FLUID MODELS FOR THE ANALYSIS AND DESIGN OF STATISTICAL MULTIPLEXING WITH LOSS PRIORITIES ON MULTIPLE CLASSES OF BURSTY TRAFFIC

Anwar I. Elwalid *Debasis Mitra*

AT&T Bell Laboratories
Murray Hill, NJ 07974, USA

ABSTRACT. The paper gives the complete solution for a stochastic fluid model of statistical multiplexing with loss priorities in ATM-based Broadband-ISDN. In this model each Markov Modulated Fluid Source generates "priority" and "marked" cell streams which are bursty (i.e., correlated in time), mutually correlated and periodic during bursts. The output of many such sources are buffered and multiplexed for transmission. The loss priority is implemented by selectively discarding marked cells when the buffer content exceeds a threshold level. The equilibrium state distribution exhibits jumps, a feature not existent in prior fluid models. The computational complexity for 2-state sources is dominated by a single system of linear equations of dimension equal to twice the number of sources; in particular, the complexity is independent of buffer size. The complete delay distribution for each traffic class is obtained. The numerical results demonstrate the manner in which (i) the threshold level controls the trade-off between delay of the priority cells and the loss probability of the marked cells, and (ii) the buffer size controls the loss probability of the priority cells. The analysis is generalized to several priority classes of traffic; this extension has significant potential for real-time services.

1. INTRODUCTION

This paper is on models and analytical techniques for statistical multiplexing with loss priorities, a central element of the future Broadband Integrated Services Digital Network (B-ISDN) which will use the Asynchronous Transfer Mode (ATM). A complementary paper [EM91(a)] treats access regulation. The two papers are of independent interest; together they provide the basis for a complete analysis of the system shown in Figure 1.1 which combines access regulation and statistical multiplexing with loss priorities.

Traffic in the emerging high speed networks is expected to be characterized by high burstiness and high variability in the bit rates, as recent studies on video [MA88, KM89, GV91], packetized voice [PD89] and facsimile [CD89] have shown. Hence, efficiency considerations force multiplexing to be the core of ATM. The cell structure of ATM is geared for efficient and standardized multiplexing. "Loss priorities" is another key concept that is built into ATM standards [CC90, CP90, EL90]. Cells which are judged to be in violation of contracts between the network and users are "marked", typically carried and dropped only in the last resort. As such it represents a "soft" violation tagging process. The motivation for this stems from the realization that burstiness, due to its very statistical nature, is hard to characterize precisely and to police. Hence loss priorities is an important mechanism for achieving the dual goals of efficiency and fairness. Yet another point of view [AS91] advocates marking to segregate traffic from services which place a premium on low loss in contrast to low delay variance.

Statistical multiplexing has been studied under various analytic frameworks. One such framework is based on Markov Modulated Poisson Processes (MMPP). The technique of Heffes and Lucantoni [HL86], which relies on

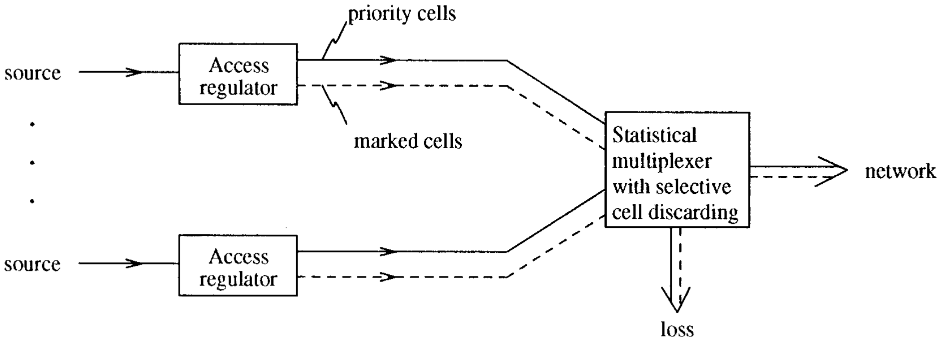


Figure 1.1: Integrated system of access regulation and statistical multiplexing

CH3133-6/92/0000-0415 $3.00 © 1992 IEEE

Matrix Geometric methods [NE81], is an example. Daigle and Lucantoni [DL90] compute the "rate matrix" from its spectral representation; they also point to the slow convergence of the conventional Matrix Geometric methods in conditions of high burstiness and traffic intensity, and also to the computational complexity in typical applications. Elwalid, Mitra and Stern [EMS91] develop algebraic techniques which give exact decompositions and efficient algorithms for computing spectral representations of solutions in the MMPP framework. See also [ID88]. A second framework relies on approximations by renewal processes and their characterizations by two moments. A prototypical work is that of Sriram and Whitt [SW86]. A third framework, which is adopted here, is based on stochastic fluid models [AM82, GL82, KO84, MI88, SE91, CI91]. The bursty traffic sources are modelled as Markov Modulated Fluid Sources in which the state of the controlling continuous time Markov chain determines the rate of fluid generation. As several authors have recently noted, these fluid models are well matched to the ATM environment at the burst level [DJ88, MA88, MG90, KO90, BC91, GH91, NR91]. There are several fundamental reasons; the small and uniform cell size (and hence constant service time) and the constant interarrival time of the cells in a burst at the time of generation ("periodicity") fit easily in the fluid framework and is difficult to handle in the queueing framework; the numerical complexity of solving fluid models with finite buffers does not depend on buffer size while with queueing model the complexity increases. The fluid approximation presumes a separation of time scales, i.e. the interarrival time of cells is small with respect to the time between changes in the rate, which is a feature of the high speed ATM environment. Several comparative evaluations of techniques for modelling and analyzing statistical multiplexing now exist [DL86, NK91, KT90]; these studies have found the approach in [AM82] which is based on stochastic fluid models, to be effective at the burst level in its accuracy and capacity to solve large systems.

The immediate precursor to this paper is [EM91(a)] which studies the access regulator (see Figure 1.1) and obtains results on (i) the 3-way trade-off between throughput, delay and burstiness of its output; (ii) the statistical characterization of the output streams of "priority" and "marked" cells. The characterization is in terms of another Markov Modulated Fluid Source which is novel in having coupling between the two traffic streams. This is the starting point of the present work. Specifically, in this model each state of the controlling Markov chain is associated with two cell generation rates, one for the priority cell stream and the other for the marked cell stream. We let G denote the generator of the controlling Markov chain, and $v_i^{(1)}$ and $v_i^{(2)}$ respectively denote the rates of generation of priority and marked cells when the source is in state i. (In our notational system classes 1 and 2 respectively refer to the priority and marked cells, and the class index is specified by the superscript in parenthesis.) Thus each Markov Modulated Fluid Source is here characterized by $(\mathbf{G}; \mathbf{v}^{(1)}, \mathbf{v}^{(2)})$, see Figure 1.2. Coupling captures correlations between streams. Such correlations typically exist; for example, in coded video during bursts both priority and marked cells are simultaneously generated at a high rate. The correlation represented by such coupling is believed to be important and new to the analysis of statistical multiplexing.

