Analysis of a Finite Buffer Queue with Heterogeneous Markov Modulated Arrival Processes: A Study of Traffic Burstiness and Priority Packet Discarding *

Jaime Bae Kim[†], Rahul Simha[⋆] and Tatsuya Suda[‡]
jkim@ms.secs.csun.edu, simha@cs.wm.edu, suda@ics.uci.edu

† *Department of Computer Science,*
California State University, Northridge,
Northridge, CA 91330

⋆ *Department of Computer Science,*
College of William and Mary,
Williamsburg, VA 23185

‡ *Department of Information and Computer Science,*
University of California, Irvine,
Irvine, CA 92717-3425

Abstract

This paper considers a single server queueing system with a finite buffer and multiple heterogeneous arrival streams. We focus on Markov Modulated Arrival processes with differing burstiness and investigate the loss of individual arrival streams when the burstiness parameters of the arrival streams are varied. In addition, we study the behavior of loss probabilities for a priority packet discarding scheme, a congestion control mechanism suitable for high-speed networks.

Several interesting numerical results are presented. We introduce a new characterization of an arrival stream, which we refer to as self-loss, and use this to qualitatively evaluate the effects of multiplexing bursty streams with non-bursty streams. The effectiveness of priority packet discarding is also investigated numerically. In addition, we study via simulation the conditional probability of packet loss given that one has occurred.

*This material is based upon work supported by the National Science Foundation under Grant No. NCR-8907909. This research is also supported in part by the University of California MICRO program, Nippon Steel Information and Communication Systems Inc. (ENICOM), Hitachi Ltd., Hitachi America, and Tokyo Electric Power Company.

1 Introduction

Most network architectures proposed for high-speed networks, such as the Asynchronous Transfer Mode (ATM) [2] and IBM's PARIS [8], are based on packet switching and explicitly permit packet loss in order to gain bandwidth efficiency. In these networks, it is important to predict whether a network can provide the necessary resources to meet grade-of-service (GOS) requirements (i.e., an acceptable level of packet loss) for each of several service classes multiplexed on a network.

In this paper, we study a queueing system wherein heterogeneous traffic streams (sessions) are multiplexed and investigate how the heterogeneity of the arrival streams, especially varying levels of burstiness, affects packet loss in individual streams. We consider the class of Markov Modulated Arrival (MMA) streams both in continuous time (a Markov Modulated Poisson Process or MMPP [14]) and in discrete-time (a Markov Modulated Bernoulli Process or MMBP) and model each arrival stream in our queueing system as an MMA. We present expressions for packet loss in individual streams both with and without a priority packet discarding control scheme.

In our numerical study, we introduce the notion of *self-loss* of a stream: the packet loss obtained when the stream is multiplexed with a parametrically identical stream. Using this self-loss notion, we observe an interesting characterization of individual packet loss when two heterogeneous streams are multiplexed: the stream with the smaller self-loss is always penalized and the bigger the difference between the two self-loss probabilities, the greater the penalty. Our numerical examples also provide a better understanding of the packet discarding control scheme. For example, our results indicate that the traffic characteristics of low priority streams have almost no effect on the effectiveness of the discarding scheme. Finally, we present some simulation results examining the effects of burstiness on correlation in packet loss.

While past research efforts have contributed to improved models of arrival processes and aggregate packet loss behavior in multiplexed streams [1, 14, 24], the practical consequences of multiplexing streams of *varying* burstiness to individual packet loss have not been sufficiently examined. We note that most past research on priority packet discarding (e.g., [4, 19, 21, 27]) is limited to a Poisson/Bernoulli arrival assumption. In [20], a similar control scheme is analyzed for M/PH/1/N and PH/M/1/N queueing systems; however, no analysis is presented for the discrete-time case, which introduces the extra complication of simultaneous arrivals. In [5], a discrete-time model, the MMBP(2)/D/1/K queue, is considered, and a priority packet discarding scheme is studied with homogeneous sources. In contrast, our aim is to assess the impact of varying burstiness on the effectiveness of packet discarding schemes.

We follow the stochastic integral approach presented in [29], a method of independent interest, to derive our individual packet loss expressions. In doing so, we re-derive expressions for the continuous-time case presented without proof in [23]. We note that in [23], the emphasis was on analyzing parcel overflow processes using a two-state MMPP approximation for a multi-state

MMPP. Numerical results presented therein focused on the accuracy of the approximation.

The rest of this paper is organized as follows. After giving an overview of the method in the next section, we derive in Section 3 packet loss probabilities for continuous and discrete time cases when two arrival streams are present. In Section 4, our analysis is extended to accommodate a priority packet discarding scheme for two arrival streams. In Section 5, our analysis is further extended to include $N(>2)$ heterogeneous input streams. Only the most complicated case, the discrete time case with priority packet discarding, is presented. (The manner of extension is similar in all the other cases. For the analyses for $N(>2)$ streams without packet discarding control in both continuous and discrete time, we refer the reader to our earlier work [3].) In Section 6, we present several numerical results and simulation results. Concluding remarks are given in Section 7.

2 Outline of Method

In this paper the stochastic integral technique proposed in [29] is used to obtain loss probabilities for Markov Modulated Arrival streams. As will be seen, the same technique uniformly provides exact packet loss probabilities for both continuous and discrete time versions of several varieties of the packet loss problem. In the following, a brief overview of this technique is presented.

Consider an arrival process to a finite buffer queueing system. Assume that the buffer size is $K-1$. Let $N(t)$ be the number of arrivals in $[0,t]$ and $Z(t)$ denote the number of packets in the queue at time t. Let $U(t)$ be an indicator function for the times at which the buffer is full, namely, $U(t)=1$ if and only if $Z(t-)=K$, and $U(t)=0$ otherwise. Then, the packet loss probability P_{loss} is given by [29, 34]:

$$P_{\text{loss}} = \lim_{t\to\infty} \frac{1}{N(t)} \int_0^t U(s)dN(s). \tag{1}$$

The analysis carried out in this paper is based on the following observation. Many stochastic processes such as MMA's have an associated *compensator* $\Lambda(t)$ such that the process $M(t) = N(t) - \Lambda(t)$ is a martingale (see [6] for a discussion). Then, under some regularity conditions, as shown in [29], the stochastic integral

$$R(t) = \int_0^t U(s)dM(s)$$

is also a martingale, with the property that $\lim_{t\to\infty} \frac{R(t)}{t} = 0$. The regularity conditions can be shown to hold for MMA's [29]. Given this limiting ratio and by rearranging terms in equation (1), we may write

$$P_{\text{loss}} = \frac{\phi}{r} \tag{2}$$

where $r = \lim_{t\to\infty} \frac{N(t)}{t}$ is the arrival rate and $\phi = \lim_{t\to\infty} \frac{1}{t} \int_0^t U(s) d\Lambda(s)$.

Let $Y(t)$ be the underlying Markov process which modulates the arrival process of an MMA. Let's assume that the process $Z(t)$ is also Markovian (by taking service times to be Markovian), and let π denote the limiting distribution of the Markov process $\{Y(t), Z(t)\}$. For an MMA, the limit ϕ depends on π. In the remainder of the paper we provide expressions for the packet loss probability P_{loss} by computing r and ϕ in both continuous and discrete time versions. Note that when the arrival process is Poisson with rate λ, the compensator is $\Lambda(t) = \lambda t$. Then $P_{\text{loss}} = \lim_{t\to\infty} \frac{1}{t} \int_0^t U(s) ds$, which gives the well-known result equating the loss probability with the probability that the buffer is full.

We will find the following two observations useful. Both are concerned with multiple arrival streams. First, consider n arrival streams multiplexed onto a finite buffer, each with associated counting process $N_i(t)$ and compensator $\Lambda_i(t)$. Then, $\Lambda(t) = \sum_{i=1}^n \Lambda_i(t)$ is the compensator for the overall counting process $N(t) = \sum_{i=1}^n N_i(t)$; this will allow us to compose expressions for individual packet loss into one for the overall packet loss. Second, the dependence of the loss probability of a particular stream on the characteristics of the other streams appears in the computation only via the the indicator function U.

