ANALYSIS, APPROXIMATIONS AND ADMISSION CONTROL OF A MULTI-SERVICE MULTIPLEXING SYSTEM WITH PRIORITIES

Anwar Elwalid *Debasis Mitra*

AT&T Bell Laboratories
600 Mountain Avenue
Murray Hill, New Jersey 07974

Abstract

We consider an ATM system with an architecture which is designed to accommodate users with very different quality of service requirements. In the base case with only two services, sources which require low loss belong to a High Priority class, and share a FCFS buffer, which has priority access to the trunk. A Low Priority class of sources with typically less stringent requirements on loss have a separate FCFS buffer, which receives the time-varying, residual bandwidth, if any, of the trunk. By administering admission control and restricting the combination of sources to an *admissible set*, the service guarantees for both classes may be satisfied. The sources are bursty and stochastic fluid models are used to handle burst-scale congestion effects. Our contributions are: (i) we develop simple, fast and robust analytic approximations for the queue distributions in the two buffers; (ii) for admission control, we calculate the admissible set by using our analytic approximations and find that it is reasonably accurate and a conservative approximation to the correct set. We developed a fluid-flow simulator for general Markov-modulated sources, and a program which implements our analytic techniques for on-off sources. Systems with over 100 sources are completely analyzed in a few seconds, while simulations take hours. The key element in our analysis is a characterization of the output of the high priority buffer as another Markov-modulated fluid source. A refinement to the now well known effective bandwidth approximation is used to calculate buffer content distributions.

Keywords: ATM, Guaranteed Service, Engineered Best Effort Service,
fluid models, congestion control, admission control.

1 INTRODUCTION

It is expected that services on emerging networks employing the Asynchronous Transfer Mode (ATM) will have diverse quality of service requirements. Users of these services will expect tariffs which are commensurate with the quality of service. Correspondingly, the network service provider will benefit from being able to provide a range of quality of service guarantees. The system we consider, see Figure 1.1, is one of the simplest implementations which segments quality of service.

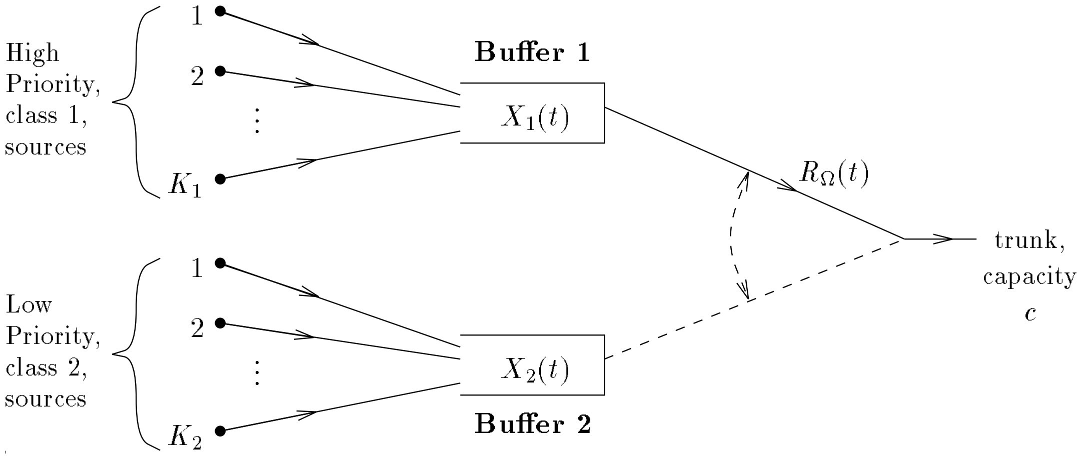


Figure 1.1: A system in which Buffer 1 has priority access to the trunk.

In this system the sources which require low loss and small delays belong to the High Priority class and their data is stored in Buffer 1, if necessary, before transmission on the trunk. A second class of sources have typically far less stringent service requirements and the data from these sources are stored in Buffer 2, if necessary, before transmission on the trunk. Buffer 1 has priority access to the trunk and only the time-varying residual bandwidth, if any, is available for servicing Buffer 2. The service in both buffers is First-Come-First-Served. Obviously class 1 sources are oblivious of the class 2 sources. By restricting the number of sources of class 1, i.e., by administering admission control, it is possible to meet stringent service requirements for that class, and hence the first service may be called "Guaranteed Service" [CSZ92]. It is not so obvious that admission control is needed for the second service. One laissez-faire approach is to provide "Best Effort Service" with no guarantees. Even in this approach it is necessary for provisioning and capacity planning to use the techniques developed in this paper. We feel however that there will be a need for "Engineered Best Effort Service" where some guarantees will be provided and admission control will be required. Now, however, admission control requires limiting combinations of sources of both classes. Note that the system shown in Figure 1.1 extends naturally to a hierarchy with several grades of service and priorities.

Our contributions are primarily two-fold: (i) we develop simple, fast and robust analytic approximations for the queue length distributions in the two buffers for bursty traffic from the sources; (ii) the admission control problem is examined, which entails investigating the admissible combinations of sources of the two classes for given buffer sizes and bounds on loss probabilities. We find that the admissible set from our analytic approximations is a reasonably accurate and conservative approximation to the correct set.

