Study of Delay Jitter With and Without Peak Rate Enforcement *

Randall Landry † Ioannis Stavrakakis ‡

†DSP R&D Center
Texas Instruments Inc.
Dallas, TX 75265

‡Dept. of Electrical and Computer Engineering
Northeastern University
Boston, MA 02115

Abstract

In this work the modification of a tagged traffic stream due to statistical multiplexing is studied, by presenting a numerical approach for the calculation of the tagged delay jitter and interdeparture processes. Both the single and multiple node cases are considered. Unlike the past work, the developed approach is applicable under both the standard FCFS policy and a peak-rate enforcing multiplexing policy. The latter policy can be adopted to re-shape the tagged traffic stream within the network. Restoring a shaped traffic profile within the network may be necessary to obtain some of the benefits for which the original shaping is carried out at the network edge. This study also provides results and insight regarding the potential gains of re-regulation within the network.

Please send correspondence to:
Prof. Ioannis Stavrakakis
Department of Electrical and Computer Engineering
Northeastern University
409 Dana Research Building
Boston, MA 02115
email: ioannis@cdsp.neu.edu
Tel: (617)373-3053/ Fax: (617)373-8970

*Research supported by the Advanced Research Project Agency under Grant F49620-93-1-0564 monitored by the Air Force Office of Scientific Research (AFOSR).

1 Introduction

The Asynchronous Transfer Mode (ATM) is presently considered to be the most promising switching and multiplexing scheme for B-ISDN, as manifested by the large number of CCITT recommendations related to ATM [1]. Although high network efficiency (utilization) can be achieved through the statistical multiplexing of bursty sources, Quality of Service (QoS) cannot – in general – be guaranteed; the latter depends on (the statistical) service demands of the other users sharing the same network resources. Since at a given time instant the cumulative peak rate of the supported ATM users may exceed the network capacity, serious congestion may occur, resulting in QoS degradation. Source traffic shaping and/or regulation has been proposed as a means of reducing traffic burstiness, thereby controlling network congestion.

It is well understood that statistical multiplexing may induce substantial distortion in an ATM cell stream [2]. Thus, characterization of a tagged traffic stream within the network seems to be necessary in order to accurately evaluate potential congestion, as well as determine the QoS delivered to the tagged stream through a per-session end-to-end study. In addition, a corrective action (that is, traffic shaping) may be necessary within the network for the purpose of reducing distortion, which is a potential cause of network congestion.

In this work, the distortion of a (regulated) source traffic stream due to a series of multiplexing operations is studied under the standard FCFS, as well as a "corrective" peak-rate enforcing, multiplexing discipline. The tagged traffic stream is characterized at the output of a multiplexing node in terms of the cell delay, delay jitter and inter-departure processes. A comparative study of the two multiplexing disciplines is carried out based on these quantities and useful insight is obtained regarding the resulting distortion of the tagged traffic stream. Finally, the node-by-node approximate description of the transformation of a tagged stream under the considered multiplexing disciplines is used for an approximate study of the end-to-end QoS delivered to the tagged traffic stream.

Past work regarding the characterization of a tagged traffic stream due to a single FCFS multiplexing operation may be found in [3], [4]. In [3] a method is developed for calculating the cell inter-departure time distribution of a general renewal stream under a FCFS policy. Statistics for the inter-departure process of an initially periodic cell stream in a FCFS multiplexer are also computed in [4] by assuming a first-order Markovian structure for the cell delay process.

The problem of measuring end-to-end performance in an ATM environment is difficult and largely open. Typically, a specific application (session) is being tagged and observed as it traverses the network. Deterministic bounds on delay have been derived in [5],[6],[7],[8], where the traffic from each session is assumed to conform to pre-defined burstiness constraints. These approaches are based on worst-case performance and the bounds are extremely loose, resulting in the under-utilization of available resources when used for call admission purposes [9]. Provable statistical bounds have been derived in [10],[11], which also appear to be too loose to be of practical use in call admission. An alternative approach is to provide approximate statistical QoS guarantees [12],[13],[14], by modeling the traffic at the network edge (or within the network itself) and analyzing the resulting queueing systems. While this approach has the advantage of simplicity, it generally suffers from the traffic assumptions which must be made at the network premises. Past work on approximate end-to-end network performance evaluation based on nodal decomposition and FCFS multiplexing may be found in [15],[16],[17],[18]. Although the end-to-end study presented in this paper is based on "standard" nodal decomposition assumptions as in [15], [18], the single node analysis approach is different, allowing for the consideration of policies which cannot be studied by employing the past approaches.

In the next section the fundamental block of this work is considered, namely a queueing system under a service policy which occasionally departs from the FCFS policy to enforce a peak output rate (minimum cell inter-departure time) for the tagged traffic stream (Class-1). It is shown (section 2.3) that the proposed service policy allows for a characterization of the tagged cell stream which is consistent with the GCRA defined in [19]. Furthermore, when the peak-rate enforcer is disabled (by properly setting a key parameter), the proposed policy becomes the standard FCFS policy considered in [20],[4]. In section 4, the approximate end-to-end study is described along with a discussion of important assumptions. Numerical results are presented in section 4 before concluding in the final section.

2 Analysis of a Single Multiplexing Stage

There are two main objectives in this work. First, the distortion in a tagged traffic stream at a single node due to the standard FCFS as well as a distortion-reducing service policy will be studied. This study will reveal the amount of distortion introduced by the policies considered, as well as provide for a description of the tagged traffic stream at the output of the multiplexing stage. The latter corresponds to the tagged stream arriving to the next multiplexing stage and, thus, the next stage can be studied in a similar manner considering the tagged traffic stream as shaped by the previous multiplexing stage. By reiterating the above over a series of multiplexing stages an approximate end-to-end study of a tagged traffic stream can be carried out, revealing the end-to-end impact of the FCFS and the distortion-reducing multiplexing disciplines on a tagged stream, which is the second major objective of this work.

2.1 The Distortion-Reducing Peak Output-Rate Enforcing (PORE) Service Discipline

The PORE service discipline is associated with a tagged traffic stream (Class-1) and treats the rest of the traffic in an indistinguishable manner (Class-2 or background traffic). As it becomes clear below, the PORE service strategy for Class-1 guarantees that cells belonging to this class are transmitted at a minimum spacing of X_{min} slots. The server operates according to the FCFS discipline except for those times when the FCFS server would place consecutive Class-1 cells on the output link at a distance of less than X_{min}. In this paper it is assumed that under the FCFS policy, Class-1 cells are served before the Class-2 cells which arrive over the same slot.

Let t_k denote the service epoch of the k-th Class-1 cell, defined as the beginning of the slot over which this cell is transmitted. Let $X_k = t_{k+1} - t_k$ denote the inter-departure time of Class-1 cells $k+1$ and k. Similarly, $\tilde{t}_k$ will mark the *potential* service epoch of the k-th Class-1 cell, which is defined to be the beginning of the slot at which all other cells that arrived prior to the k-th Class-1 cell have been served. Note that under the FCFS discipline, $\tilde{t}_k = t_k$, $\forall k \geq 1$.

The PORE service discipline can be described in terms of the following rules:

(a) Whenever $\tilde{t}_k - t_{k-1} < X_{min}$, the server delays the initiation of service to the k-th Class-1 cell by exactly $H_k = X_{min} - (\tilde{t}_k - t_{k-1})$ slots, which will be referred to as the *holding time* of the k-th cell for the remainder of the paper.

(b) During this holding time of length H_k, the server will provide FCFS service to Class-2 cells that may be present in the buffer. This implies that during the holding interval, up to H_k Class-2 cells can be served before the k-th Class-1 cell which arrived at an earlier time.

(c) When the holding time expires, the k-th Class-1 cell is served; thus, $t_k = \tilde{t}_k + H_k$, thereby ensuring that $X_{k-1} = X_{min}$.

Note that according to rule (b), the server abandons the FCFS discipline whenever a Class-1 cell is held ($H_k \geq 1$), and thus, it provides improved service to Class-2 compared to that under the FCFS policy. The PORE service policy reduces to the FCFS policy when $X_{min} = 1$, in which case $H_k = 0$, $\forall k \geq 1$.

Let the Class-1 cell inter-arrival process be denoted by $\{A_k\}_{k\geq 1}$, where A_k represents the inter-arrival time (in slots) between Class-1 cells k and $k{+}1$. It is assumed that $A_k \in [A_{min}, \infty) \subseteq \mathbf{Z}^+$, where $\mathbf{Z}^+$ denotes the set of positive integers. Class-2 traffic is considered to be the background traffic whose presence stands to corrupt $\{A_k\}_{k\geq 1}$ through the queueing process.

