Analysis of a Markov-Modulated Fluid Model for Multimedia Traffic with Loss and Delay Priorities

Marwan Krunz† and Herman Hughes‡
† Department of Electrical Engineering
‡ Department of Computer Science
Michigan State University
East Lansing, MI 48824
e-mail: [krunz,hughes]@cps.msu.edu

Abstract

This paper presents a complete fluid-flow analysis of a traffic control mechanism in an ATM multiplexer, known as Nested Threshold Cell Discarding with Multiple Buffers. NTCD/MB is a buffer management scheme that implements loss and delay priority queueing strategies in order to satisfy a wide range of quality-of-service requirements. We consider a heterogeneous mix of traffic which consists of real-time traffic with a stringent cell delay requirement, and nonreal-time traffic with a stringent cell loss requirement. Moreover, a proportion of real-time cells has an additional stringent loss requirement. Cell arrivals are modeled using a fluid-flow modeling approach. The system we analyze consists of two buffers interacting via a common server. The steady-state probability distribution for the buffers' content as well as other performance statistics are obtained. Our approach is based on computing the *conditional* probability distributions of the buffers and using this information to find the desired unconditional stationary distributions. This approach can be used, in general, to analyze multiple *dependent* systems of buffers. Numerical examples based on the analysis are used along with simulations to study the performance of NTCD/MB. It is shown that NTCD/MB can be used effectively to provide performance guarantees on different cell loss and delay requirements of the traffic.

1 Introduction

Future Broadband Integrated Services Digital Network (B-ISDN) promises to provide transport for a wide range of user applications that vary in their minimum acceptable level of performance known as *Quality-of-Service* requirements (**QoS**). Traffic generated by various applications is packetized into fixed length ATM-based cells, and carried over the optical fiber media of B-ISDN. At intermediate nodes, cells from various sources are statistically multiplexed and switched to their proper destinations. To efficiently utilize the link capacity, statistical multiplexing must realize different QoS requirements for different types of traffic. Such requirements are commonly measured by cell loss rate and end-to-end cell delay. The relative significance of these two measures varies from one application to another. Real-time applications, which require timely and synchronous delivery of cells (e.g., interactive voice and video), have stringent delay requirements. Other applications (e.g., common file transfer) require stringent loss requirements with little demand on cell delay performance. In addition to stringent delay requirements, certain applications are very sensitive to the cell loss of a selected portion of their traffic. An example of this is a certain type of compressed video traffic which consists of two types of cells; vital cells with a severe loss requirement, and normal cells with moderate loss requirement [8].

To guarantee delay requirements, buffers at ATM nodes implement time (or delay) priority mechanisms. Such mechanisms decrease the queueing delay of high priority cells [11]. Space (or loss) priority mechanisms are used to achieve different cell loss rates [2, 5, 21, 22]. Due to their simplicity, space priorities are preferred over time priorities in high-speed switching. However, a traffic mix with heterogeneous sources could be best managed using a combination of both types of priorities. The focus of this paper is to propose and analyze a new buffer management scheme which implements both types of priority queueing strategies.

The remainder of the paper is organized as follows: in section 2, we review the common space priority mechanisms. We introduce the NTCD/MB buffer management scheme and the traffic model used to analyze it in section 3. Section 4 provides a complete fluid analysis for NTCD/MB. Numerical results and computational considerations are presented in section 5. Finally, we summarize the major contributions of the paper.

2 Space Priority Mechanisms

In space priority mechanisms, cells are tagged according to their priority classes. If congestion occurs, cells of a lower priority class are dropped to give space to higher priority cells. Most space priority mechanisms are variations of two basic schemes; the *push-out* and the *partial buffer sharing*. In the push-out mechanism, all cells enter one shared buffer up to the maximum buffer size [5]. If a high priority cell arrives at a saturated buffer that contains low priority cells, a low priority cell is dropped and its place is given to the high priority cell. Despite its efficiency, push-out adds more complexity to buffer management. In [8], Kroner *et. al* analyzed the performance of another scheme, called the *partial buffer sharing* with *threshold discarding*. Generalization of this method is known as the *Nested Threshold Cell Discarding* (**NTCD**) mechanism (Figure 1). Here, the buffer is logically partitioned by a number of threshold levels, n, which corresponds to the number of priority classes. Cells of priority class i enter the buffer up to threshold level T_i. If the buffer occupancy is above T_i, arriving cells of class i are dropped. NTCD with shared buffer was studied in [8, 21, 3] for different traffic arrival processes. Its simplicity makes it suitable for implementing loss priorities in ATM switching nodes. However, NTCD with one buffer may be insufficient if both loss and delay priorities are to be realized.

In such situations, a buffer management scheme that combines loss and delay priority queueing strategies seems to be an attractive solution. In [9], we proposed one such scheme which was referred to as *NTCD with multiple buffers* (**NTCD/MB**), and studied its performance via simulations using

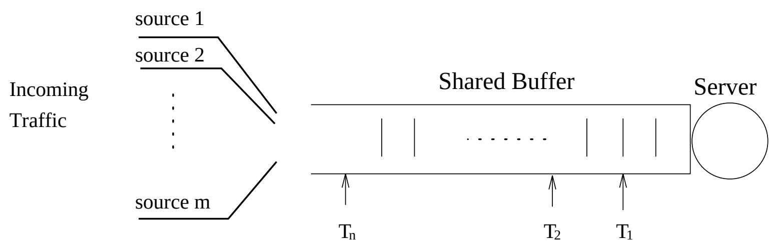

Figure 1: NTCD with a shared buffer and n thresholds.

a Markov-modulated Poisson arrival processes. An observation from this study indicated that the overall performance of NTCD/MB (measured by the amount of buffer resources used to satisfy certain QoS requirements) is better than that of the common-buffer NTCD, particularly when the traffic mix has a wide range of requirements.

Performance evaluation of buffer management schemes is based on a presumed traffic model that encapsulates the main stochastic characteristics of the actual traffic. To obtain tractable results, common analyses of buffer systems used simple traffic models such as Poisson, or more generally, renewal processes. The actual traffic arriving at ATM nodes is a complex nonrenewal stochastic process that exhibits correlations between arrivals and variability in the arrival rates. Studies have shown that such correlations can significantly degrade the network performance [18]. Hence, the use of simple traffic models (e.g., Poisson process) would probably result in crude estimates of performance (e.g., cell loss rate and average waiting times) that can be quite optimistic.

To provide more accurate results, several modeling frameworks have been recently developed among which, two have received considerable attention, namely; the *Markov-modulated Poisson process* (**MMPP**) approach [6, 4, 21], and the Fluid Flow approach [1, 7, 3, 15]. An MMPP is a doubly stochastic Poisson process whose arrival rate is a random variable which is modulated (i.e., controlled) by the state of a continuous-time Markov chain (see [4] for details of the properties of MMPP). The MMPP traffic model is suitable to approximate the *aggregate* arrival process of computer-data sources. Its major advantage is that it captures some of the important correlations between packet (i.e., cell) arrivals while remaining analytically tractable. However, like any queueing approach its computational complexity is proportional to the buffer size, which makes the MMPP model computationally impractical for systems with large buffers. In addition, it remains questionable if the MMPP model can provide a good approximation for video traffic in B-ISDN/ATM environments.

The other modeling approach, which is adopted here, is based on stochastic fluid models [1, 3, 7, 15]. Here, the traffic is viewed as a stream of fluid which is described by a flow rate. The notion of discrete arrivals of individual packets is lost as packets are assumed to be infinitesimally small. Fluid approach has been found particularly appropriate to model traffic in high-speed ATM-based networks for a number of reasons [3]. One reason is the bursty nature of ATM networks which is well captured in fluid models. Another reason is due to the standardization of small (53 bytes) fixed ATM cells which, along with high transmission speeds (in the order of hundreds of megabits and above), makes the impact of individual cell arrivals insignificant. This provides a strong justification for separating time scales which is the basis of the fluid approach. One last reason is related to the computational complexity involved in fluid modeling which is, unlike queueing analysis, independent of the buffer size.

In this paper, we use fluid approximation to analytically evaluate the performance of NTCD/MB. The fluid approach (first introduced by Anick *et. al.* [1]) was used in [3] to analyze NTCD mechanism with one buffer and a threshold. In [23], Zhang analyzed a two-buffer system with complete preemptive priority for one buffer, i.e., the multiplexer dedicates up to its full capacity to the high priority buffer, with the low priority buffer served only when the high priority buffer is empty. Other variations are found in [13, 20]. *The particular significance of our analysis is that it does not assume a preemptive priority to be implemented in any of the two buffers; it guarantees a fractional capacity to each buffer which can be used to provide performance guarantees; and additionally it assumes a work-conserving service discipline.* The last requirement implies that if one buffer is not fully utilizing its fractional guaranteed capacity, the difference is not wasted and instead is given to the other buffer. Such extensions are by no means trivial, as they imply dependency between the content of the two buffers. The approach we used to analyze the system is based on finding the conditional probability distribution for each buffer given the content of the other buffer, then solving for the unconditional probability distributions. This approach can be used, in general, to analyze any system of dependent multiple buffers.