3 Computation of Packet Loss Probabilities - Two Streams

In this section, we derive packet loss probabilities for continuous and discrete time versions of the problem in which two MMA streams are multiplexed onto a single finite buffer. Arriving packets that find the buffer full are lost. Priority-based packet discarding schemes are analyzed later in sections 4 and 5. For the continuous-time case, each arrival stream is modeled by a 2-state MMPP, and for the discrete-time case, it is modeled by a 2-state MMBP. Note that the 2-state MMPP or MMBP is a fairly general process – by selecting appropriate parameter values, it can represent a Poisson or geometric arrival process (suitable for data arrivals) and an Interrupted Poisson Process (IPP) or an Interrupted Bernoulli Process (IBP) (suitable for on-off traffic sources such as voice and still picture) [10, 13, 16, 22, 24, 33]. 2-state MMPP's have been used to represent a superposition of several identical sources [14], and thus, each arrival stream in our model can be viewed as a single source or as a superposition of multiple identical sources.

3.1 Continuous-Time Case

In this subsection, we consider the multiplexing of two heterogeneous streams (with different parameter values). Several interesting numerical results are derived from this simple case (presented in Section 7.1)

Consider a single first-come-first-served queueing system driven by two 2-state MMPP arrival processes (streams A and B). The queueing system has a finite buffer space of size $K-1$ packets (thus, the maximum system size is K). Service times of packets from streams A and B are exponentially distributed with rate μ. A 2-state MMPP is characterized by a 'driving' Markov process [14] that alternates between two states, spending an exponentially distributed amount of time in each. Packets are generated in each state according to a Poisson processes with a rate that is state-dependent. Additional notation associated with each arrival stream are (refer to Figure 1):

- *Stream A.* Let $N_A(t)$ denote the counting process and $Y_A(t)$ the driving Markov process for stream A. The states of $Y_A(t)$ are labeled 1 and 2; the transition rate from state 1 to state 2 is denoted by α_1, and the transition rate from state 2 to state 1 is denoted by β_1. Thus, for $Y_A(t)$, the sojourn times in states 1 and 2 are exponentially distributed with the mean $1/\alpha_1$ and $1/\beta_1$, respectively. Let r_A be the rate of arrivals from A.

- *Stream B.* Similarly, let $N_B(t)$ be the counting process, $Y_B(t)$ the driving Markov process, and r_B the arrival rate for stream B. The states of $Y_B(t)$ are labeled 3 and 4. α_2 and β_2 are the state transition rates of Y_B.

The generation of packets when the MMPP is in state i follows a Poisson process with rate λ_i. For $i = 1, 2$ and $j = 3, 4$, define the following indicator functions:

$$I_i(t) = \begin{cases} 1, & \text{if } Y_A(t) = i \\ 0, & \text{otherwise} \end{cases} \quad \text{and} \quad I_j(t) = \begin{cases} 1, & \text{if } Y_B(t) = j \\ 0, & \text{otherwise.} \end{cases}$$

Then, the aggregate arrival rate at time t is $\sum_{i=1}^{2} I_i(t)\lambda_i + \sum_{j=3}^{4} I_j(t)\lambda_j$. Let $Z(t)$ $(0 \leq Z(t) \leq K)$ denote the system state (the number of packets in the system) at time t. Define the following indicator function for system state q:

$$U_q(t) = \begin{cases} 1, & \text{if } Z(t-) = q \\ 0, & \text{otherwise.} \end{cases}$$

Let $N(t) = N_A(t) + N_B(t)$ be the cumulative number of arrivals in the time interval $[0, t]$, and let $\Lambda_A(t)$ and $\Lambda_B(t)$ denote the *compensators* for the processes $N_A(t)$ and $N_B(t)$, respectively. It is well-known that the compensators for $N_A(t)$ and $N_B(t)$ are given by [29]:

$$\Lambda_A(t) = \int_0^t (I_1(s)\lambda_1 + I_2(s)\lambda_2)ds, \qquad \text{and} \quad \Lambda_B(t) = \int_0^t (I_3(s)\lambda_3 + I_4(s)\lambda_4)ds. \tag{3}$$

Finally, define the following limiting probabilities. Let $\pi(i, j, q)$ $(i = 1, 2,\ j = 3, 4,\ 0 \leq q \leq K)$ be the limiting distribution for the Markov process $\{Y_A(t), Y_B(t), Z(t)\}$. Let $\pi(i, q)$ be the limiting distribution for the Markov process $\{Y_A(t), Z(t)\}$, and $\pi(j, q)$ be the limiting distribution for the Markov process $\{Y_B(t), Z(t)\}$. Note that $\sum_j \pi(i, j, q) = \pi(i, q)$ and $\sum_i \pi(i, j, q) = \pi(j, q)$.

In this analysis, we obtain the following probabilities:

- the long term probability $P_A(q)$ that an arrival from stream A sees the system in state q,
- the long term probability $P_B(q)$ that an arrival from stream B sees the system in state q, and
- the long term probability $P(q)$ that an arbitrary arrival sees the system in state q.

From these probabilities, we can easily obtain the loss probabilities for stream A ($P_{\text{loss}}(A) = P_A(K)$) and stream B ($P_{\text{loss}}(B) = P_B(K)$). Furthermore, the overall packet loss probability $P_{\text{loss}}(O)$ for the aggregated arrival process (i.e., the loss probability of packets, not distinguishing between streams A and B) is given by $P_{\text{loss}}(O) = P(K)$.

First, let us calculate the long term probability $P_A(q)$ for an arrival from stream A to see the system state q. From equation (2),

$$P_A(q) = \frac{1}{r_A} \lim_{t\to\infty} \frac{1}{t} \int_0^t U_q(s) d\Lambda_A(s). \tag{4}$$

Here,

$$r_A = \lim_{t\to\infty} \frac{N_A(t)}{t} = \frac{\lambda_1 \frac{1}{\alpha_1} + \lambda_2 \frac{1}{\beta_1}}{\frac{1}{\alpha_1} + \frac{1}{\beta_1}} = \frac{\lambda_1\beta_1 + \lambda_2\alpha_1}{\alpha_1 + \beta_1}. \tag{5}$$

Next,

$$\begin{aligned}
\lim_{t\to\infty} \frac{1}{t} \int_0^t U_q(s) d\Lambda_A(s) &= \lim_{t\to\infty} \frac{1}{t} \int_0^t U_q(s)(I_1(s)\lambda_1 + I_2(s)\lambda_2) ds \\
&= \lim_{t\to\infty} \frac{1}{t}\lambda_1 \int_0^t U_q(s) I_1(s) ds + \lim_{t\to\infty} \frac{1}{t}\lambda_2 \int_0^t U_q(s) I_2(s) ds \\
&= \lambda_1\pi(1,q) + \lambda_2\pi(2,q).
\end{aligned} \tag{6}$$

From Eqs. (4), (5) and (6), we obtain

$$P_A(q) = \frac{(\alpha_1 + \beta_1)(\lambda_1\pi(1,q) + \lambda_2\pi(2,q))}{\lambda_1\beta_1 + \lambda_2\alpha_1}. \tag{7}$$

Using the same argument for stream B, we obtain

$$P_B(q) = \frac{(\alpha_2 + \beta_2)(\lambda_3\pi(3,q) + \lambda_4\pi(4,q))}{\lambda_3\beta_2 + \lambda_4\alpha_2}. \tag{8}$$