2.2 Jitter (Cell Delay Variation) and the PORE Policy

An important measure of distortion of a real-time application's cell stream is delay jitter, or cell delay variation. When a session experiences a large amount of delay jitter, or distortion to its traffic profile, extensive buffering may be required at the receiver in order for the original data stream to be recreated before being played back. Clearly, this implies that network-induced delay jitter can be removed at the receiver at the expense of larger buffers and potentially increased delays. This can be accomplished by temporarily storing cells in a delay jitter removal buffer so that the inter-arrival times of cells at the receiving end are "close" to the inter-arrival times at the originating node (Fig. 1). In view of this, it seems reasonable to expect that the end user buffering requirements can potentially be significantly reduced by controlling delay jitter within

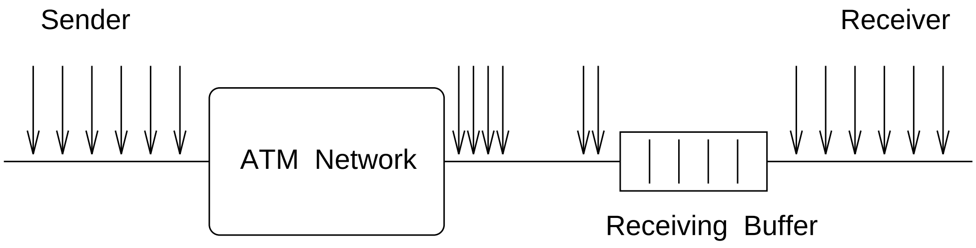

Figure 1: Jitter removal at the receiver.

the network. The implication here is that some of the buffers can be displaced from the receiving end to the network premises where resources are shared by more than one user. Finally, besides creating the need for increased buffering at the receiver, an initially well-behaved cell stream can induce queueing problems and QoS degradation for existing sessions as it becomes distorted and potentially more bursty within the network.

A number of definitions for delay jitter have appeared in the literature, including those which compute the probability distribution of cell delays and measure the difference between lower and upper quantiles [4],[21]. In this paper a cell-level definition for delay jitter, like the one considered in [4], will be adopted to describe the amount of distortion introduced into the Class-1 cell stream.

Definition 1 *Let D_k denote the total queueing delay (in slots) of the k-th Class-1 cell, including the transmission slot. The delay jitter of the k-th cell is defined as*

$$J_k = D_{k+1} - D_k \quad \textit{for } k \geq 1. \quad \Box \tag{1}$$

In terms of the quantities defined in the previous subsection, an equivalent definition of delay jitter is given by

$$J_k = X_k - A_k \quad \text{for } k \geq 1. \tag{2}$$

Note that J_k measures the degree to which the inter-arrival time A_k of Class-1 cells k and $k+1$ has been distorted, and it can take both positive and negative values. *Negative* delay jitter applies to those instances when Class-1 cells k and $k+1$ are transmitted at a spacing of less than A_k (cell clustering), while *positive* delay jitter applies to those times when cells k and $k+1$

are transmitted at a spacing of greater than A_k (cell spreading). Note that when the process $\{A_k\}_{k\geq 1}$ has a finite region of support (i.e. $A_k \in [A_{min}, A_{max}]$), the PORE policy completely eliminates a portion of the negative delay jitter by guaranteeing that $J_k \geq X_{min} - A_{max}$.

Since providing a minimum spacing of X_{min} eliminates some of the negative delay jitter, it would be reasonable to investigate whether the PORE policy induces probabilistically larger positive delay jitter than the FCFS policy. This is addressed in the following theorem.

Theorem 1 *Consider any Class-1 arrival stream with minimum cell inter-arrival time A_{min}. Let J_k^{FCFS} (J_k^{PORE}) denote the delay jitter of the k-th Class-1 cell under the FCFS (PORE) service policy. If $X_{min} < A_{min}$, then for all $k \geq 2$,*

$$Pr\left\{J_k^{PORE} \geq j\right\} \leq Pr\left\{J_k^{FCFS} \geq j\right\} \quad \forall j \geq 0,$$

with equality when $X_{min} = 1$.

<u>Proof</u>: Recall that if the k-th Class-1 cell is held, it is guaranteed to be served at time t_k such that $X_k = X_{min}$. This implies that a held cell will never be placed on the output link at a distance less than X_{min} from the previous Class-1 cell. Since X_{min} is always less than A_{min}, then $X_k < A_k$. From (2), this implies that $J_k = X_k - A_k$ cannot be greater than or equal to zero, or positive delay jitter is never induced through this output-rate-enforcement action. In fact, the positive delay jitter is decreased, since while the k-th Class-1 cell is being held, the server is attending to Class-2, and consequently, the service time of the $(k+1) - st$ cell is possibly moving closer to t_k. □

Theorem 1 establishes the fact that regulating Class-1 by enforcing a peak output rate, does not increase the positive delay jitter of this traffic class. The PORE service policy actually reduces the two types of distortion most commonly induced by multiplexers, namely spreading and clustering.

2.3 The Generic Cell Rate Algorithm (GCRA) for ATM

The negative impact of delay jitter on an ATM connection (as well as other supported sessions in the network) has been recognized by the ATM Forum [19], and one of the consequences has been the development of the GCRA, which establishes a relationship between Peak Cell Rate R_p and Cell Delay Variation Tolerance (CDVT). R_p is defined as the inverse of the minimum cell inter-arrival time at the user; assuming that A_{min} is the minimum inter-arrival time contracted by the user, then $R_p = \frac{1}{A_{min}}$. On the other hand, CDVT represents a bound on the cell clustering phenomenon previously described and it is defined in relation to R_p according to the GCRA as explained below.

The basic function of the GCRA is to update a Theoretical Arrival Time (TAT), which is simply the expected arrival time of a cell assuming that cells are equi-spaced when the source is active. Let t_k^a denote the actual arrival time of the k-th cell at some observation point within the network. If t_k^a is not "too early" relative to the TAT, or if $t_k^a \geq \mathrm{TAT} - \tau$ for some small $\tau > 0$, then the cell is said to be *conforming*; otherwise the cell is *non-conforming*. In order to ensure that all cells belonging to a traffic stream are conforming, non-conforming cells can either be held until they conform, or dropped. Such a cell stream is said to conform to GCRA(A_{min}, τ), where τ is the value of CDVT and is a mandatory parameter in any Connection Traffic Descriptor [19]. In this paper, it will be assumed that τ takes only integer values.

A cell stream conforming to GCRA(A_{min}, τ) may be considered to have a contracted peak rate of $\frac{1}{A_{min}}$, and a "tolerated" peak rate of $\frac{1}{X_{min}}$, where

$$X_{min} = A_{min} - \tau.$$

Clearly, the PORE service policy described above can be used to ensure that the Class-1 cell stream conforms to GCRA(A_{min}, τ) at the output of any multiplexer or switching site within the network.

The appropriate value of τ, which is to be chosen by the user from a set of values supported by the network, is the topic of ongoing research [19], but it is expected that under certain operating conditions the impact of CDVT on network resources will be significant. Clearly, a τ decreases the more well-behaved the cell stream becomes. In terms of the quantities associated with the PORE policy, when $X_{min} = A_{min}$, τ becomes zero and the Class-1 cell stream conforms to GCRA$(A_{min}, 0)$ which happens to be the function of the Virtual Shaper defined in [19]. Providing this strict conformance with the contracted peak rate will undoubtedly have a positive impact on the QoS delivered to other traffic streams along the same Virtual Path. As previously mentioned, the present study will focus on the impact that CDVT has on the conforming (or tagged) cell stream in terms of delay jitter, cell inter-arrival times and cell delays.

2.4 Queueing Analysis

In this section, an ATM multiplexer operating under the PORE service policy (Fig. 2) is studied in terms of performance measures associated with the Class-1 cell stream. The Class-1 inter-arrival process $\{A_k\}_{k\geq 1}$ is assumed to be i.i.d. with distribution $f_a(m)$, $A_{min} \leq m \leq A_{max}$. Note that a periodic process with period T can be considered when $A_{min} = A_{max} = T$ and $f_a(T) = 1$.

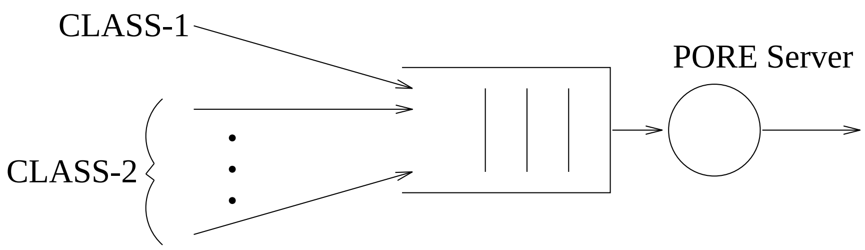

Figure 2: The discrete-time queueing system

Class-2 traffic is modeled in terms of i.i.d. batches of cells of arbitrary distribution. Let the random variable B_j—with probability mass function (PMF) $f_{b_j}(k)$, $k \geq 0$—denote the number of Class-2 cells arriving over an interval of length j (in slots). Clearly, $f_{b_j}(\cdot)$ can be derived as the j-fold convolution of $f_{b_1}(\cdot)$; let N^b be the maximum value of B_1.