There is little prior work on the analysis and admission control of priority queues subject to bursty traffic. In [ROB92] there are results on priority queueing, but mostly on space priorities. In [ZHA93] a two queue priority system is analyzed by Laplace transform techniques in which each of several on-off, 2-state fluid sources produces high and low priority streams.

We use stochastic fluid models in the analysis here since our focus is on burst-level congestion effects for which these models have been found to be accurate in various prior studies, and also tractable. We should point out here that certain large deviations phenomena, which are important in motivating and justifying our central approximations, are most clearly visible in renditions of burst scale effects in fluid models.

We now give the sample path description of the fluid model in Figure 1.1. Let Ω_t denote the output process of Buffer 1 and $R_\Omega(t)$ the rate of the process at time t. Also let $R_{\Sigma 1}(t)$ and $R_{\Sigma 2}(t)$ denote the aggregate, i.e. sum of, rates of fluid generation at time t by the sources of classes 1 and 2, respectively. The trunk bandwidth or capacity is a constant rate c. The contents of the two buffers are given by $X_1(t)$ and $X_2(t)$, as indicated in the figure. Now,

$$\begin{aligned} \frac{d}{dt} X_1(t) &= R_{\Sigma 1}(t) - c \qquad \text{if} \quad X_1(t) > 0 \\ &= [R_{\Sigma 1}(t) - c]^+ \quad \text{if} \quad X_1(t) = 0 \end{aligned} \tag{1.1}$$

where generally $[x]^+$ denotes the positive part of x. Also,

$$\begin{aligned} R_\Omega(t) &= c \qquad\qquad \text{if} \quad X_1(t) > 0 \\ &= R_{\Sigma 1}(t) \quad \text{if} \quad X_1(t) = 0 \ . \end{aligned} \tag{1.2}$$

Now $\{c - R_\Omega(t)\}$, a nonnegative quantity, is the channel bandwidth available for servicing Buffer 2. Hence,

$$\begin{aligned} \frac{d}{dt} X_2(t) &= R_{\Sigma 2}(t) - \{c - R_\Omega(t)\} && \text{if} \quad X_2(t) > 0 \\ &= [R_{\Sigma 2}(t) - \{c - R_\Omega(t)\}]^+ && \text{if} \quad X_2(t) = 0 \ . \end{aligned} \tag{1.3}$$

One of the features of stochastic fluid models is that only a change of sign of a rate distinguishes a producer from a consumer of a buffer [MIT88]. In (1.3) the consumption or service rate of Buffer 2 is $\{c - R_\Omega(t)\}$; an equivalent view is that $R_\Omega(t)$ supplies Buffer 2 and the service rate is c. Hence, either by this argument or by directly comparing sample path equations it follows that the priority system in Figure 1.1 is identical to the system in Figure 1.2 in which there is no overt priority in trunk access. Here, as before, the service in both buffers is First-Come-First-Served. Note that the output trunks of both buffers are of capacity c. This configuration belongs to the class of in-tree networks [CHJ93]. We will find the equivalence of the configurations in the two figures to be useful.

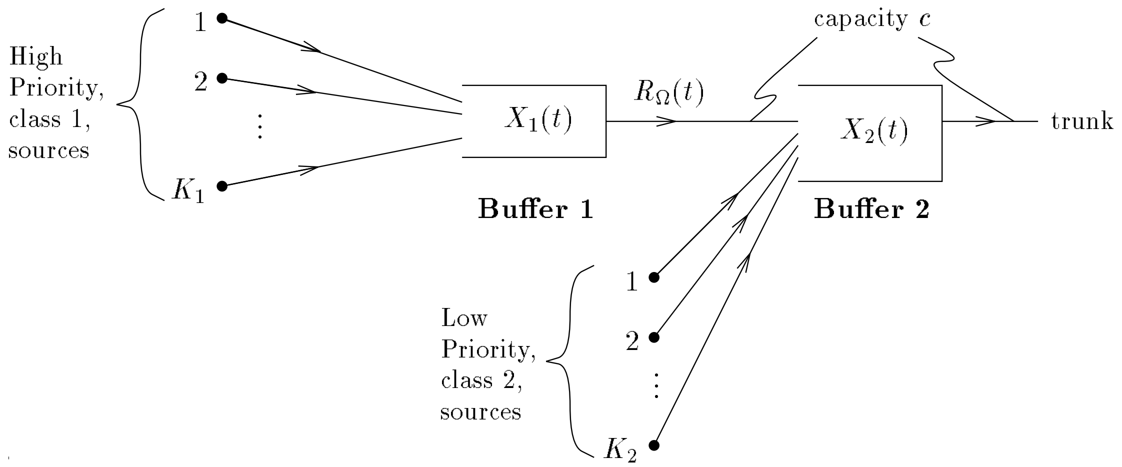


Figure 1.2: An equivalent view of the priority system in Figure 1.1.

The sources of each class are considered to be independent, statistically homogeneous. The homogeneity is by no means necessary; however, it eases the computational burden and makes the description of the techniques simpler. Each source of class j ($j = 1, 2$) is a Markov-modulated fluid source characterized by $(\boldsymbol{M}^{(j)}, \boldsymbol{\lambda}^{(j)})$, where $\boldsymbol{M}^{(j)}$ is the generator of the Markov chain and $\boldsymbol{\lambda}^{(j)}$ is the vector of state-dependent rates of fluid generation by the source.