3 NTCD with Multiple Buffers: NTCD/MB

3.1 Traffic Description

We consider two general types of traffic; *real-time* (**RT**) traffic which is delay-sensitive, and *nonreal-time* (**NRT**) traffic which is loss-sensitive. RT traffic consists of cells with two levels of priority; *real-high* (**RH**) with a stringent loss requirement, and *real-low* (**RL**) with a moderate loss requirement. The analysis to be presented is valid for a general traffic source whose arrival process can be modeled by a *Markov modulated fluid flow* (**MMFF**) process (see [19] for details). For the numerical results, we shall use NRT traffic to represent N_d number of identical data sources, and RT traffic to represent N_{vo} voice sources. Video traffic can be easily incorporated in the analysis using, for example, the fluid model proposed by Maglaris *et. al.* [14]. In that model, the *aggregate* arrival rate (or equivalently, bit rate) of several video sources was modeled using an MMFF process with a birth-death modulating process that consists of M states. In order to obtain good results, the authors in [14] suggested a value for M which is equal to 20 times the number of video sources. Their model was developed to characterize traffic of video sources that use a conditional replenishment compression scheme. Such scheme takes advantage of the temporal dependency between successive video frames. It should be noted that the video model of Maglaris *et. al.* can be adequately used to model compressed video streams that are based only on interframe coding (e.g., DPCM). Video streams that make use of both interframe and intraframe redundancy require more involved fluid models whose modulating chain may not follow a birth-and-death model (one example is MPEG-based video traffic). The analysis presented in this paper is general and can be applied to any video traffic that can be modeled by an MMFF process.

The computations involved in fluid analysis (for the finite buffer case) become numerically unstable as the number of states (i.e., arrival rates) per source increases (see section 6.1) which is often the case with video sources. Hence, special numerical procedures must be used to obtain credible results for video traffic. The computational procedures that are needed to obtain reliable results from fluid analysis of video traffic will be reported in a future paper [10].

Each data or voice source is modeled as an ON/OFF source. ON and OFF periods are exponentially distributed with respective average durations; μ_d^{-1} and r_d^{-1}, for a data source and μ_{vo}^{-1}, and r_{vo}^{-1} for a voice source. During ON periods, cells arrive at a constant rate of α_d for a data source, and α_{vo} for a voice source. No cells arrive during OFF periods. Each voice source delivers RH and RL types of cells, with correlated arrivals, i.e., arrivals of both types are modulated by the same Markov chain.

Priority classes within RT traffic can serve different purposes. Such classes can be used to increase the total admissible load for a given user whose traffic contains cells with different loss requirements. This would be useful, for example, for video sources that generate different types of frames, as in the case of MPEG-based sources which generate three types of video frames; I, P, and B. I-frames are more important than the other two with respect to image construction. Thus, I frames could be assigned RH, while P and B frames could be assigned RL. Another use of priorities is in traffic policing. Traffic sources that violate their "contracted" QoS requirements can be penalized by tagging a proportion of their cells with low priority.

3.2 The Mechanism of NTCD/MB

NTCD/MB is a buffer management scheme designed to guarantee several levels of loss and delay requirements of the traffic. Common buffer mechanisms implement loss priorities or delay priorities but not both. Push-out and NTCD with shared buffer are examples of management schemes that are based on loss priorities. Multiple buffers were used to implement delay priorities [12, 23]. A traffic mix that contains various types of traffic with different loss and delay requirements demands a buffer control that implements loss as well as delay priority queueing strategies. Accordingly, we proposed NTCD/MB mechanism in [9], which combines NTCD loss priority mechanism with the use of multiple buffers to achieve delay priorities. In this mechanism, an ATM cell arriving at a node enters one of several buffers, depending on its delay requirements. The number of buffers used correspond to the number of levels of cell delay to be supported by the network. For our purposes, two delay levels (and thus two buffers) are assumed, corresponding to RT and NRT cells. Extension of this scheme to more than two levels of delay is straightforward. NRT cells accommodate one buffer, denoted by **BNR**, with size B_1. RT cells accommodate another buffer, denoted by **BR**, with size B_2. Since RT cells have two loss requirements, buffer BR implements an NTCD mechanism with threshold T, which represents the fractional buffer capacity common to both RH and RT cells.

If a RL cell arrives at some instant when the content of BR is above T, it is discarded. An arriving RH cell is discarded only if the content of BR was B_2. Both buffers are served by a multiplexer whose capacity C is divided such that a fraction of the capacity, γC, $0 \leq \gamma \leq 1$, is guaranteed for BR, while $(1-\gamma)C$ is guaranteed for BNR. The server is work-conserving; if the traffic in one buffer does not completely utilize its reserved portion of the capacity, the residual capacity is available to the traffic in the other buffer. The scheme is shown in Figure 2.

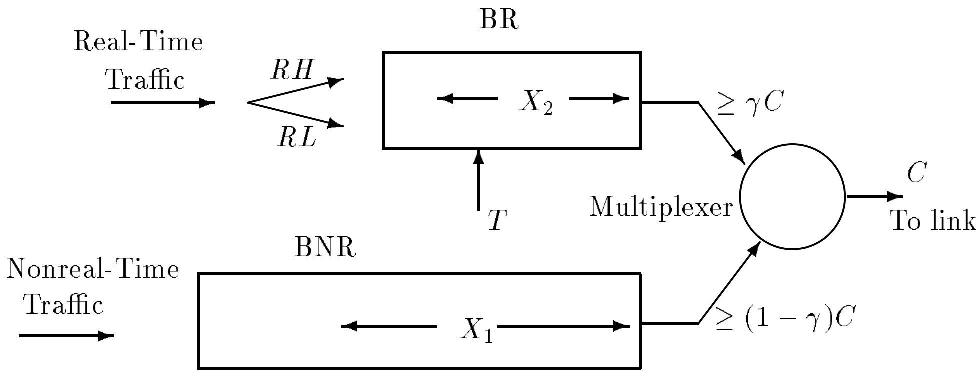

Figure 2: NTCD/MB with two buffers and a threshold.

4 Fluid-Flow Analysis of NTCD/MB Mechanism

In this section, we use fluid approximation to analyze NTCD/MB mechanism with two buffers and a threshold as in Figure 2. To simplify notation, we use subscript 1 and 2 to qualify quantities related to NRT, and RT traffic types, respectively.

4.1 Traffic Analysis

Consider first buffer BNR, which accommodates cells from N_d data sources. Let $S_1(t)$ indicate the state of the aggregate arrival process at BNR at time t. $S_1(t) \in \{0, 1, \ldots, N_d\}$ denotes the number of ON (i.e., active) data sources at time t. The total arrival rate of NRT traffic when $S_1(t) = i$ is $\lambda_1(i) = i\alpha_d$. Following the development in, say, [1], the aggregate arrival process of NRT traffic is a $(N_d + 1)$-state MMFF, whose state transition diagram follows a birth-death model, as shown in Figure 3.

Let $\Pi_1 = \left[\pi_0^{(1)}\ \pi_1^{(1)}\ \ldots\ \pi_{N_d}^{(1)}\right]^T$ be the stationary probability vector for the modulating process of Figure 3. Π_1 is a binomial distribution whose j'th element is given by [15]

$$\pi_j^{(1)} \doteq \lim_{t\to\infty} \Pr\{S_1(t) = j\} = \begin{pmatrix} N_d \\ j \end{pmatrix} \cdot \left(\frac{r_d}{r_d+\mu_d}\right)^j \left(\frac{\mu_d}{r_d+\mu_d}\right)^{N_d-j}, \quad j = 0, 1, \ldots, N_d \tag{1}$$

The effective arrival rate of NRT traffic is $\overline{\lambda}_1 = \sum_j \lambda_1(j)\pi_j^{(1)}$. Let Q_1 denote the generator matrix associated with the Markov chain in Figure 3.

