Output Traffic Analysis of a Leaky Bucket Traffic Shaper Fed by a Bursty Source

Sabine Wittevrongel and Herwig Bruneel
Laboratory for Communications Engineering, University of Ghent
Sint-Pietersnieuwstraat 41, B-9000 Gent, Belgium.

ABSTRACT A leaky bucket traffic shaper, consisting of an infinite data buffer and a finite token pool, and fed by a bursty ON/OFF source, is considered in this paper. Tokens are generated periodically. Assuming geometrically distributed ON and OFF periods, the system is analyzed by means of a generating functions approach. A method for calculating the exact steady-state distributions of the data-buffer and the token—pool occupancies as well as the cell waiting times in the leaky bucket is presented. Also, a technique to derive the joint probability generating function of two consecutive interdeparture times at the output of the traffic shaper is described. From this pgf, closed-form expressions can be obtained for the variance of the interdeparture times and the correlation between two successive interdeparture times. The results are illustrated by means of some numerical examples and the influence of both the input-traffic characteristics and the leaky—bucket parameters on the output-traffic characteristics is investigated.

1. Introduction

The asynchronous transfer mode (ATM) has been chosen as the transport method for the broadband integrated services digital network (BISDN). Since an ATM network supports a large number of bursty sources, statistical multiplexing can be used to gain bandwidth efficiency. However, if a large number of traffic sources become active simultaneously, severe network congestion can result. Therefore, congestion control is a necessity for ATM networks. One of the methods proposed to reduce network congestion is traffic shaping. The purpose of traffic shaping is to throttle cell inputs into the network in order to reduce the burstiness of the sources.

In this paper, we consider the leaky bucket traffic shaping mechanism, consisting of a data buffer and a token pool. A number of studies have investigated the behavior and the performance of the leaky bucket traffic shaper in recent years. Examples include [5], [6] in the continuous—time domain and [1]—[4] in the discrete—time domain. Usually, the quantities studied in these investigations are the data—buffer occupancy and the waiting times experienced by cells in the data buffer. Only few studies focus on the characteristics of the output traffic of the leaky bucket. In a previous paper [4], the authors have derived an approximate analytic expression for the tail distribution of the data buffer occupancy, for a leaky bucket traffic shaper fed by an ON/OFF source with geometrically distributed ON and OFF periods. The main contribution of the present paper is an analytical discrete—time study of the shaped traffic characteristics in terms of both the leaky—bucket and the input—traffic parameters. The numerical evaluation of the obtained formulas is simple and not time consuming, even for large values of the token generation period.

The outline of the paper is as follows. In section 2, we describe the system and the traffic model under study. A new analytic technique to calculate the exact steady—state distributions of the data—buffer and token—pool occupancies is presented in section 3. In section 4, the waiting time distribution is derived. A technique to derive the joint probability generating function of two consecutive interdeparture times at the output of the traffic shaper is described in section 5. Finally, in section 6, some numerical results are discussed.

2. System and traffic model

We consider a leaky bucket traffic shaper, consisting of a data buffer with one input line, a token pool for at most C tokens and one output line (see Figure 1). Time is slotted and one slot is the fixed—length time interval required to transmit exactly one fixed—length ATM cell. Cells generated by a source are stored in an infinite data buffer until they can be transmitted. Whenever a cell is transmitted, one token is removed from the token pool. When the token pool is empty, no cells can be transmitted. The cells queued in the data buffer are served on a first—come first—served basis (FCFS). Every M slots, a new token is generated. Time can thus be divided into frames of length M (slots), such that one token is generated at the end of the Mth slot of each frame. Tokens are added to the finite token pool if the latter is not full. If the token pool contains C tokens at the beginning of the token generation slot, the newly generated token is discarded, unless the data buffer inlet was active during the token generation slot. From the above description we note that there are several parameters associated with the leaky bucket system. The token generation period M controls the average number of cells transmitted to the network. The capacity of the token pool C determines the burstiness of the output traffic (in terms of the number of cells that can pass the leaky bucket back—to—back).

In the analysis, we consider a bursty traffic source on the input link of the data buffer, that alternates between two states called ON and OFF. In the ON state, exactly one
Figure 1 Queueing model of the leaky bucket traffic shaper

ATM cell is generated during each slot, whereas no cells are generated in the OFF state. It is assumed that state transitions can only occur at slot boundaries. The lengths of the ON and the OFF periods are modeled as two independent sets of independent geometrically distributed random variables, with mean lengths $1/(1-\alpha)$ and $1/(1-\beta)$ respectively. The special case of an uncorrelated Bernoulli arrival process corresponds to $\alpha + \beta = 1$.