Next, we compute the long term probability $P(q)$ of an arbitrary arrival seeing the system state q. From equation (2),

$$P(q) = \frac{1}{r_A + r_B} \lim_{t\to\infty} \frac{1}{t} \int_0^t U_q(s) d\Lambda(s) \tag{9}$$

where $\Lambda(s) = \Lambda_A(s) + \Lambda_B(s)$. Thus,

$$\begin{aligned}
\lim_{t\to\infty} \frac{1}{t}\int_0^t U_q(s)d\Lambda(s) &= \lim_{t\to\infty} \frac{1}{t}\int_0^t U_q(s)d(\Lambda_A(s)+\Lambda_B(s)) \\
&= \lim_{t\to\infty} \frac{1}{t}\int_0^t U_q(s)d\Lambda_A(s) + \lim_{t\to\infty} \frac{1}{t}\int_0^t U_q(s)d\Lambda_B(s) \\
&= \lambda_1\pi(1,q) + \lambda_2\pi(2,q) + \lambda_3\pi(3,q) + \lambda_4\pi(4,q) \qquad (10)
\end{aligned}$$

From Eqs. (9), (10) and (5), we obtain

$$P(q) = \frac{(\alpha_1+\beta_1)(\alpha_2+\beta_2)(\lambda_1\pi(1,q)+\lambda_2\pi(2,q)+\lambda_3\pi(3,q)+\lambda_4\pi(4,q))}{(\lambda_1\beta_1+\lambda_2\alpha_1)(\alpha_2+\beta_2)+(\lambda_3\beta_2+\lambda_4\alpha_2)(\alpha_1+\beta_1)}. \qquad (11)$$

3.2 Discrete-Time Case

In this subsection, the derivation technique used in the previous subsection for continuous time is applied to the discrete-time case. Again, we consider multiplexing two heterogeneous streams (streams A and B).

In discrete time, we must first decide the order in which arrivals and services take place and the times at which they occur. Without loss of generality, we assume the late arrival system with immediate access [17]. In such a system, arrivals occur just prior to the end of a time slot, and the packet in service is ejected from the service facility immediately after the beginning of a time slot (see Figure 2). An arriving packet enters the service facility if it is free, with the possibility of it being ejected almost instantaneously. Note that in this model, a packet's service time is counted as the number of slot boundaries from the point at which it enters the service facility to the point at which it departs. Therefore, even though we allow the arriving packet to be ejected almost instantaneously, its service time is counted as 1, not 0.

Consider a single first-come-first-served queue driven by two 2-state MMBP arrival processes. As in the continuous time case, the driving states of stream A are labeled 1 and 2 and the driving states of stream B are labeled 3 and 4. Without loss of generality, it is assumed that change in the states of the arrival processes occur just prior to the end of a time slot. The sojourn times in states 1 and 2 are geometrically distributed with means $1/\alpha_1$ and $1/\beta_1$ slots, and those of states 3 and 4 are geometrically distributed with means $1/\alpha_2$ and $1/\beta_2$. When in state i ($i = 1,2,3,4$), a MMBP generates packets according to a Bernoulli process with the probability of an arrival in a slot being p_i. Service times are geometrically distributed, and the probability of service completion in a slot, provided the server is busy, is s ($0 < s < 1$) for both streams.

Let $n = 0, 1, 2, 3, \ldots$ denote the slot boundary numbers and t continue to denote (continuous) time. Define the step function $\lfloor t \rfloor$, where $\lfloor t \rfloor = n$, if $n \le t < n+1$. For simplicity, we assume the slot length is equal to unit time in the system. We observe the system just prior to the end of time slots, i.e., at $n-$ (refer to Figure 2).

As in the continuous-time case, we let $Y_A(t)$ and $Y_B(t)$ be the states of the MMBP of streams A and B at time t. As before, $Z(t)$ $(0 \leq Z(t) \leq K)$ denotes the system state at time t and, $\pi(i, q)$, $\pi(j, q)$ and $\pi(i, j, q)$ are defined as the limiting probabilities of the Markov processes $\{Y_A(n-), Z(n-)\}$, $\{Y_B(n-), Z(n-)\}$, and $\{Y_A(n-), Y_B(n-), Z(n-)\}$, respectively. Note that $i = 1, 2$, $j = 3, 4$, and $0 \leq q \leq K$.

In discrete-time, a packet arriving to the system in state K is lost. Loss of packets can also happen when simultaneous arrivals occur from streams A and B when the system state is $K-1$. In this case, we must decide which packet to drop. The following scheme will be used: drop the packet from stream A with probability ϵ_A, and the packet from stream B with probability $\epsilon_B = 1 - \epsilon_A$. We will take $\epsilon_A = \epsilon_B = 0.5$.

In the following analysis, we focus on stream A and obtain its packet loss probability. Define the following indicator functions:

- $J(n)$: $J(n) = 1$ iff an arrival has taken place from stream B in the n^{th} slot.
- $V(n)$: $V(n) = 1$ iff in the n^{th} slot, a stream A packet is discarded when a stream B packet arrives (along with a stream A packet) at a time when the system is in state $K-1$.

Note that $V(1), V(2), \ldots$ is an *iid* sequence. Using $J(n)$ and $V(n)$, we can obtain the indicator function $U(n)$ for the state in which a stream A packet is discarded:

$$U(n) = U_K(n) + U_{K-1}(n)J(n)V(n) \tag{12}$$

where $U_q(n)$ is the indicator function for the system state q, i.e., $U_q(n) = 1$ iff $Z(n-) = q$.

First, let us derive the packet loss probability $P_{\text{loss}}(A)$ for stream A. From the definition of $P_{\text{loss}}(A)$, we have

$$P_{\text{loss}}(A) = \frac{1}{r_A} \lim_{t\to\infty} \frac{1}{t} \int_0^t U(s) d\Lambda_A(s). \tag{13}$$

Similar to the continuous time case, we have

$$r_A = \lim_{t\to\infty} \frac{N_A(t)}{t} = \frac{p_1\beta_1 + p_2\alpha_1}{\alpha_1 + \beta_1}. \tag{14}$$

In order to obtain the term $\lim_{t\to\infty} \frac{1}{t} \int_0^t U(s) d\Lambda_A(s)$, we will use the following manner of writing the compensator for a MMBP:

$$\Lambda_A(t) = \int_0^t (I_1(s)p_1 + I_2(s)p_2) d\lfloor s \rfloor. \tag{15}$$

(The following integral and sum are the same: $\int_0^t I_1(s)d\lfloor s\rfloor = \sum_{i=0}^{\lfloor t\rfloor} I_1(i)$.) Note that 'discreteness' is accounted for via the compensator. Thus,

$$\begin{aligned}
\lim_{t\to\infty}\frac{1}{t}\int_0^t U(s)d\Lambda_A(s) &= \lim_{t\to\infty}\frac{1}{t}\int_0^t U(s)(I_1(s)p_1 + I_2(s)p_2)d\lfloor s\rfloor \\
&= p_1\lim_{m\to\infty}\frac{1}{m}\sum_{n=0}^{m} U(n)I_1(n) + p_2\lim_{m\to\infty}\frac{1}{m}\sum_{n=0}^{m} U(n)I_2(n) \\
&= p_1\lim_{m\to\infty}\frac{1}{m}\sum_{n=0}^{m} U_K(n)I_1(n) + p_1\lim_{m\to\infty}\frac{1}{m}\sum_{n=0}^{m} U_{K-1}(n)J(n)V(n)I_1(n) \\
&\quad +p_2\lim_{m\to\infty}\frac{1}{m}\sum_{n=0}^{m} U_K(n)I_2(n) + p_2\lim_{m\to\infty}\frac{1}{m}\sum_{n=0}^{m} U_{K-1}(n)J(n)V(n)I_2(n) \\
&= p_1\pi(1,K) + p_1\lim_{m\to\infty}\frac{1}{m}\sum_{n=0}^{m} V(n)\lim_{m\to\infty}\frac{1}{m}\sum_{n=0}^{m} U_{K-1}(n)J(n)I_1(n) \\
&\quad +p_2\pi(2,K) + p_2\lim_{m\to\infty}\frac{1}{m}\sum_{n=0}^{m} V(n)\lim_{m\to\infty}\frac{1}{m}\sum_{n=0}^{m} U_{K-1}(n)J(n)I_2(n). \quad (16)
\end{aligned}$$

