# Performance Analysis of an ATM MUX with a New Space Priority Mechanism under ON-OFF Arrival Processes

Jongho Bang, Nirwan Ansari, and Sirin Tekinay

*Abstract:* **We propose a new space priority mechanism, and analyze its performance in a single Constant Bit Rate (CBR) server. The arrival process is derived from the superposition of two types of traffics, each in turn results from the superposition of homogeneous ON-OFF sources that can be approximated by means of a two-state Markov Modulated Poisson Process (MMPP). The buffer mechanism enables the Asynchronous Transfer Mode (ATM) layer to adapt the quality of the cell transfer to the Quality of Service (QoS) requirements and to improve the utilization of network resources. This is achieved by "Selective-Delaying and Pushing-In" (SDPI) cells according to the class they belong to. The scheme is applicable to schedule delay-tolerant non-real time traffic and delay-sensitive real time traffic. Analytical expressions for various performance parameters and numerical results are obtained. Simulation results in term of cell loss probability conform with our numerical analysis.**

*Index Terms:* **ATM, buffer management, priority mechanism, SDPI.**

## I. INTRODUCTION

Asynchronous Transfer Mode (ATM) networks provide a great variety of services with widely differing bandwidth and quality of service (QoS) requirements. The major characteristics of an ATM-based Broadband Integrated Service Digital Network (BISDN) include: high flexibility of network access, dynamic bandwidth allocation on demand with a fine degree of granularity, flexible bearer capacity allocation, and independence of the means of transmission at the physical layer. However, diverse traffic types, and hence different QoS requirements make traffic control of ATM networks an essential and critical challenge. ATM provides the cell transfer for all services, and the ATM adaptation layer (AAL), sitting on top of the ATM layer, provides service-dependent functions to higher layers. Much research has been concerned with the problem of effectively adapting the quality of the ATM bearer service to the diverse user QoS requirements. If all services are treated similarly, dimensioning of the ATM network would have to employ the QoS requirement for the most demanding service, thus limiting efficiency. Moreover, providing a single grade of bearer service not only limits the utilization of network resources, but also leads to a lack of flexibility in accommodating the QoS requirements of future services [1]. The incorporation of two bearer services with different levels of cell loss probability QoS requirements has been proposed for ATM networks. The low priority traffic, which has a less stringent cell loss probability constraint than the high priority traffic, can be accommodated in the network at an efficient resource utilization level.

Several special mechanisms for buffer access have been proposed. They have been used to adapt the cell loss probability of a given class of traffic to the restriction of the QoS needs of the corresponding service. These mechanisms allow a selective access to the buffer depending on the traffic class. In [2]–[6], the authors proposed a mechanism, called Push-Out, which guarantees the buffer access to a certain class of traffic if the queue is not full, and when it is full, the arriving cell can replace one with a lower priority. The selection of the lowest priority cell to be rejected is done according to the chosen replacement algorithm. Other proposed mechanisms have lower performance but simpler buffer management, called Partial Buffer Sharing [7]–[11], which guarantees the buffer access to a class $i$ cell if the buffer occupancy is less than a threshold, say, $S_i$. In general, these schemes are more flexible and more protective of high priority cells. However, this performance gain is always achieved only at the cost of a significant performance degradation of low priority cells.

The higher bandwidth promised by BISDN have made applications with real-time constraints possible, such as control, command, and interactive voice and video communications. Excessive delay renders real-time traffic useless, but a certain degree of loss can be tolerated without objectionable degradation in the grade of service. Real-time packets are lost for several reasons. The packet may arrive at the receiver after the end-to-end deadline has expired after having suffered excessive waiting times in intermediate nodes. Also, intermediate nodes may shed load by dropping packets as an overload control measure. It is natural to engineer communication networks that support real-time traffic, so that delays are bounded at the expense of some loss. However, the magnitude of this loss determines the quality of service and, hence, it is critical to predict this loss accurately in order to provide an acceptable grade of service. Given the fixed length packets and First-Come First-Serve (FCFS) principle at a multiplexer, imposing a buffer size of $K$ is essentially equivalent to imposing a time constraint of $Kd$, where $d$ is the fixed transmission time of a packet. A broadband network has to guarantee end-to-end delay. The network, in order to meet the delay requirements, forces each node to bound its maximum cell delay.

Our simple consideration suggests that the traffic can be cat-

Manuscript received July 16, 2001; approved for publication by Hussein Mouftah, Division I Editor, November 7, 2001.

The authors are with New Jersey Center for Wireless Telecommunications, Department of Electrical and Computer Engineering, New Jersey Institute of Technology University, Heights Newark, NJ 07102, USA.

egorized into two basic classes: real time traffic (RTT) and non-real time traffic (NRTT). Our model is based on the partial buffer sharing scheme. The buffer is partitioned by a threshold, set according to the maximum cell delay of the real time traffic. In order to compensate for the disadvantage of the partial buffer sharing scheme, we can give priority to the real time traffic over non-real time traffic selectively. We call such a proposed scheme, Selective-Delay Push-In (SDPI). In this paper, we make a thorough study of the proposed space priority mechanism for the case of bursty traffic. The bursty source is modeled by the Markov Modulated Poisson Process (MMPP), because it is analytically tractable and possesses properties suitable for the approximation of complicated non-renewal processes. The rest of the paper is organized as follows. Section II describes the modeling and analysis of the space priority mechanism; Section III presents performance results; finally, some conclusions are drawn in Section IV.

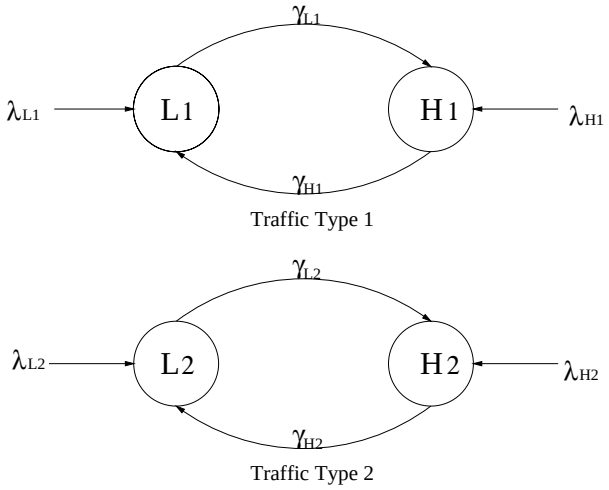


Fig. 1. 2-state MMPP models for traffic type 1 (real time traffic) and traffic type 2 (non-real time traffic).