In this paper we focus on the simple class of sources with two states and let

$$\mathbf{G} = \begin{bmatrix} -\alpha & \alpha \\ \beta & -\beta \end{bmatrix} \qquad (1.1)$$

With appropriate selection of the state-dependent fluid generation rates we obtain the on-off sources with exponentially distributed on and off periods [AM82]; for such sources the off and on states are respectively indexed 1 and 2. The case of coupled Markov Modulated Fluid Sources with a larger number of states is considered in

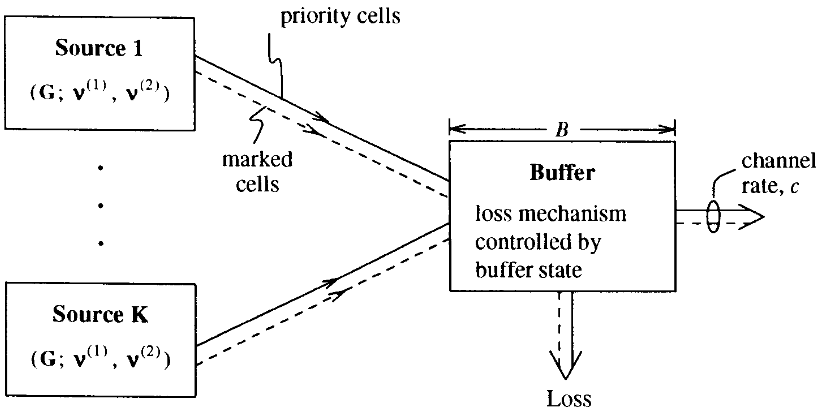


Figure 1.2: Statistical multiplexing with loss priorities of K Markov Modulated Fluid Sources. The case of 2 priority classes is depicted here. The extension to multiple priority classes is considered in Section 3.

[EM91(b)]; the treatment there is considerably more extensive than the one here.

The loss priority mechanism considered in Section 2 discards marked cells when the buffer occupancy exceeds the threshold B_1 $(B_1 \leq B)$. All incoming cells are accepted if the buffer occupancy is less than B_1 and priority cells are lost only if the buffer is full. When the buffer occupancy is exactly B_1, the rate at which marked cells are accepted equals the residual capacity, if any, of the output channel after carrying the priority cells. Service to cells already in the buffer is provided on a FCFS basis. Hence resequencing is not necessary. Note that the presence of marked cells in the buffer influence the delay seen by the priority cells. This influence is controlled by the choice of B_1. Thus the threshold B_1 controls the important trade-off between delay of the priority cells and the loss probability of the marked cells. The buffer size B determines the trade-off between the loss probability of the priority cells and their maximum delay. The choice of B will typically be sufficiently large for the loss probability of priority cells to be very small, and in this range the choice of B will have an insignificant effect on the performance of the marked cells.

Several recent papers have dealt with priority mechanisms for ATM [LP90, BF91, LB91, KH91]. The models and the analyses in these papers are quite different from the one considered here. Both Bonomi et al [BF91] and Le Boudec [LB91] work with discrete-time, finite state Markov processes and in each case the algorithms are computationally intensive. For instance, the exact algorithm in [LB91] requires a large matrix to be inverted at each value of n, where n indexes the buffer content in cells. In [BF91] priority is assigned by a Bernoulli process. Kroner et al [KH91] present only loss probabilities.

The loss priority mechanism considered here may be viewed as the analog of "trunk reservations", which is a key concept in circuit-switched communications [AK84].

In Section 3 we consider the natural extension of the loss priority mechanism from two traffic classes, priority and marked, to an arbitrary number. The motivation is as follows: if the traffic offered to nodal switches come with additional information (coded in the priority levels) on their relative importance in service quality, then during congestion periods, when cells have to be dropped, the switch can be precise in dropping only the least important. The degree of precision depends on the number of priority levels. Garrett and Vetterli [GV91] have provided evidence that a small number of priority classes is adequate for video and voice applications; the increased implementational complexity and diminishing returns does not justify many priority classes. Such forms of multiple priorities and loss control mechanisms constitute a shift from "source based" to "switch based" congestion control [JA91, GV91]. The former is reliant on feedback while the latter is open-loop and better matched to wide area networks with large propagation delays. Thus loss priorities with the number of priority classes small but typically more than two offers major advantages in real-time applications.

2. STATISTICAL MULTIPLEXING WITH TWO PRIORITY CLASSES

In this Section we consider the multiplexing of the output of K Markov Modulated Fluid Sources each with two states. The controlling Markov chain of each source is described by the generator **G** given in (1.1). When in state i $(i=1, 2)$ the source generates cells of the priority class at rate $v_i^{(1)}$, and marked cells at rate $v_i^{(2)}$. A simple special case of this model is obtained when $v_1^{(1)} = v_1^{(2)} = 0$ in which case state 1 may be called the "off" state and state 2 the "on" state.

We let Σ_t denote the state at time t of the aggregate of the K sources. In this paper it suffices to define Σ_t to be the number of sources in state 2. Thus in the particular case of on-off sources Σ_t denotes the number of on sources at time t. We denote the space of aggregate-source states by $\mathcal{S}$; hence $\mathcal{S} = \{0, 1, ..., K\}$ and $\Sigma_t \in \mathcal{S}$. It is convenient to define, in general,

$$\lambda_i^{(k)} \triangleq \text{rate of generation of class } k \text{ traffic, given } \Sigma_t = i$$
$$= i v_2^{(k)} + (K-i)\, v_1^{(k)} \qquad (2.1)$$

Let X_t denote the total fluid content of the buffer at time t. Since marked cells are selectively discarded when X_t exceeds B_1,

$$\frac{d}{dt} X_t = \left[v_2^{(1)} v_2^{(2)}\right] \Sigma_t + \left[v_1^{(1)} + v_1^{(2)}\right](K - \Sigma_t) - c \quad (0 < X_t < B_1)$$
$$= v_2^{(1)} \Sigma_t + v_1^{(1)} (K - \Sigma_t) - c \quad (B_1 < X_t < B) \qquad (2.2)$$

where c is the channel capacity. The behavior at the boundaries is especial. When $X_t = 0$,

$$\frac{d}{dt} X_t = \left[\left[v_2^{(1)} + v_2^{(2)}\right] \Sigma_t + \left[v_1^{(1)} + v_1^{(2)}\right](K - \Sigma_t) - c\right]^+ \qquad (2.3)$$

and when $X_t = B$,

$$\frac{d}{dt} X_t = \left[v_2^{(1)} \Sigma_t + v_1^{(1)} (K - \Sigma_t) - c\right]^- \qquad (2.4)$$

The behavior of dX_t/dt at $X_t = B_1$ is different and noteworthy. If

$$\left[v_2^{(1)} \Sigma_t + v_1^{(1)} (K - \Sigma_t) - c\right] \cdot \qquad (2.5)$$
$$\left[\left[v_2^{(1)} + v_2^{(2)}\right] \Sigma_t + \left[v_1^{(1)} + v_1^{(2)}\right](K - \Sigma_t) - c\right] > 0$$

i.e., the terms in brackets have the same sign, then (2.2) implies the important fact that dX_t/dt is nonzero and has the same sign for both $X_t < B_1$ and $X_t > B_1$. The sign is easily obtained:

$$\frac{d}{dt} X_t > 0 \ \text{ if } \ \left[\nu_2^{(1)} \Sigma_t + \nu_1^{(1)} (K - \Sigma_t) - c\right] > 0 \qquad (2.6,\text{i})$$

and,

$$\frac{d}{dt} X_t < 0 \text{ if} \qquad (2.6,\text{ii})$$

$$\left[\left[\nu_2^{(1)} + \nu_2^{(2)}\right] \Sigma_t + \left[\nu_1^{(1)} + \nu_1^{(2)}\right] (K - \Sigma_t) - c\right] < 0$$

On the other hand, if instead of (2.5),

$$\left[\nu_2^{(1)} \Sigma_t + \nu_1^{(1)} (K - \Sigma_t) - c\right] < 0 \qquad (2.7)$$

$$0 < \left[\left[\nu_2^{(1)} + \nu_2^{(2)}\right] \Sigma_t + \left[\nu_1^{(1)} + \nu_1^{(2)}\right] (K - \Sigma_t) - c\right]$$

then we have the following extraordinary situation implied by (2.2):

