{v_2}\right)^n \frac{c^{n-1}}{(n-1)!}\,e^{-c(\lambda+\nu)/v_2}\,.$$

Next, let $S_f|(s,u,c)$ denote the frame delay given that an arriving frame of size s finds the shaper in state (u,c). If $(u+s)/v_1 < c/(v_1 - v_2)$, then (cf. (8)–(9))

$$S_f|(s,u,c) = \frac{u+s}{v_1}$$

and otherwise,

$$S_f|(s,u,c) = \frac{c}{v_1 - v_2} + \frac{1}{v_2}\left(u + s - v_1\frac{c}{v_1 - v_2}\right) = \frac{u+s-c}{v_2}\,. \tag{11}$$

By conditioning on the state seen on arrival of a frame and using that the frame size is exponentially distributed with parameter $\kappa = \mu/v_0$ we find that

$$\begin{aligned} \Pr[S_f > t] &= \int_{(s,u,c):S_f|(s,u,c)>t} \kappa e^{-\kappa s} b(u,c) ds du dc \\ &\quad + \int_{(s,u):(u+s)/v_2>t} \kappa e^{-\kappa s} b(u,0) ds du \,, \end{aligned} \tag{12}$$

Substitution of the expressions for $b(u,c)$ and $f_0(u,n)$ into (12) yields after some algebra a finite sum expression for $\Pr[S_f > t]$ (that resembles the one for $\Pr[S_c > t]$).

Mean burst duration: The output of the shaper shows a cyclic pattern which starts with an idle period of the shaper during which the output rate is 0. Then follows a busy period during which the shaper starts to send at peak rate v_1 and after a while it possibly changes to sustainable rate v_2 (when the token buffer becomes empty) until the busy period ends. The duration of a peak-rate period will be called the burst duration, denoted by the random variable B. Clearly, the duration of an idle period is exponentially distributed with mean $1/\lambda$ (due to the memoryless property of the exponential distribution). From standard arguments from renewal theory, it follows that the mean burst duration $E[B]$ satisfies $E[B] = q_1/(q_0\lambda)$, where q_0, q_1 and q_2 denote the fractions of time the output rate equals 0, v_1 and v_2, resp. These fractions are given by

$$q_0 = \sum_{n=1}^{N} F_0(0,n) \,, \quad q_2 = \int_{u=0}^{\infty} (f_0(u,0) + f_1(u,0)) du \,, \quad q_1 = 1 - q_0 - q_2 \,.$$

Of course, the mean duration of a sustainable-rate period can be obtained along the same lines (notice that the duration of such a period may be zero, whereas the burst duration is always greater than zero).

Distribution of the burst duration: Let the random variable T denote the time the shaper is maximally allowed to send at peak rate at the beginning of a busy period, and let B_∞ denote the length of a busy period when the shaper is always allowed to send at peak rate (i.e., when the number of available tokens would be infinite). Then the burst duration B can be expressed as $B = \min\{B_\infty, T\}$, where the random variables B_∞ and T are independent. So $\Pr[B > t] = \Pr[B_\infty > t] \Pr[T > t]$. Hence, it remains to find the distributions of B_∞ and T. The distribution of T can be found by conditioning on the number of quanta at the beginning of a busy period, i.e.,

$$\Pr[T > t] = \sum_{n=1}^{N} a_n \Pr[E_n(\eta) > t] \,,$$

where a_n is the probability that at the beginning of a busy period there are n quanta in the token buffer and $E_n(\eta)$ denotes an Erlang-distributed random variable with parameters n and η, representing the time the shaper may send at peak rate given that there are n quanta available. Since the number of times per time unit that one leaves state $(0,n)$ is proportional to $f_1(0,n)$, it follows that