We briefly review the salient features of our approach. We obtain an approximate characterization of the output process Ω_t as another Markov-modulated fluid source $(\boldsymbol{M}^{(\Omega)}, \boldsymbol{\lambda}^{(\Omega)})$. The number of states of this source is typically high. Procedures for calculating $(\boldsymbol{M}^{(\Omega)}, \boldsymbol{\lambda}^{(\Omega)})$ are explicitly stated. The idea in the construction is rooted in the observation that in equilibrium Buffer 1 is typically empty for an overwhelming fraction of time. For instance, if the cell loss probability for the higher priority service is bounded by 10^{-9}, then the probability that the buffer is not empty can be as small as $10^{-3} - 10^{-5}$. This is not to say that throughput is necessarily small while the buffer is empty. On the contrary, in a fluid model if the buffer is empty then it remains so as long as the aggregate source rate is not more than the outgoing channel rate. Thus there is no necessary contradiction between the aforementioned numbers and the observation that the channel utilization by traffic from class 1 sources is moderately high, say, in the range 0.4–0.7. Such examples may be found in [AMS82], for instance. In our characterization of the output process Ω_t we model finely the rate dynamics in the empty periods of the buffer, and lump all activity during the busy, i.e. non-empty, period into just one state. That is, the buffer's busy period distribution is approximated by an exponential distribution. We calculate the correct mean busy period and, importantly the distribution of states at the onset of the buffer's empty period. These quantities are calculated quite simply. We should point out that other approaches, as exemplified by Bonomi et al. [BMM94], take the opposite course and emphasize the fine modelling of the activity during the buffer's busy period. However, such approaches are typically not based on fluid models.

Another salient feature of the analytic approximations in this paper is that in the calculation of the distribution of contents in Buffer 2, $\Pr(X_2 \geq x)$, we refine the pure exponential form e^{zx} used in effective bandwidth analyses [GAN91, GHU91, EMI93, WHI93, VCW94] by $L_0 e^{zx}$, where the prefactor L_0 is simply the loss in bufferless multiplexing. The latter quantity is very well approximated by the Chernoff large deviations approximation [HUI90, KEL91, ROB92] and, moreover, it is easy to calculate for even high dimensional sources. Note that L_0 is our approximation for the probability that the buffer is not empty and, as stated earlier, it is quite small in typical ATM applications. Hence the prefactor L_0 adds significantly to the accuracy of the effective bandwidth approximations, which can sometimes be overly conservative [CLW94].

The admission control problem which we address requires a characterization of the admissible set of sources of both classes such that, for given buffer sizes, the overflow probabilities do not exceed prespecified values. That is, for given B_1 and B_2, and bounds on the buffer overflow probabilities L_1 and L_2 for Buffers 1 and 2, respectively, the admissible set is

$$\mathcal{A}(B_1, B_2; L_1, L_2) \triangleq \{\boldsymbol{K} : G_1(B_1) \leq L_1 \text{ and } G_2(B_2) \leq L_2\} \ , \tag{1.4}$$

where $G_j(x) = \Pr(X_j \geq x)$ is the equilibrium distribution of contents in Buffer j. Following standard practice, we let buffers be of infinite capacity and let the probability that the contents exceed B_j be the conservative approximation to the overflow probability in a finite buffer of size B_j. In (1.4) typically $L_1 \ll L_2$. Of considerable interest is a simple characterization of $\mathcal{A}$.

2 PRELIMINARIES AND BACKGROUND FACTS

2.1 A General Non-Prioritized Multiplexing System

We consider a statistical multiplexing system without priorities and with various, possibly heterogeneous, statistically independent Markov-modulated fluid sources which supply a buffer of infinite capacity. The buffer is serviced by a channel of constant capacity, i.e., rate, c and the discipline is First-Come-First-Served. In this subsection it suffices to lump the source description into a single aggregate Markov-modulated fluid source with state space $\mathcal{S}$ and irreducible generator $\boldsymbol{M}$. This aggregate source generates fluid at the constant rate λ_s when in state s ($s \in \mathcal{S}$). Let $\boldsymbol{\lambda} = \{\lambda_s \mid s \in \mathcal{S}\}$. The aggregate source is characterized by $(\boldsymbol{M}, \boldsymbol{\lambda})$. Let the rate matrix $\boldsymbol{\Lambda} = \text{diag}\,(\boldsymbol{\lambda})$.

The stationary state distribution of the multiplexing system is denoted by $\boldsymbol{\pi}(x) = \{\pi_s(x) \mid s \in \mathcal{S}\}$, where

$$\pi_s(x) \triangleq \Pr(S = s,\ X \leq x) \quad (s \in \mathcal{S},\ 0 \leq x \leq \infty)\ . \tag{2.1}$$

Here S and X respectively denote the aggregate source state and buffer content. The governing system of differential equations is

$$\frac{d}{dx}\,\boldsymbol{\pi}(x)\boldsymbol{D} = \boldsymbol{\pi}(x)\boldsymbol{M} \quad (0 \leq x \leq \infty) \tag{2.2}$$

where $\boldsymbol{D} \triangleq \boldsymbol{\Lambda} - c\boldsymbol{I}$, and $\boldsymbol{I}$ is the identity matrix. The diagonal element $D_{ss} = (\lambda_s - c)$ is the drift, or rate of change, in the buffer content when the source is in state s.