Let $Q(n;j)$ denote the number of Class-2 cells which arrived before time epoch j and are still in the system at time n, $n \geq j$. The following proposition determines the evolution of $\{Q(n;j)\}_{n\geq j}$ under certain conditions, and will be useful for the study of the queueing system.

Proposition 1 *The process $\{Q(j+m;j)\}_{m=0}^{n}$ evolves as the occupancy process of a Geo/D/1 queue with batch arrivals distributed according to $f_{b_1}(\cdot)$, provided that* <u>*no*</u> *Class-1 cells are served* *over the interval* $[j, j+n]$*. In this case,*

$$Pr\left\{Q(j+m;j)=\ell \mid Q(j;j)=k\right\} = P^{(m)}_{k,\ell} = \sum_{i} P^{(m-1)}_{k,i} P_{i,\ell}, \tag{3}$$

where $P^{(m)}_{k,\ell}$ $(P_{i,\ell})$ *is the m-step (1-step) transition probability for the* Geo/D/1 *queue;* $P_{i,\ell} = f_{b_1}(\ell - i + 1) + f_{b_1}(0)\mathbf{1}_{\{\ell+i=0\}}$. $\square$

The queueing system will be described in terms of the Markov chain $\{D_k, H_k\}_{k\geq 1}$, embedded at the sequence of time instants $\{t_k\}_{k\geq 1}$; t_k marks the service epoch of the k-th Class-1 cell as defined earlier; D_k and H_k denote the total delay and holding time of the k-th cell, respectively.

The transition probabilities for the Markov chain $\{D_k, H_k\}_{k\geq 1}$ will be derived by first establishing expressions for D_k and H_k given that $(D_{k-1}, H_{k-1}) = (i, j)$, where $i \geq 1$ and $0 \leq j \leq X_{min} - 1$. In order to simplify notation, the quantity $Q(t_{k-1}; t_{k-1} - i + j)$, which appears frequently in the following expressions, will be denoted by Q^*. Also, let Q_k denote the number of Class-2 cells in queue at the arrival epoch (time $t_{k-1} - i + A_{k-1}$) of the k-th Class-1 cell. That is, $Q_k = Q(t_{k-1} - i + A_{k-1}; t_{k-1} - i + A_{k-1})$.

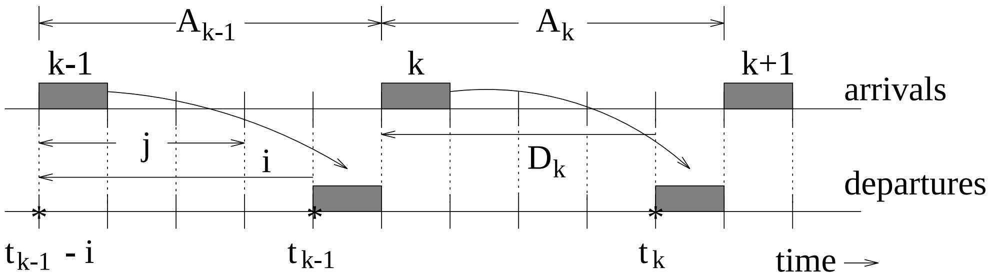

Figure 3: Realization for the derivation of D_k and H_k when $i < A_{max}$.

When $i < A_{max}$, the expressions for D_k and H_k can be derived by referring to Fig. 3, where $i < A_{k-1}$ for the realization shown there. Note that i can also take values greater than or equal to A_{k-1} while still being less than A_{max}, and these cases are also accounted for in the expressions given below:

$$D_k = \begin{cases} 1 & \text{if } Q_k = 0 \text{ and } A_{k-1} \geq X_{min} + i - 1 \text{ and } i < A_{k-1} \\ Q_k & \text{if } Q_k \geq X_{min} + i - A_{k-1} \text{ and } i < A_{k-1} \\ Q_k + H_k & \text{if } Q_k < X_{min} + i - A_{k-1} \text{ and } i < A_{k-1} \\ B_{A_{k-1}-j} + Q^* + 1 + i - A_{k-1} & \text{if } Q^* \geq X_{min} - B_{A_{k-1}-j} - 1 \text{ and } i \geq A_{k-1} \\ B_{A_{k-1}-j} + Q^* + 1 + i - A_{k-1} + H_k & \text{if } Q^* < X_{min} - B_{A_{k-1}-j} - 1 \text{ and } i \geq A_{k-1} \end{cases} \tag{4}$$

and

$$H_k = \begin{cases} 0 & \text{if } Q_k = 0 \text{and } A_{k-1} \geq X_{min} + i - 1 \text{ and } i < A_{k-1} \\ 0 & \text{if } Q_k \geq X_{min} + i - A_{k-1} \text{ and } i < A_{k-1} \\ X_{min} + i - A_{k-1} - Q_k & \text{if } Q_k < X_{min} + i - A_{k-1} \text{ and } i < A_{k-1} \\ 0 & \text{if } Q^* \geq X_{min} - B_{A_{k-1}-j} - 1 \text{ and } i \geq A_{k-1} \\ X_{min} - B_{A_{k-1}-j} - Q^* - 1 & \text{if } Q^* < X_{min} - B_{A_{k-1}-j} - 1 \text{ and } i \geq A_{k-1} \end{cases} \tag{5}$$

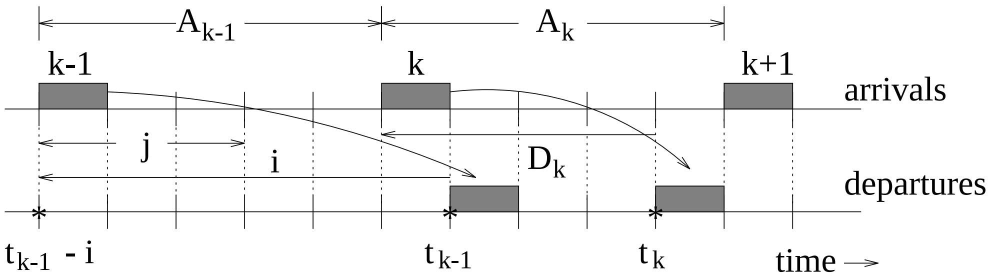

Figure 4: Realization for the derivation of D_k and H_k when $i \geq A_{max}$.

When $i \geq A_{max}$ (Fig. 4),

$$D_k = \begin{cases} B_{A_{k-1}-j} + Q^* + 1 + i - A_{k-1} & \text{if } Q^* \geq X_{min} - B_{A_{k-1}-j} - 1 \text{ and } i \geq A_{k-1} \\ B_{A_{k-1}-j} + Q^* + 1 + i - A_{k-1} + H_k & \text{if } Q^* < X_{min} - B_{A_{k-1}-j} - 1 \text{ and } i \geq A_{k-1} \end{cases} \tag{6}$$

and

$$H_k = \begin{cases} 0 & \text{if } Q^* \geq X_{min} - B_{A_{k-1}-j} - 1 \text{ and } i \geq A_{k-1} \\ X_{min} - B_{A_{k-1}-j} - Q^* - 1 & \text{if } Q^* < X_{min} - B_{A_{k-1}-j} - 1 \text{ and } i \geq A_{k-1} \end{cases} \tag{7}$$

Let $S^L = \{(i,j) : L(A_{max} - 1) + 1 \leq i \leq (L+1)(A_{max} - 1), 0 \leq j \leq X_{min} - 1\}$ represent "level" L of the Markov chain $\{D_k, H_k\}_{k\geq 1}$. Given the relationships in (4)-(7), the transition probabilities for $\{D_k, H_k\}_{k\geq 1}$, which will be denoted by $p(i,j,i',j')$, can be computed in terms of the quantities previously defined. The quantities $p(i,j,i',j')$, for $(i,j) \in S^0$ (transitions from level 0) are given by equations (8) and (9), which are derived from (4) and (5) above, while the transition probabilities for $(i,j) \in S^1$ are given by equations (10) and (11), derived from (6) and (7) above.