## II. THE SPACE PRIORITY MECHANISM

We shall first discribe the source model, and then the SDPI mechanism, followed by the analysis.

### *A. The Source Model*

The MMPP has been extensively used for modeling arrival rates of point processes because it can qualitatively model the time-varying arrival rate, capture some of the important correlations between the interarrival times, and is analytically tractable. The accuracy of MMPP in modeling an arrival process depends on which statistics of the actual process are used to determine its parameters. 2-state MMPP models [12]–[15] and 4-state MMPP models [16] have been used to approximate the superposition of ON-OFF sources. In [17], the superposition of ON-OFF sources is approximated by means of a 2-state MMPP using the Average Matching Technique. This technique provides good accuracy as compared to simulation results. In particular, the method weakly depends on the number of sources.

Consider the superposition of $N$ independent and homogeneous sources, each characterized by: 1) the peak bit rate, $F_p$; 2) the activity factor, $p$; 3) the mean burst length, $L_B$. With reference to the ATM MUX, denote C as the net output capacity, and thus $M = \lfloor C/F_p \rfloor$ indicates the maximum number of sources that can be accommodated in the MUX, assuming a peak bandwidth assignment. The superposition of $N$ such sources results in a birth-death process. The states of this process are divided into two subsets [14]: 1) an overload (OL) region, comprising the states $M$+1,$\cdots$, $N$, where the cell emission rate exceeds the capacity C; 2) an underload (UL) region, consisting of the remaining states 0, $\cdots$, $M$. Therefore, the two states of the approximated MMPP can be chosen so that one of them, called OL state, corresponding to the OL region, and the other, called UL state, associated with the UL region. Let $\pi_j$ be the limiting probability that the number of active sources is $j$. Then $\pi_j$ is given by the binomial distribution.

$$\pi_j = \binom{N}{j} p^j (1-p)^{N-j},$$

where $p$ is the activity factor of a source. Using the average matching procedure, the expression for the four parameters characterizing the MMPP can be determined.

We can adopt this Average Matching Technique for the superposition of independent heterogeneous ON-OFF sources, consisting of the real time traffic and non-real time traffic. In our case, the finite capacity can be shared by two kinds of traffics. A threshold is defined to separate the two states (Low and High) for each class of traffic. Let $N_1$ be the set of the real time traffic with peak bit rate, $F_p(1)$, and $N_2$ be the set of the non-real time traffic with peak bit rate, $F_p(2)$. $M_1$ denotes the threshold which distinguishes the two states (low and high load) for the real time traffic, and similarly, $M_2$ denotes the threshold which distinguishes the two states (low and high load) for the non-real time traffic.

$$M_1 = \Big\lfloor \frac{N_1 C}{N_1 F_p(1) + N_2 F_p(2)} \Big\rfloor, \tag{1}$$

$$M_2 = \Big\lfloor \frac{N_2 C}{N_1 F_p(1) + N_2 F_p(2)} \Big\rfloor. \tag{2}$$

Thus, each traffic can be divided into two states. That is,

- For real time traffic
  low load region (Low(1)): [0, 1, $\cdots$, $M_1$]
  high load region (High(1)): [$M_1$+1,$\cdots$, $N_1$]
- For non-real time traffic
  low load region (Low(2)): [0, 1, $\cdots$, $M_2$]
  high load region (High(2)): [$M_2+1$,$\cdots$, $N_2$]

Four parameters are required to represent the 2-state MMPP source of each traffic, as shown in Fig. 1, where $\gamma_{L1}(\gamma_{H1})$ is defined as the mean transition rate out of the Low load 1 (High load 1) state, and $\lambda_{L1}(\lambda_{H1})$ is the mean arrival rate of the Poisson process in the Low load 1 (High load 1) state for the real time traffic, respectively. Similarly, $\gamma_{L2}(\gamma_{H2})$ is defined as the mean transition rate out of the Low load 2 (High load 2) state, and $\lambda_{L2}(\lambda_{H2})$ is the mean arrival rate of the Poisson process in

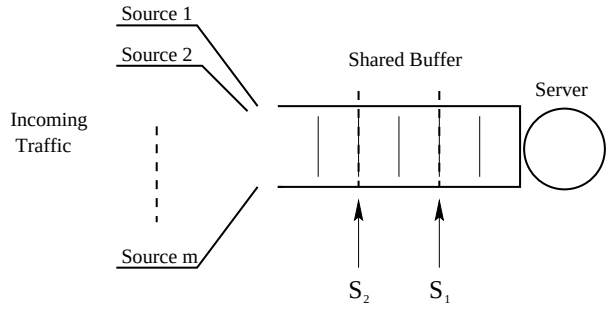


Fig. 2. The threshold-based discard (TBD) scheme operation.

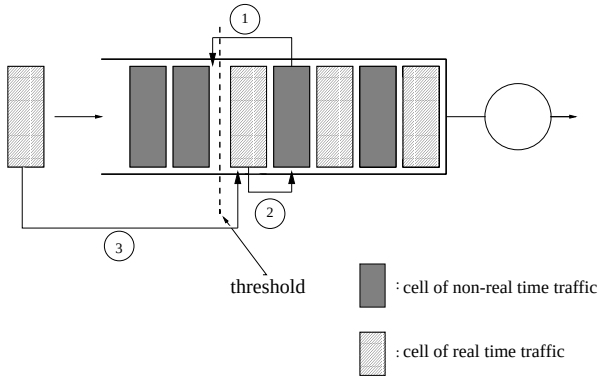


Fig. 3. Selective-delay push-in scheme operation.

the Low load 2 (High load 2) state for the non-real time traffic, respectively.