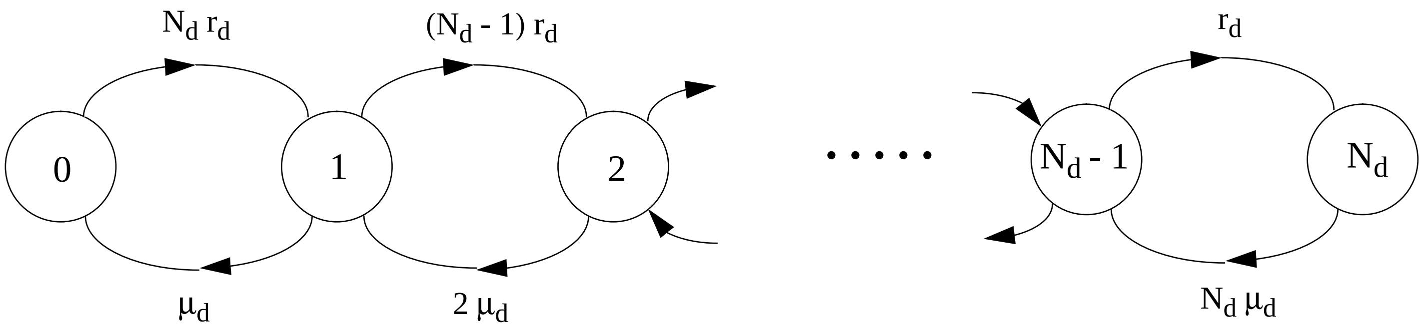

Figure 3: The state transition diagram for $N_d + 1$-state MMFF that models nonreal-time traffic.

Except for the choice of parameters, the same traffic analysis used for data sources applies as well to voice sources. Hence, the aggregate RT (i.e., voice) traffic is described by a $(N_{vo} + 1)$-state MMFF with generator matrix Q_2, and equilibrium probability vector Π_2 (defined similar to Π_1). The aggregate arrival rate of RT traffic at state $S_2(t) = i$ $(i = 0, 1, \ldots, N_{vo})$ is $\lambda_2(i) = i\alpha_{vo}$. As mentioned before, voice cells consist of two priority classes (high and low). Let $\xi = \lambda_2^H(i)/\lambda_2(i)$ denote the load ratio for RT traffic ($\lambda_2^H(i)$ is the arrival rate of RH cells when the RT arrival process is in state i). Note that ξ is constant and does not depend on the state of the process. We assume that both priority classes are modulated by the same Markov chain.

The combined total arrival rate of NRT and RT traffics can be described by a global K-state MMFF process $(K = (N_d + 1)(N_{vo} + 1))$ whose state at time t is $S(t) \doteq (S_1(t), S_2(t)) \in \mathcal{L}$, where $\mathcal{L} \doteq \{(i, j) \,:\, i = 0, 1, \ldots, N_d,\ j = 0, 1, \ldots, N_{vo}\}$. The set of pairs in $\mathcal{L}$ are ordered lexicographically. We say $S(t) = (i, j) = k$ where $k = 1, 2, \ldots, K$, to mean that the integer k represents the k'th ordered pair in $\mathcal{L}$. The global generator matrix, $\boldsymbol{Q}$, and the stationary probability vector Π for the global arrival process are [19]:

$$\boldsymbol{Q} = \boldsymbol{Q}_1 \oplus \boldsymbol{Q}_2 \tag{2}$$

$$\Pi = \Pi_1 \otimes \Pi_2 = [\pi_1 \ \pi_2 \ \ldots \ \pi_K]^T \tag{3}$$

where for every global state $k = (i, j) \in \mathcal{L}$, $\pi_k = \pi_i^{(1)} \pi_j^{(2)}$. The symbols "$\oplus$" and "$\otimes$" above indicate the Kronecker sum and the Kronecker product, respectively.

4.2 The Conditional Equilibrium Probability Distributions

In this section, we derive a set of ordinary differential equations which are solved using a common eigenvalue/eigenvector approach, to obtain conditional equilibrium distributions for the buffers. Let $X_1(S(t), t)$ and $X_2(S(t), t)$ (or simply $X_1(t)$ and $X_2(t)$) denote the instantaneous content of BNR and BR, respectively, when the system is at the global state $S(t) = k = (i, j) \in \mathcal{L}$. The instantaneous *drift* rate for the content of BNR and BR is given by $dX_1(t)/dt$ and $dX_2(t)/dt$, respectively. In addition to the state of the arrival process, the drift rate at each buffer depends on the content of the other buffer and (for BR only) the threshold T. The buffers dependency is caused by implementing a work-conserving service discipline. Accordingly, six cases (two for BNR and four for BR) are considered, as shown in Table 1. Note that in the fourth column in Table 1, the quantity $d_\alpha^{(l)}(k)$ where $l = 1, 2$, and $\alpha = a, b$ (if $l = 1$) or $\alpha = a, b, c, d$ (if $l = 2$) is the drift rate $dX_l(t)/dt$ for the case l-α at some global state $k = (i, j) \in \mathcal{L}$. Without loss of generality, we

Case	Buffer	Condition	Drift at $S(t) = k = (i, j)$
1-a	BNR	$0 < X_1 < B_1$ and $X_2 = 0$	$d_a^{(1)}(k) = \lambda_1(i) - C + \min\{\gamma C,\ \lambda_2(j)\}$
1-b	BNR	$0 < X_1 < B_1$ and $X_2 > 0$	$d_b^{(1)}(k) = \lambda_1(i) - C(1-\gamma)$
2-a	BR	$0 < X_2 < T$ and $X_1 = 0$	$d_a^{(2)}(k) = \lambda_2(j) - C + \min\{(1-\gamma)C$
2-b	BR	$T < X_2 < B_2$ and $X_1 = 0$	$d_b^{(2)}(k) = \lambda_2^H(j) - C + \min\{(1-\gamma)C$
2-c	BR	$0 < X_2 < T$ and $X_1 > 0$	$d_c^{(2)}(k) = \lambda_2(j) - \gamma C$
2-d	BR	$T < X_2 < B_2$ and $X_1 > 0$	$d_d^{(2)}(k) = \lambda_2^H(j) - \gamma C$

Table 1: The different cases considered in the analysis.

assume that for all the states in $\mathcal{L}$ and all the cases considered above, the drift rate are nonzero[1]. To use matrix notation, for each of the above cases define a drift matrix

$$\boldsymbol{D}_\alpha^{(l)} \doteq diag\left\{d_\alpha^{(l)}(k)\ :\ k = (i, j) \in \mathcal{L}\right\} \tag{4}$$

where $l = 1, 2$, and $\alpha = a, b$ (if $l = 1$), or $\alpha = a, b, c, d$ (if $l = 2$). The elements along the diagonal of $\boldsymbol{D}_\alpha^{(l)}$ are ordered lexicographically. Define the following time-dependent conditional probabilities

$$P_\alpha^{(l)}(k, x, t) \doteq \Pr\left\{S(t) = k,\ X_l(t) \le x\ /\ \text{condition of case } l\text{-}\alpha\right\},\quad \text{for all possible } l \text{ and } \alpha \tag{5}$$

In the limit, the asymptotic probabilities corresponding to Eq. 5 are:

$$F_\alpha^{(l)}(k, x) \doteq \lim_{t\to\infty} P_\alpha^{(l)}(k, x, t) \tag{6}$$

for all l and α. Let $\boldsymbol{F}_\alpha^{(l)}(x)$ be the column vector from $\{F_\alpha^{(l)}(k, x)\}_{k=1}^K$ for all l and α. Using the same development of, say, [1], the system is governed by the following set of differential equations:

$$\boldsymbol{D}_\alpha^{(l)} \frac{d\boldsymbol{F}_\alpha^{(l)}(x)}{dx} = \boldsymbol{Q}\boldsymbol{F}_\alpha^{(l)}(x),\quad \text{for all } l \text{ and } \alpha \tag{7}$$

where $\boldsymbol{Q}$ is the global generator matrix given in Eq. 2. Each case in (7) is a set of K linear differential equations whose solution takes the form:

$$\boldsymbol{F}_\alpha^{(l)}(x) = \sum_{i=0}^{K} A_\alpha^{(l)}(i)\, \boldsymbol{\Psi}_\alpha^{(l)}(i)\, \exp\left\{z_\alpha^{(l)}(i)x\right\},\quad \text{for all } l \text{ and } \alpha \tag{8}$$

In Eq. 8, for each l and α that point to one of the six cases considered in Table 1, the set of pairs $\left\{(z_\alpha^{(l)}(i),\ \boldsymbol{\Psi}_\alpha^{(l)}(i))\ :\ i = 1, \ldots, K\right\}$ is the K eigenvalues-eigenvector solutions for the generalized eigenvalue problem $z_\alpha^{(l)}(i)\, \boldsymbol{D}_\alpha^{(l)} \boldsymbol{\Psi}_\alpha^{(l)}(i) = \boldsymbol{Q}\, \boldsymbol{\Psi}_\alpha^{(l)}(i),\ \forall\ i \in \mathcal{L}$. The coefficients $A_\alpha^{(l)}(i)$ are unknown constants to be obtained from boundary conditions.