$$a_n = \frac{f_1(0,n)}{\sum_{m=1}^{N} f_1(0,m)} \,.$$

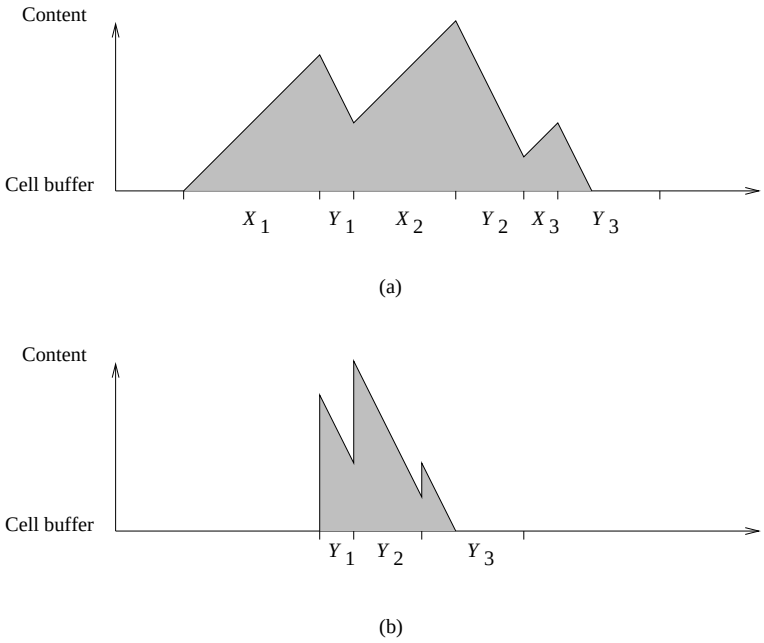

FIG. 4. *The cell buffer content of a shaper which always works at maximum rate v_1 for a realization of on-periods X_i and off-periods Y_i, where in figure (b) the on-periods have been removed*

To find the distribution of B_∞ let us consider the realization in Fig. 4(a) of a shaper which always works at rate v_1, where the random variables X_i denote the on-periods and the random variables Y_i denote the off-periods of the on-off source. Notice that during an on-period the content of the cell buffer increases with rate $v_0 - v_1$ and during an off-period it decreases with rate v_1. Hence, for the realization in Fig. 4(a) we have

$$B_\infty = X_1 + X_2 + X_3 + \frac{v_0 - v_1}{v_1}(X_1 + X_2 + X_3) = \frac{v_0}{v_0 - v_1} Z\,,$$

where

$$Z = \frac{v_0 - v_1}{v_1}(X_1 + X_2 + X_3)\,.$$

If we remove the on-periods from that realization (see Fig. 4(b)), then it is easily seen that the resulting period Z can be interpreted as a realization of a busy period of an $M/M/1$ queue with interarrival times Y_i and service times $X_i \cdot (v_0 - v_1)/v_1$. Hence, for the density $f_Z(t)$ we have that

$$f_Z(t) = \frac{1}{t\sqrt{\alpha/\lambda}}\, e^{-(\lambda+\alpha)t} I_1(2t\sqrt{\lambda\alpha})$$

where $\alpha = v_1\mu/(v_0 - v_1)$ and $I_1(\cdot)$ denotes the modified Bessel function of the first kind of order one. From the density $f_Z(t)$ we can calculate $\Pr[B_\infty > t]$ and hence also $\Pr[B > t]$.

Burstiness of the output stream: A simple measure for the burstiness of the output stream is the variance σ_C^2 of the output rate of the shaper, i.e.,

$$\sigma_C^2 = q_0(0 - \delta)^2 + q_1(v_1 - \delta)^2 + q_2(v_2 - \delta)^2\,,$$

where q_0, q_1 and q_2 are the fractions of time the output rate equals 0, v_1 and v_2, respectively, and δ is the average output rate (= average input rate).

6. Validation of the approximative model. To investigate how well the discretization technique works we compare the cell delay in the original (continuous) model with the cell delay in the approximative (discrete) model. The results for the original model have been obtained by simulation, while the results for the approximative model have been calculated via the procedure outlined in the previous sections.