### *B. The SDPI Mechanism*

First, we consider the threshold-based discarding (TBD) scheme, which is called the partial buffer sharing scheme. Priority cell discarding is a popular congestion control technique in high-speed networks that allows network resources to be used more efficiently, thereby making it easier to satisfy QoS requirements of different classes of traffics. As shown in Fig. 2, the buffer is partitioned by $n$ thresholds, $S_1, \cdots, S_n$, corresponding to $n$ priority classes, where $S_n$ is the buffer size.

Priority class $i$ cells can be buffered up to threshold level $S_i$. Once the buffer level exceeds $S_i$, arriving class $i$ cells are dropped. Note that only new arrivals are dropped; class $i$ cells that are already in the buffer are never dropped and are eventually served. In the case that two kinds of traffics (i.e., real time and non-real time traffic) are considered, the non-real time traffic such as data is allowed to access more buffer space than the real time traffic such as voice and video because of the delay limitation of the real time traffic in this scheme. It is assumed in this paper that the buffer size and the threshold are decided according to the QoS requirement of the non-real time traffic (i.e., cell loss probability) and the QoS requirement of the real time traffic (i.e., maximum cell delay), respectively. Thus, real time traffic cells are dropped from a buffer when the buffer level exceeds the threshold, decided according to its maximum cell delay.

Second, we modify the TBD scheme by giving priority to the real time traffic over the non-real time traffic selectively, and thus called selective-delay push-in (SDPI) scheme. With this scheme, non-real time traffic cells can be delayed in favor for real time traffic cells. As illustrated in Fig. 3, when the buffer level is less than the threshold, the SDPI scheme operates just like the TBD scheme. However, when the buffer level is above the threshold, if there exist non-real time traffic cells within the threshold, an arriving real time traffic cell pushes out the latest arrived non-real time traffic cell and positions itself at the end of the buffer within the threshold. At this moment, the expelled non-real time traffic cell buffers up right after the threshold. If no non-real time traffic cell is within the threshold, an arriving real time traffic cell is discarded. When the buffer is full, arriving real time or non-real time traffic cells are just discarded. The threshold is set according to the maximum cell delay of the real time traffic to satisfy its delay requirement, just like the TBD scheme. When the buffer level is above the threshold, if there exist non-real time traffic cells within the threshold, an arriving real time traffic cell is survived in the SDPI scheme, but not in the TBD scheme.

### *C. The SDPI Analysis*

The multiplexer is modeled as a finite capacity single server queue where the arrival process is MMPP, and the service is deterministic. In our analysis, we make the similar assumptions as in [16], which deals with the analysis of only one traffic type, that significantly reduce the computational complexity involved in obtaining the stationary distributions at departure points: 1) the probability that the MMPP goes through multiple state transitions between successive departures is negligible, and 2) the state transitions occur at departure epochs, i.e., if a departure leaves the MMPP in state $i$, the cell arrival rate until the next departure is $\lambda_i$. Consider a queue using SDPI where the MMPP consists of $K$ states denoted by $i$ ($0 \leq i \leq K$-1), and the arrival rates and mean state durations are denoted by $\lambda_i$ and $\mu_i$, respectively. The characteristics of this system will be determined using an imbedded Markov chain approach. As in the ordinary M/G/1 queueing system, the service completion instants are the imbedded points of the underlying Markov chain. Therefore, a probability vector $\mathbf{\Pi}$ consists of $\pi_i(n_1, n_2)$ ($0 \leq n_1 \leq S_1, 0 \leq n_2 \leq S_2$, where $S_2$ is the buffer size) which is defined by the probability that a departing cell leaves $n_1$ real time traffic cells and $n_2$ non-real time traffic cells in the system while the MMPP is in state $i$. The total transition probability matrix of the imbedded Markov chain, denoted by $\mathbf{Q}$, is formed with $K$ MMPP finite states and $F$ finite buffer states. For example, consider the traffic shown in Fig. 1, where the real time traffic and non-real time traffic can be aggregated resulting in a 4-state MMPP process (in this case, $K$=4). The $K$=4 states are $\{(L_1,L_2), (L_1,H_2), (H_1,L_2), (H_1,H_2)\}$. For a buffer with $S_1$=3 and $S_2$=6, there are $F$=22 finite buffer states corresponding to $\{\{n_1, n_2\} \mid n_1 + n_2 \leq 6 \text{ and } n_1 \leq 3\}$. Thus,

$$\mathbf{Q} = \begin{bmatrix} Q_{0,0} & Q_{0,1} & \cdots & Q_{0,K-1} \\ Q_{1,0} & Q_{1,1} & \cdots & Q_{1,K-1} \\ \vdots & \vdots & \cdots & \vdots \\ Q_{K-1,0} & Q_{K-1,1} & \cdots & Q_{K-1,K-1} \end{bmatrix}, \tag{3}$$

where $Q_{j,i}$ is a submatrix, and each element of the submatrix,

$Q_{j,i}((n_1,n_2),\ (n_1',n_2'))$ ($0 \le j,i \le K$-1, $0 \le n_1,n_1' \le S_1$, $0 \le n_2,n_2' \le S_2$) corresponds to a state transition probability. That is,

$$Q_{j,i}((n_1,n_2),\ (n_1',n_2')) = P\{(n_1',n_2'),j \mid (n_1,n_2),i\},$$

where $i$ is the present MMPP state, $j$ is the next MMPP state, $(n_1,n_2)$ is the present buffer state, and $(n_1',n_2')$ is the next buffer state. The submatrix $Q_{j,i}$ can be obtained as follows. Denote $A_i$ as the buffer state transition probability matrix of the departure point of our system at MMPP state $i$ (with arrival rate $\lambda_i$ and service time $\Delta t$). The transition probability submatrix $Q_{j,i}$ can be simply obtained by multiplying $A_i$ by the probability that the MMPP will not change its state in $\Delta t$ if $j = i$, or by the probability that the MMPP will change its state from $j$ to $i$ in $\Delta t$ if $j \ne i$. Define $q_i(k,l)$ as the transition probability that $k$ real time traffic cells and $l$ non-real time traffic cells can be positioned in the buffer during the service time ($\Delta t$) while the MMPP is in state $i$. Denote $q_i^1(k)$ as the probability of $k$ arrivals of traffic type 1 (i.e., real time traffic) and $q_i^2(l)$ as the probability of $l$ arrivals of traffic type 2 (i.e., non-real time traffic) during the service time, respectively. Define $q_i^*(k,l)$ as the transition probability that more than $k$ real time traffic cells and more than $l$ non-real time traffic cells are inserted to the buffer, but only $k$ real time traffic cells and only $l$ non-real time traffic cells can be positioned in the buffer during the service time ($\Delta t$) due to the SDPI mechanism. Thus,