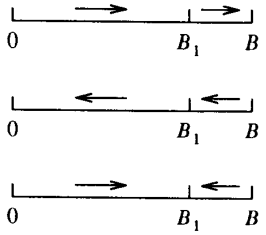
$$\frac{d}{dt} X_t > 0 \quad \text{if} \quad X_t < B_1 \qquad (2.8,\text{i})$$

$$< 0 \quad \text{if} \quad X_t > B_1 \qquad (2.8,\text{ii})$$

Hence, importantly,

$$\frac{d}{dt} X_t = 0 \ \text{ if } \ X_t = B_1 \qquad (2.8,\text{iii})$$

That is, for the particular aggregate-source states satisfying (2.7), there is a **confluence of drifts** at $X = B_1$. Notice that once $X_t = B_1$ and (2.7) hold, this condition persists until the aggregate-source state makes a transition to where (2.7) does not hold. As we shall see, the phenomenon of confluence of drifts and its consequence, the persistence of the buffer content at B_1, have an important effect on the stationary distribution of the system. In the following pictorial summary the drifts dX_t/dt in the three cases respectively identified in (2.6,i), (2.6,ii) and (2.7) are shown.



The state distribution of the system in equilibrium is given by

$$\pi_i(x) = \lim_{t \to \infty} \Pr\left[\Sigma_t = i,\ X_t \le x\right]. \quad (i \in \mathcal{S};\ 0 \le x \le B) (2.9)$$

(In what follows we drop the subscript t when specifying stationary distributions.)

Let $\boldsymbol{\pi}(x) = [\pi_0(x)\ \pi_1(x)\ \cdots\ \pi_K(x)]$. Following the procedure in, say, [AM82], the governing differential equations are readily obtained:

$$\boxed{\begin{aligned} \frac{d}{dx} \boldsymbol{\pi}(x) \mathbf{D}^{(0)} &= \boldsymbol{\pi}(x) \mathbf{M} \quad (0 < x < B_1) \\ \frac{d}{dx} \boldsymbol{\pi}(x) \mathbf{D}^{(1)} &= \boldsymbol{\pi}(x) \mathbf{M} \quad (B_1 < x < B) \end{aligned}} \qquad \begin{aligned} (2.10,\text{i}) \\ (2.10,\text{ii}) \end{aligned}$$

where,

$$\mathbf{D}^{(0)} = \text{diag}\left\{\lambda_0^{(1)} + \lambda_0^{(2)} - c,\ \lambda_1^{(1)} + \lambda_1^{(2)} - c,\ \ldots,\ \lambda_K^{(1)} + \lambda_K^{(2)} - c\right\} \qquad (2.11,\text{i})$$

and,

$$\mathbf{D}^{(1)} = \text{diag}\left\{\lambda_0^{(1)} - c,\ \lambda_1^{(1)} - c,\ \ldots,\ \lambda_K^{(1)} - c\right\}. \quad (2.11,\text{ii})$$

The i^{th} diagonal element of the matrix $\mathbf{D}^{(j)}$ is the drift or rate of change of the buffer content (away from the boundaries) when the aggregate-source state is i ($i = 0, 1, \ldots, K$) and the buffer occupancy level is j ($j = 0, 1$). (The convention we have adopted for the buffer occupancy level index is: $j = 0$ if $0 < X < B_1$, and $j = 1$ if $B_1 < X < B$. Note that the superscript in parenthesis may denote either the class or the buffer occupancy level; the context will make clear which.)

On substituting the expressions for $\{\lambda_i^{(k)}\}$ in (2.1) we obtain ($j = 0, 1$)

$$\mathbf{D}^{(j)} = \text{diag}\ \{-c^{(j)},\ \omega^{(j)} - c^{(j)},\ 2\omega^{(j)} - c^{(j)},\ \ldots,\ K\omega^{(j)} - c^{(j)}\} \qquad (2.12)$$

where,

$$\omega^{(j)} \triangleq \sum_{k=1}^{2-j} \left\{\nu_2^{(k)} - \nu_1^{(k)}\right\}, \quad c^{(j)} \triangleq c - K \sum_{k=1}^{2-j} \nu_1^{(k)}. \quad (2.13)$$

The form in (2.12) is important since it exposes the linear growth property of the diagonal elements of $\mathbf{D}^{(j)}$ for each j.

The matrix $\mathbf{M}$ in (2.10) is the tridiagonal generator of the birth and death process on $\{0, 1, \ldots, K\}$ which describes the aggregate-source process.

$$\begin{aligned} M_{ij} &= i\beta && (j = i-1) \\ &= -\{i\beta + (K-i)\alpha\} && (j = 1) \qquad (2.14) \\ &= (K-i)\alpha && (j = i+1) \end{aligned}$$

The system of differential equations in (2.10) for each individual level of buffer occupancy has been treated in [AM82]. Following this treatment we obtain

$$\boxed{\begin{aligned} \boldsymbol{\pi}(x) = \boldsymbol{\pi}^{(0)}(x) &= \sum_{i=0}^{K} a_i^{(0)}\ \boldsymbol{\phi}_i^{(0)}\ \exp(z_i^{(0)} x) \quad (0 \le x < B_1) \\ & \qquad\qquad\qquad\qquad\qquad\qquad (2.15) \\ = \boldsymbol{\pi}^{(1)}(x) &= \sum_{i=0}^{K} a_i^{(1)}\ \boldsymbol{\phi}_i^{(1)}\ \exp(z_i^{(1)} x) \quad (B_1 < x \le B) \end{aligned}}$$

Here $\{z_i^{(j)}, \boldsymbol{\phi}_i^{(j)}\}$ are solutions to two $(j=0, 1)$ sets of eigenvalue problems:

$$z_i^{(j)}\boldsymbol{\phi}_i^{(j)}\mathbf{D}^{(j)} = \boldsymbol{\phi}_i^{(j)}\mathbf{M} \quad (i = 0, 1, \ldots, K) \quad (2.16)$$

Closed-form solutions for these eigenvalues and eigenvectors have been given in [AM82] and we will assume that these are known. These results exploit the linear growth property in the diagonal elements of $\mathbf{D}^{(j)}$ and the linearity of the birth and death rates in $\mathbf{M}$, see (2.14). These results have been extensively extended in [KO84, MI88, SE91, EMS91, EM91(b)].

In (2.15) the coefficients $\{a_i^{(j)}\}$ are obtained from the boundary conditions. We consider next the composition of the equations which express the boundary conditions. Let $\mathcal{S}_D^{(j)}$ denote the set of aggregate-source states which give a downward drift to the buffer content when the level of buffer occupancy is j; similarly, let $\mathcal{S}_U^{(j)}$ be the set of aggregate-source states giving an upward drift. That is, for $j = 0, 1$,

$$\mathcal{S}_D^{(j)} = \{i|D_{ii}^{(j)} < 0\}, \quad \mathcal{S}_U^{(j)} = \{i|D_{ii}^{(j)} > 0\} \quad (2.17)$$

(We make the inessential but simplifying assumption that there exists no i such that $D_{ii}^{(j)} = 0$; [MI88] has shown how exceptions may be handled.) A little thought shows that $\Sigma_t \in \mathcal{S}_U^{(0)} \cap \mathcal{S}_D^{(1)}$ is equivalent to the condition in (2.7). As shown in (2.8), for such aggregate-source states there is a confluence of drifts at $X = B_1$. As a consequence there is probability mass accumulation at B_1 which causes $\pi_i(x)$ to have a discontinuity, i.e. a jump, at $x = B_1$ for $i \in \mathcal{S}_U^{(0)} \cap \mathcal{S}_D^{(1)}$. In contrast, as (2.6) shows,

$$\Sigma_t \in \mathcal{S}_D^{(0)} \quad \text{implies} \quad \frac{d}{dt}X_t < 0, \quad (2.18,\text{i})$$

$$\Sigma_t \in \mathcal{S}_U^{(1)} \quad \text{implies} \quad \frac{d}{dt}X_t > 0, \quad (2.18,\text{ii})$$

for all values of X_t. Hence in these cases there is no confluence of drifts; consequently, $\pi_i(x)$ is continuous at $x = B_1$ for i in either $\mathcal{S}_D^{(0)}$ or $\mathcal{S}_U^{(1)}$.