[1]The degenerative situation in which the drift in some states is zero was treated in [15].

4.3 Boundary Conditions

To obtain the constants $\{A_\alpha^{(l)}(i)\}$, a number of $6\,K$ boundary conditions is required. First, define the following sets of states:

$$\begin{aligned} \mathcal{U}_\alpha^{(l)} &\doteq \left\{ m \in \mathcal{L} \;:\; d_\alpha^{(l)}(m) < 0 \right\} \\ \mathcal{O}_\alpha^{(l)} &\doteq \left\{ m \in \mathcal{L} \;:\; d_\alpha^{(l)}(m) > 0 \right\} \end{aligned}$$

where the values of l and α correspond to the several cases defined in Table 1. Note that $\mathcal{U}_\alpha^{(l)}$ represents the set of global states in $\mathcal{L}$ that cause an *underload* drift for the case l-α. Similarly, $\mathcal{O}_\alpha^{(l)}$ refers to the states causing an *overload* drift. Following the development in [3], we have:

- *Boundary conditions for BNR.*

$$F_\alpha^{(1)}(k,0) = 0 \quad \forall\, k \in \mathcal{O}_\alpha^{(1)} \quad \text{and } \alpha = a,\, b \tag{9}$$

$$F_\alpha^{(1)}(k,B_1) = \pi_m \quad \forall\, k \in \mathcal{U}_\alpha^{(1)} \quad \text{and } \alpha = a,\, b \tag{10}$$

which provides $2\,K$ equations that can be used to solve for $\{A_\alpha^{(1)}(i)\}$ in (8), where $\alpha = a$ (for BNR when $X_2 = 0$), and $\alpha = b$ (for BNR when $X_2 > 0$).

- *Boundary conditions for BR.*

$$F_\alpha^{(2)}(k,0) = 0 \quad \forall\, k \in \mathcal{O}_\alpha^{(2)} \text{ and } \alpha = a,\, c \tag{11}$$

$$F_\alpha^{(2)}(k,B_2) = \pi_k \quad \forall\, k \in \mathcal{U}_\alpha^{(2)} \text{ and } \alpha = b,\, d \tag{12}$$

$$F_a^{(2)}(k,T) = F_b^{(2)}(k,T) \quad \forall\, k \in \mathcal{U}_a^{(2)} \bigcup \mathcal{O}_b^{(2)} \tag{13}$$

$$F_c^{(2)}(k,T) = F_d^{(2)}(k,T) \quad \forall\, k \in \mathcal{U}_c^{(2)} \bigcup \mathcal{O}_d^{(2)} \tag{14}$$

Equations (11) - (14) provide $4K$ equations which are used to solve for $\{A_\alpha^{(2)}(i)\}$, where $\alpha = a, b, c, d$. Justification of the above equations can be found in [3] and the references cited there.

4.4 The Steady-State Solution

In this section, we determine the (unconditional) equilibrium distribution of each buffer, independent of the other buffer. Consider a countable mutually exclusive class of events $\{B_i\}$ that partitions an event A. Let $\Pr\{B_i\} > 0$ for all i. By the *law of total probability* $\Pr\{A\} = \sum_i \Pr\{A/B_i\}\Pr\{B_i\}$. We apply the same concept to find the steady-state distribution of the buffer contents. Let $H(i,x) \doteq \lim_{t\to\infty} \Pr\{S(t)=i,\ X_1(t) \le x\}$ for $0 \le x < B_1$. Similarly, let $G_1(i,x) \doteq \lim_{t\to\infty} \Pr\{S(t)=i,\ X_2(t) \le x\}$ for $0 \le x < T$, and $G_2(i,x) \doteq \lim_{t\to\infty} \Pr\{S(t)=i,\ X_2(t) \le x\}$ for $T < x < B_2$. In the following, $H(i,x)$, $G_1(i,x)$ and $G_2(i,x)$ are to be found from $F_\alpha^{(l)}(k,x)$. Consider first the nonreal-time buffer, BNR:

$$\begin{aligned} \Pr\{S(t)=i,\ X_1(t) \le x\} &= \Pr\{S(t)=i,\ X_1(t) \le x \,/\, X_2(t)=0\} \cdot \Pr\{X_2(t)=0\} \\ &\quad + \Pr\{S(t)=i,\ X_1(t) \le x \,/\, X_2(t)>0\} \cdot \Pr\{X_2(t)>0\} \end{aligned}$$

Hence,

$$H(i,x) \quad = \quad F_a^{(1)}(i,x)\cdot \Pr\{X_2=0\} + F_b^{(1)}(i,x)\cdot(1-\Pr\{X_2=0\})$$

(note that the parameter t that indicates time has been dropped to indicate asymptotic quantities). But, $\Pr\{X_2=0\} = \sum_j \Pr\{S=j,\ X_2=0\} = \sum_{j=1}^{K} G_1(j,0)$. Thus,

$$H(i,x) \quad = \quad \left[F_a^{(1)}(i,x) - F_b^{(1)}(i,x)\right]\left[\sum_{j=1}^{K} G_1(j,0)\right] + F_b^{(1)}(i,x) \tag{15}$$

which is true for $0 \le x < B_1$, and for all $i \in \mathcal{L}$. Next consider BR.

$$\begin{aligned}\Pr\{S=i,\ X_2\le x\} \quad &= \quad \Pr\{S=i,\ X_2 \le x \ /\ X_1=0\}\cdot\Pr\{X_1=0\} \\ &\quad + \Pr\{S=i,\ X_2\le x\ /\ X_1>0\}\cdot \Pr\{X_1>0\}\end{aligned}$$

which reduces to

$$G_1(i,x) \quad = \quad F_a^{(2)}(i,x)\cdot\left[\sum_{j=1}^{K} H(j,0)\right] + F_c^{(2)}(i,x)\cdot\left[1-\sum_{j=1}^{K} H(j,0)\right], \quad \text{for } 0\le x<T \tag{16}$$

and

$$G_2(i,x) \quad = \quad F_b^{(2)}(i,x)\cdot\left[\sum_{j=1}^{K} H(j,0)\right] + F_d^{(2)}(i,x)\cdot\left[1-\sum_{j=1}^{K} H(j,0)\right], \quad \text{for } T<x<B_2 \tag{17}$$

Since Equation (16) is valid for $0 \le x < T$, at $x=0$

$$G_1(i,0) = F_a^{(2)}(i,0)\cdot\left[\sum_{j=1}^{K} H(j,0)\right] + F_c^{(2)}(i,0)\cdot\left[1-\sum_{j=1}^{K} H(j,0)\right] \tag{18}$$

Substituting $G_1(i,0)$ from (18) in (15) and evaluating at $x=0$, we have

$$\begin{aligned}H(i,0) \quad &= \quad F_b^{(1)}(i,0) + \left[F_a^{(1)}(i,0) - F_b^{(1)}(i,0)\right] \\ &\quad \cdot\left[\sum_{j=1}^{K}\left\{F_a^{(2)}(j,0)\cdot\left[\sum_{p=1}^{K} H(p,0)\right] + F_c^{(2)}(j,0)\left[1-\sum_{p=1}^{K} H(p,0)\right]\right\}\right]\end{aligned} \tag{19}$$

This is a set of K linear equation of K unknowns; $H(i,0)$, $i\in\mathcal{L}$, and can be easily solved. With $H(i,0)$ evaluated, $G_1(i,0)$ is obtained from (18), and thereafter, $H(i,x)$ from (15). Also, $G_1(i,x)$ and $G_2(i,x)$ are found from (16) and (17), which completes the solution.