Observe that $\lim_{m\to\infty}\frac{1}{m}\sum_{n=0}^{m} V(n) = \epsilon_A$. Also note that $\lim_{m\to\infty}\frac{1}{m}\sum_{n=0}^{m} U_{K-1}(n)J(n)I_1(n)$ represents the limiting probability that system state is $K-1$, an arrival from stream B occurs, and stream A is in state 1. It is thus equal to $p_3\pi(1,3,K-1) + p_4\pi(1,4,K-1)$. Similarly, $\lim_{m\to\infty}\frac{1}{m}\sum_{n=0}^{m} U_{K-1}(n)J(n)I_2(n)$ is equal to $p_3\pi(2,3,K-1) + p_4\pi(2,4,K-1)$. Making these substitutions in equation (13), we obtain

$$P_{\text{loss}}(A) = \frac{(\alpha_1+\beta_1)}{(p_1\beta_1+p_2\alpha_1)}[p_1\pi(1,K) + p_2\pi(2,K) + \epsilon_A\sum_{i=1}^{2}\sum_{j=3}^{4} p_ip_j\pi(i,j,K-1)] \quad (17)$$

Similarly,

$$P_{\text{loss}}(B) = \frac{(\alpha_2+\beta_2)}{(p_3\beta_2+p_4\alpha_2)}[(p_3\pi(3,K) + p_4\pi(4,K) + \epsilon_B\sum_{i=1}^{2}\sum_{j=3}^{4} p_ip_j\pi(i,j,K-1)] \quad (18)$$

The overall loss probability $P_{\text{loss}}(O)$, the loss probability seen by an arbitrary arrival, is easily computed from equation (9).

Note that for a discrete-time analysis, the extra complication of simultaneous arrivals is accounted for quite easily with the stochastic integral approach used in our paper. Our analysis can easily be extended to accommodate a wide range of circumstances in which packets are lost according to the state of the system at the time of a packet's arrival. One such case, a priority packet discarding scheme [36, 35], a technique frequently proposed for high-speed networks, is analyzed in the following section.

4 A Threshold-Based Priority Packet Discarding Scheme - Two Streams

The analysis presented in Section 3 is now extended to accommodate a priority packet discarding scheme. In this section, we consider two arrival streams, and in the following section, our analysis is extended to include $N(>2)$ heterogeneous input streams.

By introducing priority levels among packets and discarding lower priority packets (during periods of congestion, for example), it is possible to trade off packet losses in different streams. We note that priority packet discarding is a popular congestion control technique for high-speed networks [18, 26, 35, 36] which allows network resources to be used more efficiently, thereby making it easier to satisfy GOS requirements of different classes of traffic. In general, loss-sensitive traffic such as data is given priority over loss-insensitive traffic such as voice. We observe that priority packet discarding is also suited for use with video traffic in conjunction with an embedded coding technique [12, 18]. With embedded coding, packets containing more important information are given higher priority than those containing less important information, and when network congestion occurs, packets containing less important information are discarded first.

In this section, a simple threshold-based discarding scheme [4] is considered. With this scheme, low priority packets are accepted only if the current system occupancy is less than a certain threshold, θ.

4.1 Continuous-Time Case

As we will see below, extending our results in Section 3 to priority-based packet discarding schemes simply involves writing down the appropriate indicator functions that capture the instances when packets are to be discarded. Recall that $P_A(q)$ ($P_B(q)$) denotes the long term probability that an arrival from stream A (B) sees the system in state q. Assume that stream A has a higher priority than stream B. Thus, packets from stream B are discarded when $Z(t) \geq \theta$. We have

$$P_{\text{loss}}(A) = P_A(K) \quad \text{and} \quad P_{\text{loss}}(B) = \sum_{q=\theta}^{K} P_B(q). \tag{19}$$

From Eqs.(7) and (19), we have

$$P_{\text{loss}}(A) = \frac{(\alpha_1+\beta_1)(\lambda_1\pi(1,K)+\lambda_2\pi(2,K))}{\lambda_1\beta_1+\lambda_2\alpha_1} \tag{20}$$

and

$$P_{\text{loss}}(B) = \frac{(\alpha_2+\beta_2)(\lambda_3\sum_{q=\theta}^{K}\pi(3,q)+\lambda_4\sum_{q=\theta}^{K}\pi(4,q))}{\lambda_3\beta_2+\lambda_4\alpha_2}. \tag{21}$$

4.2 Discrete-Time Case

Assume that stream A has a higher priority than stream B. To obtain the loss probability of high (low) priority stream, we need to compute $U(n)$, the indicator function for the state in which the high (low) priority packet is discarded. For the high priority stream A,

$$U(n) = U_K(n) \tag{22}$$

since arriving high priority packets are discarded only when the system size reaches K. For the low priority stream B,

$$U(n) = \sum_{q=\theta}^{K} U_q(n) \tag{23}$$

since arriving low priority packets are always discarded when the system occupancy is greater than or equal to θ. Thus, following the technique used earlier, we have

$$P_{\text{loss}}(A) = \frac{(\alpha_1 + \beta_1)(p_1\pi(1, K) + p_2\pi(2, K))}{p_1\beta_1 + p_2\alpha_1} \tag{24}$$

and

$$P_{\text{loss}}(B) = \frac{(\alpha_2 + \beta_2)(p_3 \sum_{q=\theta}^{K} \pi(3, q) + p_4 \sum_{q=\theta}^{K} \pi(4, q))}{p_3\beta_2 + p_4\alpha_2}. \tag{25}$$

5 $N(> 2)$ Arrival Streams with Priority Packet Discarding

In this section, our analysis of priority packet discarding is further extended to $N(> 2)$ heterogeneous input streams. Since different streams may have different priorities, we define the thresholds for stream 1, stream 2, stream 3, ..., to be $\theta_1, \theta_2, \theta_3, \ldots$, respectively. Without loss of generality, we assume that $1 \le \theta_1 \le \theta_2 \le \cdots \le \theta_N = K$, where K is the maximum system size. Further, it is assumed that if the current system occupancy is less than some threshold, no lower priority packets are discarded even if accepting all the lower priority packets exceeds the threshold, as opposed to only accepting lower priority packets up to the threshold.

Let q denote the system occupancy, and $\mathcal{N}(q)$ be the number of streams whose thresholds are greater than q. For the stream i, let $U(n)$ denote the indicator function for the state in which the stream i packet is discarded in the n^{th} slot, and define the following indicator functions for each slot n:

- $U_q(n)$: $U_q(n) = 1$ iff the system occupancy is q in the n^{th} slot.

- $J_l^H(n)$: $J_l^H(n) = 1$ iff there are H streams whose priorities are higher than or equal to that of stream i (including the stream i), and l arrivals have taken place from the remaining $H-1$ high priority streams. Note that, when $J_l^H(n) = 1$, a total of $l+1$ (including the arrival from the i^{th} stream) number of arrivals have taken place, and $J_l^H(n) = 0$ if $l \geq H$.

- $V_{l,q}(n)$: $V_{l,q}(n) = 1$ iff in the n^{th} slot, a stream i packet is discarded when a total of l number of arrivals occur at the system state q.

Therefore, when $\theta_i \leq K - N + 1$,

$$U(n) = \sum_{q=\theta_i}^{K} U_q(n). \tag{26}$$

In this case, no packet will be discarded when $q < \theta_i$.

When $\theta_i \geq K - N$, the i arriving packets will be discarded if $q \geq \theta_i$. However, even if $q < \theta_i$, the arriving stream i packets may be discarded if the number of arrivals from the streams whose thresholds are greater than q exceeds the available buffer space. Therefore, we have

$$U(n) = \sum_{q=\theta_i}^{K} U_q(n) + \sum_{q=K-N+1}^{\theta_i - 1} U_q(n) \sum_{l=K-q}^{\mathcal{N}(q)-1} J_l^{\mathcal{N}(q)}(n) V_{l+1,q}(n). \tag{27}$$

We assume that when the number of arrivals exceeds the available buffer space, the packets are discarded randomly. Therefore,

$$P[V_{l,q}(n) = 1] = \lim_{m \to \infty} \frac{1}{m} \sum_{n=0}^{m} V_{l,q}(n) = \frac{l - (K - q)}{l} \tag{28}$$

since out of l arriving packets, only $K - q$ number of packets will be accepted. The rest of the analysis follows the stochastic integral approach discussed earlier.