We can now give the complete set of boundary conditions.

$$\boxed{\begin{array}{ll} \pi_i^{(0)}(0) = 0 & \left[i \in \mathcal{S}_U^{(0)}\right] \\ \pi_i^{(0)}(B_1) = \pi_i^{(1)}(B_1) & \left[i \in \mathcal{S}_D^{(0)} \cup \mathcal{S}_U^{(1)}\right] \\ \pi_i^{(1)}(B) = p_i & \left[i \in \mathcal{S}_D^{(1)}\right] \end{array}} \quad \begin{array}{l} (2.19,\text{i}) \\ (2.19,\text{ii}) \\ (2.19,\text{iii}) \end{array}$$

where $\mathbf{p}$ is the stationary aggregate-source distribution, i.e. $\mathbf{pM} = \mathbf{0}$, $\langle \mathbf{p}, \mathbf{1} \rangle = 1$:

$$p_i = \begin{bmatrix} K \\ i \end{bmatrix} \frac{\alpha^i \beta^{K-i}}{(\alpha+\beta)^K} . \quad (0 \le i \le K) \quad (2.20)$$

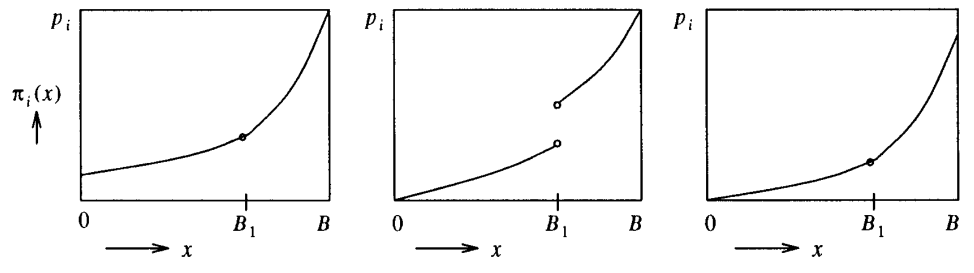
The boundary conditions in (2.19,i) and (2.19,iii) follow from familiar arguments, see for example [MI88], derived from (2.3) and (2.4) describing the physical behavior when the buffer is empty and full, respectively.

The boundary conditions in (2.19) form a system of $2(K+1)$ equations. On substituting the expressions for $\{\pi_i^{(0)}(x)\}$ and $\{\pi_i^{(1)}(x)\}$ in (2.15) we obtain a system of linear equations in the $2(K+1)$ coefficients $\{a_i^{(j)}\}$ which has to be solved numerically. Figure 2.1 sketches the resulting three categories of distributions.

Since the design of the statistical multiplexer typically aims to give very low loss probabilities for the priority packets, it is quite reasonable to assume that $B = \infty$ and then to estimate this loss probability by the probability that an arriving priority cell finds the buffer content greater than B. This procedure is further explained at the end of this Section. It has the advantage of giving a smaller set of linear equations to be solved. This is because $z_i^{(1)} > 0$ implies at once that $a_i^{(1)} = 0$. However no such reduction applies for $\{a_i^{(0)}\}$.

The following are probabilistic interpretations of the jumps in the distribution $\pi_i(x)$ at $x = B_1$:

$$\Pr(\Sigma = i, X = B_1) = \pi_i^{(1)}(B_1) - \pi_i^{(0)}(B_1) \quad (2.21)$$



$i \in \mathcal{S}_D^{(0)}$ $\quad$ $i \in \mathcal{S}_U^{(0)} \cap \mathcal{S}_D^{(1)}$ $\quad$ $i \in \mathcal{S}_U^{(1)}$

Figure 2.1: Sketches of state distributions obtained for two classes and levels of buffer occupancy.

Note that

$$\Pr(\Sigma=i,\ X=B_1) > 0 \qquad (i\in \mathcal{S}_U^{(0)} \cap \mathcal{S}_D^{(1)})$$
$$= 0 \qquad (i\in \mathcal{S}_D^{(0)} \cup \mathcal{S}_U^{(1)}) \qquad (2.22)$$

The probability atom at $X=B_1$ is due to the confluence of drifts there. Note also that
$\Pr$ $(\Sigma=i$, buffer empty$)$ $= \pi_i^{(0)}(0)$, and
$\Pr$ $(\Sigma=i$, buffer full$)$ $= p_i-\pi_i^{(1)}(B)$.

Once the steady state distributions have been computed, it is straightforward to obtain throughput and delay statistics. First consider $T^{(1)}$, the throughput of priority cells:

$T^{(1)}$, throughput of priority cells
= (rate of generation of priority cells)
− (rate of lost priority cells) .

Now, the rate of generation of priority cells = $\sum_{i=0}^{K} p_i\lambda_i^{(1)}$. Also,

rate of lost priority cells = $\sum_{i=0}^{K} \{\lambda_i^{(1)}-c\}\{p_i-\pi_i^{(1)}(B)\}$.

Hence,

$$T^{(1)} = \sum_{i=0}^{K} \lambda_i^{(1)}\pi_i^{(1)}(B) + c\left[1 - \sum_{i=0}^{K} \pi_i^{(1)}(B)\right] \qquad (2.23)$$

Next consider $T^{(2)}$, the throughput of the marked cells. The rate of lost marked cells has two components: the first accounts for $X_t > B_1$, in which case all marked cells which are generated are lost, and the second is due to $X_t=B_1$ in which condition the rate of loss of marked cells is $(\lambda_i^{(1)}+\lambda_i^{(2)}-c)$ for aggregate-source state i.

rate of lost marked cells

$$= \sum_{i=0}^{K} \{p_i-\pi_i^{(1)}(B_1)\}\,\lambda_i^{(2)}$$
$$+ \sum_{i\in \mathcal{S}_U^{(0)}\cap \mathcal{S}_D^{(1)}} \{\pi_i^{(1)}(B_1)-\pi_i^{(0)}(B_1)\}\{\lambda_i^{(1)}+\lambda_i^{(2)}-c\}$$

Hence,

$$T^{(2)} = \sum_{i=0}^{K} \lambda_i^{(2)}\pi_i^{(1)}(B_1) \qquad (2.24)$$
$$- \sum_{i\in \mathcal{S}_U^{(0)}\cap \mathcal{S}_D^{(1)}} \{\lambda_i^{(1)}+\lambda_i^{(2)}-c\}\{\pi_i^{(1)}(B_1)-\pi_i^{(0)}(B_1)\}$$

The loss probabilities for the different classes are easily obtained from their respective throughputs:

$$L^{(j)}, \text{ loss probability of class } j \text{ cells} = 1-T^{(j)}\Big/\sum_{i=0}^{K} \lambda_i^{(j)} p_i \qquad (2.25)$$

Finally consider the distribution of delay seen by arriving cells of the priority class. The delay distribution in the fluid model is easily obtained from the buffer content distribution seen by arriving priority cells.