5 Performance Statistics

Once the steady-state distributions for the content of BNR and BR are determined, it is straightforward to obtain the throughput and loss probabilities for the various types of traffic. First, the following quantities are computed:

$$\Pr\{S=i,\ \text{BNR full}\} = \pi_i - H(i,B_1),\ \Pr\{S=i,\ \text{BNR empty}\} = H(i,0)$$

$$\Pr\{S=i,\ \text{BR full}\} = \pi_i - G_2(i,B_2),\ \Pr\{S=i,\ \text{BR empty}\} = G_1(i,0)$$
$$\Pr\{S=i,\ X_2=T\} = G_2(i,T) - G_1(i,T)$$

The throughput for a given class of cells is the difference between the generation rate for that class and its loss rate. For NRT cells, the generation rate is $\sum_{i=1}^{K}\lambda_1(i)\pi_i$ (since i here is a global state, the expression for λ_1 must be slightly modified to take this into account). Loss rate for NRT cells is the positive drift rate into BNR when BNR is full. Therefore,

$$\text{loss rate for nonreal cells } = \sum_{i=1}^{K}[\lambda_1(i) - \max\{(1-\gamma)C,\ C-\lambda_2(i)\}]\cdot(\pi_i - H(i,B_1))$$

Let T_1 denote the throughput for NRT cells, T_2^H and T_2^L denote the throughput for high and low priority RT cells, respectively. (Note that we will continue to use the letter T to denote the threshold in buffer BR). Then,

$$\begin{aligned} T_1 &= \sum_{i=1}^{K}\lambda_1(i)\pi_i - \left[\sum_{i=1}^{K}(\lambda_1(i) - \max\{(1-\gamma)C,\ C-\lambda_2(i)\})(\pi_i - H(i,B_1))\right] && (20)\\ &= \sum_{i=1}^{K}\lambda_1(i)H(i,B_1) + \sum_{i=1}^{K}(\pi_i - H(i,B_1))\cdot\max\{(1-\gamma)C,\ C-\lambda_2(i)\} && (21)\end{aligned}$$

$$T_2^H = \sum_{i=1}^{K}\lambda_2^H(i)\pi_i - \left[\sum_{i=1}^{K}\left(\lambda_2^H(i) - \max\{\gamma C,\ C-\lambda_1(i)\}\right)(\pi_i - G_2(i,B_2))\right] \qquad (22)$$

where $\lambda_2^H(i) = \xi\lambda_2(i)$. Low priority real-time (RL) cells are lost whenever $X_2 > T$ in which case all generated cells are lost, or when $X_2 = T$ which results in a partial loss rate $\lambda_2(i) - \max\{\gamma C,\ C-\lambda_1(i)\}$. Therefore, the throughput for RL cells is

$$\begin{aligned} T_2^L &= \text{throughput}\mid_{X_2>T} + \text{throughput}\mid_{X_2=T}\\ &= \left\{\sum_{i=1}^{K}\lambda_2^L(i)\pi_i - \sum_{i=1}^{K}\lambda_2^L(I)(\pi_i - G_2(i,T))\right\}\\ &\quad - \left\{\sum_{i\in\mathcal{O}_a^{(2)}\bigcap\mathcal{U}_b^{(2)}}(G_2(i,T) - G_1(i,T))(\lambda_2(i) - \max\{\gamma C,\ C-\lambda_1(i)\})\right\} && (23)\end{aligned}$$

In the expression above, $\pi_i - G_2(i,T)$ is the probability of $\{X_2 > T \text{ and } S = i\}$. Also, $\lambda_2^L(i) = (1-\xi)\lambda_2(i)$.

Let L_1 denote the loss probability for nonreal cells, L_2^H and L_2^L denote the loss probability for high and low priority real cells, respectively. Then,

$$\begin{aligned} L_1 &= 1 - \frac{T_1}{\sum_{i=1}^{K}\pi_i\lambda_1(i)} && (24)\\ L_2^H &= 1 - \frac{T_2^H}{\sum_{i=1}^{K}\pi_i\lambda_2^H(i)} && (25)\\ L_2^L &= 1 - \frac{T_2^L}{\sum_{i=1}^{K}\pi_i\lambda_2^L(i)} && (26)\end{aligned}$$

In conventional fluid analysis, the distribution for the waiting time of a cell is readily obtained from the distribution for the buffer content. However, in the model considered, this is not trivial since the service rate for a given class of cells is not constant. For nonreal cells, the service rate at state $i \in \mathcal{L}$ is $C_1(i) = \max\{(1-\gamma)C,\ C - \lambda_2(i)\}$. Similarly, the service rate for high and low priority real cells is $C_2(i) = \max\{\gamma C,\ C - \lambda_1(i)\}$ which, for both cases, is a function of the state i. An approximate expression for the waiting time distribution of cells can be obtained using average values of C_1 and C_2 given by $\overline{C_1} = \sum_{i=1}^{K} C_1(i)\pi_i$ and $\overline{C_2} = \sum_{i=1}^{K} C_2(i)\pi_i$, respectively. Thus,

$$\begin{aligned}
\Pr\{\text{delay of nonreal cell} \leq t\} &\approx \frac{1}{T_1}\sum_{i=1}^{K} \lambda_1(i) H(i, \overline{C_1}t) \quad \text{for } 0 \leq t \leq B_1/\overline{C_1} \\
\Pr\left\{\text{delay of nonreal cell} = B_1/\overline{C_1}\right\} &\approx \frac{\overline{C_1}}{T_1}\left[1 - \sum_{i=1}^{K} H(i, B_1)\right]
\end{aligned}$$

For high priority real cells,

$$\Pr\{\text{delay of high real cell} \leq t\} \approx \begin{cases} \frac{1}{T_2^H}\sum_{i=1}^{K} \lambda_2^H(i) G_1(i, \overline{C_2}t) & \text{for } 0 \leq t \leq T/\overline{C_2} \\ \frac{1}{T_2^H}\sum_{i=1}^{K} \lambda_2^H(i)\left(G_2(i,T) - G_1(i,T)\right), & \text{for } t = T/\overline{C_2} \\ \frac{1}{T_2^H}\sum_{i=1}^{K} \lambda_2^H(i) G_2(i, \overline{C_2}t), & \text{for } T/\overline{C_2} < t < B_2/\overline{C_2} \end{cases}$$

$$\Pr\left\{\text{delay of high real cell} = B_2/\overline{C_2}\right\} \approx \frac{\overline{C_2}}{T_2^H}\left[1 - \sum_{i=1}^{K} G_2(i, B_2)\right]$$

Similar expressions can be obtained for low priority real-time cells.

6 Numerical Investigations

6.1 Computational Considerations

The numerical results for the analysis developed in previous sections were obtained using a C program supported by Matlab library functions. Solutions for the eigenvalue/eigenvector sets were based on the decomposition technique of [19]. In this technique, an aggregate traffic that consists of several independent and separable rate-modulated input processes arriving at a buffer system is analyzed by decomposing the system into smaller subsystems to be treated separately, then the final solution is obtained by superposing the individual results. This approach may considerably reduce the number of required computations. The other major numerical problem which was encountered is the numerical instability of the system. Such problem is known to exist in fluid approximation of finite buffers [20]. The eigenvalues for such systems can take positive as well as negative values. When solving for the constants $\left\{A_\alpha^{(l)}(i)\right\}$, the exponentiation of positive and negative terms produces a badly scaled matrix whose values may approach the precision of the machine. Typically, researchers dealt with such problem by using the complementary infinite buffer distribution as an approximation to the actual finite buffer case, since the former case does not contain positive eigenvalues. In our case, however, positive eigenvalues cannot be avoided because of the finite threshold, T. We used several scaling methods to condition the matrices involved. Row as well as column scaling was required with block diagonal matrices to obtain acceptable results.

6.2 Numerical Results and Simulations

To study the performance of NTCD/MB, numerical examples based on the previous analysis as well as simulations are used. As before, the traffic consists of voice and data sources with voice sources generating RH and RL cells.

In our simulations, an ON/OFF source model similar to the one used in the fluid approach (see Section 4.1) is used. However, an ON period in the simulation model is an exponentially distributed time interval, which is then truncated to provide an integer number of time slots. A time slot corresponds to the transmission time of a cell (with respect to the input line rate), or equivalently the cell interarrival time during a burst. Cells from one source arrive one at a time at the beginning of a slot within an ON period. OFF periods in the simulation model are identical to those of the fluid model. To provide a fractional guaranteed bandwidth (equivalent to γC, and $(1-\gamma)C$ used in the analysis), a cyclic service discipline is implemented in the simulation model in which the server alternates between the two buffers forming cycles of service. Each cycle consists of a maximum of C_R slots for BR followed by a maximum of C_N slots for BNR such that $\gamma = \frac{C_R}{C_R+C_N}$ with smallest possible integer-valued C_R and C_N (e.g., if $\gamma = 0.4$, then $C_R = 2$, $C_N = 3$). The service discipline is work-conserving, making the length of the cycles vary depending on the availability of cells in the two buffers. Buffers used in the simulations are of a finite size, whereas the numerical results based on the analysis are mostly obtained assuming infinite buffers, in order to avoid numerical instability and machine precision overflow (note, however, that BR still uses a finite threshold).

It should be noted that because of the several differences between the simulation setup and the assumptions in the analytical approach (namely, the use of finite versus infinite buffers, the truncation of ON periods in the simulations, the separation of time scales in the fluid approach, and the server arbitration), results from both approaches are not expected to be identical. However, for most cases these results were sufficiently close, and more importantly, in all cases the same trend in performance was noticed from both techniques.

