The Leaky Bucket as a Policing Device: Transient Analysis and Dimensioning *

Dimitris Logothetis
GTE Laboratories Inc.
40 Sylvan Rd.
Waltham MA 02174

Kishor Trivedi†
Department of Electrical Engineering
Duke University Box 90291
Durham NC 27708-0291

Abstract

Policing, the mechanism that the network uses to ensure that the user conforms with his negotiated contract, is an important aspect of the forthcoming broadband networks based on ATM. The leaky bucket algorithm has been adopted by ITU (formerly CCITT) as the means to police ATM connections. The performance of the leaky bucket (i.e., how well can the leaky bucket algorithm enforce the parameters of the negotiated contract) is a topic of major interest. In this paper we derive expressions for the time-dependent state probabilities and the time-averaged state-probabilities for the leaky bucket rate control scheme. Our model is based on the theory of Markov regenerative processes. In particular, we show that the leaky bucket state probabilities are periodic when studied in continuous time. We also compute time-averaged measures and show that our results specialize to those obtained by Sidi et al. [24] for the long-run behavior of leaky bucket. Our results are more general, however, in that they apply to the transient regime and to more general arrival processes. Finally, using our time-dependent analysis, we dimension the leaky bucket parameters and quantify trade-offs on these.

1 Introduction

Due to ATM’s inherent property to support any bandwidth requirement, an access control scheme is needed to ensure that a user does not exceed his negotiated parameters. One aspect of Usage / Network Parameter Control (UPC/NPC) is policing. The role of policing

*Part of this paper has appeared in the Proceedings of INFOCOM ’94.
†This work was supported in part by the National Science Foundation under Grant EEC-94-18765.
is to ensure that non-conforming cells are marked as low priority cells or they are lost. ITU (formerly CCITT) has standardized two equivalent mechanisms for policing; one of them being the leaky bucket [8]. In this paper we will present the time-dependent analysis of the leaky bucket for general point arrival processes. In contrast to previous works we allow for non-stationary arrival processes, i.e., processes where their parameters change over time since such processes can capture overload behavior. We will show that the state distribution of the leaky bucket when studied in continuous time and assuming a point process as general as the Batch Markovian Arrival Process (BMAP) is periodic. This behavior, although reported earlier [3, 18], is not widely known.

We then compute the time-averaged state distribution and show that its limits exist. Since we treat the leaky bucket as a policing mechanism, i.e., as a mechanism to filter incoming traffic, we concentrate on loss measures. We identify two types of loss measures. The first one, that we will refer to as tagged cell loss is defined as the probability that an incoming cell is lost (or marked with low priority). As we shall see later, this measure possesses a periodic limiting behavior. The other one, that we will refer to as fractional loss, is defined as ratio of the (expected) number of lost cells over the (expected) number of arrived cells in a given interval of length $t$. We shall see that this measure has a limiting distribution.\(^1\)

Despite the fact that ITU has standardized only peak rate as a traffic descriptor [8], the ATM Forum [2] considers two other optional traffic descriptors, namely, burst tolerance (or peak rate duration) and sustainable cell rate (or mean rate). The second part of the paper will be devoted to dimensioning the leaky bucket enforcing traffic parameters such as peak rate, mean rate and peak rate duration. The mathematical tool that we use is called the Markov regenerative or semi-regenerative process.

The remaining of the paper is organized as follows: In Section 2 we describe the leaky bucket rate control scheme and present the necessary background on Markov regenerative processes and the BMAP process. In Section 3 we develop the model based on the BMAP input process, while in Section 4 we present the model for non-stationary input processes and study the overload behavior of the leaky bucket. In Section 5 we study the ability of the leaky bucket to control the desired traffic parameters. Finally, in Section 6 we present our conclusions.

\section{Background}

In this section we review existing work in the analysis of the leaky bucket rate control scheme and we present some mathematical preliminaries needed to study the time-dependent behavior of the leaky bucket, namely, the delayed Markov regenerative theory and the batch Markovian arrival process.

\(^1\)Fractional loss, directly related to the cell loss ratio (CLR) Quality of Service (QoS) parameter of ATM, is probably more suitable for dimensioning but more difficult to compute when compared with the tagged cell loss.
2.1 The Leaky Bucket Rate Control Scheme

In high speed networks conventional window flow control is not feasible, since propagation delays are the dominating factor in overall delay rather than transmission or queueing delays. In simple words, if we force the transmitter to wait for an acknowledgment we will force it to remain idle for a long period of time since a time equal to the round-trip delay elapses before a new message can be transmitted. For this reason, alternative schemes for flow control need to be found. One of these, is the so called rate-control scheme where the arrival rate to a buffer is filtered through a policing device. One such device, is the leaky bucket rate control scheme [27].

The leaky bucket (see Figure 1) consists of a token generator, a pool of tokens and an optional buffer. Tokens are generated every $\tau$ seconds and stored in the pool if the pool contains less than $L$ tokens. Otherwise the token is lost. Packets/cells arrive to the buffer according to a traffic process that needs to be controlled. The arrival process can be as general as the Batch Markovian Arrival Process (BMAP) [19] although most of our numerical investigations consider special cases of the BMAP such as Markov Modulated Poisson Process (MMPP), the Interrupted Poisson Process (IPP) and the Poisson Process (PP). Furthermore, the input process can be time-inhomogeneous (i.e. its parameters may change with time), as we shall see in Section 4. An arriving packet/cell that finds the token pool non-empty departs the system immediately while removing one token from the token pool. An arriving packet/cell that finds the token pool empty joins the queue if the buffer is not full. Now when the queue is not empty (which means that the token pool is empty) and a token arrives, one packet/cell departs immediately taking the token with it.

![Figure 1: The leaky bucket rate control scheme](image)

Existing models in the open literature are concerned with steady-state measures. We can characterize these models as follows: fluid-flow model for arrival sources are considered in [7, 13, 15], point process models with continuous-time in [4, 10, 12, 16, 24], point processes

---

2It should be noted here that this definition of the leaky bucket is slightly different from the one adopted by ITU and ATM Forum.
with discrete-time in [1, 25, 28], Brownian motion models [5] and analysis based on sample paths in [6].

In this paper we will describe the (continuous) time-dependent behavior of the leaky bucket adopting a point process to characterize cell arrivals.