All that remains is to obtain the limiting probability π. For the continuous time case, the method presented in [23] can be used. For the discrete time case, we present an efficient computation method for a multiple arrival stream case. In the Appendix, our computation method is explained using four arrival streams (as an example).

6 Numerical Results

6.1 Characterizing Burstiness

In this subsection, the effects of traffic characteristics on the individual packet loss probabilities are investigated through numerical examples. Unless otherwise stated, our investigation focuses on two heterogeneous streams and the manner in which the burstiness of one stream affects the packet losses of both.

Burstiness is one of the most critical parameters determining the network performance. A number of ways have been proposed to describe the burstiness of a traffic source (see [2] for a discussion), with consensus yet to be reached concerning the most appropriate way. In keeping with our focus on Markov Modulated Arrivals, we examine the following three intuitive ways to vary the burstiness of a stream. In all three we keep the mean arrival rate of the stream *constant*.

Method 1: Keep the average sojourn times in the two driving states constant, and vary the arrival rates in two states. In this case, as the difference between the arrival rates in two states increases, the burstiness of the stream also increases.

Method 2: For an IPP (or IBP) stream, keep the ratio of the average active period (during which packets are generated) to the average idle period (during which no packets are generated) constant, and vary both active and idle periods. In this case, as the average active and idle periods increase, the burstiness of the stream also increases.

Method 3: For an IPP (or IBP) stream, keep the sum of the average active and idle periods constant, and vary the average active and idle periods. Since we keep the mean arrival rate constant, a smaller active period means a greater arrival rate during an active period. In this case, as the average active period decreases (equivalently, as the average idle period increases), the burstiness of the stream increases.

Note that Method 1 and Method 3 are two ways to vary peak-to-mean ratio. The peak-to-mean ratio is the most commonly used definition of burstiness [7, 11]. Method 2 varies the average active period. The average active period is also a widely used parameter to measure the degree of burstiness (see, for example [11, 15]).

In the following numerical examples, in order to characterize the effects of multiplexing bursty streams with non-bursty streams, we introduce the concept of "self-loss" for a (single) stream. *Self-loss* is the packet loss incurred when a stream is multiplexed with a with another parametrically identical but independent stream.

In the figures, the following notation is used to represent the loss probabilities:

- $P_{\mathrm{loss}}(A)$: the packet loss probability of stream A when it is multiplexed with stream B,
- $P_{\mathrm{loss}}(B)$: the packet loss probability of stream B when it is multiplexed with stream A,
- $P_{\mathrm{loss}}(O)$: the overall packet loss probability (i.e., the loss probability for all streams) when streams A and B are multiplexed,

- self-loss(A): the self-loss probability of stream A, and
- self-loss(B): the self-loss probability of stream B.

In the following, only results for the discrete time case are presented. (Similar results are obtained in continuous time.) The model discussed in section 3 is used. (Priority control will be treated later.) Recall that r_A and r_B denote the arrival rates of the two streams. The aggregated offered load of streams A and B, ρ, is given by

$$\rho = \frac{r_A + r_B}{s} \tag{29}$$

Throughout the numerical examples in this subsection, we assume the maximum system size $K = 10$ (i.e., buffer size = 9), the service rate $s = 0.8$ (see Section 3.2), and the offered load $\rho = 0.1$. The offered load is low in order to produce realistic loss probabilities (i.e., values of the order of 10^{-9}). We further assume that the mean arrival rates of two streams are the same. This allows us to investigate the effect of burstiness in isolation. Since $r_A = r_B$, $\rho = 0.1$ and $s = 0.8$, from Eq.(29) r_A and r_B become 0.04. In all the figures presented in this subsection, stream A is fixed, and we vary the burstiness of stream B according to the three different methods described earlier. Packets are randomly discarded when $Z(t) = K - 1$.

In Figures 3 through 5, stream A is assumed to follow a geometric arrival process (with $p_1 = p_2 = 0.04$), and the burstiness of stream B is increased using the the three methods. For Figure 3, p_3 and p_4 (arrival rates in the two driving states of stream B) are varied, keeping $\frac{1}{\alpha_2}$ and $\frac{1}{\beta_2}$ (the average sojourn times in the two states) constant. The bigger the difference between p_3 and p_4, the greater the burstiness of stream B. In this figure, the values of α_2 and β_2 are 0.01 and 0.19, respectively. At the leftmost point (i.e., when $p_4 - p_3 = 0$), stream B is geometric. In Figure 4, an IBP is used for stream B (i.e., $p_3 = 0$) in which p_4 is equal to 0.8. We vary $\frac{1}{\alpha_2}$ (the average idle period of stream B) and $\frac{1}{\beta_2}$ (the average active period of stream B), keeping their ratio $\frac{\beta_2}{\alpha_2}$ constant: $\frac{\beta_2}{\alpha_2} = 19$. In Figure 5, an IBP is used again for stream B (i.e., $p_3 = 0$), and the burstiness of stream B is varied using Method 3. In other words, we vary $\frac{1}{\beta_2}$ and $\frac{1}{\alpha_2}$, keeping $\frac{1}{\alpha_2} + \frac{1}{\beta_2}$ constant. Here, $\frac{1}{\alpha_2} + \frac{1}{\beta_2} = 100$. In all three figures burstiness increases along the x-axis.

Several observations can be made from Figures 3, 4 and 5. As the burstiness of stream B increases, the loss probabilities of both streams increase. Next, compare the $P_{\text{loss}}(A)$ curve with the self-loss(A) curve. We see that self-loss(A) is always smaller than $P_{\text{loss}}(A)$. Furthermore, stream B gains (i.e., $P_{\text{loss}}(B) <$ self-loss(B)) by sharing a buffer with a geometric stream (i.e., stream A), as opposed to sharing a buffer with another bursty stream (i.e., a stream identical to itself). Most importantly, the stream with the smaller self-loss probability is penalized. In this case, self-loss(A) is smaller than self-loss(B), and stream A is penalized (i.e., $P_{\text{loss}}(A) >$ self-loss(A)). Moreover, the bigger the difference between two self-loss probabilities, the greater the penalty.

In Figures 6 through 8, stream A is assumed to be a non-geometric, bursty stream. For stream A, we use $\alpha_1 = 0.01$, $\beta_1 = 0.19$, $p_1 = 0.02$ and $p_2 = 0.42$ in Figure 6; $\alpha_1 = 0.01$, $\beta_1 = 0.19$, $p_1 = 0$ and $p_2 = 0.8$ in Figure 7; and $\alpha_1 = \beta_1 = 0.02$, $p_1 = 0$, and $p_2 = 0.08$ in Figure 8. Again, stream A is fixed, and the burstiness of stream B is varied using the three methods described earlier and the parameter values in Figures 3 through 5. Figures 6 through 8 substantiate the observations made about Figures 3 through 5. An increase in the burstiness of one stream negatively affects both the stream itself and the other stream multiplexed. In particular, in Figure 6, when $p_4 - p_3 = 0.4$, the packet loss probabilities for stream A and stream B are the same since both streams become identical. When $p_4 - p_3 < 0.4$, stream A is more bursty than stream B, and self-loss(A) is larger than self-loss(B); the reverse is true when $p_4 - p_3 > 0.4$. In both cases, the stream with the smaller self-loss is penalized. Again, the bigger the difference between the two self-loss probabilities, the greater the penalty.

Note that the differences between individual loss probabilities are significant in all of the Figures (i.e., Figures 3 through 8). In our numerical examples, we assumed that the mean arrival rates of streams A and B are the same. However, the difference between the packet loss probabilities for stream A and B is often an order of magnitude or greater. This shows that the overall packet loss probability may not provide sufficient insight when heterogeneous traffic sources are multiplexed.