Cells arrive during a slot and are taken out synchronously from the data buffer at slot boundaries only. Tokens are generated at the end of a slot, just before the potential departure of a cell from the data buffer, and hence a token can be consumed at the end of its generation slot. The queueing system is investigated at the beginning of every slot, i.e., just after the possible transmission of a cell from the data buffer.

3. Distributions of the data-buffer and token-pool occupancies in the steady state

Let $a_{r,k}$ ($1 \leq r \leq M$) represent the number of cell arrivals during the slot preceding the $r$th slot of frame $k$. Let $q_{r,k}$ and $t_{r,k}$ denote the number of cells stored in the data buffer and the tokens in the token pool respectively (excluding the "couple" which is possible under transmission) at the beginning of the $r$th slot of frame $k$ and define $d_{r,k} = q_{r,k} - t_{r,k}$. Since during a slot, at most one cell can arrive and at most one token can be generated, it is impossible for the token pool and the cell buffer to be nonempty simultaneously at the beginning of a slot. Hence, the single variable $d_{r,k}$ suffices to describe the contents of the data buffer and the token pool at the beginning of the $r$th slot of frame $k$. The following system equations hold for $d_{r,k}$:

$$d_{r,k} = d_{r-1,k} + a_{r,k}, \quad 2 \leq r \leq M,$$

$$d_{1,k} = d_{M,k-1} + a_{1,k-1}, \quad d_{M,k-1} > -C,$$

$$d_{1,k} = d_{M,k-1} - C.$$  \hspace{1cm} (3)

Now let us define the partial probability generating functions (pgfs) $Q_{r,n,k}(z)$, $0 \leq n \leq 1$, $1 \leq r \leq M$, as

$$Q_{r,n,k}(z) = \sum_{j=-C}^{\infty} \text{Prob}[a_{r,k} = n, d_{r,k} = j] z^j. \hspace{1cm} (4)$$

Then, due to equations (1)–(3) and the above described traffic model, the partial pgfs $Q_{r,0,k}(z)$ and $Q_{r,1,k}(z)$ can be expressed in terms of $Q_{r-1,0,k-1}(z)$ and $Q_{r-1,1,k-1}(z)$. After some calculations, we find, for $1 \leq r \leq M$,

$$z^{Q_{r,0,k}(z)} - Q_{r,1,k}(z) = F \left[ \begin{array}{c} Q_{r-1,0,k-1}(z) \\ Q_{r-1,1,k-1}(z) \\ \end{array} \right]$$

$$+ \text{Prob}[a_{1,k} = 0, d_{M,k-1} = -C] z^{-C} (z-1)^{F-1} \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \hspace{1cm} (5)$$

where

$$F^{-1} = \begin{bmatrix} f_{0}(z) & f_{1}(z) \\ f_{0}(z) & f_{1}(z) \end{bmatrix} = \begin{bmatrix} \beta & 1-\alpha \\ 1-\beta & \alpha \end{bmatrix}.$$  \hspace{1cm} (6)

The functions $f_{1}(z)$ can be derived in terms of the 2 eigenvalues of the matrix $F$. When the queuing system has reached its steady state, $Q_{r,n,k}(z)$ becomes independent of $k$, for all $n$ ($0 \leq n \leq 1$) and all $r$ ($1 \leq r \leq M$). This implies that in the steady state the distribution of the difference between the data-buffer occupancy and the token-pool occupancy varies periodically, with period $M$. From (5) and (6), the steady-state partial pgfs $Q_{r,0}(z)$ and $Q_{r,1}(z)$ associated to the $r$th slot of a frame, can then be derived explicitly as

$$Q_{r,n}(z) = \frac{P_{0,n,r}(z)}{[z-f_{0}(z)]^{r} - f_{0}(z)} - \frac{P_{1,n,r}(z)}{[z-f_{0}(z)]^{r} - f_{0}(z)}$$

$$= \frac{1}{[z-f_{0}(z)]^{r} - f_{0}(z)} \begin{bmatrix} \beta & 1-\alpha \\ 1-\beta & \alpha \end{bmatrix}^{r-1} p_{c}, \hspace{1cm} (7)$$

where

$$P_{r,n}(z) = f_{0}(z) \{z-f_{0}(z)\}^{r} + f_{0}(z) \{z-f_{0}(z)\}^{r} n_{0}$$

and

$$p_{c} = \lim_{k \to \infty} \text{Prob}[a_{1,k} = 0, d_{M,k} = -C]. \hspace{1cm} (8)$$

is the steady-state probability of having an empty data buffer and a full token pool at the beginning of the $M$th slot of an arbitrary frame and a passive inlet during this slot. The quantity $p_{c}$ can be derived from the normalization equation $Q_{r,0}(1) + Q_{r,1}(1) = 1$. Using de l'Hospital's rule, we obtain $p_{c} = 1 - M p$, where $p$ is the average traffic load offered by the source, i.e., $p = (1-\beta)/(2\alpha - \beta)$.