λ	μ	v_0	v_1	v_2	ρ	C	$E[S_N]$				$E[S]$
						N	5	10	25	50	
0.144	0.72	14.4	10	3	0.8	7	19.81	19.80	19.80	19.79	19.80
						14	17.91	17.87	17.85	17.84	17.84
						28	14.76	14.64	14.56	14.54	14.52
						70	8.74	8.37	8.14	8.06	7.98
0.162	0.81	16.2	10	3	0.9	7	43.01	43.01	43.00	43.00	43.01
						14	40.91	40.88	40.87	40.87	40.87
						28	37.07	37.00	36.95	36.93	36.93
						70	27.98	27.65	27.45	27.38	27.32
0.171	0.855	17.1	10	3	0.95	7	89.32	89.32	89.32	89.32	89.27
						14	87.11	87.10	87.09	87.09	87.03
						28	82.88	82.84	82.81	82.80	82.74
						70	71.63	71.41	71.27	71.23	71.13
0.144	0.72	14.4	5	3	0.8	4	20.74	20.74	20.73	20.73	20.74
						8	19.63	19.61	19.60	19.60	19.60
						16	17.72	17.66	17.63	17.62	17.61
						40	13.60	13.41	13.29	13.25	13.22
0.162	0.81	16.2	5	3	0.9	4	43.98	43.97	43.97	43.97	43.98
						8	42.76	42.75	42.74	42.74	42.75
						16	40.49	40.46	40.44	40.43	40.44
						40	34.74	34.59	34.51	34.48	34.45
0.171	0.855	17.1	5	3	0.95	4	90.31	90.30	90.30	90.30	90.25
						8	89.03	89.03	89.02	89.02	88.97
						16	86.57	86.55	86.54	86.54	86.49
						40	79.78	79.69	79.64	79.62	79.56

TABLE 1
Convergence of $E[S_N]$ to its limit $E[S]$

The rationale behind the parameter settings in Table 1 is the following. The time to transmit 30 cells over the output link is used as unit of time. It is assumed that the bandwidth dedicated to the connection is 10% of the link capacity, so $v_2 = 3$. For the peak rate v_1 we consider the values $v_1 = 5$ and $v_1 = 10$. The load ρ of the shaper, defined as

$$\rho = \frac{v_0}{v_2} \cdot \frac{\lambda}{\lambda + \mu},$$

is varied from 0.8, 0.9 to 0.95. In each example we set $\mu = 5\lambda$ and the mean frame size, i.e. v_0/μ, is set to 20 cells. Then, depending on the load, v_0 varies from 14.4, 16.2 to 17.1, and μ from 0.72, 0.81 to 0.855. To determine appropriate values for C, first note that, when the content of the token buffer is equal to K, where

$$K = \frac{1}{\mu} \cdot \frac{v_0}{v_1} \cdot (v_1 - v_2),$$

the shaper is capable of sending on average one frame at peak rate. For $v_1 = 10$ and $v_1 = 5$, this yields $K = 14$ and $K = 8$, resp. In the examples in Table 1 the size C of the token buffer is varied from $0.5K$, K, $2K$ to $5K$.

In the last column of Table 1 we list $E[S_c]$, the mean cell delay in the original model. The simulation results are obtained from 40 runs of approximately 10^7 frames, yielding an accuracy of 1% for $\rho = 0.8, 0.9$ and 2% for $\rho = 0.95$. Other columns in Table 1 list the corresponding values of $E[S_N]$, the mean cell delay in the approximative model when the maximum number of quanta in the token buffer is equal to N. Computation of $E[S_N]$ requires a fraction of a second. This is negligible compared to the effort required to obtain $E[S_c]$ by simulation. We can conclude from the results in Table 1 that $E[S_N]$ is an excellent approximation for $E[S_c]$ already for small N.

For one of the parameter settings in Table 1 we show in Fig. 5 results for the distribution of the cell delay and the frame delay. The simulated distributions have been obtained from one long run of approximately 10^7 frames. The two graphs demonstrate that the stochastic discretization produces accurate approximations. Note that near $t = 2$ the direction of the curve of the cell delay suddenly changes. This is due to the fact that the shaper switches from the fast transmission rate v_1 to the slow rate v_2.