6.2 Effectiveness of Priority Packet Discarding

We now turn our attention to the effectiveness of the threshold-based priority packet discarding scheme. Figures 9 through 15 focus on two streams of varying characteristics. In these figures, we assume $K = 20, s = 0.8$ and $\rho = 0.75$, unless otherwise stated. We further assume that the mean arrival rates of the two streams are the same. Now, $r_A = r_B = 0.3$ since the offered load $\rho = 0.75$. Here, stream A is assumed to be a geometric arrival process (i.e., $p_1 = p_2 = 0.3$); stream B is bursty with $\alpha_2 = \beta_2 = 0.01$, $p_3 = 0.2$ and $p_4 = 0.4$, unless otherwise stated. It is assumed that stream A models loss-insensitive traffic such as voice and that stream B consists of loss-sensitive traffic such as video. Priority is thus given to stream B, and stream A packets are discarded when the system size is greater than or equal to a given threshold value θ. Where relevant, the GOS requirement of stream A is taken to be a packet loss probability of 10^{-3} and that of stream B is taken to be 10^{-9}. Packet loss probabilities of 10^{-3} and 10^{-9} are chosen since they represent GOS requirements of voice and video respectively [28, 35].

Figure 9 shows the individual loss probabilities of streams A and B for various values of K (system size). The horizontal axis represents θ values. Note that a system size of 20 can satisfy the GOS requirements of both streams A and B by setting the threshold value θ to 12. To meet the same

GOS requirements without priority packet discarding, the required value of K is 48, illustrating the effectiveness of using priority packet discarding. Figure 10 supports the same observation. In this figure, the bursty stream B is fixed, and the mean arrival rate of the geometric stream A is varied. The threshold value θ is set to 12. With priority packet discarding, we can accommodate stream A up to the mean arrival rate of 0.3 without violating GOS requirements of both streams A and B, whereas without priority packet discarding, the acceptable mean arrival rate of stream A is less than 0.13. Thus, with priority packet discarding, the required buffer space decreases or the acceptable offered load of low priority stream increases.

Figures 11 and 12 illustrate the effect of traffic characteristics (burstiness and offered load) of the high priority stream on the effectiveness of priority packet discarding. In Figure 11, the geometric stream A is fixed, and the burstiness of stream B is varied using Method 1 described in subsection 5.1; p_3 and p_4 are varied, keeping $\frac{1}{\alpha_2}$ and $\frac{1}{\beta_2}$ constant. In Figure 12, the mean arrival rate of the low priority stream (stream A) is fixed to 0.3, and the mean arrival rate of the high priority stream (stream B) is varied. $\alpha_2 = \beta_2 = 0.01$ and $\frac{p_4}{p_3} = 3$ in Figure 12. In both figures, the vertical axis shows $\frac{P_{\text{loss}}(B)}{P_{\text{loss}}(A)}$, the ratio of stream B packet loss probability to stream A packet loss probability. We see that this packet loss probability ratio is very sensitive to the traffic characteristics (burstiness and offered load) of high-priority stream; priority packet discarding becomes most effective when the high priority stream is not bursty and the offered load of high priority stream is the smallest.

Figures 13 and 14 illustrate the effect of traffic characteristics (burstiness and offered load) of low priority stream on the effectiveness of priority packet discarding. In Figure 13, both streams are bursty. The high priority stream (stream B) is fixed, and the burstiness of the low priority stream (stream A) is varied using Method 1. In Figure 14, the mean arrival rate of the low priority stream (stream A) is varied. Here, for stream B, we take $\alpha_2 = \beta_2 = 0.01$, $p_3 = 0.2$ and $p_4 = 0.4$. Figures 13 and 14 show that the traffic characteristics of low priority stream have almost no effect on the effectiveness of priority packet discarding.

Figure 15 shows the maximum acceptable (without violating the GOS requirements) offered load as a function of the GOS requirement of the high priority stream (stream B). In this figure, the same geometric stream is used for both stream A and stream B (i.e., $p_1 = p_2 = p_3 = p_4$). The GOS requirement of stream A is assumed to be a 10^{-3} packet loss probability, and the GOS requirement of stream B is varied. When no priority packet discarding is employed, the most stringent packet loss requirement needs to be satisfied. For the system with priority packet discarding, a threshold value which gives the maximum offered load is used. We see that the performance gain increases as the difference between the GOS requirements of stream A and stream B increases.

Figure 16 shows a numerical example for a system with 4 heterogeneous streams and a buffer size of 100. For each stream i $(i = 1, 2, 3, 4)$, we choose $\alpha_i = 0.01$ and $\beta_i = 0.1$. We let each stream have different burstiness: $p_1 = p_2$, $p_3 = 0.18$, $p_4 = 0.4$, $p_5 = 0.16$, $p_6 = 0.6$, $p_7 = 0.14$ and $p_8 = 0.8$.

The greater the burstiness of the stream, the higher the priority. Therefore, the stream 4 has the highest priority. We take $\theta_1 = 60$, $\theta_2 = 80$, and $\theta_3 = 96$. The streams 2, 3 and 4 are fixed, with the constant mean arrival rate 0.2 for each, and the average arrival rate of the geometric stream 1 is varied. The GOS requirement of stream 1 is taken to be a packet loss probability of 10^{-3}, that of stream 2, 10^{-5}, that of stream 3, 10^{-7} and that of stream 4, 10^{-9}. The same observation obtained in Figure 10 can be made; with priority packet discarding, the acceptable offered load increases. Without priority packet discarding, since the most stringent packet loss probability (i.e., 10^{-9}) has to be satisfied, the acceptable mean arrival rate of stream 1 is less than 0.07. On the other hand, with priority packet discarding, the acceptable mean arrival rate of stream 1 upto 0.23 can be accommodated without violating the GOS requirement of each stream.

6.3 Consecutive Packet Loss - A Simulation Study

For certain types of traffic, obtaining the long-term time-averaged loss probability may not be sufficient to predict the quality of service. Even when the long-term time-averaged value of loss rate is kept small, it is possible that a large number of packets are lost during congestion periods. In voice communication, for example, this burst loss of voice packets may cause noticeable performance degradation (clicks) at the destination. Therefore, it is important to investigate how losses are correlated.

We define the conditional loss probability as the probability of a packet being lost in the current slot given that one was lost in the previous slot. A loss period is a sequence of successive slots in each of which a packet is lost. The average loss period, defined as the average length of a loss period, is easily obtained from the conditional loss probability:

$$\text{avg. loss period} = \frac{1}{1 - \text{conditional loss probability}}. \tag{30}$$

The conditional loss probability (or the average loss period) is a measure of loss correlation.

We extend previous results [31, 32] on loss correlation to cases in which heterogeneous streams are present, one with the use of a packet discarding scheme and the other without. In this section, simulation results are presented to investigate the dependence of the loss correlation on various parameter values. As before, we consider two heterogeneous arrival streams, each modeled by MMBP.

First, loss correlation is investigated for a system without priority packet discarding. We assume that $s = 0.9$ and $r_A = r_B = 0.4$. Thus, the offered load ρ is approximately 0.89. This value is high in order to obtain meaningful simulation data without severe computational requirements.

In Figure 17, stream A is assumed be a geometric process, and the burstiness of stream B is increased using Method 1, namely, p_3 and p_4 are varied, keeping $\frac{1}{\alpha_2}$ and $\frac{1}{\beta_2}$ constant (here,

$\alpha_2 = \beta_2 = 0.1$. The system size K is taken to be 20. Figure 17 shows both conditional and unconditional loss probabilities for streams A and B as a function of the burstiness of stream B. We see that the effect of burstiness on the conditional loss probability is similar to that on the unconditional loss probabilities: an increase in the burstiness of one stream negatively affects both the stream itself and the other stream with which it is multiplexed. However, burstiness affects the unconditional loss probabilities more significantly than than it does the conditional loss probabilities. This suggests that an increase in the traffic burstiness does not really affect the average length of loss periods (see eq.(27)); rather, it creates more loss periods of approximately the same length.