$$q_i(k,l) = q_i^1(k)q_i^2(l),$$

where

$$q_i^1(k) = \frac{(\lambda_i^1\Delta t)^k}{k!}e^{-(\lambda_i^1\Delta t)},$$

$$q_i^2(l) = \frac{(\lambda_i^2\Delta t)^l}{l!}e^{-(\lambda_i^2\Delta t)},$$

and $\lambda_i^1$, $\lambda_i^2$ are the arrival rates for traffic type 1 and 2, respectively, and $\lambda_i$=$\lambda_i^1$+$\lambda_i^2$.

Since at most one cell is served between successive imbedded points, transitions from $n_1$ to $n_1' < n_1 - 1$, from $n_2$ to $n_2' < n_2 - 1$, and from $n_1 + n_2$ to $n_1' + n_2' < n_1 + n_2 - 1$ are not possible.

*Transitions to* $n_1' + n_2' < S_2$ *and* $n_1' < S_1$:

$$q_i(k,l) = q_i^1(k)q_i^2(l) \tag{4}$$

*Transitions to boundaries*:

1. at $n_1' + n_2' < S_2$ and $n_1' = S_1$,

$$q_i^*(k,l) = \sum_{n=k}^{\infty} q_i^1(n)q_i^2(l). \tag{5}$$

2. at $n_1' + n_2' = S_2$ and $n_1' = S_1$,

$$q_i^*(k,l) = \sum_{n=k}^{\infty} q_i^1(n)q_i^2(l) + \sum_{n=k}^{\infty}\sum_{m=l+1}^{\infty} q_i^1(n)q_i^2(m)\frac{\binom{n}{k}\binom{m}{l}}{\binom{n+m}{n}}. \tag{6}$$

3. at $n_1' + n_2' = S_2$ and $n_1' < S_1$,

$$q_i^*(k,l) = \sum_{n=k}^{\infty}\sum_{m=l}^{\infty} q_i^1(n)q_i^2(m)\frac{\binom{n}{k}\binom{m}{l}}{\binom{n+m}{n}}. \tag{7}$$

The transition probability (4) denoted by $q_i(k,l)$ implies that exactly $k$ arrivals of traffic type 1 and exactly $l$ arrivals of traffic type 2 occur in any order during the service time. The transition probability (5) implies that more than $k$ arrivals of traffic type 1 and exactly $l$ arrivals of traffic type 2 occur in any order during the service time. Since the present state $n_1' = S_1$, even though there are more than $k$ arrivals of traffic type 1, only $k$ cells can be positioned in the buffer. According to the SDPI mechanism, an arriving cell is dropped when the buffer is full. Thus, the transition probability (6) consists of two terms. The first term represents that more than $k$ arrivals of traffic type 1 and exactly $l$ arrivals of traffic type 2 occur. The second term means that more than $k$ arrivals of traffic type 1 and more than $l$ arrivals of traffic type 2 occur. The fraction in the second term represents the probability that $k$ out of $n$ traffic type 1 and $l$ out of $m$ traffic type 2 are the first arrivals. The transition probability (7) represents that more than $k$ arrivals of traffic type 1 and more than $l$ arrivals of traffic type 2 occur, just like the second term of the probability (6).

Define the stationary probability vector $\mathbf{\Pi}$ as

$$\mathbf{\Pi} = \{\pi_0(0,0),...,\pi_0(S_1,S_2-S_1),\pi_1(0,0),...,\pi_1(S_1,S_2-S_1),...,\pi_{K-1}(0,0),...,\pi_{K-1}(S_1,S_2-S_1)\}.$$

Then, these stationary probabilities can be obtained as follows:

$$\mathbf{\Pi} = \mathbf{\Pi}\mathbf{Q}, \qquad \sum_{i=0}^{K-1}\sum_{n_1}\sum_{n_2}\pi_i(n_1,n_2) = 1.$$

To derive the loss probabilities, it is necessary to determine the probability distribution of the system length ($n_1$+$n_2$+1, including the server) from the arrival viewpoint, which is equivalent to the steady-state probability distribution $p_i(n_1,n_2)$ [18]. The probabilities must be different from the former departure-point probabilities $\pi_i(n_1,n_2)$, because the state space is enlarged by the state $G = S_2 + 1$, where the "1" accounts for the server. Asymptotically, the number of arriving ATM cells equals the number of departing cells. Hence, the departure rate must be equal to the effective arrival rate of ATM cells which are able to join the system.

$$\begin{aligned}\frac{1-p_i(0,0)}{\Delta t} &= \lambda_i^2\{1-\sum_{n_1+}\sum_{n_2=G}p_i(n_1,n_2)\}\\ &\quad+\lambda_i^1\{1-\sum_{n_2=0}^{S_2-S_1}p_i(S_1+1,n_2)\\ &\quad-\sum_{n_2=0}^{S_2-S_1}p_i(S_1,n_2+1)\frac{1}{S_1+1}\}\end{aligned} \tag{8}$$

where $p_i(n_1,n_2)$ is the steady state probability that an arriving cell sees $n_1$ real time traffic cells and $n_2$ non-real time traffic

cells in the system while the MMPP is in state $i$ (i.e., from an arrival point of view). $\frac{1}{S_1+1}$ is the probability that the non-real time traffic cell is being served, when $S_1+1$ cells (i.e., $S_1$ real time cells and 1 non-real time cell) are within the threshold including the server).

In general, the arrival point queue length distribution of a single server queue is identical to the departure point queue length distribution, given that arrivals and departures occur singly, i.e., $\pi_i(n_1,n_2)$ is the state probability seen by a cell who joins the queueing system [19], [21]. Therefore, the following equation holds for the state probabilities just after a departure.