2.2 Delayed Markov regenerative processes

In the development of BMAP/G/1 and GI/BMAP/1 type of queueing models [9, 17], it is commonly assumed that the time origin coincides with a service completion or an arrival instant. When studying overload behavior of queueing systems, as we shall see in this paper, we need models where the time origin does not necessarily coincide with service completion or arrival instant. For this reason, in this section we generalize the well-known Markov regenerative process (MRGP) [9, 17] to the delayed Markov regenerative process (DMRGP) and develop the underlying equations that govern the conditional state probabilities. Before we formally define the delayed Markov regenerative process, we give some short of loose definition of the Markov regenerative process. An MRGP can be thought of as a process where there exist time epochs where the process regenerates in the Markovian sense, i.e., the stochastic evolution of the marking process does not depend on the history of the process but rather on the current state. The time epochs that the process regenerates together with the value that process has at that point form what we call a Markov renewal sequence (MRS). The stochastic behavior of the Markov regenerative process can be described by two matrices, namely, $K(t)$ (global kernel), that captures the behavior of the process assuming the regeneration has occurred and $E(t)$ (local kernel) that considers the behavior between two regenerations. It has been shown [17] that matrix $V(t)$ with entries $V_{ij}(t)$ the time dependent state probabilities of an MRGP $P(Z(t) = j | Z(0) = i)$, satisfies a (matrix) generalized Markov renewal equation: $V(t) = E(t) + \int_0^t dK(x)V(t-x)$. Now for the development of the delayed Markov regenerative process, we first define the delayed Markov renewal sequence, the generalization of the Markov renewal sequence in Markov regenerative theory.

**Definition 1** A sequence of bivariate random variables $\{(Y_n, T_n)\}$ is called a time homogeneous Delayed Markov Renewal Sequence (DMRS) if:

1. $T_0 = 0, T_{n+1} \geq T_n; Y_n \in \Omega$

2. $P\{Y_1 = j, T_1 \leq x \mid Y_0 = i\} = F_{ij}(x)$

3. $\forall n \geq 1,$

$$P\{Y_{n+1} = j, T_{n+1} - T_n \leq x \mid Y_n = i, T_n, Y_{n-1}, T_{n-1}, \cdots, Y_0, T_0\}$$

$$= P\{Y_{n+1} = j, T_{n+1} - T_n \leq x \mid Y_n = i\}$$

$$= P\{Y_2 = j, T_2 - T_1 \leq x \mid Y_1 = i\} = K_{ij}(x)$$

Then, the DMRGP is defined as follows:
Definition 2 A stochastic process \( \{Z(t), t \geq 0\} \) is called a Delayed Markov Regenerative Process (DMRGP) if there exists a \( T_1 \geq 0 \) such that the process \( \{Z(t+T_1), t \geq 0\} \) is a Markov Regenerative process.

Theorem 1 Let \( V^D(t), V(t) \) be the matrices with entries \( V^D_{i,j}(t) \) and \( V_{i,j}(t) \) the conditional state probabilities of a Delayed Markov Regenerative Process (DMRGP) (i.e., \( P(Z(t) = j|Z(0) = i) \)) and a standard Markov regenerative process (i.e., the process \( \{Z(t+T_1), t \geq 0\} \)), respectively. Then, \( V^D(t) \) satisfies the following integral equation:

\[
V^D(t) = \mathbf{M}(t) + \int_0^t d\mathbf{F}(s) \mathbf{E}(t-s) + \int_0^t d\mathbf{F}(s) \int_0^{t-s} d\mathbf{K}(x) \mathbf{V}(t-s-x)
\]

where \( \mathbf{M}(t) \) represents the local kernel for the first regeneration instant with entries:

\[
M_{i,j}(t) = P\{Z(t) = j, T_1 > t|Y_0 = i\}
\]

\( \mathbf{E}(t) \) the local kernel for all other regeneration instants, with entries:

\[
E_{i,j}(t) = P\{Z(t) = j, T_2 - T_1 > t|Y_1 = i\}
\]

(Proof) Conditioning on \( T_1 = s \), the first (Markovian) regeneration instant, it is obvious that for any \( t > s \) matrix \( V^D(t) \) satisfies the following equation:

\[
V^D(t | T_1 = s) = \mathbf{E}(t-s) + \int_0^{t-s} d\mathbf{K}(x) \mathbf{V}(t-s-x), \quad t > s
\]

Unconditioning we obtain Equation (1). □

Note that Equation (1) is not a generalized Markov renewal Equation, since on the left hand side we have \( V^D(t) \) and on the right hand side we have \( V(t) \). Loosely speaking, in a delayed Markov process we allow for the statistics of the first regeneration to be different that the remaining ones. That is the reason why we define two local kernels \( (\mathbf{E}(t) \) and \( \mathbf{M}(t)) \) and two global kernels \( (\mathbf{K}(t) \) and \( \mathbf{F}(t)) \).

2.3 The Batch Markovian Arrival Process

The Batch Markovian Arrival Process (BMAP) [19] is a point process where times between arrivals are governed by a Continuous Time Markov Chain (CTMC) with infinitesimal generator \( \mathbf{D} \) over finite state space \( S \). For every state \( i \in S \) we associate an (exit) rate \( \lambda_i \) and a set of probabilities \( p_i(k, j), k = 0, 1, \ldots \) and \( j \in S - \{i\} \).

The BMAP process determines arrivals based on state transitions; given that the CTMC was in state \( i \) after an exponentially distributed sojourn time (with parameter \( \lambda_i \)), when a transition from state \( i \) to state \( j \) occurs it will give rise to \( k \) arrivals and the corresponding probability is \( p_i(k, j) \).

The evolution of the BMAP can be described by a sequence of matrices \( \{\mathbf{D}_k, k \geq 0\} \) with entries \( D_{0,i} = -\lambda_i, \ 1 \leq i \leq m, \ D_{0,j} = \lambda_i p_i(0, j), \ 1 \leq i, j \leq m, \ j \neq i \) and \( D_{k,i} = \lambda_i p_i(k, j), \ k \geq 1, \ 1 \leq i, j \leq m \).
Let \( N(t) \) be the number of arrivals in \((0, t]\) and \( J(t) \) represent the state of the arrival process at time \( t \). The pair \((N(t), J(t))\) is a CTMC on the state space \( \{(i, j) : i \geq 0, 1 \leq j \leq m \} \) and infinitesimal generator \( Q_{\infty} \). As the state space is infinite along the first dimension, we truncate it at \( N(t) = n \) for the purpose of representation. The corresponding generator matrix \( Q_n \) is given by:

\[
Q_n = \begin{bmatrix}
D_0 & D_1 & D_2 & \cdots & \sum_{k=n}^{\infty} D_k \\
0 & D_0 & D_1 & \cdots & \sum_{k=n-1}^{\infty} D_k \\
0 & 0 & D_0 & \cdots & \sum_{k=n-2}^{\infty} D_k \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
0 & 0 & 0 & \cdots & \sum_{k=0}^{\infty} D_k
\end{bmatrix}
\]

Note that \( Q_{\infty} = \lim_{n \to \infty} Q_n \). Matrix \( e^{Q_{n}t} \) is given by:

\[
e^{Q_{n}t} = \begin{bmatrix}
A(0, t) & A(1, t) & A(2, t) & \cdots & \sum_{k=n}^{\infty} A(k, t) \\
0 & A(0, t) & A(1, t) & \cdots & \sum_{k=n-1}^{\infty} A(k, t) \\
0 & 0 & A(0, t) & \cdots & \sum_{k=n-2}^{\infty} A(k, t) \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
0 & 0 & 0 & \cdots & \sum_{k=0}^{\infty} A(k, t)
\end{bmatrix}
\]

\( A(i, t) \) is an \( m \times m \) matrix with entries, \( a_{jk}(i, t) \), the probability of having \( i \) arrivals in \( t \) time units given that the auxiliary process was in state \( j \) at time 0 and is in state \( k \) at time \( t \). The entries of \( e^{Q_{n}t} \), with few exceptions, cannot be computed in closed-form. In the Appendix we develop a method to compute \( A(i, t) \) for the Markovian arrival process (MAP) (i.e., \( D_k = 0, \ k > 1 \)). For large size problems, however, uniformization \cite{uniformization} should then be used to numerically evaluate the elements \( A(i, t) \).