The stationary probability vector for the aggregate source is $\boldsymbol{w}$, where $\boldsymbol{wM} = \boldsymbol{0}$ and $\langle \boldsymbol{w}, \boldsymbol{1} \rangle = 1$. The symbol $\langle \cdot, \cdot \rangle$ denotes the inner product of vectors, and $\boldsymbol{1}$ is the vector in which all elements are unity. We assume ergodicity,

$$\overline{\lambda} \triangleq \langle \boldsymbol{w}, \boldsymbol{\lambda} \rangle < c \ , \tag{2.3}$$

and, to avoid trivialities, $c < \hat{\lambda}$, where $\hat{\lambda}$ is the peak source rate, i.e., $\hat{\lambda} \triangleq \max_s \lambda_s$.

Let the stationary buffer overflow distribution be given by $G(x)$, i.e.,

$$G(x) \triangleq \Pr\,(X \geq x) = 1 - \langle \boldsymbol{\pi}(x), \boldsymbol{1} \rangle \ . \tag{2.4}$$

Since $\boldsymbol{\pi}(x)$ is the solution of the linear system (2.2) and is bounded, $G(x)$ has the following spectral representation,

$$G(x) = \sum_{i \geq 1} A_i e^{z_i x} \quad (0 \leq x \leq \infty) \ . \tag{2.5}$$

Here $\{z_i\}$ are eigenvalues with negative real parts, which are obtained as solutions of the eigenvalue problem $z\boldsymbol{\phi D} = \boldsymbol{\phi M}$. From the structure of $(\boldsymbol{D}, \boldsymbol{M})$ it is known [EMI93] that one of the eigenvalues, say z_1, is real and dominates the other eigenvalues in the sense that

$$0 > z_1 > \max_{i>1} \ \mathrm{Re}(z_i) \ . \tag{2.6}$$

Hence z_1 is called the *dominant eigenvalue.*

From (2.5) it follows that the tail behavior of the buffer distribution is given thus,

$$G(x) \sim Ae^{zx} \quad \text{as} \quad x \to \infty \ , \tag{2.7}$$

where $A = A_1$ and $z = z_1$.

2.2 The Dominant Eigenvalue

The dominant eigenvalue and its calculation has been a topic of central importance in almost all studies of statistical multiplexing based on stochastic fluid models [AMS82, KOS84, KOS86, MIT88, NRS91, SEL91, ROB92]. Here we give two facts from [EMI93] which will be useful in the sequel.

First, observe that for z real and negative, $(\boldsymbol{\Lambda} - \frac{1}{z}\, \boldsymbol{M})$ is an irreducible matrix with non-negative off-diagonal elements, and hence is called "essentially nonnegative". Such matrices are of course closely related to irreducible nonnegative matrices. In particular, such matrices have a real eigenvalue, called the "maximal real eigenvalue" (MRE), which is greater than the real part of all other eigenvalues. Let

$$g(z) \triangleq \mathrm{MRE}\left(\boldsymbol{\Lambda} - \frac{1}{z}\,\boldsymbol{M}\right) \ . \tag{2.8}$$

FACT 1: The dominant eigenvalue of the multiplexing system (2.2) with source $(\boldsymbol{M}, \boldsymbol{\lambda})$ and channel bandwidth c is obtained by solving

$$g(z) = c \ . \tag{2.9}$$

∎

It turns out that the burden of numerically solving (2.9) is relatively light because $g(z)$ is monotonic decreasing for $z < 0$, and c is sandwiched between $\hat{\lambda} = g(-\infty)$ and $\overline{\lambda} = \lim_{z\to 0} g(z)$.

Now introduce the feature that the aggregate source $(\boldsymbol{M}, \boldsymbol{\lambda})$ is composed of K independent sources $(\boldsymbol{M}^{(k)}, \boldsymbol{\lambda}^{(k)})$, $k = 1, 2, \ldots, K$. In this case we have

FACT 2: The dominant eigenvalue of the multiplexing system (2.2) is obtained by solving

$$\sum_{k=1}^{K} g^{(k)}(z) = c \ , \tag{2.10}$$

where $g^{(k)}(z) \triangleq \mathrm{MRE}(\boldsymbol{\Lambda}^{(k)} - \frac{1}{z}\,\boldsymbol{M}^{(k)})$. ∎

This fact is extraordinarily useful since it decomposes the computation of the MRE of a typically very large matrix into that of only K much smaller computations.

In the special case of on-off sources with exponentially distributed on and off periods,

$$\boldsymbol{M}^{(k)} = \begin{bmatrix} -\alpha & \alpha \\ \beta & -\beta \end{bmatrix} \quad \text{and} \quad \boldsymbol{\lambda}^{(k)} = [0 \ \ \lambda] \ , \tag{2.11}$$

$$g^{(k)}(z) = \frac{1}{2z}\left[\lambda z + \beta + \alpha - \sqrt{(\lambda z + \beta - \alpha)^2 + 4\alpha\beta}\right] \ . \tag{2.12}$$

2.3 Busy Periods of the Multiplexing Buffer

For the sequel we will need two key facts on the empty and busy periods of the multiplexing buffer in the system described in Section 2.1. These results are from [EMI91], which draws on earlier work by Howard [HOW71], Li [LI89] and Neuts [NEU81].