For $(i,j) \in S^0$,

$$\begin{aligned} p(i,j,i',0) &= \sum_{n=A_{min}}^{A_1} \sum_{m=0}^{N_1} f_a(n) b_{n-j}(m) P^{(j)}_{0,i'-i+n-m-1} + \\ & \sum_{n=A_2}^{A_{max}} \sum_{l=0}^{(N^b-1)j} \sum_{m=0}^{N^b(i-j+1)} f_a(n) P^{(j)}_{0,l} b_{i-j+1}(m) \left[P^{(n-i-1)}_{m+l,i'} + P^{(n-i-1)}_{m+l,0} \mathbf{1}_{\{i'=1\}} \right], \end{aligned} \tag{8}$$

where $A_1 = min\{i, A_{max}\}$, $A_2 = max\{A_{min}, i+1\}$ and $N_1 = min\left\{(n-j)N^b, i'-1-i+n\right\}$.

$$\begin{aligned} p(i,j,i',j') &= \sum_{n=A_{min}}^{A_1} \sum_{m=0}^{N_2} f_a(n) b_{n-j}(m) P^{(j)}_{0,X_{min}-1-m-j'} + \\ & \sum_{n=A_2}^{A_{max}} \sum_{l=0}^{(N^b-1)j} \sum_{m=0}^{N^b(i-j+1)} f_a(n) P^{(j)}_{0,l} b_{i-j+1}(m) \left[P^{(n-i-1)}_{m+l,i'-j'} + P^{(n-i-1)}_{m+l,0} \right] \cdot \\ & \mathbf{1}_{\{i'=X_{min}-n+i\}}, \end{aligned} \tag{9}$$

where $N_2 = min\left\{(n-j)N^b, X_{min}-1-j'\right\}$.
For $(i,j) \in S^1$,

$$p(i,j,i',0) = \sum_{n=A_{min}}^{A_1} \sum_{m=0}^{N_1} f_a(n) b_{n-j}(m) P^{(j)}_{0,i'-i+n-m-1}, \tag{10}$$

and

$$p(i,j,i',j') = \sum_{n=A_{min}}^{A_1} \sum_{m=0}^{N_2} f_a(n) b_{n-j}(m) P^{(j)}_{0,X_{min}-1-m-j'} \mathbf{1}_{\{i'=X_{min}-n+i\}}. \tag{11}$$

By lexicographically mapping the quantities $p(i,j,i',j')$ into $p(\ell,\ell')$ through the relations $\ell = (i-1)X_{min}+j$ and $\ell' = (i'-1)X_{min}+j'$, the transition probability matrix $\mathbf{P}$ of the Markov chain $\{D_k, H_k\}_{k\geq 1}$ can be written as a stochastic matrix of the M/G/1 type:

$$\mathbf{P} = \begin{bmatrix} \mathbf{B_0} & \mathbf{B_1} & \mathbf{B_2} & \mathbf{B_3} & \cdots \\ \mathbf{A_0} & \mathbf{A_1} & \mathbf{A_2} & \mathbf{A_3} & \cdots \\ 0 & \mathbf{A_0} & \mathbf{A_1} & \mathbf{A_2} & \cdots \\ 0 & 0 & \mathbf{A_0} & \mathbf{A_1} & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix} \tag{12}$$

where the matrices $\mathbf{B_n}$ and $\mathbf{A_n}$, $n \geq 0$, have dimensions $(A_{max} - 1)X_{min} \times (A_{max} - 1)X_{min}$. Note that the matrices $\mathbf{B_n}$ consist of the transition probabilities for which $(i, j) \in S^0$ and $(i', j') \in S^n$, while the matrices $\mathbf{A_n}$ consist of the transition probabilities for which $(i, j) \in S^1$ and $(i', j') \in S^n$. The system in (12) can now be solved as outlined in appendix A based on the matrix analytic techniques presented in [22]. The stationary probability vector $\boldsymbol{\pi}$ for $\mathbf{P}$ is mapped back into the space (i, j, i', j') to obtain the stationary probabilities for $\{D_k, H_k\}_{k\geq 1}$, given by $\pi(i, j)$, for $1 \leq i < \infty$, $0 \leq j \leq X_{min} - 1$.

The PMF for cell delay D of Class-1 is easy to calculate in terms of $\pi(i, j)$, and it is given by

$$Pr\{D = i\} = \sum_{j=0}^{X_{min}-1} \pi(i, j) \quad . \tag{13}$$

By considering the definition in (1), the delay jitter distribution is easily determined to be

$$Pr\{J = k\} = \sum_{i=1}^{\infty} \sum_{j=0}^{X_{min}-1} \sum_{j'=0}^{X_{min}-1} p(i, j, i + k, j')\pi(i, j) \tag{14}$$

Computing the PMF of the Class-1 cell inter-departure time X is not as straight-forward as computing that of J. From equations (1) and (2), the inter-departure time of cells $k+1$ and k is given as $X_k = D_{k+1} - D_k + A_k$. Since the Markov chain $\{D_k, H_k\}_{k\geq 1}$ does not provide adequate information to describe the inter-departure process, the system will be described by the Markov chain $\{D_k, H_k, A_k\}_{k\geq 1}$ in order to compute the PMF of X. Let $\tilde{p}(i, j, m, i', j', m')$ denote the transition probabilities and $\tilde{\pi}(i, j, m)$ denote the stationary probabilities for this Markov chain. It is easy to show that the PMF for the cell inter-departure process is given by

$$Pr\{X = n\} = \sum_{i=1}^{\infty} \sum_{j=0}^{X_{min}-1} \sum_{m=A_{min}}^{A'} \sum_{j'=0}^{X_{min}-1} \sum_{m'=A_{min}}^{A_{max}} \tilde{p}(i, j, m, n + i - m, j', m')\tilde{\pi}(i, j, m), \tag{15}$$

where $A' = min\{n+i-1, A_{max}\}$.

The quantities $\tilde{\pi}(i,j,m)$ can be computed in the same manner as the quantities $\pi(i,j)$ with the exception that the dimensionality of each level in $\mathbf{P}$ has increased by a factor of A_{max}. Note, however, that upon arrival of the k-th Class-1 cell, D_k is determined. That is, D_k is equal to the queue content at the arrival instant plus the amount of holding time, H_k, to be suffered by the cell. The latter quantity is completely determined by the past (i.e. the service time of the $(k-1)-st$ cell) and neither D_k nor H_k are dependent upon the next Class-1 arrival. Consequently, the random variables D_k and H_k are both independent of the random variable A_k, and the stationary probabilities for the Markov chain $\{D_k, H_k, A_k\}_{k\geq 1}$ can be computed as $\tilde{\pi}(i,j,m) = \pi(i,j)f_a(m)$, where $\pi(i,j)$ and $f_a(m)$ are as given above.

3 Approximate End-to-End Study

The N-node tandem queueing system used for the approximate end-to-end performance evaluation of the tagged session is shown in Fig. 5. The PORE service policy will be adopted at each node so as to allow for the consideration of peak-rate enforcement within the network. In order to simplify notation, the Class-1 cell inter-arrival (inter-departure) process at node i will be denoted as A^i (X^i). The arrival process at node i ($1 \leq i \leq N$) consists of the tagged (Class-1) inter-departure process from node $i-1$, and some background traffic (Class-2) delivering batches of cells distributed according to the PMF $f_{b_1}(\cdot)$. X^0 is defined to be the tagged cell inter-generation distnce at the source; A^1 may be slightly different than X^0 due to potentila truncation (as discussed below).

The set of possible integer values which can be taken by the Class-1 inter-arrival processes at nodes 1 through N must be truncated in order to facilitate a numerical solution at each node. The minimum (maximum) Class-1 inter-arrival time at node i will be denoted by A^i_{min} (A^i_{max}). That is, $A^i \in [A^i_{min}, A^i_{max}]$ for $i = 1, 2, \ldots, N$. The minimum inter-arrival time for $i \geq 1$ is simply $A^i_{min} = X^{i-1}_{min}$. The choice of A^i_{max} for $i \geq 1$ will be made according to the rule

$$A^i_{max} = \max\left\{k : Pr\left\{X^{i-1} = k\right\} \geq \epsilon\right\}, \tag{16}$$

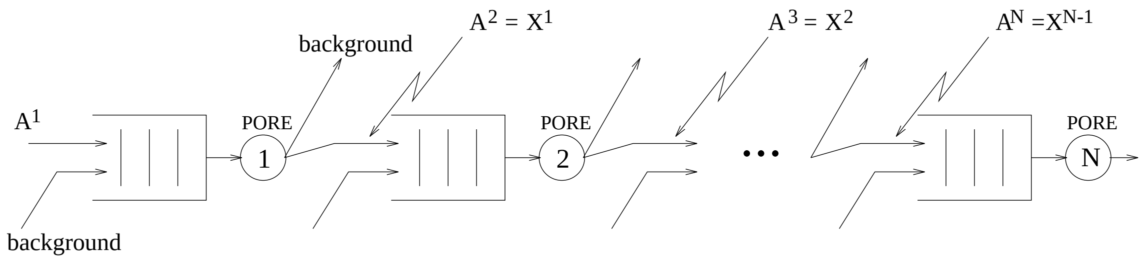

Figure 5: Tandem queueing network for end-to-end performance evaluation.

where ϵ can be any positive real number less than 1. The PMF for the process A^i, $i \geq 1$, is then given by

$$Pr\left\{A^i = k\right\} = \begin{cases} Pr\left\{X^{i-1} = k\right\} & \text{if } 0 < k < A^i_{max} \\ 1 - \sum_{j=1}^{A^i_{max}-1} Pr\left\{X^{i-1} = j\right\} & \text{if } k = A^i_{max} \\ 0 & \text{otherwise} \end{cases} \tag{17}$$

It is clear from (16) and (17) that as ϵ approaches zero, A^i_{max} increases and the truncated Class-1 inter-arrival time PMF at a node approaches the actual Class-1 inter-departure time PMF from the previous node. The cost of improving the accuracy of the adopted Class-1 inter-arrival process by increasing A^i_{max}, is a corresponding increase in numerical complexity. Recall that the dimensionality of the square matrices B_n and A_n, $n \geq 0$, which make up the transition probability matrix $\mathbf{P}$ for a single queue, increases linearly with the maximum Class-1 inter-arrival time (see section 2.4), which in turn affects the tractability of the solution technique outlined in appendix A.