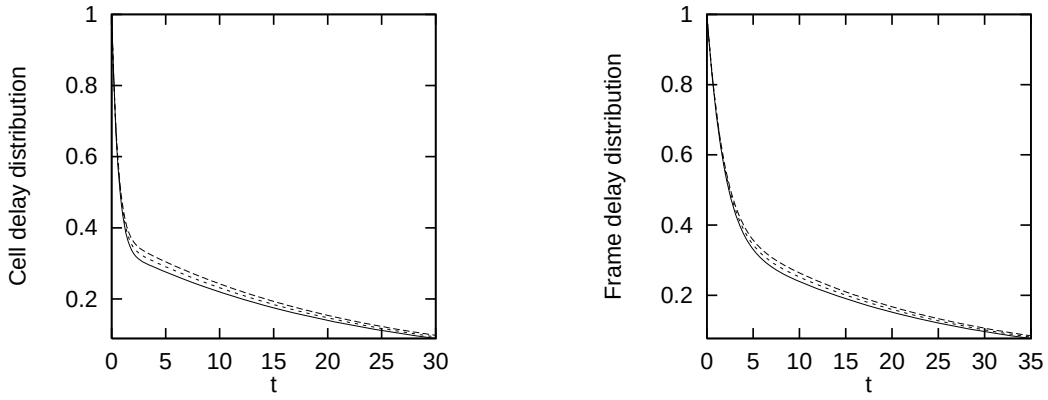

FIG. 5. *The cell delay distribution* $\Pr[S_c > t]$, *and the frame delay distribution* $\Pr[S_f > t]$ *for* $\lambda = 0.144$, $\mu = 0.72$, $v_0 = 14.4$, $v_1 = 10$, $v_2 = 3$ *and* $C = 70$. *The solid line is the simulated distribution, the dashed and dotted line are the approximate distributions for* $N = 5$ *and* $N = 10$

Finally, we pay attention to the burst duration. In Table 2 we compare the mean burst duration in the approximative model, $E[B_N]$, with the one in the original model, $E[B]$. In each setting, $E[B]$ is estimated from 40 runs of approximately 10^7 frames. The accuracy of the simulation results is 0.1%. The results show fast convergence of $E[B_N]$ to its limit $E[B]$.

In the left graph of Fig. 6 we show results for $\Pr[B > t]$, the distribution of the burst duration. The simulated distribution has been obtained from one long run of approximately 10^7 frames. Fig. 6 clearly illustrates the approximation of the *maximum* burst duration, $t_{\max} = C/(v_1 - v_2)$, by an Erlang-N distribution. Substantial improvement of the approximation near $t = t_{\max}$ requires very large values of N. However, the symmetry of the approximative curve around $t = t_{\max}$ can be exploited to improve the approximation without enlarging N. Let the random variable B_N denote the burst duration in the approximative model, then $\Pr[B > t]$ for $0 \le t \le t_{\max}$ may be approximated by $\Pr[B_N > t] + \Pr[B_N > t_{\max} + (t_{\max} - t)]$. The right graph of Fig. 6 demonstrates that this simple approximation is very accurate around $t = t_{\max}$.

λ	μ	v_0	v_1	v_2	ρ	C	$E[B_N]$				$E[B]$
						N	5	10	25	50	
0.144	0.72	14.4	10	3	0.8	7	0.699	0.702	0.704	0.705	0.706
						14	1.076	1.084	1.089	1.090	1.092
						28	1.478	1.491	1.499	1.501	1.504
						70	1.901	1.919	1.931	1.935	1.939
0.162	0.81	16.2	10	3	0.9	7	0.692	0.694	0.696	0.696	0.697
						14	1.063	1.068	1.070	1.071	1.072
						28	1.460	1.467	1.471	1.473	1.474
						70	1.894	1.904	1.910	1.912	1.914
0.171	0.855	17.1	10	3	0.95	7	0.689	0.690	0.691	0.692	0.692
						14	1.056	1.059	1.061	1.062	1.063
						28	1.450	1.454	1.456	1.457	1.458
						70	1.885	1.890	1.893	1.894	1.895
0.144	0.72	14.4	5	3	0.8	4	1.469	1.481	1.489	1.492	1.494
						8	2.353	2.377	2.393	2.398	2.404
						16	3.416	3.456	3.480	3.488	3.497
						40	4.814	4.871	4.906	4.918	4.930
0.162	0.81	16.2	5	3	0.9	4	1.464	1.474	1.481	1.483	1.485
						8	2.349	2.368	2.380	2.384	2.388
						16	3.434	3.462	3.479	3.485	3.491
						40	4.936	4.972	4.994	5.001	5.009
0.171	0.855	17.1	5	3	0.95	4	1.462	1.471	1.476	1.478	1.480
						8	2.346	2.363	2.373	2.377	2.380
						16	3.440	3.462	3.476	3.480	3.485
						40	4.983	5.007	5.022	5.026	5.032