The following steady-state probabilities can be obtained by combining (8) and (9)

The cell loss probabilities are then given as follows:

a) CLP for non-real time traffic (NRTT)

$$CLP_{NRTT} = \sum_{n_1+} \sum_{n_2=G} p(n_1,n_2). \tag{11}$$

b) CLP for real time traffic (RTT).

$$\begin{aligned} CLP_{RTT} = & \sum_{n_2=0}^{S_2-S_1-1} p(n_1 = S_1+1, n_2) \\ & + \sum_{n_2=0}^{S_2-S_1-1} p(n_1 = S_1, n_2+1)\frac{1}{S_1+1} \\ & + CLP_{NRTT}. \end{aligned} \tag{12}$$

## III. PERFORMANCE RESULTS

The performance of the SDPI scheme is evaluated for two kinds of traffics. We choose source parameters which are characterized by the peak bit rate $F_p$, the activity factor $p$, and the mean burst length $L_B$. Assume that the superposition of such heterogeneous ON-OFF sources are offered to an ATM MUX with the net output link capacity $C$. The performance of

---

$$\pi_i(n_1,n_2) = \begin{cases} \dfrac{p_i(n_1,n_2)}{1-\dfrac{\lambda_i^2}{\lambda_i}\displaystyle\sum_{n_1+}\sum_{n_2=G} p_i(n_1,n_2) - \dfrac{\lambda_i^1}{\lambda_i}\{\displaystyle\sum_{n_2=0}^{S_2-S_1} p_i(S_1+1,n_2) + \sum_{n_2=0}^{S_2-S_1} p_i(S_1,n_2+1)\dfrac{1}{S_1+1}\}}, \\ \quad \text{for } n_1+n_2 \le S_1 \text{ or } n_1 < S_1 \text{ and } n_1+n_2 \le S_2 \\ \dfrac{\dfrac{\lambda_i^2}{\lambda_i}p_i(n_1,n_2)}{1-\dfrac{\lambda_i^2}{\lambda_i}\displaystyle\sum_{n_1+}\sum_{n_2=G} p_i(n_1,n_2) - \dfrac{\lambda_i^1}{\lambda_i}\{\displaystyle\sum_{n_2=0}^{S_2-S_1} p_i(S_1+1,n_2) + \sum_{n_2=0}^{S_2-S_1} p_i(S_1,n_2+1)\dfrac{1}{S_1+1}\}}, \\ \quad \text{for } n_1 = S_1 \text{ and } n_1+n_2 \le S_2 \end{cases} \tag{9}$$

$$p_i(n_1,n_2) = \begin{cases} \dfrac{\pi_i(n_1,n_2)}{\pi_i(0,0)+\lambda_i\Delta t}, \quad for \quad n_1+n_2 \le S_1 \; or \; n_1 < S_1 \; and \; n_1+n_2 \le S_2 \\ \dfrac{\lambda_i}{\lambda_i^2}\dfrac{\pi_i(n_1,n_2)}{\pi_i(0,0)+\lambda_i\Delta t}, \quad for \quad n_1 = S_1 \; and \; n_1+n_2 \le S_2 \\ 1 - \displaystyle\sum_{\{n_1,n_2\}\in B_1}\sum \frac{\pi_i(n_1,n_2)}{\pi_i(0,0)+\lambda_i\Delta t} - \sum_{\{n_1,n_2\}\in B_2}\sum \frac{\lambda_i}{\lambda_i^2}\frac{\pi_i(n_1,n_2)}{\pi_i(0,0)+\lambda_i\Delta t}, \\ \quad for \;\; n_1+n_2 = G, \\ \quad where \; B_1 = \{n_1,n_2 \mid n_1+n_2 \le S_1 \; or \; n_1 < S_1 \; and \; n_1+n_2 \le S_2\}, \\ \qquad B_2 = \{n_1,n_2 \mid n_1 = S_1 \; and \; n_1+n_2 \le S_2\} \end{cases} \tag{10}$$

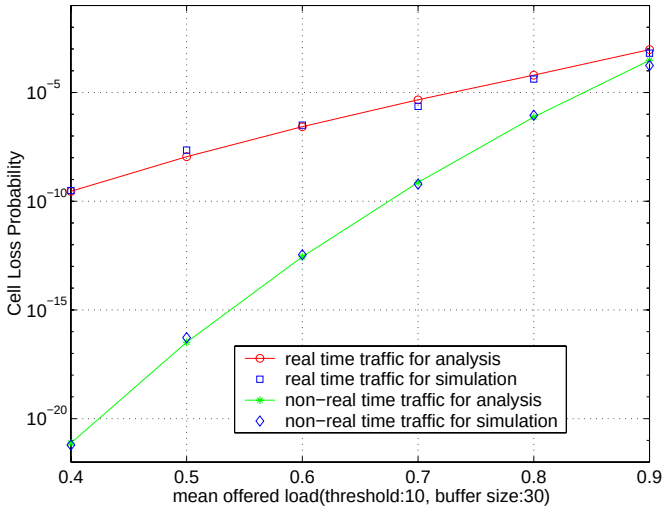


Fig. 4. Cell loss probability versus mean offered load (comparison between numerical and simulation results).

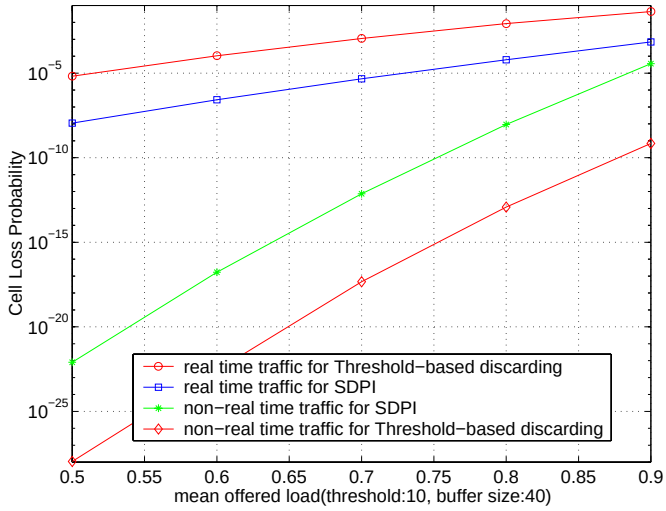


Fig. 5. Cell loss probability versus mean offered load (comparison between TBD and SDPI schemes).