Let $\mathcal{S}_{UL}$ denote the set of underload source states, i.e.,

$$\mathcal{S}_{UL} \triangleq \{s : s \in \mathcal{S} \ \text{ and } \ \lambda_s < c\} \ . \tag{2.13}$$

The complementary set is the set of overload source states, $\mathcal{S}_{OL}$. Let $q_s(s \in \mathcal{S}_{UL})$ denote the stationary probability that the source state is s at the onset of the buffer's empty period. Let $\boldsymbol{q} = \{q_s \mid s \in \mathcal{S}_{UL}\}$; clearly,

$$\langle \boldsymbol{q}, \mathbf{1} \rangle = 1 \ . \tag{2.14}$$

Then, from (5.9) in [EMI91],

FACT 3:

$$\boldsymbol{q} = \frac{1}{g} \ \boldsymbol{\pi}_{UL}(0) \boldsymbol{M}_{UL,UL} \ , \tag{2.15}$$

where $\boldsymbol{\pi}_{UL}(0) = \{\pi_s(0) \mid s \in \mathcal{S}_{UL}\}$ and, similarly, $\boldsymbol{M}_{UL,UL} = \{M_{ss'} \mid s \in \mathcal{S}_{UL}, s' \in \mathcal{S}_{UL}\}$ is the principal submatrix of $\boldsymbol{M}$ associated with the underload source states. Finally, g is the normalizing constant which gives (2.14). ∎

The key observation on the formula (2.15) is that it is straightforward to obtain $\boldsymbol{q}$ once the stationary state distribution $\boldsymbol{\pi}(x)$ has been calculated. Also, from (5.17) of [EMI91],

FACT 4:

$$\overline{b} = \text{mean busy period of buffer} = -G(0)/g \ . \tag{2.16}$$

∎

Recall that $G(0)$ = Pr (buffer not empty). Hence, $G(0) = 1 - \pi(0)$, where $\pi(0) \triangleq \langle \boldsymbol{\pi}(0), \mathbf{1} \rangle$.

2.4 Loss in Bufferless Multiplexing

Here we give a result, which is used later in the paper, for estimating L, the stationary probability that the sum of the instantaneous rates of fluid generation by all sources exceeds the channel rate. Thus L is the loss probability in a multiplexing system without a buffer. The basic procedure given here is due to Hui [HUI90] and Kelly [KEL91], and is based on the Chernoff bound.

Suppose there are J source classes, and K_j sources of class j, all of which are characterized by $(\boldsymbol{M}^{(j)}, \boldsymbol{\lambda}^{(j)})$; let $\boldsymbol{w}^{(j)}$ denote the stationary probability vector. The stationary rate of fluid generation $R_{ji}(t)$ from source i of class j has the moment generating function

$$M_j(s) \triangleq E[e^{sR_{ji}}] = \sum_i^j w_i^{(j)} e^{s\lambda_i^{(j)}} \ . \tag{2.17}$$

Now,

$$L = \Pr\ (R \geq c)\ , \tag{2.18}$$

where the total source rate, $R = \sum_{j=1}^J \sum_{i=1}^{K_j} R_{ji}$. Chernoff's bound gives

$$\log L \leq -F(s^*)\ , \tag{2.19}$$

where

$$F(s) = sc - \sum_j K_j \log M_j(s) \tag{2.20}$$

and $F(s^*) = \sup_{s \geq 0} F(s)$. It is easily verified that $F(s)$ is a strictly concave function with an unique maximum at $s^*(s^* > 0)$, which is obtained by solving $F'(s) = 0$.

If $K_j = \gamma_j c$ with $\gamma_j = O(1)$, and $c \to \infty$, then [CSE93],

$$\log L = -F(s^*) \left[1 + O\left(\frac{\log c}{c}\right)\right] \ . \tag{2.21}$$

Hence, in addition to the bound in (2.19), the asymptotic, large deviations approximation is $L \sim \exp(-F(s^*))$.

For our numerical procedures we use a refinement to the above large deviations approximation [PET65, CSE93]. For the same asymptotic scaling which gives (2.21),

FACT 5:

$$L = \frac{\exp\{-F(s^*)\}}{s^*\sigma(s^*)\sqrt{2\pi}}\ [1 + o(1)] \tag{2.22}$$

where $\sigma^2(s) = \partial^2 \log E(e^{sR})/\partial s^2$. ■

More specifically,

$$\sigma^2(s) = \sum_{j=1}^J K_j \left[\frac{M_j''(s)}{M_j(s)} - \left\{\frac{M_j'(s)}{M_j(s)}\right\}^2\right] \ . \tag{2.23}$$

In the sequel we let L_0 denote the leading term on the right hand side of the expression for L in (2.22).

2.5 Approximate Buffer Distribution of General Multiplexing System

We give a formula for the buffer distribution in the general non-prioritized multiplexing system described in Section 2.1, which is used later in the paper. The prior results in this section motivate the formula, and also give complete numerical procedures for computing its elements. The formula is

$$G(x) \approx L_0 e^{zx} \ , \tag{2.24}$$

where, see (2.7) and (2.22), z is the dominant eigenvalue and L_0 is the Chernoff large deviations approximation to the instantaneous source rate exceeding the channel bandwidth. The intuition underlying (2.24) is that large buffer (x) behavior in the logarithmic scale is dominated by z and at the other extreme of bufferless multiplexing the loss is given by L_0. Also note that in typical ATM applications where cell loss probabilities are $\approx 10^{-9}$, a substantial contribution is derived from the mechanism underlying bufferless multiplexing; it is not atypical for L_0 to be in the range $10^{-3} - 10^{-5}$.