Next, we investigate the loss correlation for the same system as in Figure 17, but with priority packet discarding. Figure 18 shows the average loss periods for streams A and B as a function of threshold value θ. Since $K = 20$, threshold value 20 corresponds to the case without priority packet discarding. For stream B, the following parameter values are used: $\alpha_2 = \beta_2 = 0.1$, $p_3 = 0$, and $p_4 = 0.8$. An interesting observation can be made from this figure: the average loss period appears to be insensitive to the threshold value chosen until a certain threshold size is reached. Therefore, even though stream A has lower priority, it may not suffer from severe consecutive losses as the threshold value decreases.

7 Summary

In this paper, we considered a queueing system with a finite buffer and heterogeneous arrival streams to investigate how heterogeneity affects packet loss in individual streams. We examined the class of MMA streams both in continuous time (a MMPP) and in discrete time (a MMBP) and presented an exact analysis of individual packet loss for MMA streams. We also derived individual packet loss probabilities for a priority packet discarding scheme. Our method of analysis is based on a stochastic integral approach, allowing both continuous- and discrete-time analyses as well as the analysis for the priority packet discarding scheme to be treated similarly.

Our first emphasis was on examining the effect of burstiness of traffic streams on packet loss. The concept of self-loss for a single stream was introduced for this purpose. The following summarizes our results:

- An increase in the burstiness of one stream results in an increase in the packet loss probabilities of the stream itself and of the other stream multiplexed together.
- When two different streams are multiplexed, the less bursty stream is always penalized, and the more bursty stream always benefits.
- When two different streams are multiplexed, the stream with the smaller self-loss probability

is penalized. The bigger the difference between two self-loss probabilities, the greater the penalty.

The importance of knowing the behavior of individual loss probabilities is underscored in applications like admission control [11, 15, 30]. Admission control decides whether to accept or reject a new call based on whether GOS requirements can be maintained. If the overall packet loss probability is used as a criterion in admission control when heterogeneous traffic sources are multiplexed, our results show that the GOS of the new call may not be guaranteed.

We also investigated the effectiveness of a priority packet discarding scheme and loss correlation. The results indicated that the high-priority stream played a more important role in characterizing loss. One interesting result on loss correlation was that the average loss period was not sensitive to the threshold for a range of threshold values.

8 Appendix

As in Section 5, we assume θ_1, θ_2, θ_3, and θ_4 $(0 < \theta_1 \leq \theta_2 \leq \theta_3 \leq \theta_4 = K\ )$ are the thresholds for stream 1, stream 2, stream 3, and stream 4 respectively. The transition probability matrix P for this system can be represented by:

In the above matrix P, the block matrix P_4 represents the case when the buffer size is less than θ_1. In this case, the arriving packets from all four streams can be accepted into the buffer. P_3 represents the case when the buffer size is greater than or equal to θ_1 but is less than θ_2. In this case, only the arriving packets from stream 2, stream 3, and stream 4 can be accepted into the buffer. P_2 and P_1 are similarly defined.

Each block matrix can further be represented by

$$\boldsymbol{P_4} = \begin{pmatrix} A_{00}B & A_1B & A_2B & A_3B & A_4B & 0 & \cdots & & & & & \\ A_{-1}B & A_0B & A_1B & A_2B & A_3B & A_4B & \cdots & & & & & \\ & & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & & & \\ & & & & & & A_{-1}B & A_0B & A_1B & A_2B & A_3B & A_4B \end{pmatrix}$$

$$\boldsymbol{P_3} = \begin{pmatrix} A'_{-1}B & A'_0B & A'_1B & A'_2B & A'_3B & 0 & \cdots & & & & \\ & A'_{-1}B & A'_0B & A'_1B & A'_2B & A'_3B & \cdots & & & & \\ & & \cdots & \cdots & \cdots & \cdots & \cdots & & & & \\ & & & & & & A'_{-1}B & A'_0B & A'_1B & A'_2B & A'_3B \end{pmatrix}$$

$$
\boldsymbol{P_2} = \begin{pmatrix}
A''_{-1}B & A''_0B & A''_1B & A''_2B & 0 & \cdots & & & & & \\
 & A''_{-1}B & A''_0B & A''_1B & A''_2B & \cdots & & & & & \\
 & & \cdots & \cdots & \cdots & \cdots & & & & & \\
 & & & & & & A''_{-1}B & A''_0B & A''_1B & A''_2B & \cdots
\end{pmatrix}
$$

$$
\boldsymbol{P_1} = \begin{pmatrix}
A'''_{-1}B & A'''_0B & A'''_1B & 0 & \cdots & & \\
 & A'''_{-1}B & A'''_0B & A'''_1B & \cdots & & \\
 & & \cdots & \cdots & \cdots & & \\
 & & & & & A''''_{-1}B & A''''_0B
\end{pmatrix}
$$

In the above matrices, the block matrix B represents the transition probabilities between the different traffic states and is given by the Kronecker product $B = B_1 \otimes B_2 \otimes B_3 \otimes B_4$, where B_i is given by

$$
B_i = \begin{pmatrix} 1-\alpha_i & \alpha_i \\ \beta_i & 1-\beta_i \end{pmatrix} \tag{31}
$$

The block matrices A_i can be written as

$$
\begin{aligned}
A_{00} &= P_{1a}P_{1d} + P_{0a} \\
A_{-1} &= P_{0a}P_{1d} \\
A_0 &= P_{0a}P_{0d} + P_{1a}P_{1d} \\
A_1 &= P_{1a}P_{0d} + P_{2a}P_{1d} \\
A_2 &= P_{2a}P_{0d} + P_{3a}P_{1d} \\
A_3 &= P_{3a}P_{0d} + P_{4a}P_{1d} \\
A_4 &= P_{4a}P_{0d} \\
A'_{-1} &= P'_{0a}P_{1d} \\
A'_0 &= P'_{0a}P_{0d} + P'_{1a}P_{1d} \\
A'_1 &= P'_{1a}P_{0d} + P'_{2a}P_{1d} \\
A'_2 &= P'_{2a}P_{0d} + P'_{3a}P_{1d} \\
A'_3 &= P'_{3a}P_{0d} \\
A''_{-1} &= P''_{0a}P_{1d} \\
A''_0 &= P''_{0a}P_{0d} + P''_{1a}P_{1d} \\
A''_1 &= P''_{1a}P_{0d} + P''_{2a}P_{1d} \\
A''_2 &= P''_{2a}P_{0d}
\end{aligned}
$$

$$
\begin{aligned}
A_{-1}''' &= P_{0a}'''P_{1d} \\
A_0''' &= P_{0a}'''P_{0d} + P_{1a}'''P_{1d} \\
A_1''' &= P_{1a}'''P_{0d} \\
A_0'''' &= P_{1d} \\
A_1'''' &= P_{0d}
\end{aligned}
$$

where P_{ia} represents the probability that i packets arrive at the system, and P_{id} represents the probability that i packets depart from the system.

We compute the π iteratively. Let $\pi^{(i)}$ be the limiting probability calculated at the i^{th} iteration.

1. Let $\pi^{(0)}$ be the initial estimation of the probability π. For simplicity, we take $\boldsymbol{\pi}^{(0)} = \frac{1}{16(K+1)}\boldsymbol{e}$, where $\boldsymbol{e}$ is the vector of all 1's and $16(K+1)$ is the number of states.

2. Start the iteration. The result of the $(i+1)^{th}$ iteration can be calculated from that of the i^{th} iteration by using $\boldsymbol{\pi}^{(i+1)} = \boldsymbol{\pi}^{(i)}\boldsymbol{P}$. Putting it into 16×16 block forms, we have

$$
\begin{aligned}
\pi_0^{(i+1)} &= (\pi_0^{(i)}A_{00} + \pi_1^{(i)}A_{-1})B \\
\pi_1^{(i+1)} &= (\pi_0^{(i)}A_1 + \pi_1^{(i)}A_0 + \pi_2^{(i)}A_{-1})B \\
\pi_2^{(i+1)} &= (\pi_0^{(i)}A_2 + \pi_1^{(i)}A_1 + \pi_2^{(i)}A_0 + \pi_3^{(i)}A_{-1})B \\
&\vdots \\
\pi_K^{(i+1)} &= (\pi_{K-1}^{(i)}A_1''' + \pi_K^{(i)}A_0'''')B
\end{aligned}
$$

   where $\pi^{(i)} = (\pi_0^{(i)}, \pi_1^{(i)}, \ldots, \pi_K^{(i)})$. Since all the matrices except B are diagonal, the number of operations required for each iteration is $O(K2^N)$, where N is the number of streams multiplexed.