Table 1. System parameters.

| class | $F_p$ | $p$ | $L_B$ |
|---|---|---|---|
| real time traffic | 32Kbps | 0.35 | 1400 |
| non-real time traffic | 128Kbps | 0.1 | 1600 |

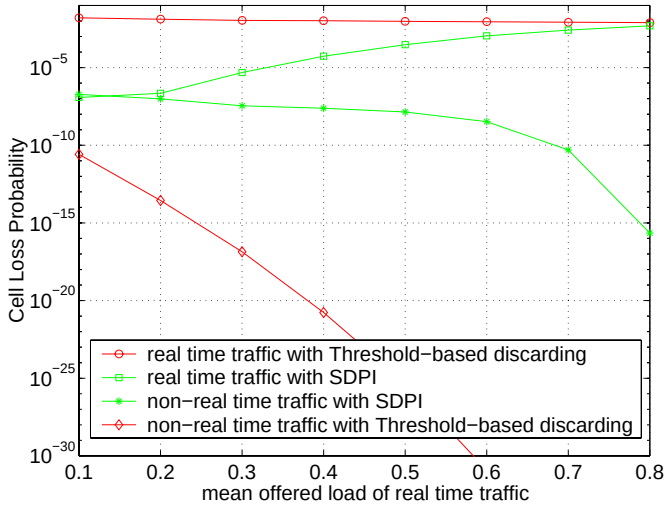


Fig. 6. Cell loss probability versus mean offered load of the real time traffic (comparison between TBD and SDPI schemes) (fixed total offered load=0.9, threshold=10, buffer size=40).

the MUX is evaluated by the queueing model with the MMPP source and the SDPI priority scheme. The constant service time of the MUX is given by $\theta$=53 bytes/$C$. The net link capacity is assumed to be 150Mbps.

Some simulation results are reported to evaluate the accuracy of the cell loss probability by using the SDPI scheme. The simulations have been performed on SUN SparcStation 60. The source parameters used in our simulations and numerical analysis, which are the same as in [22], are tabulated in Table 1. These source parameters are used for each user.

In Fig. 4, cell loss probabilities are plotted as functions of the mean offered load (real time traffic and non-real time traffic). Note that the simulation results are sufficiently reliable, since the 95% confidence intervals range within 10% of the estimated cell loss probability. The threshold and buffer size are assumed to be 10 and 30, respectively. Fig. 5 shows the comparison between the SDPI and TBD scheme. It is intuitive to see that SDPI achieves the performance improvement for the real time traffic (which is more critical) at the expense of the non-real time traffic. As we mentioned before, when the occupancy is above the threshold, if there exist non-real time traffic cells within the threshold, an arriving real time traffic cell is survived in the SDPI scheme, but not in the TBD scheme. At this point, we have the improvement for the real time traffic with the SDPI scheme; that is, the SDPI scheme compensates for the disadvantage of the real time traffic using the TBD scheme, under the circumstance that the threshold is fixed due to the maximum cell delay of the real time traffic.

Fig. 6 shows the cell loss probabilities as functions of the real time traffic offered load with a fixed total offered load at 0.9. Note the improvement for the real time traffic using SDPI, as compared to the TBD scheme, just like Fig. 5. As the real time traffic offered load increases, improvement for the real time traffic using SDPI diminishes. As the real time traffic increases and non-real time traffic decreases, the possibility that the non-real time traffic is within the threshold decreases and the possibility that arriving real time traffic cells are dropped increases when the buffer occupancy exceeds the threshold. In Fig. 7, the cell loss probabilities are plotted against the offered load of the non-real time traffic. The offered load of the real time traffic is fixed at 0.3. As the offered load of the non-real time traffic increases, the performance of the real time traffic using the SDPI scheme is improving, but the performance of the non-real time traffic is getting worsening, as compared to the TBD scheme, for the same reason as in Fig. 6.

In Fig. 8, cell loss probabilities are plotted as functions of the buffer size. As the buffer size increases while holding the threshold fixed, cell loss probabilities for the real time traffic remain constant, but cell loss probabilities for the non-real time traffic

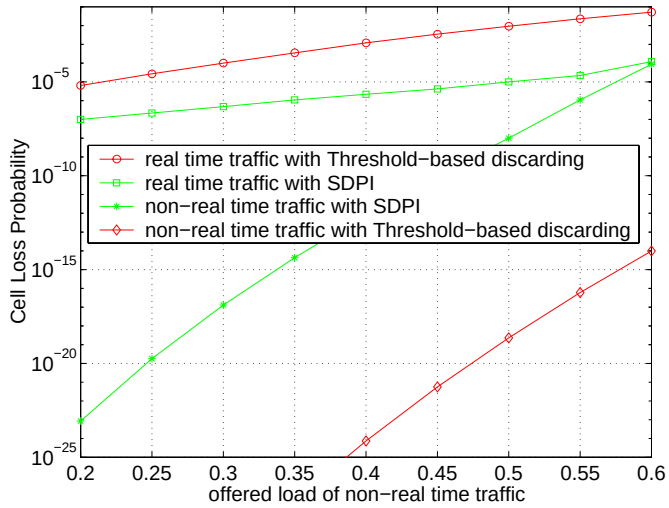


Fig. 7. Cell loss probability versus mean offered load of non-real time traffic (comparison between TBD and SDPI schemes) (offered load of the real time traffic is fixed at 0.3, threshold=10, buffer size=40).