The antecedents of (2.24) are now noted. Coffman et al. [CIK91] give numerical evidence to support the claim $G(0) \approx L$. Note that $G(0) \geq L$. Simonian and Guibert [SGU94], while treating on-off sources with general on and off distributions, give a related approximation in which the exponent in the exponential term is complicated by the possible non-Markovian character of their general system. Recently in [EHL94] complete and rigorous results based on large deviations theory have been obtained for the case of time-reversible Markovian sources of arbitrary dimension. Suppose that there are K such sources of the same class (related results hold for multiple source classes) and channel bandwidth $c = \xi K$ in which ξ is held fixed as K is varied.

FACT 6: For every $x > 0$,

$$\lim_{K \to \infty} \frac{1}{K} \ \log G(Kx) \leq zx + \ell \tag{2.25}$$

where ℓ is from the Chernoff large deviations approximation. The constant z and ℓ are the best possible since

$$\lim_{x \downarrow 0} \ \lim_{K \to \infty} \ \frac{1}{K} \ \log G(Kx) = \ell \ , \tag{2.26}$$

and

$$\lim_{x\to\infty} \lim_{K\to\infty} \frac{1}{Kx} \log G(Kx) = z \ . \tag{2.27}$$

■

Although time reversibility is used crucially in the proof, it does not seem to be necessary. Also, in [EHL94] the approximation is used in the admission control of video teleconferencing sources. Each source has about 60 states and comparisons are made with simulations using data from real teleconferences. Extensive data supports the claim that the approximation (2.24) is effective.

Other approaches for improving the exponential bound on the buffer distribution are in [CLW94] and [DUF93].

3 MAIN RESULTS

3.1 Characterization of Output Process of High Priority Buffer: General Case

We characterize the output process Ω_t of Buffer 1 as another Markov-modulated fluid process $(\boldsymbol{M}^{(\Omega)}, \boldsymbol{\lambda}^{(\Omega)})$. The idea of the construction is, as described in Section 1, rooted in the large deviations-based argument that for the overwhelming fraction of time Buffer 1 is empty. Hence in the characterization we model finely the rate dynamics in the empty periods of the buffer, and lump all activity during the busy, i.e. non-empty, period into just one state. We do take care to correctly calculate the mean busy period and the distribution of states at the onset of the buffer's empty period.

Let $(\boldsymbol{M}^{(\Sigma 1)}, \boldsymbol{\lambda}^{(\Sigma 1)})$ denote the aggregate of high priority, i.e., class 1, sources. Let $\mathcal{S}_{UL}^{(1)}$ denote the set of underload states and $d_{UL}^{(1)}$ the number of such states. Let $\boldsymbol{M}_{UL,UL}^{(\Sigma 1)}$ denote the principal submatrix of $\boldsymbol{M}^{(\Sigma 1)}$ corresponding to the underload states (see Section 2.3) and denote its dimension by $d_{UL}^{(1)}$.

Our characterization $(\boldsymbol{M}^{(\Omega)}, \boldsymbol{\lambda}^{(\Omega)})$ has dimension $d_\Omega = d_{UL}^{(1)} + 1$:

$$\boldsymbol{M}^{(\Omega)} \triangleq \left[\begin{array}{c|c} \boldsymbol{M}_{UL,UL}^{(\Sigma 1)} - \mathrm{diag}\ (\boldsymbol{a}) & \boldsymbol{a} \\ \hline \boldsymbol{b} & -(1/\overline{b}) \end{array}\right] \tag{3.1}$$

$$\boldsymbol{\lambda}^{(\Omega)} \quad \triangleq \quad \left[\boldsymbol{\lambda}_{UL}^{(\Sigma 1)}, c\right] \ .$$

In (3.1),

$$a_i \triangleq \sum_{j \in \mathcal{S}_{OL}^{(1)}} M_{ij}^{(\Sigma 1)} \quad (i = 1, 2, \ldots, d_{UL}^{(1)}) \tag{3.2}$$

i.e., a_i is the sum of rates from underload state i to all overload states. Also,

$$b_i \triangleq q_i / \overline{b} \quad (i = 1, 2, \ldots, d_{UL}^{(1)}) \tag{3.3}$$

where q_i is the stationary probability that the source state is i at the onset of the empty period of Buffer 1, and $\overline{b}$ is the mean busy period of the buffer.

Specifically, $\boldsymbol{q}$ and $\boldsymbol{b}$ are obtained as follows from $\boldsymbol{\pi}_{UL}^{(\Sigma 1)}(0)$, the vector of equilibrium probabilities that Buffer 1 is empty and the aggregate source state is i, $i \in \mathcal{S}_{UL}^{(1)}$: from (2.15),

$$\boldsymbol{q} = \frac{1}{g} \ \boldsymbol{\pi}_{UL}^{(\Sigma 1)}(0) \ \boldsymbol{M}_{UL,UL}^{(\Sigma 1)} \ , \tag{3.4}$$

and g is the normalizer to ensure that $\langle \boldsymbol{q}, \mathbf{1} \rangle = 1$. Also, from (2.16),

$$\overline{b} = -G^{(1)}(0)/g \ , \tag{3.5}$$

where $G^{(1)}(0) = 1 - \langle \boldsymbol{\pi}_{UL}^{(\Sigma 1)}(0), \mathbf{1} \rangle$ is the equilibrium probability that Buffer 1 is busy, i.e., not empty.