$W^{(1)}(t)$, $\Pr$ (priority cell delay$\le t$)

$$= \frac{1}{T^{(1)}} \sum_{i=0}^{K} \lambda_i^{(1)}\pi_i^{(0)}(ct) \qquad (0\le t<B_1/c) \quad (2.26,\text{i})$$

$$= \frac{1}{T^{(1)}} \sum_{i=0}^{K} \lambda_i^{(1)} \Pr(\Sigma=i,\ X=B_1) \quad (t=B_1/c) \quad (2.26,\text{ii})$$

$$= \frac{1}{T^{(1)}} \sum_{i=0}^{K} \lambda_i^{(1)}\pi_i^{(1)}(ct) \qquad (B_1/c<t<B/c) \quad (2.26,\text{iii})$$

Finally,

$$\Pr\left[\text{priority cell delay} = \frac{B}{c}\right] = \frac{c}{T^{(1)}}\left[1-\sum_{i=0}^{K} \pi_i^{(1)}(B)\right] \qquad (2.26,\text{iv})$$

The expression for $\Pr(\Sigma=i,\ X=B_1)$ has been given earlier in (2.21). The distribution of delay experienced by the marked cells, $W^{(2)}(t)$, is similarly obtained.

Now consider the modification to the above formula for the case of an infinite buffer. As explained earlier, this case is of interest since its solution is easier to compute and it can typically be used to obtain sharp estimates of the loss and delay statistics for given, finite B. In (2.26) the throughput $T^{(1)}$ is now given by the rate at which priority cells are generated, $\Sigma\, p_i\lambda_i^{(1)}$. Also note that the expression for the estimate of $\Pr$ $(\Sigma=i$, buffer full) is the same as in the case of a finite buffer, namely, $\{p_i-\pi_i^{(1)}(B)\}$.

Let $\rho^{(j)}$ denote the traffic intensity at buffer occupancy level j, i.e.,

$$\rho^{(0)} = \frac{1}{c}\sum_{i=0}^{K} \{\lambda_i^{(1)}+\lambda_i^{(2)}\}\, p_i \qquad (2.27,\text{i})$$
$$= \frac{1}{c}\,\frac{K}{\alpha+\beta}\left[\beta\{\nu_1^{(1)}+\nu_1^{(2)}\} + \alpha\{\nu_2^{(1)}+\nu_2^{(2)}\}\right]$$

and,
$$\rho^{(1)} = \frac{1}{c}\sum_{i=0}^{K} \lambda_i^{(1)}\, p_i \qquad (2.27,\text{ii})$$
$$= \frac{1}{c}\,\frac{K}{\alpha+\beta}\left[\beta\nu_1^{(1)}+\alpha\nu_2^{(1)}\right] .$$

In the case of an infinite buffer stability requires that $\rho^{(1)}<1$.

3. SEVERAL LOSS PRIORITIES

Here the results in the preceding Section are extended to the case where each source generates traffic of an arbitrary number, say J, of priority classes. As before, each source has two states with generator G given in (1.1) and, also, the traffic streams are coupled, i.e. mutually correlated. Loss priority is implemented by partitioning the buffer into as many levels of occupancy as there are priority classes; cells of a particular priority class are either admitted or discarded depending upon the prevailing level of buffer occupancy. Each source when in state i $(i=1,2)$ generates traffic of class j $(j=1,2,\ldots,J)$ at rate $\nu_i^{(j)}$. Hence each of the K

sources is specified by $(\mathbf{G};\, \mathbf{v}^{(1)}, \ldots, \mathbf{v}^{(J)})$ where $\mathbf{v}^{(j)} = [v_1^{(j)}, v_2^{(j)}]$. The model in the preceding Section is obtained when $J=2$ with traffic classes 1 and 2 respectively denoting priority and marked cell streams.

The levels of buffer occupancy are denoted by $L_0, L_1, \ldots, L_{J-1}$ where, as Figure 3.1 shows, the thresholds are $B_1, B_2, \ldots, B_{J-1}$. By convention $B_0=0$ and $B_J=B$, the buffer capacity.

Loss priority is implemented as follows: when the buffer content is in the buffer occupancy level L_j, i.e. $B_j < X_t < B_{j+1}$, the only cells admitted to the buffer belong to classes $\{1, 2, \ldots, J-j\}$. Thus at level L_0 cells of all classes are admitted, while at level L_{J-1} only cells of class 1 are admitted. Moreover, at the boundary B_{j+1} $(0 \le j \le J-1)$ cells of class $(J-j)$ are admitted only if their presence does not cause the buffer occupancy to exceed B_{j+1}. That is, at the upper boundary of each level, cells of the class with the lowest priority, among those admitted at the level, may be lost and the loss rate is determined by the residual capacity of the channel.

As in Section 2, the state, state space and the generator of the Markov aggregate-source process are respectively given by Σ_t, $\mathcal{S}$ and $\mathbf{M}$ (see (2.14)). Also, (Σ_t, X_t) is Markov and the equilibrium distribution is given by $\pi_i(x)$, see (2.9). The differential equations governing $\{\pi_i(x)\}$ are piece-wise linear: for $j = 0, 1, \ldots, J-1$,

$$(\gamma_i^{(j)} - c)\, \frac{d}{dx}\, \pi_i(x) = \sum_{i' \in \mathcal{S}} \pi_{i'}(x) M_{i',i} \quad (i \in \mathcal{S};\, x \in L_j). \tag{3.1}$$

Here $\gamma_i^{(j)}$ is the sum of traffic rates of all classes admitted to the buffer at time t, given that $\Sigma_t = i$ and $X_t \in L_j$. Hence, by virtue of the implementation of loss priority, $\gamma_i^{(j)}$ is the sum of the rates at which cells of classes $1, 2, \ldots, J-j$ are generated when $\Sigma_t = i$. That is,

$$\gamma_i^{(j)} = \sum_{k=1}^{J-j} \lambda_i^{(k)} \quad (i \in \mathcal{S};\ j = 0, 1, \ldots, J-1) \tag{3.2}$$

where $\{\lambda_i^{(k)}\}$ have been defined in (2.1). On substituting for $\lambda_i^{(k)}$, we obtain

$$\gamma_i^{(j)} - c = i\omega^{(j)} - c^{(j)} \tag{3.3}$$

where,

$$\omega^{(j)} = \sum_{k=1}^{J-j} (v_2^{(k)} - v_1^{(k)}), \text{ and } c^{(j)} = c - K \sum_{k=1}^{J-j} v_1^{(k)}. \tag{3.4}$$

The following uses vector notation to represent (3.1) and (3.3): for $j = 0, 1, \ldots, J-1$,

$$\boxed{\frac{d}{dx}\, \boldsymbol{\pi}(x)\mathbf{D}^{(j)} = \boldsymbol{\pi}(x)\mathbf{M} \quad (x \in L_j)} \tag{3.5}$$

where,

$$\mathbf{D}^{(j)} = \operatorname{diag}\,\{-c^{(j)}, \omega^{(j)} - c^{(j)}, 2\omega^{(j)} - c^{(j)}, \ldots, K\omega^{(j)} - c^{(j)}\}\,.$$

Note the linear growth property of the diagonal elements of $\mathbf{D}^{(j)}$ which is the drift matrix for the j^{th} buffer occupancy level.