The parameters used in the following experiments are normalized with respect to a time and a data units. The average length of ON state for a voice source was chosen as the time unit (i.e, $\mu_{vo}^{-1} = 1$), which corresponds to 352 msec. Based on this time unit, it is assumed that $r_{vo}^{-1} = 1.85$ (or 650 msec), $r_d^{-1} = 4$ and $\mu_d^{-1} = 2$. As noted, a data source is more bursty than a voice source. The size of a cell (i.e., 53 bytes) is used as the data unit. In Figure 4, the effect of the size of BNR on the loss probability of NRT cells is shown for two levels of nonreal-time offered load; $\rho_{non} = 98\%$ (heavy load), and $\rho_{non} = 64\%$ (light load). Here, $\rho_{non} \triangleq \overline{\lambda_1}/(1-\gamma)C$. These results were obtained from the analytical model using both finite and infinite buffers (parameters are shown in the caption of Figure 4). As noted, the infinite buffer approximation provides a lower bound on the finite buffer performance. The complementary buffer distribution for BR, $\Pr\{X_2 > x\}$, is shown in Figure 5 as a function of x for different values of T. Notice that $\Pr\{X_2 > x\}$, which is based on infinite buffers, approximates the loss probability of RH cells. The confluence of drifts at the threshold level is reflected by the discontinuity of $\Pr\{X_2 > x\}$ at $x = T$.

The effect of T on the loss probability for RL cells is shown in Figure 6 for different values of γ. Results in this figure are obtained using the analytical model with finite buffers. Note that the parameter γ is a key factor is determining the trade-off between the performance for NRT and RT traffic and should be determined according to the desired QoS requirements. At larger values of γ, the value of T has more impact on loss probability of RL cells. For the same experiment, the average waiting times of RH cells are shown in Figure 7. At lower values of γ, the value of T has more effect on the delay performance. Compared to the previous figure, at smaller values of γ the waiting time of RH cells is the significant factor in determining T, while at large γ, it is the loss probability of RL cells that determines T.

In Figure 8, the fluid approach with infinite buffers is used along with simulations to obtain the loss probability of RL cells and the average waiting time of RH cells, for different values of T. Note that the performance of the system as obtained from the simulations is slightly worse than that obtained from the analysis. For the same experiment, the probability that the waiting time of a real-time cell exceeds some fixed number t is shown in Figure 9 as a function of T. Again, the analytical results indicate a slightly better performance than the one demonstrated via simulations. The discrepancy between the two approaches, however, is minor. Also, the figure shows that Pr{waiting time of a RH cell $> t$} $>$ Pr{waiting time of a RL cell $> t$} which is an intuitive result. The dotted line in the figure indicates the value of T below which the waiting time of an RH or an RL cell will be less than t, with probability one.

The effect of the load ratio, ξ, of RT traffic on performance is illustrated in Figures 10 and 11. Three values of T are used; $T = 20, 40, 60$. Figure 10 shows that the loss probability of RL cells increases with the load ratio. This is verified by the analytical results (with infinite buffer) and simulation experiments. The same trend can be seen in Figure 11, where the average waiting time of an RH cell is shown to increase with ξ (only analytical results are shown here). The implication of Figures 10 and 11 is that a threshold-based cell discarding mechanism (e.g., NTCD) is most effective with small values of ξ, i.e., high priority cells consist a relatively small fraction of the traffic.

7 Summary

In this paper, we studied the performance of a buffer management scheme for ATM nodes known as Nested Threshold Cell Discarding with Multiple Buffers. NTCD/MB was designed to guarantee several levels of cell loss and delay requirements for the traffic. Such diverse range of requirements is expected to exist in B-ISDN as a result of applications that simultaneously generate data, voice, and video traffic types, i.e., multimedia applications. We used the fluid approach to analyze the performance of NTCD/MB. Our analytical approach was based on computing the probability distributions for each buffer conditioned on the content of the other buffer, from which the unconditional stationary probability distributions were obtained. This approach can be used to analyze dependent multiple-buffer systems that often exist in actual buffer systems. The throughput and loss probabilities for all types of traffic were found. Moreover, we obtained approximate expressions for the waiting time distributions for real-time and nonreal-time traffic. Numerical examples based on the analysis as well as simulation experiments were used to study the performance of NTCD/MB mechanism. Both simulation and analytical results demonstrated a similar trend in performance. These results were used to show the tradeoff that exists between the loss rate and cell delay for various buffer parameters.

References

[1] D. Anick, D. Mitra, and M. M. Sondhi, "Stochastic theory of a data-handling system with multiple sources," *Bell System Tech. J.*, vol. 61, no. 8, pp. 1871-1894, Oct. 1982.

[2] J. J. Bae and T. Suda, "Survey of traffic control schemes and protocols in ATM networks," *Proc. of the IEEE*, Vol. 79, No. 2, Feb. 1991.

[3] A. I. Elwalid and D. Mitra, "Fluid models for the analysis and design of statistical multiplexing with loss priorities on multiple classes of bursty traffic," *Proc. INFOCOM '92*, pp. 3c.4.1-3c.4.11, 1992.

[4] W. Fischer and K. Meier-Hellstern, "The Markov-modulated Poisson process (MMPP) cookbook", in Performance Evaluation Journal, vol. 18, pp. 149-171, 1993.

[5] G. Hebuterne and A. Gravey, "A space priority queueing mechanism for multiplexing ATM channels," *ITC Specialist Seminar*, paper 7.4, Adelaide, 1989.

[6] H. Heffes and D. M. Lucantoni, "A Markov modulated characterization of packetized voice and data traffic and related statistical multiplexer performance," *IEEE JSAC*, Vol. SAC-4, No. 6, pp. 856-868, Sep. 1986.

[7] L. Kosten, "Stochastic theory of data-handling systems with groups of multiple sources," in *Performance of Computer-Communication Systems*, (Ed. H. Rudin and W. Bux), pp. 321-331, Elsevier, Amsterdam, 1984.

[8] H. Kroner, G. Hebuterne, P. Boyer, and A. Gravey, "Priority management in ATM switching nodes," *IEEE JSAC*, 1991.

[9] M. Krunz, H. Hughes, and P. Yegani, "Traffic control in ATM switching using multiple buffers and nested threshold cell discarding," in *Proc. of 2'nd Int. Conf. on Computer Communications and Networks*, pp. 121-127, San Diego, CA, June 1993.

[10] M. Krunz, H. Hughes, and P. Yegani, "A fluid flow analysis of video traffic," manuscript in preparation.

[11] Y. Lim and J. Kobza, "Analysis of a delay-dependent priority discipline in a multiclass traffic packet switching node," *Proc. IEEE INFOCOM '88*, pp. 9A.4.1-9A.4.1.10.

[12] 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 JSAC*, vol. 9, no. 9, pp. 1524-1536, Dec. 1991.

[13] M. L. Luhanga, "A fluid approximation model of an integrated packet voice and data multiplexer," *Proc. of IEEE INFOCOM '88*, pp. 7A.4.1-7A.4.6, 1988.

[14] B. Maglaris, D. Anastassiou, P. Sen, G. Karlsson, and J. Robbins, "Performance models of statistical multiplexing in packet video communications," *IEEE JSAC*, vol. 36, no. 7, pp. 834-844, July 1988.

[15] D. Mitra, "Stochastic theory of a fluid model of producers and consumers coupled by a buffer," *Adv. Appl. Prob.*, vol. 20, pp. 646-676, 1988.

[16] D. W. Petr, L. A. DaSliva, Jr., and V. S. Frost, "Priority discarding of speech in integrated packet networks," *IEEE JSAC*, vol. 7, no. 5 pp. 644-659, June 1989.

[17] D. W. Petr and V. S. Frost, "Nested threshold cell discard for ATM overload control: optimization under cell loss constraints," *Proc. IEEE INFOCOM '91*, pp. 12A.4.1-10, Bal Harbour, 1991.

[18] K. Sriram and W. Whitt, "Characterizing superposition arrival processes in packet multiplexers for voice and data," *IEEE JSAC*, Vol. SAC-4, No. 6, pp. 833-846, Sep. 1986.

[19] T. E. Stern and A. I. Elwalid, "Analysis of separable Markov-modulated rate models for information-handling systems," *Adv. Appl. Prob.*, vol. 23, pp. 105-139, 1991.

[20] Roger C. F. Tucker, "Accurate method for analysis of a packet-speech multiplexer with limited delay," *IEEE Trans. on Communications*, vol. 36, no. 4, pp. 479-483, April 1988.

[21] P. Yegani, "Performance models for ATM switching of mixed continuous bit rate and bursty traffic with threshold based discarding," *ICC 1992*.

[22] N. Yin, S-Q. Li, and T. E. Stern, "Congestion control for packet voice by selective packet discarding," *Proc. IEEE GLOBECOM '87*, pp. 45.3.1-45.3.4.