An important property of the above characterization $(\boldsymbol{M}^{(\Omega)}, \boldsymbol{\lambda}^{(\Omega)})$ of the output process of the high priority buffer is that certain key equilibrium mean values from it exactly equal corresponding mean values of the output process Ω_t. To see this note that by construction the mean dwell time in the state d_Ω of the Markovian characterization exactly equals $\overline{b}$, the mean busy period of the buffer, and the corresponding rates in both cases have the common value c. Also, in the characterization, the conditional probability of entering state i $(i = 1, 2, \ldots, d_{UL}^{(1)})$ after exiting state d_Ω, is identical to the probability that in the output process of the buffer the empty period begins in state i. Finally, all the transition rates from state i $(i = 1, 2, \ldots, d_{UL}^{(1)})$ to other states j $(j = 1, 2, \ldots, d_{UL}^{(1)})$ are identical in the characterization to the output process's transition rates; moreover the transition rate from state i to state d_Ω in the characterization equals the inverse of the dwell time in state i just prior to the onset of the buffer's busy period. Hence, we have the following important property:

Proposition 3.1 *Let* $\boldsymbol{w}^{(\Omega)} = (w_1^{(\Omega)}, \ldots, w_{d_\Omega}^{(\Omega)})$ *denote the stationary state distribution of the source* $(\boldsymbol{M}^{(\Omega)}, \boldsymbol{\lambda}^{(\Omega)})$, *i.e.,*

$$\boldsymbol{w}^{(\Omega)}\boldsymbol{M}^{(\Omega)} = \mathbf{0} \tag{3.6}$$

and $\langle \boldsymbol{w}^{(\Omega)}, \mathbf{1}\rangle = 1$. *Then*

$$\begin{aligned} w_i^{(\Omega)} &= \pi_i^{(\Sigma 1)}(0), \quad i = 1, 2, \ldots, d_{UL}^{(1)} \\ &= G^{(1)}(0), \quad i = d_\Omega \ . \end{aligned} \tag{3.7}$$

■

Since all the rates also match, a corollary to the above result is that *all* the stationary moments of the actual rate process $R_\Omega(t)$, see Figure 1.1, are *identical* to the corresponding moments of the rate process characterized by $(\boldsymbol{M}^{(\Omega)}, \boldsymbol{\lambda}^{(\Omega)})$.

A particular desirable, but initially not obvious, implication of the above corollary is that

$$\langle \boldsymbol{w}^{(\Omega)}, \boldsymbol{\lambda}^{(\Omega)}\rangle = K_1 \langle \boldsymbol{w}^{(1)}, \boldsymbol{\lambda}^{(1)}\rangle \ , \tag{3.8}$$

i.e., the mean rate of the source $(\boldsymbol{M}^{(\Omega)}, \boldsymbol{\lambda}^{(\Omega)})$ equals the sum of the mean rates of the high priority sources, which is also the mean of the output rate process $R_\Omega(t)$.

3.2 Characterization of Output Process of the High Priority Buffer: Special Case

Upon this point in this section the sources have been general Markov modulated fluid sources. Here we specialize to the case of on-off sources with exponentially distributed on and off distributions. The numerical results in the next section are for this case. The mean on and off periods for the high priority, i.e., class 1, sources are given by $1/\alpha_1$ and $1/\beta_1$, respectively, and the rates of fluid generation in the on and off states are given by λ_1 and 0. Thus each source is described by $(\boldsymbol{M}^{(1)}, \boldsymbol{\lambda}^{(1)})$, where

$$\boldsymbol{M}^{(1)} = \begin{bmatrix} -\alpha_1 & \alpha_1 \\ \beta_1 & -\beta_1 \end{bmatrix} \quad \text{and} \quad \boldsymbol{\lambda}^{(1)} = (0 \ \lambda_1) \ . \tag{3.9}$$

In aggregating the K_1 on-off sources of class 1, we let the aggregate state denote the number of on sources at time t, with the state space, $\mathcal{S}^{(1)} = \{0, 1, \ldots, K_1\}$. The aggregate source process is birth-and-death. The set of underload states, $\mathcal{S}_{UL}^{(1)} = \{0, 1, \ldots, \lfloor c/\lambda_1 \rfloor\}$, so that the number of such states, $d_{UL}^{(1)} = \lfloor c/\lambda_1 \rfloor + 1$. The stationary state distribution $\boldsymbol{\pi}^{(\Sigma 1)}(x) = \{\pi_i^{(\Sigma 1)}(x) \mid i \in \mathcal{S}^{(1)}\}$ has been given explicitly in [AMS82].