The piece-wise linear form of (3.5) gives the following structure to the solution: for $j = 0, 1, \ldots, J-1$,

$$\boxed{\boldsymbol{\pi}(x) = \boldsymbol{\pi}^{(j)}(x) = \sum_{i \in \mathcal{S}} a_i^{(j)} \boldsymbol{\phi}_i^{(j)} \exp(z_i^{(j)} x) \quad (x \in L_j)} \tag{3.6}$$

Here $\{z_i^{(j)}, \boldsymbol{\phi}_i^{(j)}\}$ are solutions to J separate sets of eigenvalue problems: for $j = 0, 1, \ldots, J-1$,

$$z_i^{(j)} \boldsymbol{\phi}_i^{(j)} \mathbf{D}^{(j)} = \boldsymbol{\phi}_i^{(j)} \mathbf{M} \quad (i = 0, 1, \ldots, K) \tag{3.7}$$

As noted in Section 2, the specific structure of $(\mathbf{D}^{(j)}, \mathbf{M})$ has allowed all the eigenvalues and eigenvectors to be obtained in closed form [AM82]. It remains to obtain the coefficients $\{a_i^{(j)}\}$ from the boundary conditions.

For each buffer occupancy level we separate the aggregate-source states which give a downward drift to the buffer content from those which give an upward drift: $(j = 0, 1, \ldots, J-1)$

$$\mathcal{S}_D^{(j)} = \{\sigma \in \mathcal{S} \mid \gamma_\sigma^{(j)} < c\}, \quad \mathcal{S}_U^{(j)} = \{\sigma \in \mathcal{S} \mid \gamma_\sigma^{(j)} > c\} \tag{3.8}$$

We make the natural assumption that the controls act to reduce the intake rate as the level of buffer occupancy increases, i.e.,

$$\gamma_i^{(0)} \ge \gamma_i^{(1)} \ge \cdots \ge \gamma_i^{(J-1)} \tag{3.9}$$

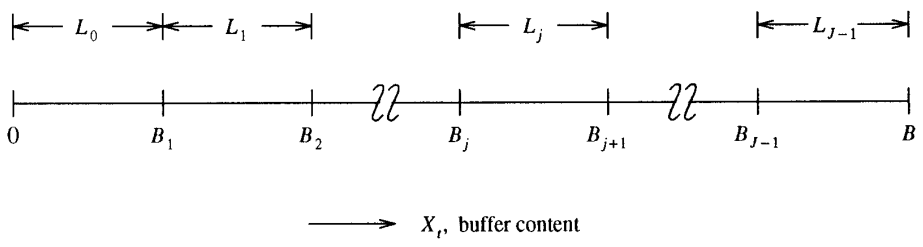


Figure 3.1: Levels of buffer occupancy $\{L_j\}$ used in implementing loss priorities.

for all $i \in \mathcal{S}$. Consequently,

$$\begin{aligned} &\mathcal{S}_D^{(0)} \subseteq \mathcal{S}_D^{(1)} \subseteq \cdots \subseteq \mathcal{S}_D^{(J-1)} \\ &\mathcal{S}_U^{(0)} \supseteq \mathcal{S}_U^{(1)} \supseteq \cdots \supseteq \mathcal{S}_U^{(J-1)} \end{aligned} \tag{3.10}$$

Arguing as in Section 2 on the basis of the behavior of $\frac{d}{dt} X_t$ at the thresholds $B_1, B_2, \ldots, B_{J-1}$, in conjunction with (3.10), we obtain the following complete system of boundary conditions:

$$\boxed{\begin{array}{lll} \pi_i^{(0)}(0) = 0 & (i \in \mathcal{S}_U^{(0)}) & (3.11,\text{i}) \\ \pi_i^{(j)}(B_{j+1}) = \pi_i^{(j+1)}(B_{j+1}) & \begin{array}{l} (i \in \mathcal{S}_D^{(j)} \cup \mathcal{S}_U^{(j+1)} \\ (j = 0, 1, \ldots, J-2) \end{array} & (3.11,\text{ii}) \\ \pi_i^{(J-1)}(B) = p_i & (i \in \mathcal{S}_D^{(J-1)}) & (3.11,\text{iii}) \end{array}}$$

The above system of $J(K+1)$ equations is the generalization of (2.19). Substitution of the expressions for $\{\pi_i^{(j)}(x)\}$ given in (3.6) yields a system of linear equations in the $J(K+1)$ coefficients $\{a_i^{(j)}\}$ which has to be solved numerically. This completes the description of the procedure for calculating $\boldsymbol{\pi}(x)$.

As shown in Figure 2.1, the distributions $\{\pi_i(x)\}$ exhibit jumps at the boundaries $\{B_j\}$ of the buffer occupancy levels and the following probabilistic interpretation applies to the jumps: $(j = 0, 1, \ldots, J-2)$

$$\Pr(\Sigma = i,\ X = B_{j+1}) = \pi_i^{(j+1)}(B_{j+1}) - \pi_i^{(j)}(B_{j+1}) \quad (i \in \mathcal{S}) \tag{3.12}$$

Note that

$$\begin{aligned} \Pr(\Sigma = i,\ X = B_{j+1}) &> 0 \quad (i \in \mathcal{S}_U^{(j)} \cap \mathcal{S}_D^{(j+1)}) \\ &= 0 \quad (i \in \mathcal{S}_D^{(j)} \cup \mathcal{S}_U^{(j+1)}) \end{aligned}$$

The throughput of cells of class $(J-j)$ $\quad (0 \le j \le J-2)$,

$$T^{(Jj)} = \sum_{i \in \mathcal{S}} \lambda_i^{(J-j)} \pi_i^{(j+1)}(B_{j+1}) - \tag{3.13}$$

$$- \sum_{i \in \mathcal{S}_U^{(j)} \cap \mathcal{S}_D^{(j+1)}} \{\gamma_i^{(j)} - c\} \Pr(\Sigma = i,\ X = B_{j+1})$$

This expression is the generalization of (2.24). The throughput of class 1, the class of highest priority, is

$$T^{(1)} = \sum_{i \in \mathcal{S}} \gamma_i^{(J-1)}\ \pi_i^{(J-1)}(B) + c \Pr(\text{buffer full}) \tag{3.14}$$

where, $\quad \Pr(\text{buffer full}) = 1 - \sum_{i \in \mathcal{S}} \pi_i^{(J-1)}(B) \quad (3.15)$

The loss probabilities for the various classes are obtained from their throughputs by the formula in (2.25).

The delay distribution of class 1 cells is a straightforward generalization of the formulas in (2.26):

$W^{(1)}(t)$, Pr (class 1 cell delay $\le t$)

$$= \frac{1}{T^{(1)}} \sum_{i=0}^{K} \lambda_i^{(1)} \pi_i^{(j)}(ct) \quad (B_j/c < t < B_{j+1}/c)$$
$$(j=0, 1, \ldots, J-1) \tag{3.16,i}$$

$$= \frac{1}{T^{(1)}} \sum_{i=0}^{K} \lambda_i^{(1)} \Pr(\Sigma = i,\ X = B_j) \quad (t = B_j/c)$$
$$(j = 0, 1, \ldots, J-1) \tag{3.16,ii}$$

Finally,

$$\Pr\left[\text{class 1 cell delay} = \frac{B}{c}\right] = \frac{c}{T^{(1)}} \Pr[\text{buffer full}] \tag{3.16,iii}$$

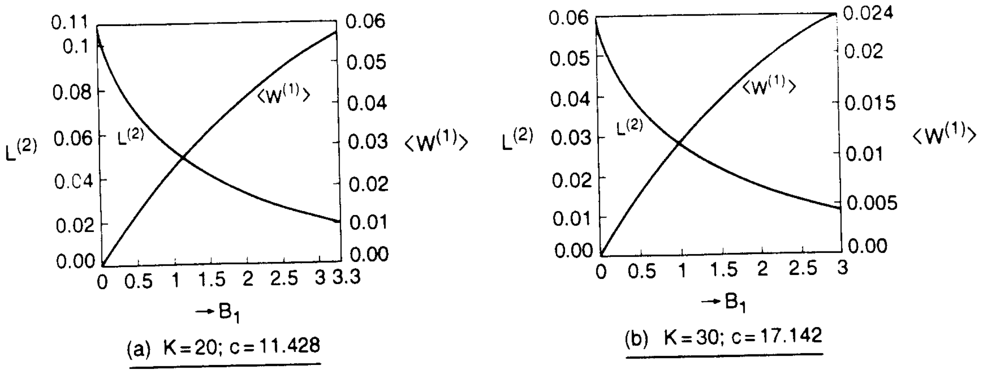
The cell delay distributions for all the other classes are similarly obtained.