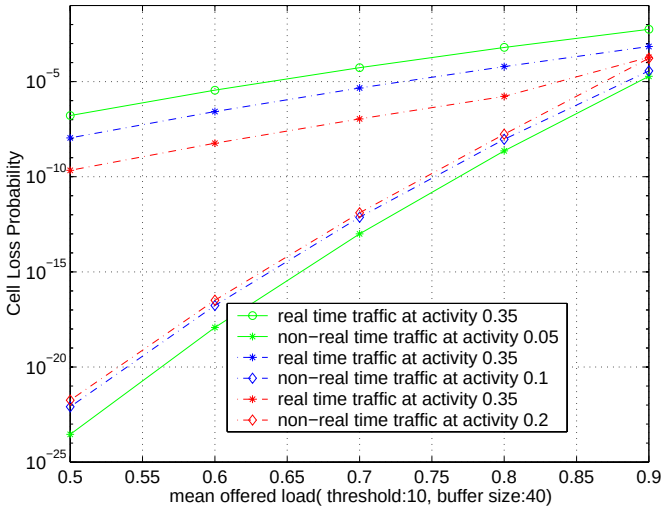


Fig. 9. Cell loss probability versus mean offered load (as activities of the non-real time traffic are varied).

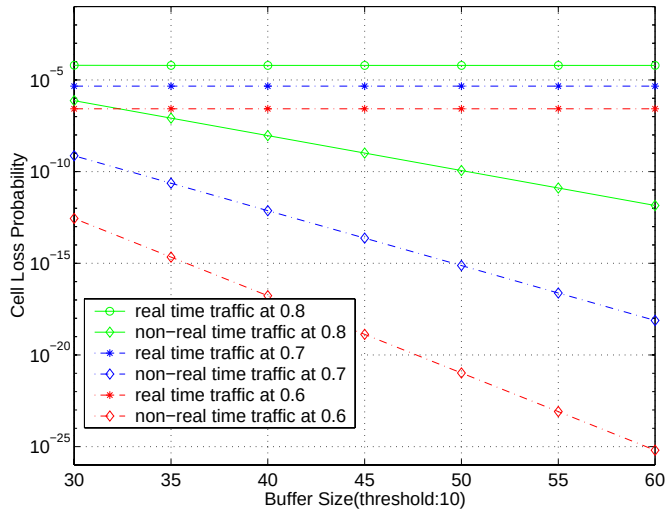


Fig. 8. Cell loss probability versus buffer size (as the mean offered load is varied).

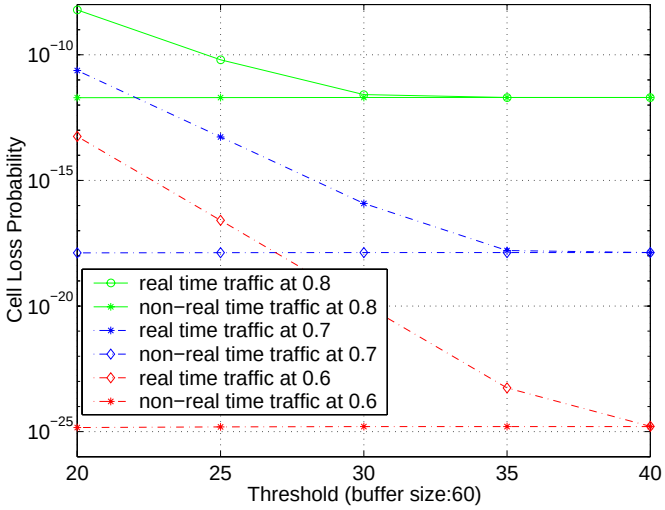


Fig. 10. Cell loss probability versus threshold (as the mean offered load is varied).

decrease. Thus, SDPI outperforms TBD for accommodating the real time traffic, and SDPI may reach comparable performance as the TBD scheme for accommodating the non-real time traffic by increasing the buffer size with the fixed threshold (due to the maximum cell delay of the real time traffic). Fig. 9 shows the effect of traffic characteristics on individual cell loss probabilities. As the activity for the non-real time traffic changes, cell loss probability for each traffic is affected. Fig. 10 shows the cell loss probabilities versus the thresholds. Cell loss probabilities for the non-real time traffic is almost unchanged, but cell loss probabilities for the real time traffic decrease as the threshold reaches the buffer size.

## IV. CONCLUSIONS

We have studied the cell loss performance of an ATM MUX loaded with a traffic stream from the superposition of multiple ON-OFF sources in the two-class environment using the proposed buffer management scheme. By modeling each type of traffic by a 2-state MMPP, we were able to derive the CLP of the respective traffics (i.e., real time traffic and non-real time traffic) using the proposed SDPI space priority scheme. This scheme is applicable to schedule delay-tolerant non-real time traffic and delay-sensitive real time traffic. That is, by delaying the non-real time traffic cells and pushing in the real time traffic cells selectively, more real time traffic can be accepted within the acceptable QoS requirement (e.g., CLP). By provisioning additional priority to the real time traffic, SDPI compensates for the disadvantage of the threshold-based discarding (TBD) scheme which favors the non-real time traffic at the expense of the real time traffic, under the circumstance that the threshold is fixed due to the maximum cell delay constraint of the real time traffic. Thus, channel utilization is improved for the real time traffic. Simulations have also validated our numerical analysis.

## REFERENCES


[1] A. Y-M Lin and J. A. Silvester, "Priority queueing strategies and buffer allocation protocols for traffic control at an ATM integrated broadband switching system," *IEEE J. Select. Areas Commun.*, vol. 9, no. 9, pp. 1524–1536, Dec. 1991.

[2] L. Tassiulas, Y. Hung, and S. Panwar, "Optimal buffer control during congestion in an ATM network node," *IEEE/ACM Trans. Networking*, vol. 2, no. 4, pp. 374–386, Aug. 1994.

[3] S. Suri, D. Tipper, and G. Meempat, "A comparative evaluation of space priority strategies in ATM networks," in *IEEE. Infocom'94*, 1994, pp. 516–523.

[4] S. Sharma and Y. Viniotis, "Optimal buffer management policies for shared-buffer ATM switches," *IEEE/ACM Trans. Networking*, vol .7, no. 4, pp. 575–587, Aug. 1999.

[5] M. Atiquzzaman, "Buffer dimensioning in ATM networks using no-priority and pushout space priorities," in *Proc. IEEE Globecom'95*, 1995, pp. 388–392.