The BMAP contains many widely used processes such as the Poisson process, the Batch Poisson process, phase-type renewal process, the Markovian arrival process (MAP) and the Markov Modulated Poisson process (MMPP) as special cases.

### 3 The Model with a BMAP input source

Let \( X(t) \) denote the number of tokens in the token pool at time \( t \), \( Y(t) \) denote the number of cells in the buffer at time \( t \) and \( J(t) \) denote the phase of the auxiliary process at time \( t \).
By observing the fact that token pool and the cell buffer cannot be simultaneously non-empty, (i.e., \( X(t) > 0 \Rightarrow Y(t) = 0 \) and \( Y(t) > 0 \Rightarrow X(t) = 0 \)) we define the process \( Z(t) = L - X(t) + Y(t) \) (i.e., we keep track of the sum of the empty token places in the pool and the number of cells in the cell buffer. It is easy to see that:

\[
X(t) = (L - Z(t))^+ \\
Y(t) = (Z(t) - L)^+
\]

where \((a)^+ = \max(a, 0)\) and \(L\) the token pool size. We will therefore study the two-dimensional stochastic process \((Z(t), J(t))\).

**Theorem 2** Consider a leaky bucket fed by a BMAP as an input and fix the time origin at \(s_0\) \((\tau_0)\) time units relative to the previous (next) token generation instant. The two-dimensional stochastic process \((Z(t), J(t))\) is a delayed Markov regenerative process.

**Proof** It is easy to see that the sequence \((F_n, T_n)\), where \(T_n\) is the time of the next token generation instant and \(F_n = Z(T_n^+)\) is a delayed Markov renewal sequence.

**Theorem 3** Let \(E(t, \tau)\) and \(K(t, \tau)\) given by:

\[
E_{ij}(t, \tau) = \begin{cases}
A(j - i, t)[u(t) - u(t - \tau)] & i \geq 0 \\
\sum_{l=M+L-i}^{\infty} A(l, t)[u(t) - u(t - \tau)] & i \leq j < M + L \\
0 & j = M + L \\
\end{cases}
\]

\[
K_{ij}(t, \tau) = \begin{cases}
[A(0, \tau) + A(1, \tau)]u(t - \tau) & i = 0 \\
A(j + 1, \tau)u(t - \tau) & j = 0 \\
A(j - i + 1, \tau)u(t - \tau) & 1 \leq j < M + L - 1 \\
\sum_{l=M+L-i}^{\infty} A(l, \tau)u(t - \tau) & i > 0 \\
0 & j = M + L - 1 \\
\end{cases}
\]

with \(A(i, t)\) as defined earlier then the local and global kernels for the first regeneration period are given by \(E(t, \tau_0)\) and \(K(t, \tau_0)\) while for any other regeneration periods are given by: \(E(t, \tau)\) and \(K(t, \tau)\) respectively.

**Proof** Straightforward from the definitions of matrices \(K(t)\) and \(E(t)\).
Note that matrix $K(t, \tau)$ (Equation (8)) depends on $t$ only through the unit step function $u(t - \tau)$. Hence, $K(t, \tau)$ can be written as:

$$K(t, \tau) = K(\infty, \tau)u(t - \tau) = P(\tau)u(t - \tau)$$

(9)

where $P(\tau)$ is the transition probability matrix of the DTMC embedded immediately after the token generation instants. Observing the fact that the derivative of $K(t, \tau)$ can be written as $P'(\tau)\delta(t - \tau)$, where $\delta(t - \tau)$ is the Dirac’s delta function (also called the unit impulse). The recursive relationship between conditional state probabilities at time $t$ and state probabilities at time $t - \tau$ is:

$$
V(t) = \begin{cases} 
E(t, \tau_0) & 0 \leq t \leq \tau_0 \\
P(\tau_0)E(t - \tau_0, \tau_0) & \tau_0 < t < \tau_0 + \tau \\
P(\tau)V(t - \tau) & t > \tau + \tau_0 
\end{cases}
$$

(10)

Figure 2, shows the transient state probabilities obtained from the solution of a three state system (i.e., $L + M = 2$) with $\tau_0 = 0$. Observe the periodic behavior of $V_{ij}(t)$ for large values of $t$ and the discontinuities that occur at regeneration points. For any time $t$ Equation (10) can be written as:

$$
V(t) = E(t, \tau_0)^{1-u(t-\tau_0)}P(\tau_0)^{u(t-\tau_0)}P^n(\tau)E(s, \tau)^{u(t-\tau_0)}
$$

(11)

with $n = [\frac{t-\tau_0}{\tau}]$ and $s = t - n\tau - \tau_0$. The above equation is useful when studying the leaky bucket behavior under non-stationary conditions (overload) in a later section. For an initial
system state probability vector \( p(0) \) the system state probability vector at time \( t \) is given by:

\[
p(t) = p(0) E(t, \tau_0)^{1-n(t-\tau_0)} p(\tau_0)^{n(t-\tau_0)} P^n(\tau) E(s, \tau)^{n(t-\tau_0)}
\]

where \( u(t - \tau_0) \) is the unit step function shifted \( \tau_0 \) units from the origin. In the limit as \( t \) approaches infinity, \( p(t) \) becomes periodic with period \( \tau \):

\[
\lim_{n \to \infty} p(\tau_0 + n\tau + s) = \pi E(s, \tau)
\]

where \( \pi = \lim_{n \to \infty} p(0) P(\tau_0) P^n(\tau) \). This periodicity has been reported in [4, 18]. Note that \( \pi \) satisfies the equation \( \pi = \pi \times P(\tau) \). The elements of \( \pi \) represent the stationary probability vector of a DTMC embedded immediately after the token generation instants.