Now, let us denote by $a_{r,k}$ and $d_{r,k}$ the steady-state versions of $a_{r,k}$ and $d_{r,k}$ respectively. The joint probability mass functions (pmfs) $p_{r}(n,j) = \text{Prob}[a_{r,k} = n, d_{r,k} = j]$, $0 \leq n \leq 1$, $1 \leq r \leq M$, can be calculated by taking the inverse $z-$transform of $Q_{r,n}(z)$. A practical method for the numerical calculation of $p_{r}(n,j)$ can be found in [8]. In the next sections, these pmfs will be used for the derivation of the pgf of the cell waiting times and the joint pgf of two consecutive interdeparture times. Also, it is clear that the pmf of $d_{r}$ can be derived as $\text{Prob}[d_{r+k} = j] = p_{r}(0,j) + p_{r+1}(1,j)$.

As in [7], it can be argued that for sufficiently large $j$, $p_{r}(n,j)$ will be dominated by the contribution of the pole of $Q_{r,n}(z)$ with the smallest absolute value. Denoting this dominant pole by $z_{0}$, it is clear that $z_{0}$ should necessarily be.
real and strictly positive, in order to ensure that \( p_r(n,j) \), 
\( 0 \leq n \leq 1 \), is nonnegative for all \( j \). Therefore, we find the following approximation for \( p_r(n,j) \), for sufficiently large:

\[
p_{r}(n,j) \approx \frac{B_{r,n} z^{-j}}{z_0}.
\]

where \( B_{r,n} \) is the residue of \( Q_{r,n}(z) \) in the point \( z = z_0 \).

Finally, let \( q \) and \( t \) denote the data-buffer occupancy and the token-pool occupancy respectively, at the beginning of an arbitrary slot in the steady state and define \( d = q - t \). The distributions of \( d, q \) and \( t \) can be derived from the above results, as follows:

\[
\text{Prob}[d=j] = \frac{1}{M} \sum_{r=1}^{M} \text{Prob}[d_r=j]; \quad (11)
\]

\[
\text{Prob}[q=0] = \sum_{j=-C}^{0} \text{Prob}[d=j]; \quad (12)
\]

\[
\text{Prob}[q=j] = \text{Prob}[d=j], \quad 0 < j ; \quad (13)
\]

\[
\text{Prob}[t=0] = \sum_{j=0}^{\infty} \text{Prob}[d=j]; \quad (14)
\]

\[
\text{Prob}[t=-j] = \text{Prob}[d=-j], \quad 0 < j \leq C . \quad (15)
\]

Furthermore, using (12) and (13), we obtain the mean data-buffer occupancy \( E[q] \) as

\[
E[q] = E[d] - \sum_{j=-C}^{0} j \text{Prob}[d=j], \quad (16)
\]

where \( E[d] \) can be calculated from (7), (8) and (11), as

\[
E[d] = -C + \frac{(M-1)p}{2(1-Mp)} \left( 1 + \frac{2(\alpha-p)}{\alpha} \right). \quad (17)
\]

The tail distributions of \( d_r, d \) and \( q \) can be derived by combining (10)–(13). It has been observed that the obtained results are in full agreement with those in [4].

4. Waiting time distribution in the steady state

Let us define the waiting time of a cell as the time period between the end of the slot during which the cell has arrived in the data buffer and the start of the slot during which the cell will be transmitted. With this definition the cell waiting time always consists of an integral number of slots and thus it can be considered as a discrete random variable. Let \( w \) with pmf \( w(n) \) and pgf \( W(z) \), denote the cell waiting time in the steady state. In order to derive an expression for \( W(z) \), we further refine our state description of the system. Specifically, we define that the system is in state \( (r,n,j) \) at the beginning of the current slot if the current slot is the \( r \)th slot of a frame, the difference between the data-buffer and the token-pool contents at the beginning of this slot equals \( j \), the number of cell arrivals during the previous slot equals \( n \) and a cell arrives during the current slot. We will refer to these states as the "arrival states". Let us further define \( p_{r}(n,j) \) as the steady-state probability of finding the system in state \( (r,n,j) \). The probabilities \( p_{r}(n,j) \) can then be expressed in terms of the joint pmf's \( p_r(n,j) \), as follows, for \( j \geq -C, 1 \leq r \leq M \):

\[
p_{r}(0,j) = (1-\beta) \frac{p_r(0,j)}{M} ; \quad (18)
\]

\[
p_{r}(1,j) = \alpha \frac{p_r(1,j)}{M} . \quad (19)
\]