[6] C. Chang and H. Tan, "Queueing analysis of explicit policy assignment push-out buffer sharing schemes for ATM networks," in *Proc. IEEE Infocom'94*,1994, pp. 500–509.

[7] H. Chiou and Z. Tsai, "Performance of ATM switched with age priority packet discarding under the ON-OFF source model," in *Proc. IEEE Infocom'98*, 1998, pp. 931–938.

[8] T. Huang, "Performance analysis of a new cell discarding scheme in ATM networks," in *Proc. IEEE ICC'97*, 1997, pp. 205–209.

[9] A. Choudhury and E. Hahne, "Dynamic thresholds for multiple loss priorities," *IEEE ATM'97 Workshop*, 1997, pp. 272–281.

[10] J. Kim, R. Simha, and T. Suda, "Analysis of a finite buffer queue with heterogeneous markov modulated arrival processes: A study of traffic burstiness and priority packet discarding," *Computer Networks and ISDN Systems 28*, pp. 653–673, 1996.

[11] M. Krunz, H. Hughes, and P. Yegani, "Design and analysis of a buffer management scheme for multimedia traffic with loss and delay priorities," in *Proc. IEEE Globecom'94*, 1994, pp.1560–1564.

[12] H. Heffs and D. M. Lucantoni, "A markov modulated characterization of packetized voice and data traffic and related statistical multiplexer performance," *IEEE J. Select. Areas Commun.*, vol. SAC-4, no. 6, pp. 856–868, Sept. 1986.

[13] R. Nagarajan, J. F. Kurose, and D. Towsley, "Approximation techniques for computing packet loss in finite-buffered voice multiplexers," *IEEE J. Select. Areas Commun.*, vol. 9, no. 3, pp. 368–377, Apr. 1991.

[14] A. Baiocchi, N. B. Melazzi, and M. Listanti, "Loss performance analysis of an ATM multiplexer loaded with high speed ON-OFF sources," *IEEE J. Select. Areas Commun.*, vol. 9, no. 3, pp. 388–393, Apr. 1991.

[15] S. Shah-Heydari and T. Le-Ngoc, "MMPP modeling of aggregated ATM traffic," in*Proc Canadian Conf. Electrical and Computer Engineering (CCECE'98)*, 1998, pp. 129–132.

[16] F. Yegenoglu and B. Jabbari, "Performance evaluation of MMPP/D/1/K queues for aggregate ATM models," in *Proc. IEEE Infocom'93*, 1993, pp. 1314–1319.

[17] S. Kim, M. Lee, and M. Kim, "$\Sigma$-Matching technique for MMPP modeling of heterogeneous ON-OFF sources," in *Proc. IEEE Globecom'94*, 1994, pp. 1090–1094.

[18] B. Wolff, "Poisson arrivals see time averages," *Operational Research*, vol. 30, no. 2, pp. 223–231, Apr. 1982.

[19] R. Cooper, *Introduction to Queueing Theory*, New York: Macmillan, 1972.

[20] H. Schulzrinne and J. Kurose, "Congestion control for real-time in high-speed networks," in *Proc. IEEE Infocom'90*, 1990, pp. 543–550.

[21] D. Gross and C. M. Harris, *Fundamentals of Queueing Theory*, New York: Wiley, 1985.

[22] S. Kowtha and D. Vaman, "A generalized ATM model and its application in bandwidth allocation," in *Proc. IEEE ICC'92*, 1992, pp. 1009–1013.


**Jongho Bang** received his B.S. and M.S. degrees, both in electrical engineering with concentration in computer network from Chung-Ang University, Seoul, Korea. He completed his Ph.D. in electrical engineering with concentration on wireless telecommunication networks at the New Jersey Center for Wireless Telecommunication (NJCWT), New Jersey Institute of Technology (NJIT), New Jersey, in 2001. His interesting research areas are resource and mobility management for wireless networks and buffer management in ATM switch.

**Nirwan Ansari** received the B.S.E.E. (summa cum laude), M.S., E.E., and Ph.D. from the New Jersey Institute of Technology, University of Michigan, and Purdue University in 1982, 1983, and 1988, respectively. He joined the Department of Electrical and Computer Engineering at NJIT in 1988, and has been Professor since 1997. He is a technical editor of the IEEE Communications Magazine, was instrumental, while serving as its Chapter Chair, in rejuvenating the North Jersey Chapter of the IEEE Communications Society which received the 1996 Chapter of the Year Award, currently serves as the Chair of the IEEE North Jersey Section, and also serves in the IEEE Region 1 Board of Directors and various IEEE committees. He was the 1998 recipient of the NJIT Excellence Teaching Award in Graduate Instruction, and a 1999 IEEE Region 1 Award. His current research focuses on various aspects of high speed networks including QoS routing, congestion control, traffic scheduling, video traffic modeling and delivery, and resource allocation. He authored with E.S.H. Hou Computational Intelligence for Optimization (1997, and translated into Chinese in 2000), and edited with B. Yuhas Neural Networks in Telecommunications (1994), both published by Kluwer Academic Publishers.

**Sirin Tekinay** has been with the department of Electrical and Computer Engineering at New Jersey Institute of Technology where she currently serves as the co-director of the New Jersey Center for Wireless Telecommunications, since 1997. Her research interests include teletraffic modeling and management, resource allocation, mobility management, wireless geolocation systems, and next generation wireless networking. She holds a Ph.D.('94) degree with concentration in telecommunications from School of Information Technology and Engineering, George Mason University. Before joining the academia, she served as a visiting scientist at CONTEL, as a senior member of scientific staff at NORTEL, and later at the Bell Labs, LUCENT Technologies. She has authored numerous publications in these areas, and given short courses and tutorials. She holds three patents involving wireless geolocation systems, and demand modeling. She is an active member of the IEEE and is involved in several IEEE technical committees, including the technical committees on personal communications, multimedia communications and vehicular technology. She has served on several major conference technical committees, organized and chaired the first Symposium on Next Generation Wireless Networks. She is on the editorial boards of the IEEE Communications Magazine, the IEEE Communications Surveys and the IEEE Journal of Selected Areas in Communications: Wireless Communications Series. She is also a member of the Eta Kappa Nu, Sigma Xi, and New York Academy of Sciences.