[23] J. Zhang, "Performance study of Markov modulated fluid flow models with priority traffic," in *Proc. INFOCOM '93*, pp. 1a.2.1-1.a.2.8, 1993.

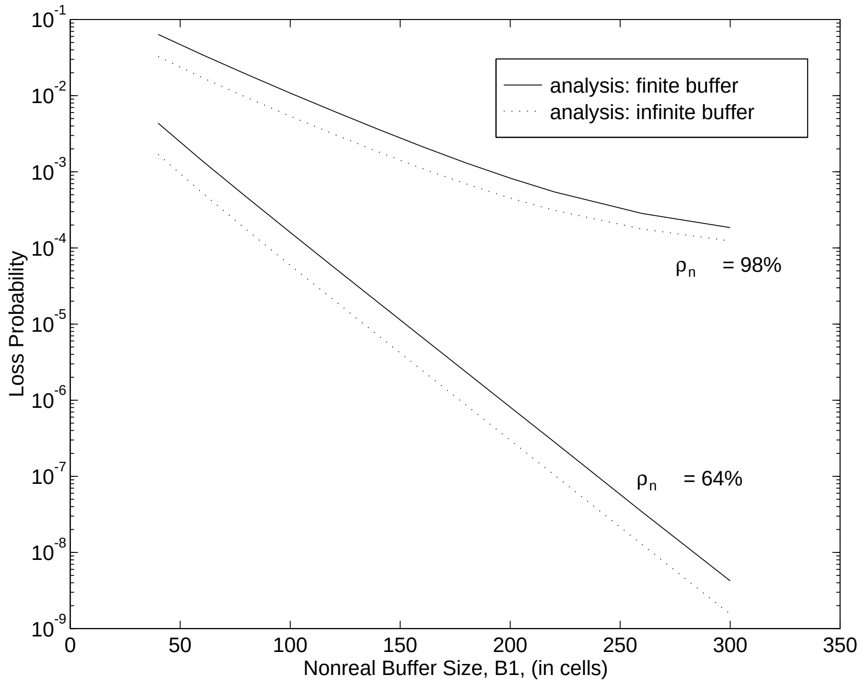

Figure 4: Loss probability of NRT cells versus the size of BNR for different ρ_{non} ($N_d = 3,\ N_{vo} = 6,\ B_2 = 100,\ T = 50,\ \gamma = 0.67,\ \xi = 0.2$).

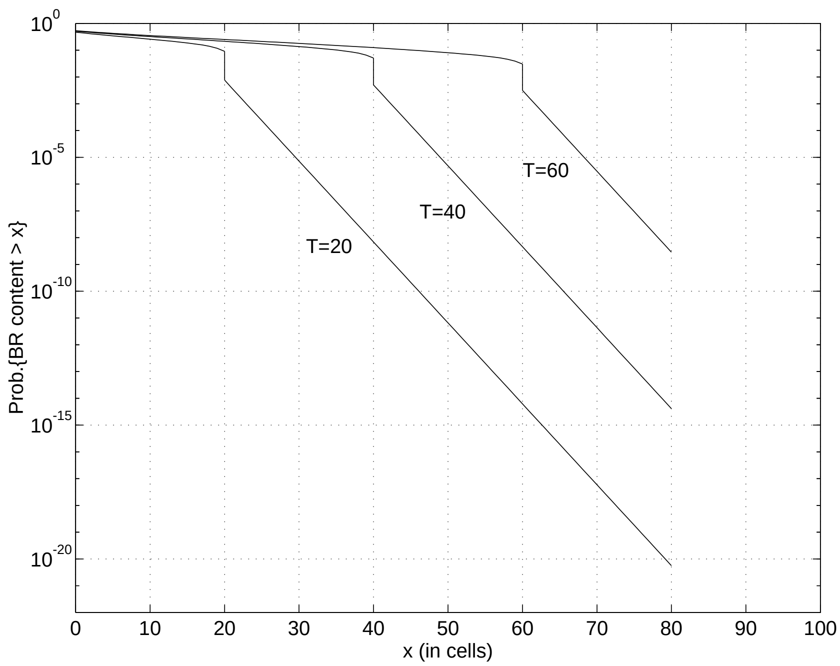

Figure 5: Complementary buffer distribution for BR for different T ($N_d = 3,\ N_{vo} = 6,\ \gamma = 0.67,\ \xi = 0.5,\ \rho_{real} = 80\%$).

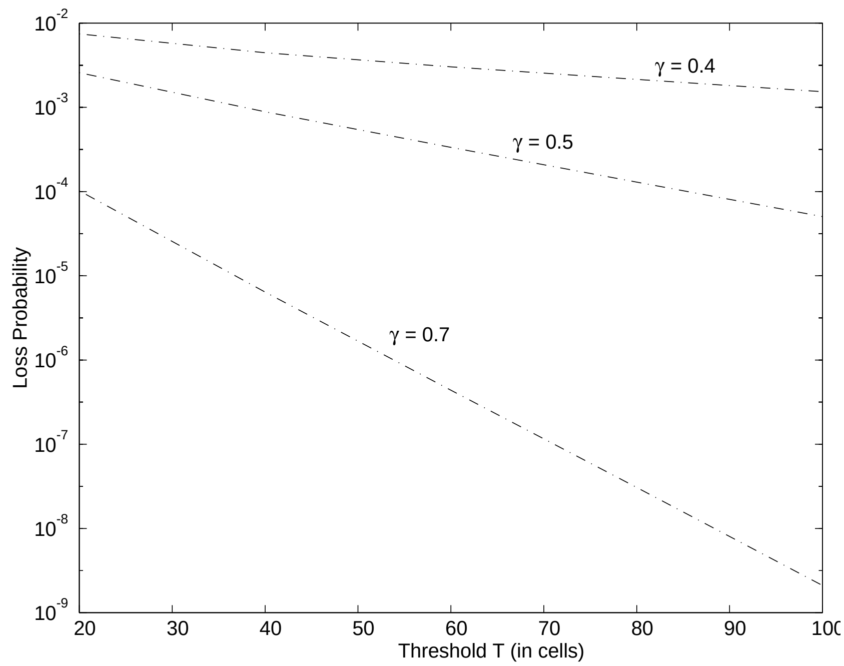

Figure 6: Loss probability of low priority RT cells versus threshold T for different γ ($N_d = 3,\ N_{vo} = 8,\ B_1 = 200,\ B_2 = 100,\ C = 218,\ \xi = 0.2$).

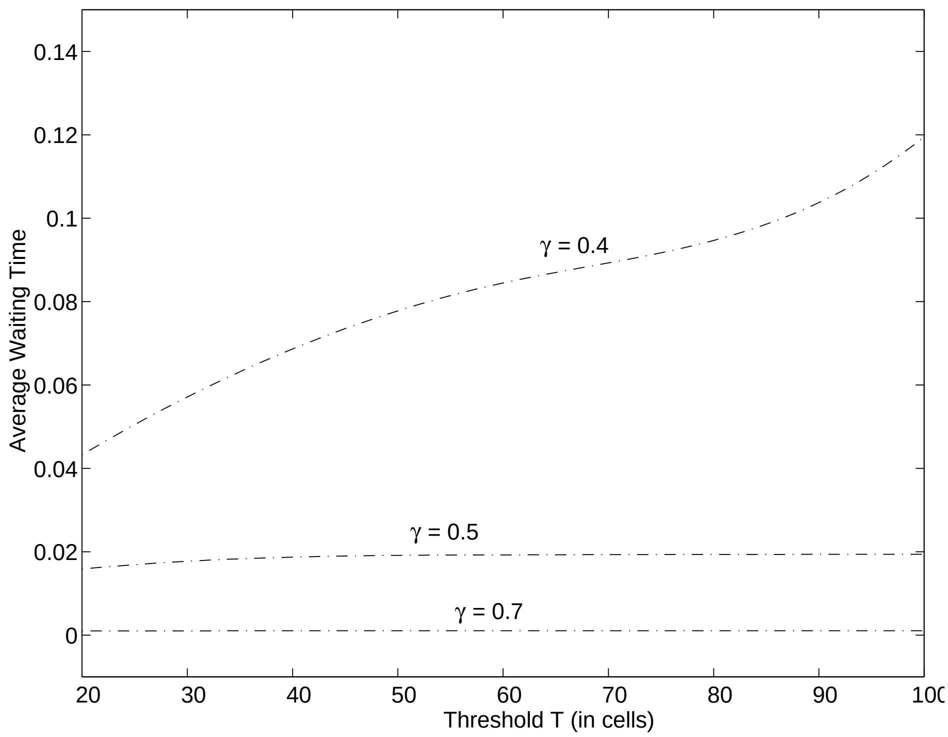

Figure 7: Average waiting time for high priority RT cells versus threshold T using different γ and same parameters of Figure 6.