TABLE 2
Convergence of $E[B_N]$ to its limit $E[B]$

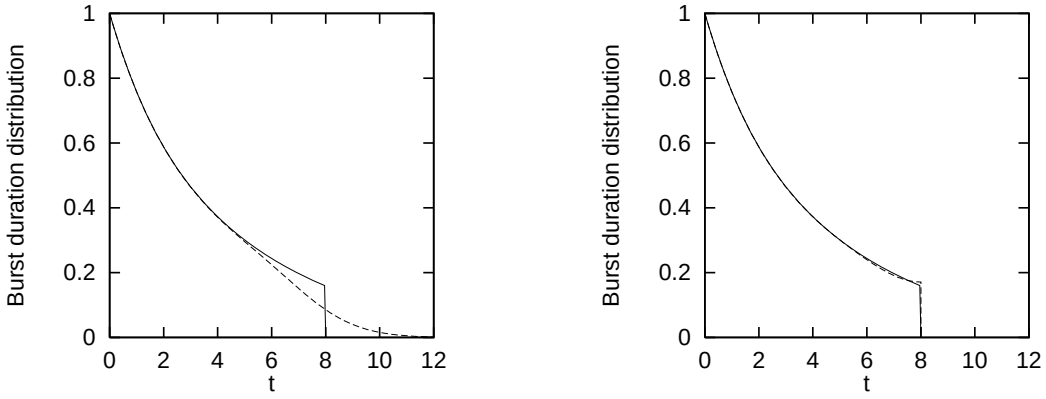

FIG. 6. *The distribution of the burst duration $P[B > t]$ for $\lambda = 0.171$, $\mu = 0.855$, $v_0 = 17.1$, $v_1 = 5$, $v_2 = 3$ and $C = 16$. The solid line is the simulated distribution, the dashed line in the left graph is the approximate distribution $\Pr[B_N > t]$, and the one in the right graph is the approximate distribution $\Pr[B_N > t] + \Pr[B_N > t_{\max} + (t_{\max} - t)]$, both for $N = 25$*

7. Numerical results. To illustrate the behaviour of the two-level traffic shaper, we first look into the influence of the size of the token buffer on the performance of the shaper. All results in this section are obtained from the approximative model with $N = 25$.

In Fig. 7, we display the mean frame delay $E[S_f]$, the frame delay distribution, the mean burst duration $E[B]$ and the burstiness σ^2 as a function of the size of the token