We can also quantify the discontinuities in state probabilities as follows: Let \( R(n) = V(n\tau^+ + \tau_0) - V(n\tau^- + \tau_0) \) the discontinuities of the elements of matrix \( V(t) \) at time \( n\tau + \tau_0 \). We can easily see that:

\[
R(n) = P(\tau_0)^{n(t-\tau_0)} P^{n-1}(\tau) [P(\tau) - E(\tau, \tau)]
\]

As \( n \) approaches infinity, \( R = \lim_{n \to \infty} R(n) \) is given by:

\[
R = \Pi [I - E(\tau, \tau)]
\]

and \( \Pi = \lim_{n \to \infty} P(\tau_0)^{n(t-\tau_0)} P(\tau)^n \). One can also see that:

\[
\begin{align*}
V_{ij}(n\tau^+ + \tau_0) &= V_{ij+1}(n\tau^- + \tau_0), & 0 \leq i \leq M + L, & 0 < j < M + L \\
V_{i,M+L}(n\tau^+ + \tau_0) &= 0, & 0 \leq i \leq M + L \\
V_{00}(n\tau^- + \tau_0) + V_{11}(n\tau^- + \tau_0) &= V_{00}(n\tau^+ + \tau_0), & 0 \leq i \leq M + L
\end{align*}
\]

For single arrivals, we define the loss probability at time \( t \) as follows:

\[
p_{loss}(t) = \text{Prob}\{\text{pool empty and buffer full} \mid \text{arrival occurs at time } t\}
\]

The above measure is basically the probability that an incoming customer will find the buffer full (and the token pool empty) at the time of the arrival. Relating the outside observer’s distribution \( p(t) \) with the arriving customer’s distribution and for single arrivals we can show (see Cooper [11]) that:

\[
p_{loss}(t) = \frac{p_{L+M}(t)\lambda_{L+M}}{\sum_{j=0}^{L+M} p_j(t)\lambda_j}
\]

For the Poisson process, arrivals are not state-dependent and the above equation becomes:

\[
p_{loss}(t, a) = p_{L+M}(t, a)
\]

Figure 3, shows the loss probability obtained as \( V_{0,L+M}(t) \) (assuming that we start with a full token pool). For MAP arrivals tagged cell loss can be expressed as a vector:

\[
p_{loss}(t, a) = \frac{p_{L+M}(t, a) D_1}{\sum_{j=0}^{L+M} p_j(t) D_1 e^j}
\]
with \(D_1\) as defined in Section 2.3 and \(p_{\text{loss}}(t)\) is a vector with entries:

\[
p_{\text{loss},j}(t) = P\{Z(t) = L + M, \Phi(t) = j\}
\]

It is interesting to note is that the denominator of Equation (20) does not depend on \(\tau\). For example, for an IPP source Equation (20) becomes:

\[
p_{\text{loss}}(t, \tau) = \frac{p_{L+M}(t, \tau)}{p_{\text{ON}}(t)}
\]  (21)

with \(p_{\text{ON}}(t)\) the probability that the source is ON at time \(t\).

---

**Figure 3:** Transient loss probabilities for different leaky bucket parameters

Generalizing this measure of loss for batch arrivals we define the probability that a batch of size \(n \geq 1\) will suffer \(m \leq n\) losses as:

\[
p_{\text{loss}}(n, m, t) = \frac{\sum_{L+M+n-m}^{L+M} p_j(t)D_n}{\sum_{j=0}^{L+M} p_j(t)D_n e^T}
\]  (22)

Another measure of loss which is very closely related to the Cell Loss Ratio (CLR) defined in ATM Forum [2] is the expected fraction of cells lost up to time \(t\). This can be evaluated as:

\[
f_{\text{loss}}(t) = \frac{\text{Expected number of losses in}(0, t)}{\text{Expected number of arrivals in}(0, t)}
\]  (23)

For Poisson arrivals the above measure is simply:

\[
f_{\text{loss}}(t) = \frac{1}{t} \int_0^t p_{\text{loss}}(x) dx
\]  (24)
The above measure dictates the need to compute time-averages of the state probabilities \( p_j(t) \). Define \( s(t) \) to be the time average of vector \( p(t) \) i.e.:

\[
s(t) = \frac{1}{t} \int_0^t p(x) \, dx
\]  

(25)

Indeed, it is not difficult to show after integrating Equation (12) for any \( t = \tau + n\tau + s \):

\[
s(t) = \begin{cases} 
\frac{\mu(t \to 0)}{t} \int_0^t E(x) \, dx & t < \tau_0 \\
\frac{\mu(t \to 0)}{t} \left[ \int_0^\tau E(x) \, dx + P(\tau_0) \sum_{i=0}^{n-1} P^i(\tau) \int_0^\tau E(x) \, dx + \right. \\
& \left. + P(\tau_0) P^n(\tau) \int_0^\tau E(x) \, dx \right] & t > \tau_0 
\end{cases}
\]  

(26)

As \( t \) approaches infinity, \( s(t) \) approaches a limit:

\[
s = \lim_{t \to \infty} s(t) = \pi \frac{1}{\tau} \int_0^\tau E(x) \, dx
\]  

(27)

Figure 4, shows the time averages corresponding to the probabilities of Figure 2 (Poisson arrivals). Notice that no discontinuities are present. It is interesting to see that for Poisson arrivals, as \( t \) tends to infinity, \( f(t) \) approaches a limit:

\[
f = \lim_{t \to \infty} f(t) = \sum_{n=0}^{M+L-1} \pi_n \left[ 1 - \frac{1}{\lambda t} \sum_{i=0}^{M+L-n-1} \left( 1 - \sum_{j=0}^{i} a_j \right) \right]
\]  

(28)
Table 1: Steady-state vs transient measures

<table>
<thead>
<tr>
<th>Bucket size</th>
<th>$t = 100000$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\rho = 1$</td>
</tr>
<tr>
<td>10</td>
<td>$5.356 \times 10^{-2}$</td>
</tr>
<tr>
<td>20</td>
<td>$2.583 \times 10^{-2}$</td>
</tr>
<tr>
<td>50</td>
<td>$1.0055 \times 10^{-2}$</td>
</tr>
</tbody>
</table>

Vectors $s_j(t)$ are computed using Equation (26).

In Table 1 we compare transient and steady-state fractional loss computed using Equations (24) and (28) respectively. The mission time $t$ was chosen as 100000 time units that corresponds to a connection of duration of 10 minutes. Note that transient and steady-state measures are very close. This indicates the fact during the duration of the connection the leaky bucket practically reaches steady-state. Time intervals where transient analysis is important should therefore be much smaller. We consider such short intervals in subsequent examples. In Figure 5, we show the time-averaged loss probabilities for different bucket sizes.

We can also define the throughput at time $t$ as:

$$T(t) = \frac{1}{t} \left[ \int_0^t \lambda [1 - p_{\text{loss}}(x)] dx - \sum_{i=L+1}^{L+M} ip_i(t) \right]$$

where the summation term represents the mean queue length of the buffer. As $t$ tends to infinity:

$$T = \lim_{t \to \infty} T(t) = \lambda \tau (1 - f)$$

Once again, this matches with Formula (4) in [24].
In Figure 5, the time averaged loss probabilities for different leaky-bucket parameters are shown. The graphs illustrate the loss probabilities for different values of $L/M$. A leaky bucket with Poisson arrivals and the same effective arrival rate is also shown for comparison.

In Figure 6, the loss probability of a leaky bucket with $L + M = 10$, and an IPP input with $\lambda_{ON} = 1$ and $\tau = 0.9$ is shown for different values of $J(0)$. A leaky bucket with Poisson arrivals and the same effective arrival rate is also shown for comparison.