For arrival states \( (r,n,j) \), \( j \leq -1 \), the waiting time of the cell that arrived during slot \( r \) is zero. On the other hand, for arrival states \( (r,n,j) \), \( j \geq 0 \), the waiting time of this cell equals \( M-r+jM \). Therefore, the pgf \( W(z) \) is given by

\[
W(z) = \frac{1}{z_0} \sum_{r=1}^{M} \sum_{n=0}^{\infty} \sum_{j=-C}^{\infty} p_{r}(n,j) z^{M-r+jM} , \quad (20)
\]

where \( a \) is the steady-state probability that a cell arrives, i.e., the cumulated probability of all the arrival states. It is clear that \( a \) equals the load \( p \) on the inlet of the data buffer. Using (18)–(19), the pgf \( W(z) \) can also be expressed in terms of the pgf's \( Q_{r,n}(z) \). The mean waiting time \( E[w] \) experienced by cells in the data buffer can then be obtained by taking the first derivative of \( W(z) \) with respect to \( z \) at \( z=1 \). It has been verified that \( E[w] = E[q]/p \), in agreement with Little's theorem.

The pmf \( w(i) \) of \( w \) can be derived in terms of the joint pmf's \( p_r(n,j) \), by taking the inverse \( z \)-transform of (20). Furthermore, using (10), we find the following approximation for the tail distribution of \( w \):

\[
w(M+r) \approx \frac{W_{r} z^{-j}}{z_0} , \quad (21)
\]

for sufficiently large values of \( M+r \), where

\[
W_{r} = \frac{1}{M} \left[ \frac{1-\beta}{1-\alpha} B_{r,0} \right] + \left[ \frac{1}{1-M} \right] \left[ \frac{1-\beta}{1-\alpha} B_{r,1} \right] , \quad 0 \leq r \leq M-1. \quad (22)
\]

It can also be shown that \( W_{r} z_{0}^{-r}/M \) is equal for all \( r, 0 \leq r \leq M-1 \). Hence, from equation (22) we note that the distribution of the waiting times has a geometric tail.

5. Joint pgf of two consecutive interdeparture times

Similarly to the input traffic, we can expect the output traffic of the leaky bucket to be bursty. However, due to the leaky bucket traffic shaping mechanism, its characteristics will be changed with respect to the input traffic. In this section, we describe a method to determine the joint pgf \( I(x,y) \) of two consecutive interdeparture times at the outlet of the leaky bucket traffic shaper. More detailed derivations can be found in [8].

Firstly, as in section 4, we refine our state description of the system. We define that the system is in state \( (r,n,j) \) (at the beginning of the current slot) if the current slot is the \( r \)th slot of a frame, the difference between the data-buffer and token-pool contents at the beginning of this slot equals \( j \), the number of cell arrivals during the previous slot equals \( n \) and a cell departs at the end of the previous slot. These (microscopic) states will be referred to as the "departure states". We also define \( p_r^{(0)}(n,j) \) as the steady-state
probability of finding the system in state \((r,n,j)^{d}\). These probabilities can be derived from the probabilities \(p_{r}(n,j)\).
For \(2 \leq r \leq M\), we have
\[
p_{d}(0,j) = 0, \quad -C \leq j \leq \infty \quad (23)
\]
\[
p_{d}(1,j) = 0, \quad 1 \leq j \leq \infty \text{ and } j = -C \quad (24)
\]
\[
p_{d}(1,j) = p_{0}(1,j)/M, \quad -C < j \leq 0 \quad (25)
\]
whereas for \(r = 1\), the following relationships hold:
\[
p_{1}(1,j) = p_{0}(1,j)/M, \quad -C \leq j \leq \infty \quad (26)
\]
\[
p_{1}(0,j) = 0, \quad -C \leq j \leq -1 \quad (27)
\]
\[
p_{1}(0,j) = p_{0}(0,j)/M, \quad 0 \leq j \leq \infty \quad (28)
\]
Next, we categorize the possible departure states into several groups, for each of which it is possible to derive the (conditional) joint pgf of the next two interdeparture times. Finally, averaging over all the possible departure states, we obtain an explicit closed-form expression for the joint pgf \(I(x,y)\), which is given in [8].