A number of other assumptions will be made in order to facilitate the proposed approximate end-to-end study. First, it will be assumed that only the tagged traffic stream is forwarded from one multiplexer to the next as shown in Fig. 5. All of the remaining traffic (background) at each multiplexer is routed elsewhere. Implicit in this assumption is the presence of a large-scale networking structure, in which there exists a large number of independent sessions and a large number of total nodes (only N of which serve the tagged session). In such a network, it is expected to be reasonable to assume that, at the output of each node, most of the un-tagged sessions (background traffic) are split, or routed along paths which are different from that of the tagged session.

It is also assumed that the inter-departure process $\{X_k\}_{k\geq 1}$ from each node is an i.i.d. process. This assumption is necessary for employment of the analysis of section 2.2 at each multiplexing stage. The validity of the i.i.d. assumption, as well as the effect of truncating the inter-arrival time PMF at each node, will be examined through a comparative simulation study.

The analysis presented in the previous section for a single queue will be applied successively to each node along the path in Fig. 5 and the PMF for the inter-departure process at node N will be used to establish the amount of delay jitter induced in the cell stream of the tagged session over all N hops.

4 Numerical Results and Discussion

In this section, some numerical results are presented to characterize the distortion induced in a tagged traffic stream under both FCFS and peak-rate enforcing service strategies. Before considering end-to-end performance, the delay jitter induced at a single ATM multiplexer will be studied.

load	$T = 5$	$T = 10$	$T = 20$
0.70	0.0236	0.0086	0.0028
0.80	0.0489	0.0209	0.0079
0.95	0.1160	0.0607	0.0296

Table 1: Squared coefficient of variation for cell inter-departure time (C_X^2).

In order to provide for a simple interpretation of the multiplexer-induced distortion, the Class-1 arrival process is assumed to be periodic with period T slots in the results presented in this section. In addition, a typical regulated traffic stream (whose re-shaping may be attempted within the network) can be considered to resemble a periodic one for some time horizon which – in some cases – may be sufficient to induce "local" stationary statistics. It should be emphasized though that the study presented in this paper is applicable to traffic streams described in terms of general interarrival processes. The Class-2 batch arrival process is assumed to be binomially distributed with mean rate λ^b and maximum batch size $N^b = 15$. Table 1 displays the squared coefficient of variation for the Class-1 cell inter-departure time process when a FCFS policy is adopted. The squared coefficient of variation will be denoted by C_X^2 and given as

$$C_X^2 = \frac{Var(X)}{E^2(X)} \quad .$$

Note that C_X^2 is equal to zero if no distortion is introduced into the periodic arrival stream. From Table 1 it is clear that delay jitter is greatest, under the FCFS policy, for small periods T and high offered loads.

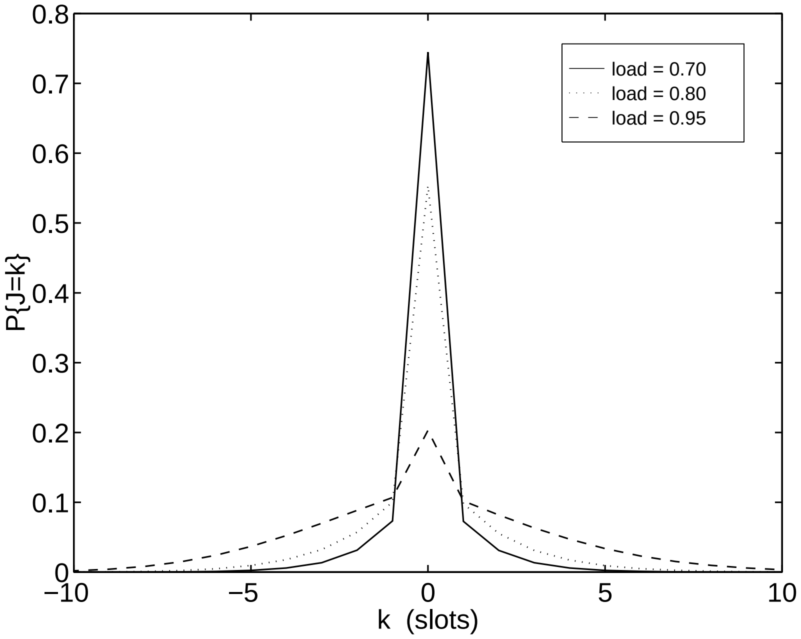

Figure 6: Jitter PMF at a FCFS multiplexer for $T = 20$ and various loads.

The multiplexer-induced distortion can also be quantified by observing the PMF for J as in Fig. 6. In this case the delay jitter PMF is plotted for $T = 20$ and the offered loads given in Table 1. Notice that when the total load is 0.7, much of the periodicity is retained in the Class-1 arrival stream, which is illustrated by the fact that most of the probability mass exists at $J = 0$. However, as the load increases to 0.95, much of this mass is redistributed and the Class-1 arrival process no longer resembles a periodic source. It is also interesting to note that the delay jitter

PMF appears to be symmetric around zero when the service policy at the multiplexer is FCFS; this symmetry disappears when a peak output rate is enforced.

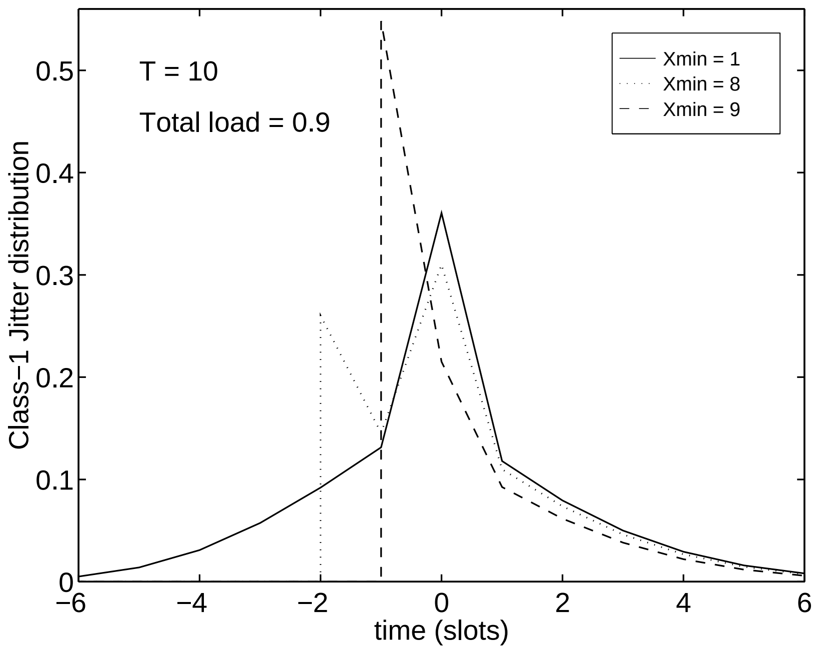

Figure 7: Jitter PMF for $T = 10$ and various values of X_{min} at a single multiplexer.

In Fig. 7, the delay jitter PMF is considered for different values of X_{min}, $T = 10$ and a total load of 0.9 ($\lambda^b = 0.8$). When X_{min} is set to $T-1$ (in this case 9), the tagged cell stream conforms to GCRA$(10, 1)$ (see section 2.3), and the delay jitter process possesses a very narrow region of support. On the other hand, under a FCFS policy a significant amount of cell clustering is observed since the cell stream now "conforms" to GCRA$(10, 9)$; cell spreading is also slightly increased.

The cost of providing a less distorted traffic stream at the multiplexer output is illustrated in Fig. 8 in terms of increased cell delays. As discussed in section 2.2, however, the effect of reducing distortion at the network premises, thereby increasing queueing delays (and buffer requirements), is essentially to distribute the buffers more uniformly throughout the network.