In Figure 7, we plot fractional losses that correspond to the loss probabilities of Figure 6. Note that in spite of the fact we matched the mean arrival rates in both arrival streams the IPP model gives much larger loss measures than the Poisson model. This strengthens
the fact that loss probability strongly depends on higher moments (not only the mean) of interarrival times. In Figure 8 we plot fractional loss for various initial bucket sizes.)
4 Response to an overload

In this section, we will consider the leaky bucket response to an overload. In particular, we will assume that the leaky bucket is observed when its limiting state probability vector is \( \pi(i_0) \). Here, \( i_0 \) is a vector whose elements are the set of parameters used to describe the arrival process. For an ON-OFF source, this vector consists of three elements, \((a_0, b_0, \lambda_0)\). At \( s_0, 0 \leq s_0 \leq \tau \) time units after a token generation, an overload occurs (so that an arrival process with a parameter vector \( i_1 \) is activated). For instance, in the case of an ON-OFF source, \( i_1 = (a, b, \lambda) \) with \( a \geq \frac{1}{s_0} \). We seek an expression for system state probabilities \( t = t + \tau \) time units after the overload occurs. Using Equation (12) and \( \pi(i_0) \) as the initial state probability vector we obtain:

\[
p(s_0, i_0, i_1, t) = \begin{cases} 
\pi(i_0)E(i_0, s_0)E(i_1, t) & t < \tau_0 \\
\pi(i_0)E(i_0, s_0)P(i_1, \tau_0)P^n(i_1, \tau)E(i_1, s) & t \geq \tau_0
\end{cases}
\]

where \( n = \lfloor \frac{t - \tau_0}{\tau} \rfloor \), \( \tau_0 = \tau - s_0 \) and \( s = t - n\tau - \tau_0 \). Let us now assume that \( t_v \) time units after the occurrence of the overload it disappears. Then generalizing Equation (32) we can write down for any time \( t \):

\[
p(s_0, i_0, i_1, t_v, t) = \begin{cases} 
\pi(i_0)E(i_0, s_0)E(i_1, t) & t < \tau_0 \\
\pi(i_0)E(i_0, s_0)P(i_1, \tau_0)P^n(i_1, \tau)E(i_1, s) & \tau_0 \leq t \leq t_v \\
p(s_0, i_0, i_1, t_v, t_v)E(i_0, t - t_v) & t_v \leq t \leq t_v + \tau_v \\
p(s_0, i_0, i_1, t_v, t_v)P(i_0, \tau_v)E(i_1, s_1) & t \geq t_v + \tau_v
\end{cases}
\]
with \( n, s \) same as in Equation (32) and \( n_1 = \lfloor \frac{t - \tau_0}{\tau_v} \rfloor \), \( s_1 = t - n_1 \tau - t_v - \tau_v \). Expressions for fractional losses can also be obtained by integrating Equation (33) as:

\[
\begin{align*}
f(s_0, i_0, i_1, t_v, t) &= \begin{cases} 
\pi(i_0)E(i_0, s_0)I_E(i_1, t) & \text{if } t < \tau_0 \\
\sum_{i=0}^{n_1-1} \pi(i_0)E(i_0, s_0)P(i_1, \tau_0)P^i(i_1, \tau)I_E(i_1, \tau) & \text{if } \tau_0 < t < t_v \\
f(s_0, i_0, i_1, t_v, t_v) + P(s_0, i_0, i_1, t_v, t_v)I_E(i_0, t - t_v) & \text{if } t_v < t \leq t_v + \tau_v \\
\sum_{i=0}^{n_1-1} P(i_0, \tau_v)P^i(i_0, \tau)I_E(i_0, \tau) & \text{if } t \geq t_v + \tau_v \\
\end{cases}
\end{align*}
\]

where \( I_E(i_1, t) = \frac{1}{\tau} \int_{0}^{t} E(i_1, x) dx \). In Figures 10 and 11 we assume a non-homogeneous Poisson arrival process with rate:

\[
\lambda(t) = \begin{cases} 
\lambda_0 & t < 0 \\
\lambda_1 & 0 \leq t < 3 \\
\lambda_0 & t \geq 3 
\end{cases}
\]

The above defined non-homogeneous process represents an overload with \( i_0 = (0, b, \lambda_0) \), \( i_1 = (0, b, \lambda_1) \) and \( t_v = 3 \). We assign \( \lambda_1 = 2 \times \lambda_0 = 2 \).

Figures 10 shows the loss probability while Figure 11 shows the fractional loss. Note that standard transient response measures, such as overshoot and relaxation times can be observed in Figure 11.

5 Dimensioning the Leaky Bucket

In the previous sections we focused on the derivation of the leaky bucket state probability vector at an arbitrary time point \( t \) assuming a given input parameter vector \( i \) and leaky bucket parameters token pool size \( L \), buffer size \( M \) and intertoken generation time \( \tau \). In this section we consider the reverse problem, for an input traffic with specific input traffic characteristics, what are the leaky bucket parameters to achieve given system state distribution. In particular, we will focus on measures of loss probability. The periodic behavior of the leaky bucket state distribution makes this problem very interesting.
It was observed in this paper and also noted previously [13, 4] that the loss probability depends on the sum of the sizes of the two buffers, cell and token, \( L + M \). Let \( T \) denote this sum. The role of the cell buffer is to smooth out the input traffic at the expense of larger delays [13]. The cell buffer becomes important when the leaky bucket is used as traffic shaper and its behavior is characterized using the cell departure process. When the leaky bucket is used as a policing device, i.e., as a means to control and protect network resources from malicious as well as unintentional misbehavior, loss probability is the primary measure.
of interest and as mentioned earlier the partition of the total buffer space into token buffer and cell buffer is not important. We will therefore dimension the leaky bucket considering the parameters $T$ and $\tau$.

The role of the leaky bucket as policing mechanism is to enforce negotiated traffic parameters at connection setup. The number of parameters needed to characterize traffic and QoS in B-ISDN is still an open question. So far, ITU [8] and ATM Forum [2] have standardized peak (cell) rate $B_p$, (cell) delay variation tolerance $v$, sustainable or mean (cell) rate and burst tolerance. Note that burst tolerance defines the maximum burst length that should be accommodated by the network. Many authors, however, such as [7] consider the mean burst size (as opposed to the worst-case burst size) as the traffic parameter. In any case, the parameters can be mapped to an ON-OFF source traffic model. ON-OFF source models have been used extensively to model ATM traffic, due to their simplicity and their ability to capture source burstiness. Figure 12 shows the ON-OFF source model: in the ON period

\[ \frac{1}{B} \]

\[ D \quad F \]

Figure 12: The ON-OFF source model

of duration $D$, a constant peak rate, $B_p$, is offered to the leaky bucket, whereas in the OFF period of duration $F$ the rate is 0. The mean rate is then given by: $B_m = \frac{D}{D+F} B_p$. When the duration of the peak rate has the interpretation of burst tolerance the ON and OFF periods are deterministic numbers. If however mean burst sizes are considered then ON and OFF periods random variables. For most of our results we will consider an IPP as our ON-OFF source model and therefore our burst size interpretation is that of a mean value.