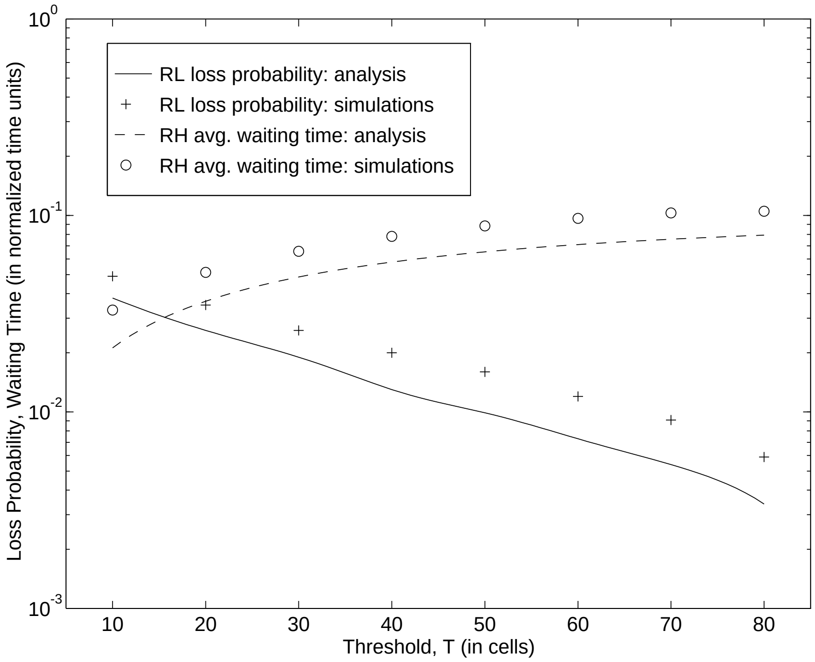

Figure 8: Loss probability for RL and average waiting time for RH versus T. Results are obtained based on analysis (infinite buffer) and simulations (finite buffer) using $N_d = 5,\ N_{vo} = 10,\ \rho_{real} = \frac{\overline{\lambda_2}}{\gamma C} = 75\%,\ B_1 = 200,\ B_2 = 80,\ \xi = 0.2$.

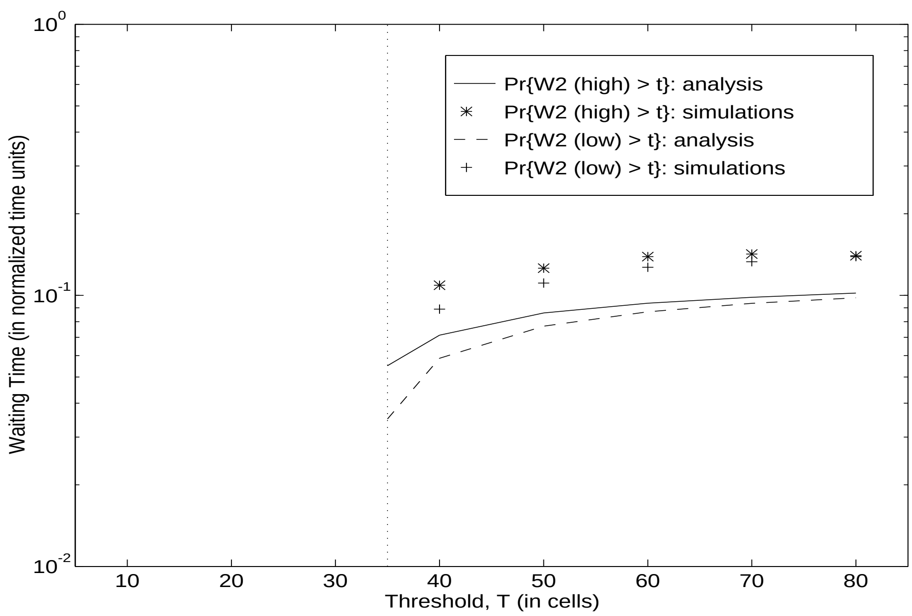

Figure 9: Probability that the waiting time for RT exceeds a fixed value $t = 0.2841$ time units (100 msec). Other parameters are as in Figure 8.

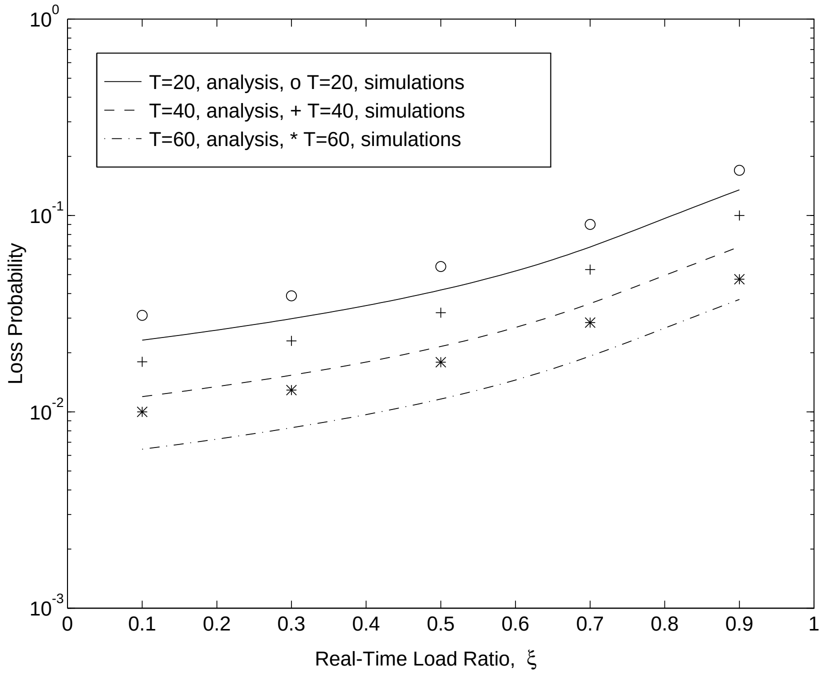

Figure 10: Loss probability of low priority RT cells versus ξ for different values of T and the same parameters as in Figure 8.

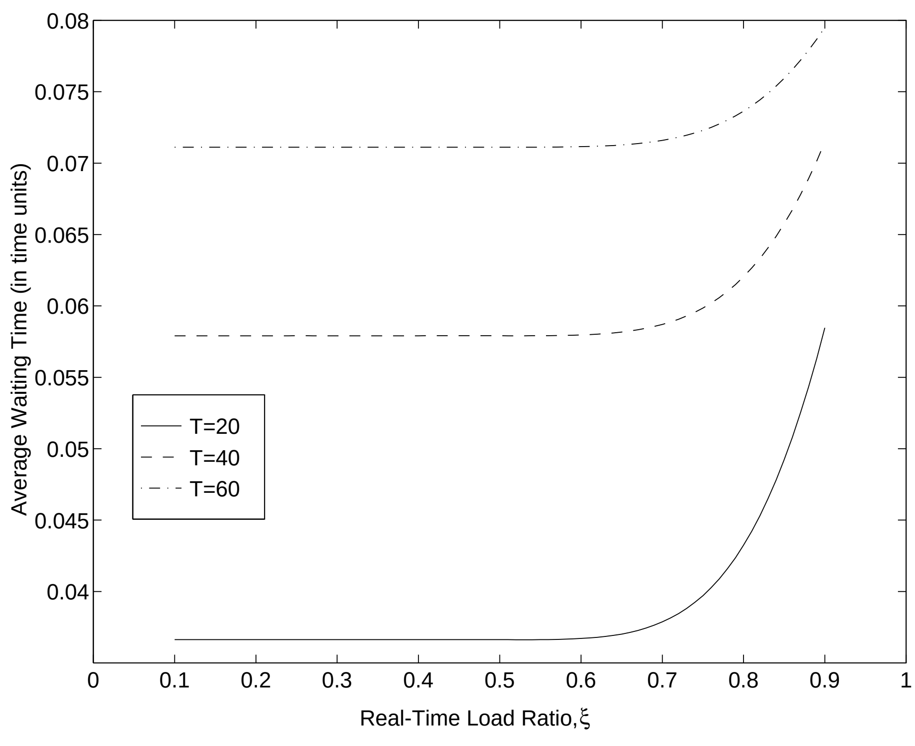

Figure 11: Average waiting time for high priority RT cells versus load ratio, ξ, for different T and the same parameters as in Figure 8 (analysical results are obtained based on infinite buffers).