The remainder of this section will deal with the approximate end-to-end performance study. A network of five nodes in tandem will be considered, where the tagged traffic stream originates as a periodic source at node 1 with period $T = 5$. The background traffic at each node is

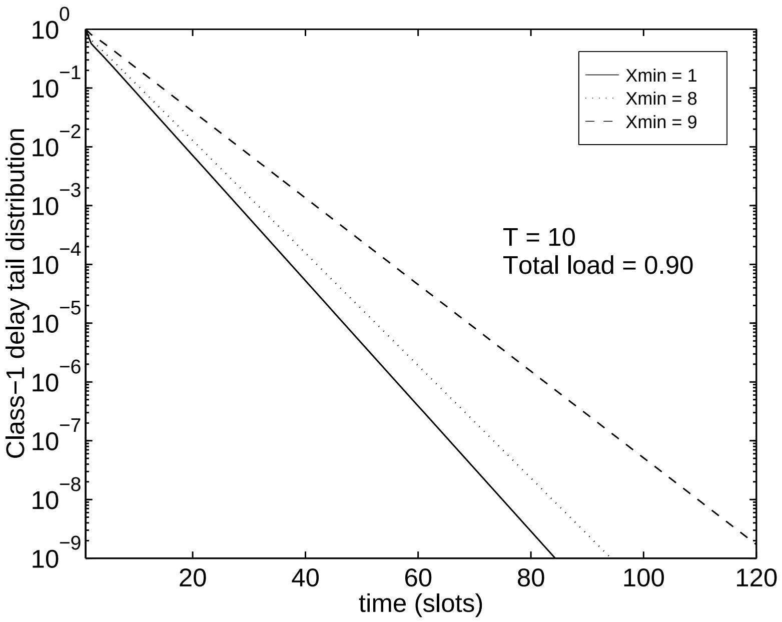

Figure 8: Delay tail distributions for $T = 10$ and various values of X_{min} at a single multiplexer.

assumed to consist of 10 identical Bernoulli streams each with rate 0.07, resulting in a total offered load of 0.9 at each multiplexer. The relatively high load at each node, together with the small period of $T = 5$, is adopted to provide a scenario in which significant delay jitter is induced into the tagged cell stream (see Table 1). The Class-1 inter-arrival time PMFs at nodes 2 through 5 will be truncated by setting ϵ equal to 10^{-6} and using the rule presented in the previous section. This choice of ϵ resulted in values of A^i_{max}, $2 \leq i \leq 5$, ranging from 20 to 30.

For the results in Fig. 9, each of the five ATM multiplexers serve arriving cells according to the FCFS policy ($X_{min} = 1$). The PMFs for inter-departure processes at nodes 1 and 5 are computed and compared with simulation results for the identical tandem queueing network. At the output of node 1, the cell inter-departure process X^1 has experienced both clustering and spreading, but over 35% of this traffic remains periodic with $T = 5$. As the tagged stream passes through additional FCFS multiplexing stages, the inter-departure times become more and more uniformly distributed, and at the output of node 5, there are no dominant terms in the PMF. This implies that, in a real-time scenario, significant buffering may be required at the end user in order to achieve an acceptable playback quality. Also, a traffic profile such as that at the output of node 5 will likely induce more severe queueing problems (and possibly QoS degradation) for other users sharing the same network resources.

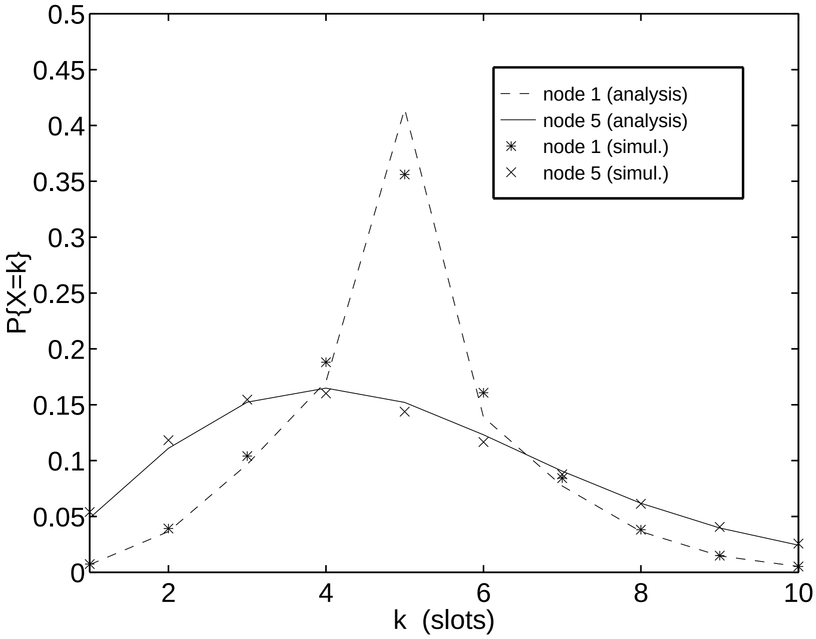

Figure 9: Cell inter-departure time PMFs at nodes 1 and 5 under FCFS policies ($X_{min} = 1$) at each node; $T = 5$.

The validity of an i.i.d. assumption for cell inter-departures (and inter-arrivals) is addressed through the simulation results plotted in Fig. 9 as discrete points. Clearly, under a FCFS policy at each node, this assumption does not significantly affect the accuracy of the end-to-end study, as the numerical results seem to be in good agreement with simulations. Similar agreement is present in Figs. 10 and 11, where the traffic has been shaped by the peak-rate enforcing policy at each node.

In Fig. 10 each multiplexer enforces a minimum inter-departure time of $X_{min} = 4$, and the resulting PMFs for X^1 and X^5 are presented. In both cases, the tagged stream conforms to GCRA$(5, 1)$ throughout the network. That is, all cell clustering (or negative delay jitter) less than -1 is eliminated, and the PMF at node 5 has a much more narrow region of support than the corresponding PMF under the FCFS policy. This can be observed in Fig. 11, where three networks of servers with X_{min} equal to 1, 3 and 4 are considered. As X_{min} increases (and the

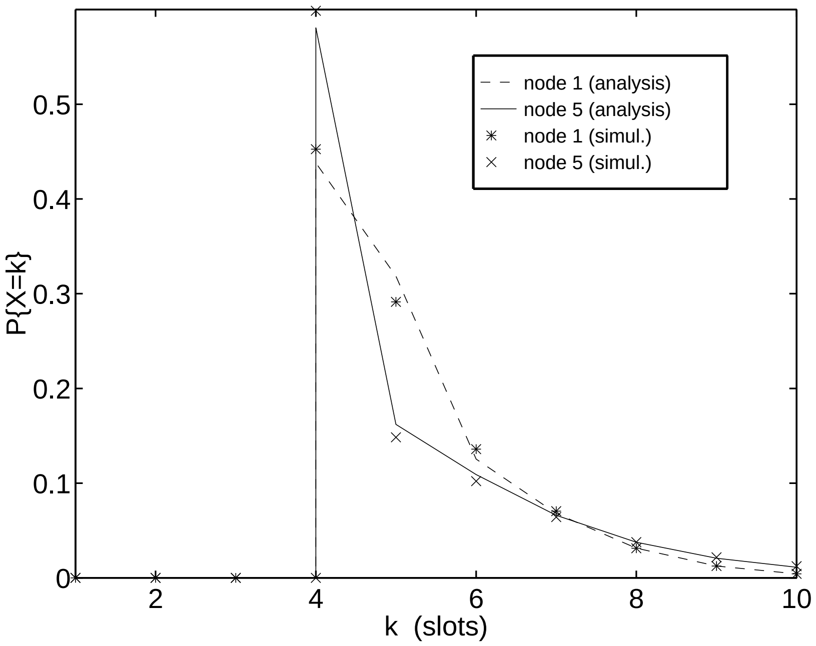

Figure 10: Cell inter-departure time PMFs at nodes 1 and 5 for $X_{min} = 4$ at each node; $T = 5$.

CDVT parameter τ decreases), the end-to-end performance of the tagged stream, in terms of the induced distortion, improves dramatically. It is also interesting to note that the approximate analysis agrees very well with simulation results, which validates the i.i.d. assumption for cell inter-departures even when the corrective peak output rate enforcing queueing discipline is adopted.

The squared coefficient of variation for the Class-1 inter-arrival process A^i is plotted in Fig. 12 as a function of the node index i for the three networks of servers considered in Fig. 11. Note that because the arrival process to node 1 is periodic, $C_A^2 = 0.0$ for node 1 in each case. As expected, C_A^2 increases with the node index for all values of X_{min}. However, depending upon the peak rate being enforced by each of the servers, the rate of increase, as well as the magnitude, of C_A^2 varies greatly. Under a FCFS policy at each server, the tagged cell stream is changing significantly as it enters node 6, and the induced delay jitter is quite large. A noticeable improvement in the variability of the tagged cell stream is observed when X_{min} increases, or the CDVT parameter τ decreases. In fact, when $X_{min} = 4$ at each of the PORE servers, C_A^2 is always smaller than 0.1. The cumulative mean queueing delay for the tagged stream is plotted as a