### 5.1 Dimensioning based on Peak Rate enforcement

In this section we will focus on peak rate dimensioning, (i.e., successive cell interarrival times should exceed a prespecified value $\frac{1}{B_p}$). Let us assume that the arrival source can be characterized as a renewal process with an interarrival distribution $F_A(t) = P(A < t)$. We will say that the source complies if two successive arrivals occur at least $b = \frac{1}{B_p}$ time units apart. Let $C$ denote the event that the source complies with the imposed constraint on the peak rate. Then:

\[ P(C) = 1 - F_A(b) \]  \hspace{1cm} (36)

and

\[ P(C^c) = F_A(b) \]  \hspace{1cm} (37)

Using the theorem of total probability [26] we can express cell loss probability as:

\[ p_{loss} = p_{loss|C} P(C) + p_{loss|C^c} P(C^c) \]  \hspace{1cm} (38)
In cases where $p_{\text{loss}C}$ has a large value, the contribution to loss probability can be due to $p_{\text{loss}C}$ and therefore dimensioning for a small $p_{\text{loss}}$ may actually mean minimizing loss when the source is not complying and therefore the leaky bucket is not actually rejecting or marking cells as low priority cells when it is supposed to do so. The way we conceive peak rate enforcement is as follows:

Given an interarrival distribution $F_A(t) = P(A < t)$ and a peak rate constraint, $B_p$, (i.e., successive cells should be spaced at least $b = \frac{1}{B_p}$ time units apart), find the values of leaky bucket parameters, $T = L + M$ and $\tau$ such that $p_{\text{loss}C}$ is minimized (ideally 0) and $p_{\text{loss}C}$ is maximized (ideally 1).

We now turn our attention to computing $p_{\text{loss}C}$ and $p_{\text{loss}C}$.

Given an interarrival distribution $F_A(t) = P(A < t)$ and a peak rate constraint, $B_p$, (i.e., successive cells should be spaced at least $b = \frac{1}{B_p}$ time units apart), find the values of leaky bucket parameters, $T = L + M$ and $\tau$ such that $p_{\text{loss}C}$ is minimized (ideally 0) and $p_{\text{loss}C}$ is maximized (ideally 1).

We now turn our attention to computing $p_{\text{loss}C}$ and $p_{\text{loss}C}$.

We will also assume that $\tau \geq b$. Let $\pi_E(s)$ denote the state probability vector at time $s$, $0 < s < \tau$. Using the arrival theorem [11] we can find the state distribution as seen by arrivals, $\pi^*(s)$. If an arriving customer sees the system in state $L + M - 1$ or in state $L + M$, with corresponding probabilities $\pi_{L+M-1}^{(a)}(s)$ and $\pi_{L+M}^{(a)}(s)$, then immediately after the arrival the state of the system is $L + M$ with the corresponding probability $\pi_{L+M-1}^{(a)}(s) + \pi_{L+M}^{(a)}(s)$. Our goal is to compute the loss probability conditional on the next arrival. Referring to Figure 13 we can express $P_{\text{loss}C}$ and $P_{\text{loss}C}$ as:

$$
\begin{align*}
\pi_{L+M-1}^{(a)}(s) + \pi_{L+M}^{(a)}(s) & \quad F_A(\tau - s) - F_A(b) \quad \tau > b + s \\
0 & \quad \tau < b + s
\end{align*}
$$

$$
\begin{align*}
\pi_{L+M-1}^{(a)}(s) + \pi_{L+M}^{(a)}(s) & \quad F_A(b) \quad \tau > b + s \\
(\pi_{L+M-1}^{(a)}(s) + \pi_{L+M}^{(a)}(s))F_A(\tau - s) & \quad s < \tau < b + s
\end{align*}
$$

$^3$p_{\text{loss}C}$ and $p_{\text{loss}C}$ are commonly referred to as probability of false alarm and probability of detection, respectively.

Figure 13: Timing relations between two arrivals in a leaky bucket

$$
P_{\text{loss}C} = \begin{cases} 
(\pi_{L+M-1}^{(a)}(s) + \pi_{L+M}^{(a)}(s)) [F_A(\tau - s) - F_A(b)] & \tau > b + s \\
0 & \tau < b + s
\end{cases} \quad (39)
$$

$$
P_{\text{loss}C} = \begin{cases} 
(\pi_{L+M-1}^{(a)}(s) + \pi_{L+M}^{(a)}(s))F_A(b) & \tau > b + s \\
(\pi_{L+M-1}^{(a)}(s) + \pi_{L+M}^{(a)}(s))F_A(\tau - s) & s < \tau < b + s
\end{cases} \quad (40)
$$
Noting the fact that an arrival instant is uniformly distributed in the interval \((0, \tau)\) and taking time averages for the interval of length \(\tau\) from Equation (38), an expression for loss probability that does not depend on the observation instant can be written down as:

\[
    f_{loss} = f_{loss|C} P(C) + f_{loss|\overline{C}} P(\overline{C})
\]

(41)

In Figure 15 we plot \(f_{loss|C}\) vs \(f_{loss|\overline{C}}\) as a function of \(\tau\) and different values of \(T\). The arrival process is assumed to be a Poisson process with parameter \(\lambda = 1\). \(b\) is chosen as 0.5 time units. The straight line that corresponds to \(f_{loss|C} = f_{loss|\overline{C}}\) is also shown. As mentioned earlier, the ideal behavior is obtained at the point \((f_{loss|C} = 0, f_{loss|\overline{C}} = 1)\). Note that by decreasing \(\tau\) we actually decrease both fractional losses and for \(\tau = b\), \(f_{loss|C}\) becomes zero. This means that the parameter \(\tau\), by itself, cannot effectively control the loss of non-complying cells. Note that by decreasing the buffer size both \(f_{loss|C}\) and \(f_{loss|\overline{C}}\) increase; therefore the choice \(b = \tau\) and \(T = 1\) will at least guarantee no loss for compliant cells and simultaneously will reject a large fraction of (but not all) the non-compliant cells. The above suggests that deterministic token generation times cannot guarantee the rejection of all non-compliant cells. Nevertheless, there exists a way to make this happen if one token is generated \(\tau = b\) time units after the last compliant cell. In fact this trivial one-buffer leaky bucket with this non-periodic token generation times is equivalent to the “continuous-state leaky bucket” standardized by ITU [8] for peak rate enforcement with the assumption that the cell delay variation tolerance is zero. An example that illustrates why periodic token generations cannot effectively enforce peak rate is shown in Figure 14. Scheme (1) will often let non-compliant arrivals enter the network, while scheme (2) will not. Note that both will let all compliant arrivals in.