The marginal pgf of the interdeparture times (between two consecutive cells) at the output of the shaper can be found by putting \(x = 1\) or \(y = 1\) in \(I(x,y)\). An expression for \(I(z,1)\) is given in [8] and it has been verified that \(I(z,1) = I(1,z)\). In general, closed-form expressions can be obtained from \(I(x,y)\) for any moment of the two-dimensional interdeparture-time distribution. It suffices to evaluate the consecutive mixed partial derivatives of \(I(x,y)\) with respect to \(x\) and \(y\), for \(x = y = 1\), and to express the desired moment as a function of these mixed partial derivatives. In [8], explicit formulas are given for the mean and the variance of the interdeparture times and the coefficient of correlation between two consecutive interdeparture times.

6. Numerical results and discussion

Let us define the "burstiness factor" \(K\) of the source as
\[
K = (1-p)/(1-\alpha) = p/(1-\beta), \quad (29)
\]
where \(p\) is the average input load. It is clear that \(p\) describes the ratio of the mean lengths of the ON and OFF periods, whereas \(K\) is a measure for the absolute lengths of these periods. The mean and variance of the interarrival times at the inlet of the traffic shaper can then be derived as
\[
E[IAT] = 1/p, \quad (30)
\]
and\ind indicated the burstiness factor \(K\), and
\[
\text{var}[IAT] = \frac{(1-\alpha)(\alpha+\beta)}{(1-\beta)^{2}} = \frac{(1-p)(2K-1)}{p^{2}}, \quad (31)
\]
which increases linearly with \(K\).

Numerical results show that the mean interdeparture time at the outlet of the shaper equals the mean interarrival time on the input link, which is expected due to the infinite data buffer size. In Figure 2, the variance of the interdeparture times var[IDT] is shown as a function of the load \(p\), for \(M = 3, C = 5\) and different values of the burstiness factor \(K\) of the source. The figure illustrates that var[IDT] is an increasing function of \(K\). Thus, the more bursty the source is, the higher var[IDT] is. We also see that the influence of \(K\) on var[IDT] decreases with increasing \(p\), which, of course, can be explained by the fact that the (mean) interdeparture times become smaller if the load \(p\) increases. In fact, the variance of the interdeparture times appears to be much the same as the variance of the interarrival times, given in (31).

In Figure 3, we have plotted the ratio of the variance of the interdeparture times to the variance of the interarrival times, var[IDT]/var[IAT], versus the load \(p\), for various values of \(K\). From this figure, we see that the ratio var[IDT]/var[IAT] decreases with increasing values of \(K\). Also we notice that var[IDT]/var[IAT] is always smaller than one, especially for higher loads. Hence, the leaky bucket traffic shaper reduces the variability of the input traffic, especially when the burstiness (\(K\)) of the sources is high. However, for low loads its effectiveness as a traffic smoother is rather limited, since var[IDT]/var[IAT] is almost equal to one. In Figure 4, the influence of the shaper's parameters \(M\) and \(C\) on var[IDT]/var[IAT] is shown, for \(p = 0.15\) and \(K = 2\). We observe that var[IDT]/var[IAT] decreases as the token generation period \(M\) increases, but increases with increasing token pool sizes \(C\). It is also clear that only small to intermediate token pool sizes \(C\) allow some reduction of the input variability, and longer token generation periods \(M\) allow a higher reduction of the variability.

However, the queue length and the waiting time of cells in the data buffer get worse with decreasing \(C\) and increasing \(M\). Specifically, in Figure 5, the mean data—buffer occupancy \(E[q]\) is shown as a function of the token pool size \(C\), for \(p = 0.15, K = 2\) and \(M = 2, 3, 4, 5\) respectively. The figure reveals that \(E[q]\) increases as tokens are generated at a slower rate (corresponding to larger values of \(M\)), whereas \(E[q]\) decreases if less tokens get lost (larger values of \(C\)), which is intuitively clear. Similar conclusions can also be drawn with respect to the evolution of the mean waiting time \(E[w]=E[q]/p\) in terms of \(M\) and \(C\). So the advantages of the traffic shaper with respect to traffic smoothing are achieved at the cost of larger cell queues and waiting times.

In Figures 6–7, we have plotted the coefficient of correlation between two consecutive interdeparture times, for various values of \(p\), \(K\), \(M\) and \(C\). From these figures it is clear that the correlation is always (slightly) negative. The leaky bucket thus transforms the uncorrelated interarrival times at its input into slightly negatively correlated interdeparture times at its output, which can also be considered as a favorable effect. The absolute value of the correlation increases as \(M\) or \(p\) increase or \(C\) decreases.

Acknowledgement

This work has been partly funded by the RACE 2061 project. The second author wishes to thank the Belgian National Fund for Scientific Research (N.W.O.) for support of this research.
References