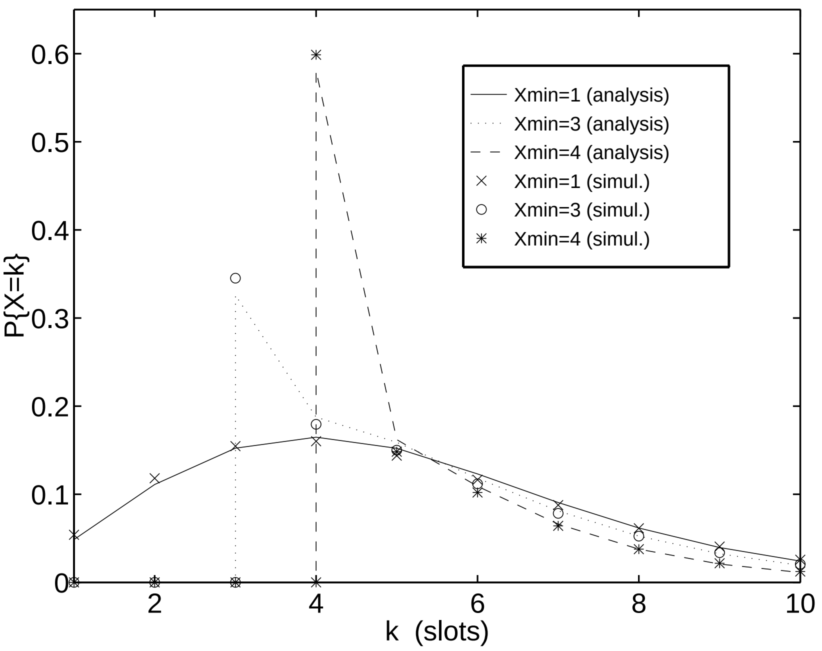

Figure 11: Cell inter-departure time PMFs at node 5 for different X_{min}.

function of the node index in Fig. 13. The total mean delay appears to increase linearly with the node index, and the rate of increase is clearly dependent upon the value of X_{min}. These results indicate that for real-time applications the tradeoff between violating deadline constraints and violating either delay jitter constraints or a declared peak rate, needs to be considered carefully.

Note that while delay distributions for the tagged stream are computed at each multiplexing stage, it is not yet clear how the end-to-end delay distribution can be computed numerically. By assuming that the cell delay process is independent from one node to the next, a simple convolution of individual delay PMFs would yield the approximate end-to-end delay distribution. Note that this assumption is not necessary for the computation of first moments (Fig. 13), where the mean end-to-end delay is simply the sum of individual mean delays. Although results are not presented here, the above approximation has been examined by comparing numerical and simulation results. It seems that under a network of FCFS servers, the assumption of inter-nodal independence for the delay process yields quite accurate results. However, when peak-rate enforcement is employed at each multiplexing stage, the inter-nodal independence assumption breaks down and the convolution of delay PMFs proves to be a poor approximation for the

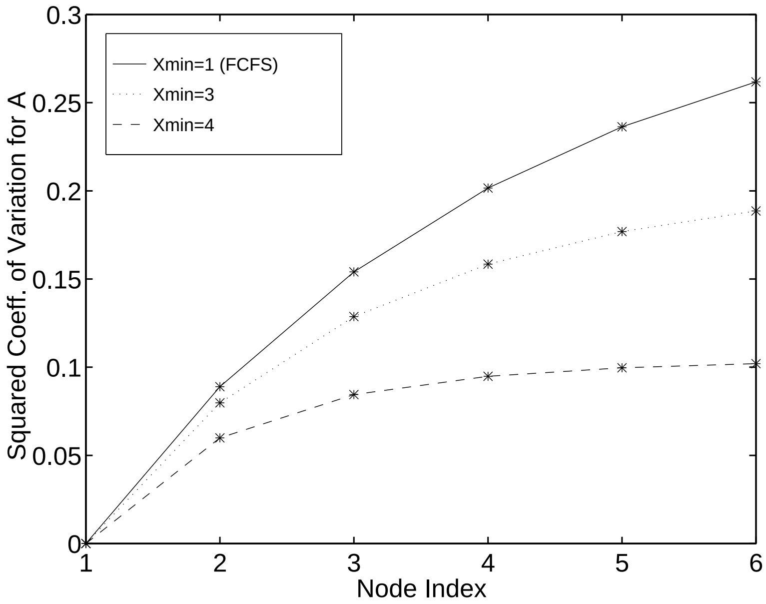

Figure 12: Squared coefficient of variation for the cell inter-arrival process to each node.

end-to-end delay distribution.

Finally, it should be noted that the main contribution of this work is to study the impact of re-regulating on a traffic stream, analyze the PORE policy, compare this policy's impact to that under the FCFS policy and present results for a perfectly regulated traffic (periodic traffic) to make the impact of re-regulation easier to interpret. This work is directly applicable to the case in which the peak rate of one traffic stream is being enforced; this would be the case when one jitter sensitive traffic stream is multiplexed with jitter insensitive traffic. If a small number of jitter sensitive traffic streams are present (or simply, peak rate is to be enforced in a small number of traffic streams), some scheduling conflicts may arise. Due to these conflicts, a greater than desirable spacing may be imposed sometimes, potentially adding to the positive jitter (more spreading), without affecting the enforced peak rate (minimum spacing). This work will still be useful in such an environment. When a large number of tagged streams are present, conflicts may dominate and although a peak rate can still be enforced, the positive jitter and the delay may be increased significantly. In this case, the relative shifting of the periods (if the streams are periodic) is expected to have a significant impact [23].

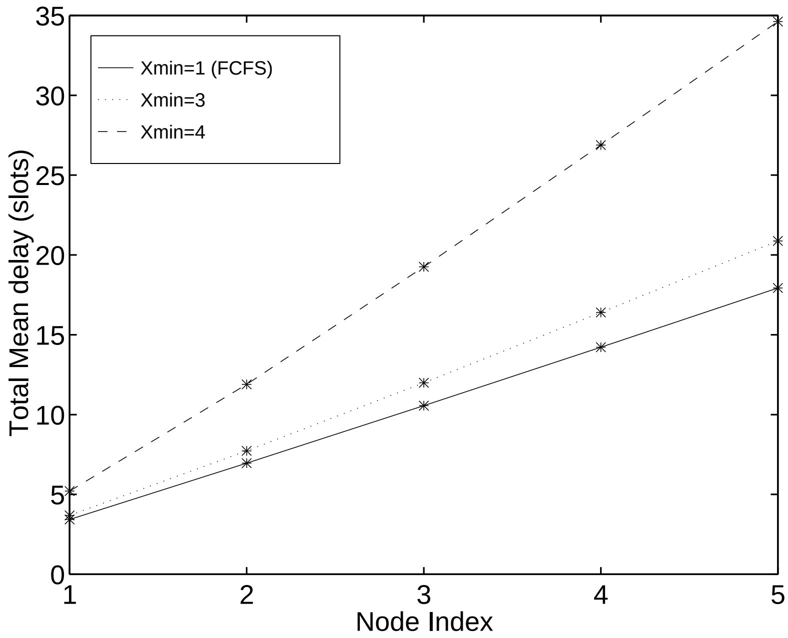

Figure 13: Total mean queueing delay as a function of node index.

5 Conclusions

In this paper, a queueing system was formulated to study the distortion induced in a tagged cell traffic stream; both FCFS and Peak Output Rate Enforcing (PORE) service disciplines were considered. In fact, the FCFS policy was shown to be a special case of the PORE discipline. At the output of the delay jitter-reducing service strategy, the tagged cell stream is easily characterized in accordance to the GCRA specified in [19]. A discrete-time analysis in the M/G/1 paradigm yielded numerical results for cell delay, delay jitter and inter-departure time probability distributions at a multiplexer served according to the PORE policy.

An approximate end-to-end performance study of the tagged cell stream was also carried out by analyzing a tandem queueing network of PORE servers. Numerical results indicate that a peak rate enforced at the network edge can be severely violated within the network and, thus, traffic re-regulation may be necessary. As intended, the traffic profile of a cell stream conforming to GCRA(T, τ) was shown to be much more regular, or less variable, when the CDVT parameter τ is small (close to zero). The tradeoff for enforcing small τ was also clearly illustrated through numerical results in the form of increased cell delays.

Simulation results for the end-to-end tandem queueing study indicate that the assumption of an i.i.d. inter-arrival process, with a PMF computed through the numerical approach, does not significantly affect the accuracy of the solution. The good agreement between simulation and numerical results was observed for a tandem network of FCFS servers as well as PORE servers. It is not yet clear how this numerical accuracy will be affected when a network of queues with cross-correlated traffic and different routing realizations is simulated.

A Solution of the stationary probability vector for P

This appendix outlines the steps involved in computing the stationary probability vector for $\mathbf{P}$, given by $\boldsymbol{\pi} = [\boldsymbol{\pi}_0\ \boldsymbol{\pi}_1\ \boldsymbol{\pi}_2\ \cdots]$. The vectors $\boldsymbol{\pi}_i$, $i \geq 0$, are the stationary probability vectors associated with level i as defined in section 2.4. The boundary probability vector $\boldsymbol{\pi}_0$ is evaluated first based on the matrix analytic techniques of Neuts [22]. The vectors $\boldsymbol{\pi}_i$, $i \geq 1$, are then recursively computed from $\boldsymbol{\pi}_0$.