![Leaky bucket with regular token generation instants (1)](image1)

![Leaky bucket with irregular token generation instants (2)](image2)

Figure 14: A sample arrival stream to the two leaky bucket algorithms
5.2 Dimensioning based on Mean Rate enforcement

We now take a different angle and dimension based on mean rate. The leaky bucket is now required to enforce a given mean rate. We first define two parameters that are important on our policing mechanisms.

- **Probability of false alarm** is the probability the leaky bucket will reject a compliant cell. For mean rate enforcement, \( f_{\text{loss}} \) may have the interpretation as fraction of cells rejected due to false alarms when the mean rate is compliant.

- **Detection time** is elapsed time before the leaky bucket takes action in the occurrence of an overload.

The problem is stated as follows:

Given a desired mean rate \( \lambda_m \), what are the values of \( \tau \) and \( T = L + M \) that guarantee minimum loss i.e., probability of false alarm is minimized while the detection time is small.

In order to answer this problem in Figure 16 we look at how the fractional loss \( f_{\text{loss}} \) behaves as a function of the normalized mean arrival rate, for different values of \( \tau \) and \( T = 10 \). The point 1 on the \( x \) axis represent a compliant source (e.g., its mean rate is the prespecified one). Ideally we would like to have \( f_{\text{loss}} \) equal to zero at that point. Also as the mean rate deviates from the prespecified one, loss probability should increase such that the throughput is reduced to the prespecified one, \( \lambda_m \), i.e.,

\[
\lambda_m = \lambda (1 - f_{\text{loss}}) \Rightarrow f_{\text{loss}} = 1 - \frac{\lambda_m}{\lambda} \tag{42}
\]

for \( \lambda > \lambda_m \).
From Figure 16 we see that the choice $\tau = \frac{1}{\lambda_m}$ gives results very close to the ideal behavior, except when $\lambda$ is very close to $\lambda_m$. Now in order to minimize the probability of false alarm we need to choose $T = L + M$ as large as possible. Nevertheless, we also minimize the quick response to an overload, or short term congestion, (i.e., we make the leaky bucket less sensitive to short term congestions. In order to quantify the detection time we use our previously developed model for overload behavior. We consider the following non-stationary arrival process:

$$
\lambda(t) = \begin{cases} 
\lambda_m & t < 0 \\
\lambda_1 & t > 0 
\end{cases}
$$

(43)

The detection time can be interpreted as the time of the fractional loss to rise from the value $\epsilon_1$ to a value $(1 - \epsilon_2) f_{\text{loss}}(\infty)$ for the arrival process defined in Equation (43). A good behavior would imply that $f_{\text{loss}}$ will achieve a value close to $\frac{\lambda_m}{\lambda_1}$ in a short period of time. In Figure 17 we illustrate the effect of the total buffer size on the detection time assuming $\epsilon_1 = 0$, $\epsilon_2 = 0.2$ and $\lambda = 2\lambda_m = 2$. We note that $t_5 < t_10 < t_20$ and we conclude as we increase the total buffer size detection time increases, where $t_i$ denotes detection time for a leaky bucket with total buffer space equal to $i$.

### 5.3 Controlling the peak rate duration

In this section we will quantify the effectiveness of the leaky bucket to control the duration of the peak rate. Assume a source that emits cells at (constant) peak rate $B_p$ and let token generation rate be $r$. Let $I(T)$ denote the sum of the number of tokens in the token buffer and the number of empty places in the cell buffer. Then for a given peak duration $D$, $D \times B_p$
cells and $D \times r$ tokens are generated. We require that no more than $\epsilon\%$ lost cells for the duration of the burst i.e.,

$$\max\{0, D \times B_p - D \times r - I(T)\} \leq \epsilon$$  \hspace{1cm} (44)$$

We can find the maximum allowable burst that satisfies the given loss requirements as:

$$D \leq \frac{I(T)}{B_p - r - \epsilon}$$  \hspace{1cm} (45)$$

Now since $I(T)$ is proportional to $T$ we can say that the burst tolerance is directly proportional to $T$ and is inversely proportional for a given $\epsilon$ to the difference between peak rate and token rate.

If the burst size parameter is defined as a mean value, an ON-OFF source model such as IPP can be used to dimension the leaky bucket. We will then find the value of $T = L + M$ such that a “typical” burst will suffer a loss less than a given threshold, $\epsilon\%$. As a “typical” burst (Figure 18) we consider an ON period that lasts for $\frac{1}{\lambda}$ time units, (the mean sojourn time in the ON state) and that occurs $\frac{1}{2\lambda}$ time units before (or after) a token generation instant. The (limiting) probability that the beginning of a burst that occurs $\frac{1}{2\lambda}$ relatively to a token generation instant will find the leaky bucket in a given state is given by:

$$Pr\{\text{State } i \mid \text{source is ON}\} = \frac{Pr\{\text{State } i \cap \text{source is ON}\}}{Pr\{\text{source is ON}\}}$$  \hspace{1cm} (46)$$

Now the long-run probability that the source is ON is simply $\frac{1}{\lambda + \epsilon}$. The above suggests that the limiting distribution at the beginning of ON periods is simply found by normalizing
according to the state probabilities that the source is ON. Fractional loss will then be given as:

$$f_{\text{loss}}(\frac{1}{a}) = \left[ \pi_{ON} \left( \frac{\tau}{2} \right) S \left( \frac{1}{a} \right) \right]_{L+M}$$

(47)

In Figure 19 we plot fractional loss for an ON-OFF source with parameters $a = 0.65$, $b = 0.35$, $\lambda = \frac{1}{0.35}$ and $\tau = 1$. Observe that if we require the fractional loss for the typical burst to be less than 9% we should choose a bucket size equal to 15.

To summarize, we observe that for mean rate enforcement the choice $\lambda_m = \frac{1}{T}$ is an essential one. It should be noted that this choice leads to large loss probabilities (false alarm) for reasonable values of $T = L + M$. Additionally, the sizes of the buffer and the token pool control the detection time (or equivalently responsiveness to short-term congestions), the burst size, and the probability of false alarm.
6 Conclusions

In this paper we presented the time-dependent behavior of the leaky bucket rate control scheme under general point arrival processes. Using the theory of Markov regenerative processes we showed that system state distribution has no pointwise limits but its time averages have limits. Due to its popularity as a policing device, we defined two time-dependent loss-related measures. We used these measures to quantify the overload response of the leaky bucket and to study the effectiveness of the leaky bucket to control standard traffic parameters such as peak rate, mean rate and peak rate duration. Transient analysis provides us with additional insight about the dynamic behavior of the leaky bucket and its responsiveness to overload conditions. Our results suggest the tradeoffs for the different quantities to be controlled.