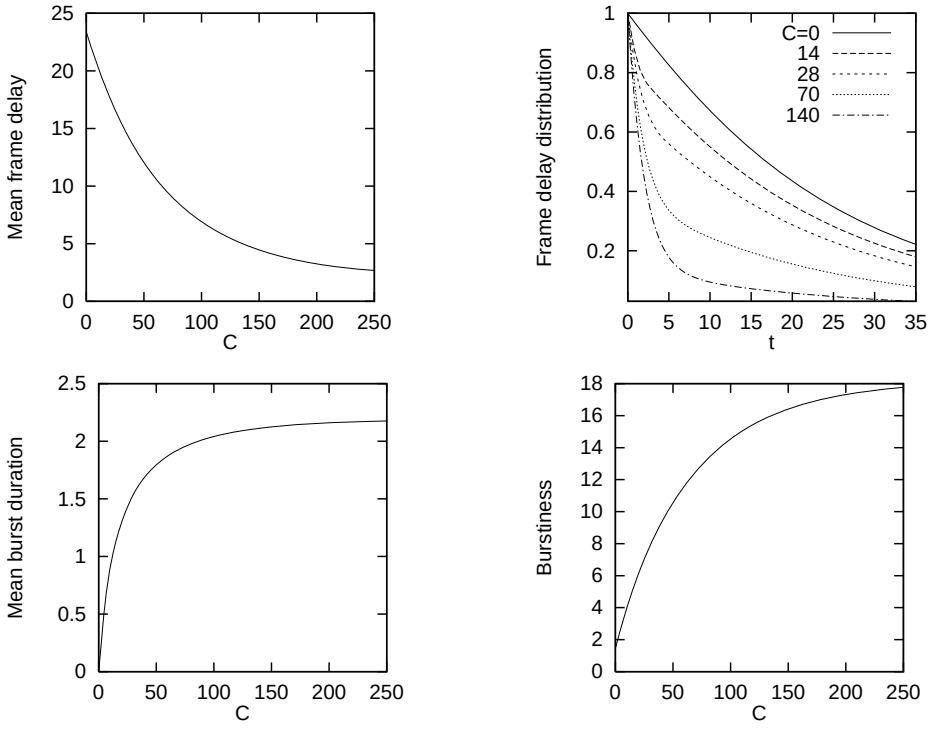

FIG. 7. *The mean frame delay* $E[S_f]$, *the frame delay distribution* $\Pr[S_f > t]$, *the mean burst duration* $E[B]$ *and the burstiness* σ^2 *as a function of* C *for the parameter setting* $\lambda = 0.144$, $\mu = 0.72$, $v_0 = 14.4$, $v_1 = 10$ *and* $v_2 = 3$

buffer C, for a fixed parameter setting. Note that a large part of the possible reduction of the mean frame delay is already achieved for small values of C. Clearly, the mean burst duration and the burstiness of the output stream increase as the size of the token buffer increases. Graphs such as the ones in Fig. 7 may be used to make the trade-off between the benefit of delay reduction versus the drawback of extra burstiness of the output stream.

Finally, in Fig. 8 we show the same performance characteristics as in Fig. 7, but now as a function of ρ, the load of the shaper. For the setting considered in Fig. 8 we keep v_1 and v_2 fixed, $v_1 = 10$ and $v_2 = 3$. The other parameters depend on the load as follows, $v_0 = 18\rho$, $\mu = 0.9\rho$ and $\lambda = 0.18\rho$. So the number of frames generated per unit of time depends on the load, but the mean frame size is constant, i.e. 20 cells. Clearly, as ρ increases, the delay also increases. The mean burst size remains fairly constant, but the burstiness of the output stream decreases as ρ increases, which is due to the fact that the fraction of the traffic transmitted at high rate decreases.

8. Summary. In this paper we studied a two-level traffic shaper for an on-off source. Using a stochastic discretization technique we were able to derive an excellent approximation for the stationary joint distribution of the content of the cell buffer and the content of the token bank. From this distribution various performance measures like the cell delay distribution, the frame delay distribution, the distribution of the burst duration and the burstiness of the output stream could be obtained. The results of this paper can be used to make the trade-off between delay reduction on the one hand and extra burstiness on the other hand.