Computation of $\boldsymbol{\pi}_0$:

A brief outline of the steps involved in computing $\boldsymbol{\pi}_0$ is presented here; a detailed description of the method can be found in [22]. To apply the method the following matrices need to be defined:

$$\mathbf{A} = \sum_{n=0}^{\infty} \mathbf{A}_n \tag{18}$$

$$\mathbf{G}(z) = z \sum_{n=0}^{\infty} \mathbf{A}_n \mathbf{G}^n(z) \quad , |z| \leq 1 \tag{19}$$

$$\mathbf{K}(z) = z \sum_{n=0}^{\infty} \mathbf{B}_n \mathbf{G}^n(z) \quad , |z| \leq 1 \tag{20}$$

The computation of matrix $\mathbf{G} \triangleq \mathbf{G}(z)|_{z=1}$, which satisfies the nonlinear matrix equation

$$\mathbf{G} = \sum_{n=0}^{\infty} \mathbf{A}_n \mathbf{G}^n, \tag{21}$$

is a key step in the application of the technique. Let $\mathbf{g}$ denote the stationary probability vector associated with $\mathbf{G}$. That is, $\mathbf{g}$ is the solution to $\mathbf{gG} = \mathbf{g}$ and $\mathbf{ge} = 1$, where $\mathbf{e}$ denotes the unit column vector.

Matrix $\mathbf{G}$ can be obtained as the limit of the monotonically increasing, convergent matrix sequence $\{\mathbf{G}_k\}_{k\geq 0}$, defined by

$$\mathbf{G}_0 = 0, \quad \mathbf{G}_{k+1} = \sum_{n=0}^{\infty} \mathbf{A}_n \mathbf{G}_k^n, \quad k \geq 0. \tag{22}$$

Matrix $\mathbf{G}$ is then used to determine matrix $\mathbf{K}$, defined by $\mathbf{K} = \mathbf{K}(1)$, and given by

$$\mathbf{K} = \sum_{n=0}^{\infty} \mathbf{B}_n \mathbf{G}^n, \tag{23}$$

with stationary probability vector $\boldsymbol{\kappa}$.

Let $\boldsymbol{\kappa}^*$ be the vector defined as

$$\boldsymbol{\kappa}^* = \left.\frac{d\mathbf{K}(z)}{dz}\right|_{z=1} \mathbf{e}.$$

The expression for $\boldsymbol{\kappa}^*$ is obtained by differentiating (20) at $z = 1$:

$$\begin{aligned} \boldsymbol{\kappa}^* &= \left.\frac{d\mathbf{K}(z)}{dz}\right|_{z=1} \mathbf{e} \\ &= \sum_{n=0}^{\infty} \mathbf{B}_n \mathbf{G}^n \mathbf{e} + \sum_{n=1}^{\infty} \mathbf{B}_n \sum_{k=0}^{n-1} \mathbf{G}^k \left.\frac{d\mathbf{G}(z)}{dz}\right|_{z=1} \mathbf{G}^{n-1-k} \mathbf{e} \\ &= \mathbf{K}\mathbf{e} + \sum_{n=1}^{\infty} \mathbf{B}_n \sum_{k=0}^{n-1} \mathbf{G}^k \boldsymbol{\mu}, \end{aligned} \tag{24}$$

where $\boldsymbol{\mu} = \left.\frac{d\mathbf{G}(z)}{dz}\right|_{z=1} \mathbf{e}$ and is given by

$$\boldsymbol{\mu} = (\mathbf{I} - \mathbf{G} + \mathbf{e}\mathbf{g}) \left(\mathbf{I} - \mathbf{A} + (\mathbf{e} - \boldsymbol{\beta})\mathbf{g}\right)^{-1} \mathbf{e}. \tag{25}$$

The vector $\boldsymbol{\beta}$ is given by

$$\boldsymbol{\beta} = \sum_{n=1}^{\infty} n\mathbf{A}_n \mathbf{e}.$$

Finally, the stationary probability vector $\boldsymbol{\pi}_0$ can be expressed in terms of the vectors $\boldsymbol{\kappa}$ and $\boldsymbol{\kappa}^*$ as

$$\boldsymbol{\pi}_0 = \frac{\boldsymbol{\kappa}}{\boldsymbol{\kappa}\boldsymbol{\kappa}^*}. \tag{26}$$

Computation of $\boldsymbol{\pi}_i$, $i \geq 1$:

The vectors $\boldsymbol{\pi}_i$, for $i \geq 1$, can now be computed using the recursive scheme attributed to Ramaswami [24],

$$\boldsymbol{\pi}_i = \left[\boldsymbol{\pi}_0 \mathbf{B}_i^* + \sum_{j=1}^{i-1} \boldsymbol{\pi}_j \mathbf{A}_{i+1-j}^*\right] (\mathbf{I} - \mathbf{A}_1^*)^{-1}, \quad i \geq 1, \tag{27}$$

where

$$\mathbf{B}_n^* = \sum_{k=n}^{\infty} \mathbf{B}_k \mathbf{G}^{k-n}, \quad \mathbf{A}_n^* = \sum_{k=n}^{\infty} \mathbf{A}_k \mathbf{G}^{k-n}, \quad n \geq 0.$$

References

[1] R. Onvural. *Asynchronous Transfer Mode: Performance Issues.* Artech House, 1993.

[2] I. Stavrakakis. Efficient modeling of merging and splitting processes in large networking structures. *IEEE J. Select. Areas Commun.*, 9(8), Oct. 1991.

[3] W. Matragi, C. Bisdikian, and K. Sohraby. Jitter calculus in atm networks: Part i, single node. In *IEEE INFOCOM'94*, Toronto, Canada, 1994.

[4] J. Roberts and F. Guillemin. Jitter in atm networks and its impact on peak rate enforcement. *Perf. Eval.*, 16(1–3), Nov. 1992.

[5] R. Cruz. A calculus for network delay, part i: Network elements in isolation. *IEEE Transactions on Information Theory*, 37(1), Jan. 1991.

[6] R. Cruz. A calculus for network delay, part ii: Network analysis. *IEEE Transactions on Information Theory*, 37(1), Jan. 1991.

[7] A. Parekh and R. Gallager. A generalized processor sharing approach to flow control in integrated services networks: The single-node case. *IEEE/ACM Trans. Networking*, 1(3), June 1993.

[8] A. Parekh and R. Gallager. A generalized processor sharing approach to flow control in integrated services networks: The multiple node case. *IEEE/ACM Trans. Networking*, 1(3), June 1993.

[9] D. Towsley. Providing quality of service in packet switched networks. *Performance Evaluation of Computer and Communication Systems*, 1993. Joint Tutorial Papers of Performance'93 and Sigmetrics'93.

[10] J. Kurose. On computing per-session performance bounds in high-speed multi-hop computer networks. *ACM SIGMETRICS'92*, June 1992.

[11] O. Yaron and M. Sidi. Calculating performance bounds in communication networks. In *IEEE INFOCOM'93*, San Francisco, CA, 1993.

[12] R. Guerin et al. Equivalent capacity and its application to bandwidth allocation in high-speed networks. *IEEE J. Select. Areas Commun.*, 9(7), Sept. 1991.

[13] D. Ferrari and D. Verma. A scheme for real-time channel establishment in wide-area networks. *IEEE J. Select. Areas Commun.*, 8(3), April 1990.

[14] J. Hyman et al. Real-time scheduling with quality of service constraints. *IEEE J. Select. Areas Commun.*, 9(7), Sept. 1991.

[15] W. Matragi, K. Sohraby, and C. Bisdikian. Jitter calculus in atm networks: Part ii, multiple nodes. In *IEEE INFOCOM'94*, Toronto, Canada, 1994.

[16] H. Kroner, M. Eberspacher, T. Theimer, P. Khun, and U. Briem. Approximate analysis of end-to-end delay in atm networks. In *IEEE INFOCOM'92*, Florence, Italy, 1992.

[17] J.F. Ren, J. Mark, and J. Wong. End-to-end performance in atm networks. In *ICC'94*, New Orleans, LA, 1994.

[18] Y. Ohba, M. Murata, and H. Miyahara. Analysis of interdeparture processes for bursty traffic in atm networks. *IEEE J. Select. Areas Commun.*, 9(3), April 1991.

[19] ATM Forum. *ATM User-Network Interface Specification.* Version 3.0, 1993.

[20] C. Bisdikian, W. Matragi, and K. Sohraby. A study of the jitter in atm multiplexers. In *Proceedings of the Fifth International Conference on Data Communication Systems and their Performance*, Raleigh, NC, 1993.

[21] F. Guillemin and W. Monin. Management of cell delay variation in atm networks. In *IEEE GLOBECOM'92*, Orlando, FL, 1992.

[22] M.F. Neuts. *Structured Stochastic Matrices of the M/G/1 Type and their Applications.* Marcel Dekker, 1989.

[23] R. Landry and I. Stavrakakis. Non-deterministic periodic packet streams and their impact on a finite-capacity multiplexer. In *IEEE INFOCOM'94*, Toronto, Canada, 1994.

[24] V. Ramaswami. A stable recursion for the steady state vector in markov chains of the m/g/1 type. *Commun. Statist. - Stochastic Models*, 4(1), 1988.