References


Appendix

Evaluation of $A(n, t)$ for a MAP process

The $(i, j)$th element of matrix $A(n, t)$ denotes the probability of $n$ arrivals in $t$ time units given that the initial state is $i$ and the state at time $t$ is $j$. Matrices $A(n, t)$ satisfy the differential equations:

$$\frac{dA(n, t)}{dt} = A(n, t)D_0 + A(n-1, t)D_1, \quad n > 0$$

with the initial condition $A(0, 0) = I$. The entries of matrices $D_0$ and $D_1$ are defined in Section 3.2. The system of differential equations can be solved in $s$ domain to give us the Laplace transform of matrix $A(n, t)$:

$$A(n, s) = (sI - D_0)^{-1}(D_1)^n$$

Inverting the Laplace transform we obtain:

$$A(n, t) = \sum_{k=0}^{n} B_k \frac{t^k}{k!} e^{-\rho_1 t} + \sum_{k=0}^{n} C_k \frac{t^k}{k!} e^{-\rho_2 t}$$

Calculating matrices, $B_k$, $C_k$ is not a trivial task. Here we will use an alternative approach. By taking Laplace transforms in both sides of Equation (48) we obtain:

$$A(n, s) = (sI - D_0)^{-1} A(n - 1, s)D_1$$


The above equation involves a matrix inversion; for the special case of a 2-state MAP we have:

$$(sI - D_0)^{-1} = \frac{1}{(s + \rho_1)(s + \rho_2)}L(s)$$

where

$$L(s) = \begin{bmatrix} s + \lambda_1 & \lambda_0p_0(0, 1) \\ \lambda_1p_1(0, 0) & s + \lambda_0 \end{bmatrix}$$

$$\rho_1, \rho_2$$ are the (real) roots of the polynomial $(s + \lambda_0)(s + \lambda_1) - \lambda_0\lambda_1p_0(0, 1)p_1(0, 0)$ and $p_0(0, 1), p_1(0, 0)$ as defined in Section 2.3.

Let $B_k^{[n]}, C_k^{[n]}, k = 0, 1, 2, \cdots, n$ denote the coefficients of $A(n, t).$ We would like to express the coefficients $B_k^{[n+1]}, C_k^{[n+1]}, k = 0, 1, 2, \cdots, n+1,$ in terms of $B_k^{[n]}, C_k^{[n]}, k = 0, 1, 2, \cdots, n.$ In s-domain we have:

$$A(n, s) = \sum_{k=0}^{n} B_k \frac{1}{(s + \rho_1)^{k+1}} + \sum_{k=0}^{n} C_k \frac{1}{(s + \rho_2)^{k+1}}$$

(54)

Now $A(n+1, s)$ can be written as:

$$A(n + 1, s) = \sum_{k=0}^{n} L(s) B_k D_1 \frac{1}{(s + \rho_1)^{k+2}(s + \rho_2)} + \sum_{k=0}^{n} L(s) C_k D_1 \frac{1}{(s + \rho_2)^{k+2}(s + \rho_1)}$$

(55)

Using partial fraction expansion the above equation becomes:

$$A(n + 1, s) = \sum_{k=0}^{n} \left[ B_{k0} \frac{1}{s + \rho_2} + \sum_{l=1}^{k+2} B_{kl} \frac{1}{(s + \rho_1)^l} \right]$$

$$+ \sum_{k=0}^{n} \left[ C_{k0} \frac{1}{s + \rho_1} + \sum_{l=1}^{k+2} C_{kl} \frac{1}{(s + \rho_2)^l} \right]$$

(56)

The coefficients $B_{kl}$ and $C_{kl}$ are given by:

$$B_{kl} = \begin{cases} B_k L(-\rho_2) D_1 \frac{1}{(\rho_1 - \rho_2)^{k+l}}, & l = 0 \\ B_k L(-\rho_2) D_1 (-1)^{k+2-l} \frac{1}{(\rho_2 - \rho_1)^{k+l}}, & 0 < l < k + 2 \\ B_k L(-\rho_1) D_1 \frac{1}{\rho_2 - \rho_1}, & l = k + 2 \end{cases}$$

(57)
and:

\[
\begin{aligned}
C_{kl} &= \begin{cases} 
C_k L(-\rho_1) D_1 \frac{1}{(\rho_2 - \rho_1)^{k+l+1}} & l = 0 \\
C_k L(-\rho_2) D_1 (-1)^{k+2-l} \frac{1}{(\rho_1 - \rho_2)^{k+l+1}} & 0 < l < k + 2 \\
C_k L(-\rho_2) D_1 \frac{1}{\rho_1 - \rho_2} & l = k + 2 
\end{cases} 
\end{aligned}
\] (58)

The above equation after some algebraic manipulation becomes:

\[
\begin{aligned}
&\left[ \sum_{k=0}^{n} B_{k1} + \sum_{k=0}^{n} \sum_{i=0}^{n} C_{i0} \right] \frac{1}{s + \rho_1} + \sum_{l=2}^{n+2} \left[ \sum_{k=0}^{n} B_{kl} \right] \frac{1}{(s + \rho_1)^l} \\
&+ \left[ \sum_{k=0}^{n} C_{k1} + \sum_{k=0}^{n} B_{k0} \right] \frac{1}{s + \rho_2} + \sum_{l=2}^{n+2} \left[ \sum_{k=0}^{n} C_{kl} \right] \frac{1}{(s + \rho_2)^l} 
\end{aligned}
\] (59)

We can then write:

\[
B_0^{(n+1)} = \sum_{k=0}^{n} (B_{k1}^{(n)} + C_{k0}^{(n)}) \\
B_l^{(n+1)} = \sum_{k=l-2}^{n} B_{kl}^{(n)}, \quad 0 < l \leq n + 1 \\
C_0^{(n+1)} = \sum_{k=0}^{n} (C_{k1}^{(n)} + B_{k0}^{(n)}) \\
C_l^{(n+1)} = \sum_{k=l-2}^{n} C_{kl}^{(n)}, \quad 0 < l \leq n + 1
\] (60)

Let’s also define the matrix \( \tilde{A}(n,t) \) as:

\[
\tilde{A}(n,t) \triangleq \sum_{k=n}^{\infty} A(k,t)
\] (61)

Matrix \( \tilde{A}(n,t), \) for \( n > 0, \) represents the probability of having more than \( n - 1 \) arrivals up to time \( t \) given that we were in state \( i \) at time 0 and we are in state \( j \) at time \( t. \) This can be computed as:

\[
\tilde{A}(n+1,t) = \tilde{A}(n,t) - A(n,t)
\] (62)

given the initial condition \( \tilde{A}(0,t):\)

\[
\tilde{A}(0,t) = \begin{bmatrix} n_1 + n_2 e^{-(a+b)} & n_2 \left(1 - e^{-(a+b)}\right) \\
2 \left(n_1 e^{-(a+b)}\right) & n_1 \left(1 - e^{-(a+b)}\right) + n_2 e^{-(a+b)} \end{bmatrix}
\] (63)

with \( n_1 = \frac{a}{a+b}, \ n_2 = \frac{b}{a+b}, \ a = \lambda_0 [p_0(0,1) + p_0(1,1)] \) and \( b = \lambda_1 [p_1(0,0) + p_1(1,0)]. \) A similar approach can be applied for an \( m \)-state MAP. Nevertheless, this method requires the roots of a polynomial of degree \( m. \) Uniformization could then be applied to compute numerically the entries \( A(n,t). \)