3. Stop the iteration when $\| \boldsymbol{\pi}^{(k+1)} - \boldsymbol{\pi}^{(k)} \| < \delta$, where δ is the given degree of accuracy.

Note that the above iterative method is reasonably efficient when executed on a parallelized or vectorized machine[1], because there is no loop dependency within an iteration. Also, since $\pi^{(i)}$ remains normalized during each iteration, it eliminates the accumulated error introduced by re-normalizing the probabilities after each iteration.

[1]We used a Convex c240 in obtaining numerical results in Section 7.

References

[1] D. Anick, D. Mitra and M. M. Sondhi, "Stochastic Theory of a Data-handling System with Multiple Source," *Bell Syst. Tech. J.*,, Vol. 61, No. 8, pp.1871-1894, October 1982.

[2] J. J. Bae and T. Suda, "Survey of Traffic Control Schemes and Protocols in ATM Networks," *Proc. IEEE*, Vol.79, No.2, pp.170-189, February 1991.

[3] J. J. Bae, T. Suda and R. Simha, "Heterogeneous Arrival Streams, Burstiness and Packet Discarding: A Study of Individual Packet Loss," Technical Report, 91-58, Dept. of Info. & Comp. Sci., Univ. of Calif., Irvine, July 1991.

[4] K. Bala, I. Cidon and K. Sohraby, "Congestion Control for High Speed Packet Switched Networks," *Proc. IEEE INFOCOM'90*, pp.520-526.

[5] J-Y. Le Boudec, "An Efficient Solution Method for Markov Models of ATM Links with Loss Priorities," *IEEE J. on Select Areas in Commun.*, Vol.9, No.3, pp.408-417, April 1991.

[6] P. Bremaud, Point Processes and Queues: Martingale Dynamics, *Springer-Verlag*, New York, 1981.

[7] T. Y. Choi, "Statistical Multiplexing of Bursty Sources in an ATM Network," Multimedia '89.

[8] I. Cidon, I. Gopal, M. Gopal and S. Kutten, "Distributed Control in PARIS", *Proc. 9th Symposium on Principles of Distributed Computing*, Quebec, Canada, 1990.

[9] E. Cinlar, Introduction to Stochastic Processes, *Prentice Hall*, 1975.

[10] J.N. Daigle, Y. Lee and M.N. Magalhaes, "Discrete Time Queues with Phase Dependent Arrivals", *IEEE Trans. Commun.*, Vol. 32, No. 2, Feb 1994, pp. 606-614.

[11] G. Gallassi, G. Rigolio and L. Fratta, "ATM: Bandwidth Assignment and Bandwidth Enforcement Policies," *Proc. IEEE GLOBECOM'89*, pp.49.6.1-49.6.6.

[12] D. J. Goodman, "Embedded DPCM for variable bit rate transmission," *IEEE Trans. Commun.*, Vol.COM-28, pp.1040-1046, July 1980.

[13] O. Hashida, Y. Takahashi and S. Shimogawa, "Switched Batch Bernoulli Process and the Discrete Time SBBP/G/1 Queue", *IEEE J. Selec. Areas Commun.*, Vol. 9, No. 3, April 1991, pp. 394-401.

[14] H. Heffes and D. 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, September 1986.

[15] M. Hirano and N. Watanabe, "Characteristics of a Cell Multiplexer for Bursty ATM Traffic," *Proc. IEEE ICC'89*, pp.13.2.1-13.2.5.

[16] S.S. Huang, "Source Modeling for Packet Video", *Proc. ICC 88*, pp. 38.7.1-38.7.5, 1988.

[17] J. J. Hunter, Mathematical Techniques of Applied Probability, Vol.II, *Academic Press*, New York, pp.193, 1983.

[18] F. Kishino, K. Manabe, Y. Hayashi and H. Yasuda, "Variable Bit-Rate Coding of Video Signals for ATM Networks," *IEEE J. Select. Areas Commun.*, Vol.7, No.5, pp.801-806, June 1989.

[19] H. Kroner, "Comparative Performance Study of Space Priority Mechanisms for ATM Networks," *Proc. IEEE INFOCOM 90*, pp.1136-1143, 1990.

[20] S. Q. Li, "Overload Control in a Finite Message Storage Buffer," *IEEE Trans. on Commun.*, Vol.37, No.12, pp.1330-1338, December 1989.

[21] D. M. Lucantoni and S. P. Parekh, "Selective Cell Discard Mechanisms for a B-ISDN Congestion Control Architecture," *Proc. 7th Intl. Teletraffic Congress Seminar*, October 1990.

[22] B. Maglaris, D. Anastassiou, P. Sen, G. Karlsson and J. Robbins, "Performance Analysis of Statistical Multiplexing for Packet Video Sources", *Proc. GLOBECOM 87*, pp. 47-8.1-47.8.10, 1987.

[23] K. S. Meier-Hellstern, "The Analysis of a Queue Arising in Overflow Models," *IEEE Trans. Commun.*, Vol.37, No.4, pp.367-372, April 1989.

[24] 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, 1991.

[25] M. F. Neuts, Structured Stochastic Matrices of M/G/1 Type And Their Applications, *Marcel Dekker Inc.*, New York, 1989.

[26] D. W. Petr, L. A. DaSilva, Jr., and V. S. Frost, "Priority discarding of Speech in Integrated Packet Networks", *IEEE J. Select. Areas Commun.*, Vol.7, No.5, pp.644-659, June 1989.

[27] D. W. Petr and V. S. Frost, "Nested Threshold Cell Discarding for ATM Overload Control: Optimization Under Cell Loss Constraints", *Proc. IEEE INFOCOM 91*, pp.12A.4.1-12A.4.10, 1991.

[28] M. de Prycker, Asynchronous Transfer Mode: Solution for Broadband ISDN, *Ellis Horwood Ltd*, England, 1991.

[29] W. A. Rosenkrantz and R. Simha, "Some Theorems on Conditional Pasta: A Stochastic Integral Approach", *Operations Research Letters*, Vol. 11, April 1992, pp. 173-177.

[30] G. M. Woodruff, R. G. Rogers, and P. S. Richards, "A Congestion Control Framework for High-Speed Integrated Packetized Transport," *Proc. IEEE GLOBECOM 88*, pp.7.1.1-7.1.5, 1988.

[31] H. Schulzrinne and J. F. Kurose, "Distribution of the Loss Period for Some Queues in Continuous and Discrete Time," *IEEE INFOCOM 91*, pp.12C.1.1-12C.1.10, 1991.

[32] H. Schulzrinne, J. F. Kurose and D. F. Towsley, "Loss Correlation for Queues with Bursty Streams," *Proc ICC'92*, Chicago, 1992.

[33] P. Sen, N. Rikli and B. Maglaris, "Queueing Analysis for Packet Switched Video", *Proc. SPIE*, Vol. 845, 1987, pp. 440-445.

[34] R.W.Wolff, Stochastic Modeling and the Theory of Queues, *Prentice Hall*, 1989.

[35] G. M. Woodruff and R. Kositpaiboon, "Multimedia Traffic Management Principles for Guaranteed ATM Network Performance," *IEEE J. Select. Areas Commun.*, Vol.8, No.3, pp.437-446, April 1990.

[36] N. Yin, S-Q. Li and T. E. Stern, "Data Performance in an Integrated Packet Voice/Data System Using Voice Congestion Control," *Proc. IEEE GLOBECOM 88* , pp.16.4.1-16.4.5, 1988.