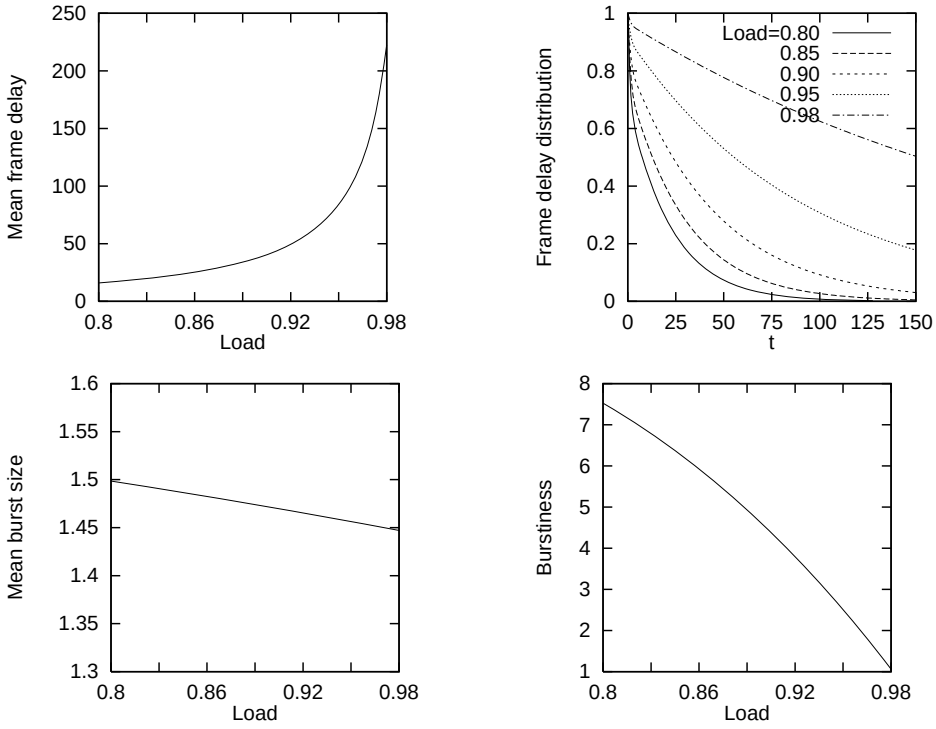

FIG. 8. *The mean frame delay* $E[S_f]$, *the frame delay distribution* $\Pr[S_f > t]$, *the mean burst duration* $E[B]$ *and the burstiness* σ^2 *as a function of the load* ρ *for the parameter setting* $\lambda = 0.18\rho$, $\mu = 0.9\rho$, $v_0 = 18\rho$, $v_1 = 10$ *and* $v_2 = 3$

REFERENCES

[1] I.J.B.F. ADAN, E.A. VAN DOORN, J.A.C. RESING AND W.R.W. SCHEINHARDT, Analysis of a single-server queue interacting with a fluid reservoir. *Queueing Systems* **29** (1998) 313–336.

[2] D. ANICK, D. MITRA AND M.M. SONDHI, Stochastic theory of a data-handling system with multiple sources. *Bell System Tech. J.* **61** (1982) 1871–1894.

[3] A.W. BERGER, Performance analysis of a rate-control throttle where tokens and jobs queue. *IEEE J. Select. Areas Commun.* **9** (1991) 165–170.

[4] A.W. BERGER AND W. WHITT, The impact of a job buffer in a token-bank rate-control throttle. *Stochastic Models* **8** (1992) 685–717.

[5] F.M. BRONCHIN, A cell spacing device for congestion control in ATM networks. *Performance Evaluation* **16** (1992) 107–127.

[6] A.I. ELWALID AND D. MITRA, Analysis and design of rate-based congestion control of high speed networks, Part I: Stochastic fluid models, access regulation. *Queueing Systems* **9** (1991) 29–64.

[7] D. FREEDMAN, Approximating countable Markov chains. Holden Day, San Francisco, 1971.

[8] D.P. KROESE AND W.R.W. SCHEINHARDT, A system of fluid queues with feedback, Memorandum 1422, University of Twente, Faculty of Applied Mathematics, The Netherlands, 1997.

[9] B.V. PATEL AND C.C. BISDIKIAN, On the performance behavior of ATM end-stations. Proc. INFOCOM '95, Boston, 188–196.

[10] M. RITTER, Performance analysis of the dual cell spacer in ATM systems. IFIP 6th International Conference on High Performance Networking, Palma de Mallorca, 1995.

[11] W.R.W. SCHEINHARDT, Markov-modulated and feedback fluid queues. Ph.D. Thesis, University of Twente, The Netherlands, 1998.

[12] M. SIDI, W.Z. LIU, I. CIDON AND I. GOPAL, Congestion control through input rate regulation. *IEEE Trans. Commun.* **41** (1993) 471–477.

[13] J. TURNER, New directions in communications (or which way to the information age?). *IEEE Communications Magazine* **24** (1986) 8–15.