4. NUMERICAL INVESTIGATIONS

In this Section we report on numerical results obtained from fluid models for the performance analysis of statistical multiplexing with loss priorities. We consider throughout the case of on-off sources with exponentially distributed on and off periods. The unit of time is selected to be the mean on period and the unit of information to be equal to the amount of priority traffic generated in an average on period. Thus the peak rate of priority traffic is 1 unit of information per 1 unit of time, i.e. $\beta = 1$ and $\nu_2^{(1)} = 1$. This convention follows from a natural normalization of system parameters. Incidentally, it considerably simplifies the study of the effects of varying the jitteriness of the sources which is caused by increasing or decreasing their mean on and off periods by the same factor. All that is necessary is to simply reinterpret the unit of information and the unit of time; in fluid models no changes in the calculations have to be made. Due to lack of space we only give results for the peak rate of marked cells ($\nu_2^{(2)}$) set to 0.5 and the mean off period $(1/\alpha)$ set to 0.4. In all the results presented here an infinite buffer is assumed and the performance results for finite buffers are inferred, as explained in Section 2.

The trade-off between the loss of marked cells and the mean delay of the priority cells is displayed in Figure 4.1, (a)-(b), for $K = 20$ and $K = 30$. These figures show clearly how the threshold B_1, as it is varied from 0 to about 3 units of information, controls the trade-off by decreasing the marked cells' loss and increasing the mean delay of the priority cells. The influence of B_1 on the trade-off is diluted considerably if B_1 is more than a small number in the range of 1.0 to 2.0.

In Figure 4.2, (a)-(b), the marked cells' loss and the priority cells' mean delay are plotted as a function of the threshold for different values of the channel capacity c. In this Figure $K = 30$. The plots show the nonlinear manner in which decreasing the channel capacity from 22.01 to 14.11 (resulting in $\rho^{(0)}$ increasing from 0.58 to 0.81, and $\rho^{(1)}$ from



(a) K=20; c=11.428

(b) K=30; c=17.142

Figure 4.1: The trade-off between $L^{(2)}$, loss probability for marked cells, and $\langle W^{(1)}\rangle$, mean delay for priority cells, as a function of B_1, threshold. $\alpha=0.4$, $\beta=1.0$; $v_1^{(1)}=v_1^{(2)}=0$; $v_2^{(1)}=1.0$; $v_2^{(2)}=0.5$. For both (a) and (b), $\rho^{(0)}=0.75$ and $\rho^{(1)}=0.50$.

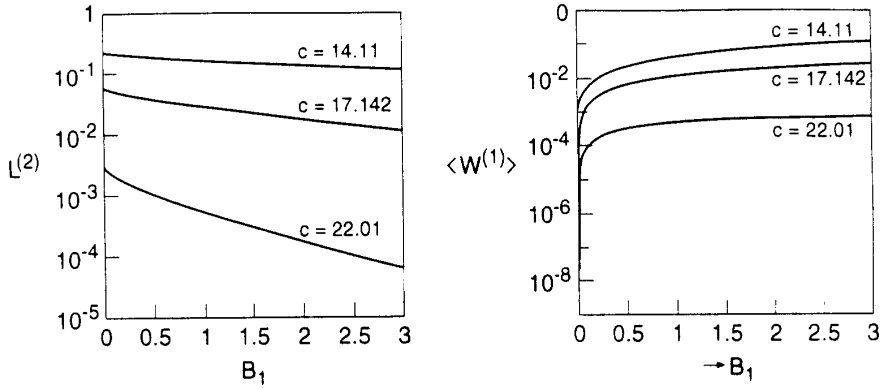


Figure 4.2: Effect of the threshold B_1 on $L^{(2)}$, loss probability for marked cells and on $\langle W^{(1)}\rangle$, mean delay for priority cells. $K=30$; $\alpha=0.4$, $\beta=1.0$; $v_1^{(1)}=v_1^{(2)}=0.0$; $v_2^{(1)}=1.0$; $v_2^{(2)}=0.5$. Observe the logarithmic scale in (b).

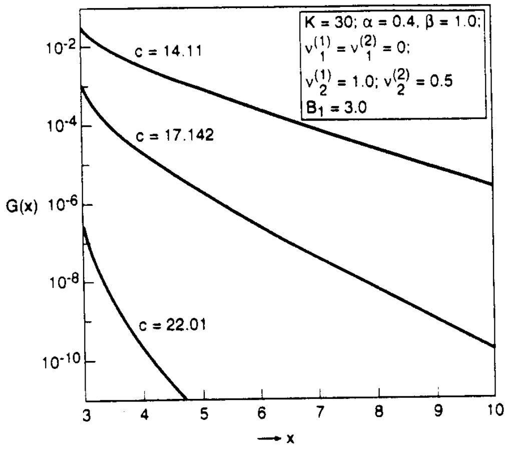


Figure 4.3: $G(x)$, the complementary buffer distribution as seen by an arriving priority cell. $G(x)$ approximates $L^{(1)}$ for buffer of size x.

0.39 to 0.61) causes both the mean delay and the loss probability to increase.

The complementary distribution of the buffer content as seen by an arriving priority cell, $G(x)$, is plotted in Figure 4.3 for different channel capacities, with B_1 held fixed at 3 units of information. We recall that $G(x)$ closely approximates the loss probability of priority cells in a system with a finite buffer of size x, provided that $G(x)$ is small.

5. CONCLUSIONS

We summarize the salient elements of the paper.

(i) A stochastic fluid model is proposed for statistical multiplexing with loss priorities. In this model Markov Modulated Fluid Sources reflect the *bursty* and *periodic* characteristics of cell generation.

(ii) Coupling between cell streams of different priorities models correlations between the streams.

(iii) The model and its analysis extend easily to multiple priority classes.

(iv) An exact analysis of the fluid model is given. It has uncovered features of the stationary distribution not present in prior fluid models. The algorithm for obtaining the solution is efficient since its computational complexity is dominated by the task of solving a single system of linear equations of dimension at most $(K+1)J$ where K is the number of sources and J is the number of priority classes. (The complexity is smaller if the infinite buffer approximation is made.) In particular, the complexity is independent of the buffer size.

(v) The analysis gives in a straightforward manner the complete delay distributions for all the cell classes.

(vi) The numerical results clearly demonstrate the trade-off between delay of the priority cells and the loss probabilities of the marked cells, as well as the influence of the threshold in the control of the trade-off.

A forthcoming work [EM91(b)] will report on

(i) higher dimensional Markov Modulated Fluid Sources [EMS91],

(ii) multiple types of sources,

(iii) approximations, and

(iv) end-to-end performance of the system in Figure 1.1.

REFERENCES

[AK84] J. M. Akinpelu, "The overload performance of engineered networks with nonhierarchical and hierarchical routing," AT&T Bell Labs. Tech. J., **63**, Sept '84, 1261-1281.

[AM82] D. Anick, D. Mitra and M. M. Sondhi, "Stochastic theory of a data-handling system with multiple sources," Bell System Tech. J., **61**, 1982, 1871-1894.

[AS91] G. A. Awater and F. C. Schoute, "Optimal queueing policies for fast packet switching of mixed traffic," IEEE JSAC, **9**, No. 3, 1991, 458-467.

[BF91] F. Bonomi, L. Fratta, S. Montagna and R. Paglino, "Priority on cell service and on cell loss in ATM switching," preprint, 1991.

[BC91] M. Butto, E. Cavallero and A. Tonietti, "Effectiveness of the "leaky bucket" policing mechanism in ATM networks," IEEE JSAC, **9**, No. 3, 1991, 335-342.

[CC90] CCITT SGXVIII, Draft Recommendation I.361, "ATM layer specification for B-ISDN," Report XVIII-R 23-E, Geneva, 1990.

[CD89] C. Chamzas and D. L. Duttweiler, "Encoding facsimile images for packet-switched networks," IEEE JSAC, **7**, No. 5, 1989, 857-864.

[CP90] J.-P. Coudreuse, G. Pays and M. Trouvat, "Asynchronous transfer mode," Commutation and Transmission, No. 3, 1990, 5-16.

[CI91] E. G. Coffman, Jr., B. M. Igelnik and Y. A. Kogan, "Controlled stochastic model of a communication system with multiple sources," IEEE Trans. Infor. Theory, **37**, 1991, 1379-1387.

[DJ88] L. Dittman and S. B. Jacobsen, "Statistical multiplexing of identical bursty sources in an ATM networks," GLOBECOM '88, 1293-1297.

[DL86] J. N. Daigle and J. D. Langford, "Models for analysis of packet voice communication systems," IEEE JSAC, **4**, No. 6, 1986, 847-855.

[DL90] J. N. Daigle and D. M. Lucantoni, "Queueing systems having phase-dependent arrival and service rates," in *Numerical Solutions of Markov Chains* (Ed. W. J. Stewart) Marcel Dekker, 1991, 161-202.

[EL90] A. E. Eckberg, D. T. Luan and D. M. Lucantoni, "An approach to controlling congestion in ATM networks," Int. J. Digital and Analog Commun. Syst., **3**, 1990, 199-209.

[EM91(a)] A. I. Elwalid and D. Mitra, "Analysis and design of rate-based congestion control for high speed networks, I: stochastic fluid models, access regulation," Queueing Systems (special issue on Communication Systems), **9**, 1991, 29-63.

[EM91(b)] A. I. Elwalid and D. Mitra, "Analysis and design of rate-based congestion control for high speed networks, II: statistical multiplexing, loss priorities,'' in preparation.

[EMS91] A. I. Elwalid, D. Mitra and T. E. Stern, ''Statistical multiplexing of Markov modulated sources: theory and computational algorithms,'' Proc. ITC-13, Copenhagen, June 1991, 495-499.

[GV91] M. W. Garrett and M. Vetterli, ''Joint source/channel coding of statistically multiplexed real time services on packet networks,'' preprint, 1991.

[GH91] R. J. Gibbens and P. J. Hunt, "Effective bandwidths for multitype UAS channel,'' Queueing Systems (QUESTA, special issue on Communication Systems), **9**, 1991, 17-28.

[GL82] D. P. Gaver and J. P. Lehoczky, "Channels that cooperatively service a data stream and voice messages,'' IEEE Trans. Commun., **COM-30**, No. 5, 1982, 1153-1162.

[HL86] H. Heffes and D. Lucantoni, ''A Markov modulated characterization of packetized voice and data traffic and related statistical multiplexer performance,'' IEEE JSAC, **4**, No. 6, 1986, 856-868.

[ID88] I. Ide, ''Superposition of interrupted Poisson processes and its application to packetized voice multiplexers,'' Proc. ITC-12, Torino, 1988.

[JA90] R. Jain, ''Myths about congestion management in high speed networks,'' Tech. Rep. DEC-TR-726, Digital Equipment Corp., Littleton, Mass. 01460, 1990.

[KO84] L. Kosten, ''Stochastic theory of data-handling systems with groups of multiple sources,'' in *Performances of Computer-Communication Systems* (Ed. H. Rudin and W. Bux), Elsevier, Amsterdam, 1984, 321-331.

[KO90] H. Kobayashi, ''Performance issues of Broadband ISDN,'' Proc. Intl. Conf. Comp. Commun., ICCC90, New Delhi, 349-361.

[KH91] H. Kroner, G. Hebuterne, P. Boyer and A. Gravey, ''Priority management in ATM switching nodes,'' IEEE JSAC, **9**, No. 3, 1991, 418-427.

[KM89] F. Kishino, K. Manabe, Y. Hayashi and H. Yasudo, ''Variable bit-rate coding of video signals for ATM networks,'' IEEE JSAC, **7**, No. 5, 1989, 801-806.

[KT90] H. Kroner, T. H. Theimer and U. Briem, ''Queueing models for ATM systems – a comparison,'' Proc. 7th ITC Specialist Seminar, Morristown, 1990, paper 9.1.

[LB91] J.-Y. Le Boudec, ''An efficient solution method for Markov models of ATM links with loss priorities,'' IEEE JSAC, 1991, **9**, No. 3, 1991, 408-417.

[LP90] D. M. Lucantoni and S. P. Parekh, ''Selective cell discard mechanisms for a B-ISDN congestion control architecture,'' Proc. 7th ITC Specialist Seminar, Morristown, 1990, paper 10.3.

[MA88] B. Maglaris, P. Anastassiou, P. Sen, G. Karlsson and J. D. Robbins, ''Performance models of statistical multiplexing in packet video communications,'' IEEE Trans. Commun., **36**, No. 7, 1988, 834-843.

[MG90] J. A. S. Monteiro, M. Gerla and L. Fratta, ''Leaky bucket input rate control in ATM networks,'' Proc. Intl. Conf. Comp. Commun., ICCC90, New Delhi, 370-376.

[MI88] D. Mitra, ''Stochastic theory of a fluid model of producers and consumers coupled by a buffer,'' Adv. Appl. Prob., **20**, 1988, 646-676.

[NE81] M. F. Neuts, ''*Matrix Geometric Solutions in Stochastic Models*,'' John Hopkins University Press, Baltimore, 1981.

[NK91] R. Nagarajan, J. F. Kurose and D. Towsley, ''Approximation techniques for computing packet loss in finite-buffered voice multiplexers,'' IEEE JSAC, **9**, No. 3, 1991, 368-377.

[NR91] I. Norros, J. W. Roberts, A. Simonian and J. T. Virtamo, ''The superposition of variable bit rate sources in an ATM multiplexer,'' IEEE JSAC, **9**, No. 3, 1991, 378-387.

[PD89] D. W. Petr, L. A. DaSilva and V. S. Frost, ''Priority discarding of speech in integrated packet networks,'' IEEE JSAC, **7**, No. 5, 1989, 644-656.

[SE91] T. E. Stern and A. I. Elwalid, ''Analysis of separable Markov-modulated rate models for information-handling systems,'' Adv. Appl. Prob., **23**, 1991, 105-139.

[SW86] K. Sriram and W. Whitt, ''Characterizing superposition arrival processes in packet multiplexers for voice and data,'' IEEE JSAC, **4**, No. 6, 1986, 833-846.