The transition rates of the Markov chain $\boldsymbol{M}^{(\Omega)}$ have the simple form shown in Figure 3.1 in the special case. The number of states $d_\Omega = d_{UL}^{(1)} + 1$. In Figure 3.1, $a = (K_1 - \lfloor c/\lambda_1 \rfloor)\alpha_1$

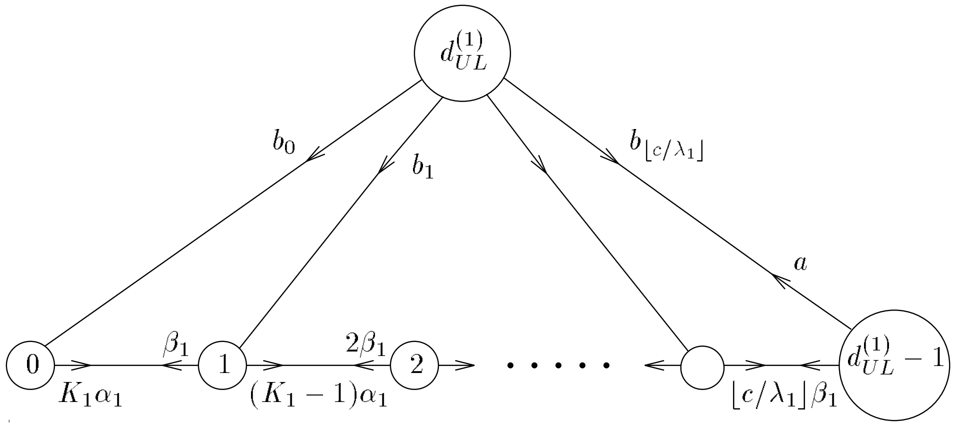


Figure 3.1: Transition diagram corresponding to $\mathbf{M}^{(\Omega)}$ for the special case.

and $b_i = q_i/\overline{b}$ $(i = 0, 1, \ldots, \lfloor c/\lambda_1 \rfloor)$, where

$$q_i = \frac{1}{g}\left[\pi_{i-1}^{(\Sigma 1)}(0)(K_1 - i + 1)\alpha_1 - \pi_i^{(\Sigma 1)}(0)\{(K_1 - i)\alpha_1 + i\beta_1\} + \pi_{i+1}^{(\Sigma 1)}(0)(i+1)\beta_1\right\} \ ; \tag{3.10}$$

it is understood that for $i = \lfloor c/\lambda_1 \rfloor$ the final term is absent and for $i = 0$ the first term is absent, and g is the normalization constant which ensures that $\sum_{i=0}^{\lfloor c/\lambda_1 \rfloor} q_i = 1$. Finally, the mean busy period in the special case,

$$\overline{b} = \frac{1 - \pi^{(\Sigma 1)}(0)}{a\pi_{\lfloor c/\lambda_1 \rfloor}^{(\Sigma 1)}(0)} \ , \tag{3.11}$$

where $\pi^{(\Sigma 1)}(0) = \sum_i \pi_i^{(\Sigma 1)}(0)$. In the special case of birth-and-death processes the normalizer g is given by a single term [EMI91], and this fact is reflected in (3.11).

3.3 Procedure for Computing the Buffer Distribution

Here we put together the elements previously discussed to obtain a complete procedure for computing the distributions of the contents of the buffers. The high priority buffer's analysis is conventional and we assume that standard techniques are available for the calculation of the joint source-buffer equilibrium distribution $\boldsymbol{\pi}^{(\Sigma 1)}(x)$. As outlined in the preceding

Sections 3.1 and 3.2, this distribution is used to synthesize $(\boldsymbol{M}^{(\Omega)}, \boldsymbol{\lambda}^{(\Omega)})$, which characterizes the output of Buffer 1.

The distribution of the contents of Buffer 2 is given by the approximation

$$G_2(x) \approx L_0^{(2)} e^{z^{(2)} x} \ , \tag{3.12}$$

where $G_2(x) = \Pr\ (X_2 \geq x)$, $L_0^{(2)}$ is the refined Chernoff large deviation approximation to the probability that the instantaneous rate from the source $(\boldsymbol{M}^{(\Omega)}, \boldsymbol{\lambda}^{(\Omega)})$ and all the low priority sources exceeds the channel bandwidth c, and $z^{(2)}$ is the dominant eigenvalue of a simple, non-prioritized system which multiplexes the source $(\boldsymbol{M}^{(\Omega)}, \boldsymbol{\lambda}^{(\Omega)})$ and K_2 independent sources $(\boldsymbol{M}^{(2)}, \boldsymbol{\lambda}^{(2)})$.

We give details of the computation of $z^{(2)}$ and $L_0^{(2)}$, which are performed as described in Sections 2.2 and 2.4, respectively. The dominant eigenvalue $z^{(2)}$ is the unique solution z in $(-\infty, 0)$ of the equation

$$g^{(\Omega)}(z) + K_2 g^{(2)}(z) = c \ , \tag{3.13}$$

where $g^{(\Omega)}(z)$ and $g^{(2)}(z)$ are the maximum real eigenvalues of $[\boldsymbol{\Lambda}^{(\Omega)} - \frac{1}{z}\ \boldsymbol{M}^{(\Omega)}]$ and $[\boldsymbol{\Lambda}^{(2)} - \frac{1}{z}\ \boldsymbol{M}^{(2)}]$, respectively. Here $\Lambda^{(\Omega)} = \text{diag}\ (\boldsymbol{\lambda}^{(\Omega)})$ and $\Lambda^{(2)} = \text{diag}\ (\boldsymbol{\lambda}^{(2)})$. Computing the maximum real eigenvalues of irreducible, essentially nonnegative matrices, such as these matrices, is almost identical to the task of computing the Perron-Frobenius eigenvalues of irreducible, nonnegative matrices, and various standard, simple techniques are available [WIL65]. In the special case where the low priority sources are on-off, and the on and off periods are exponentially distributed, $g^{(2)}(z)$ in (3.13) is given by the expression in (2.12).

As pointed out in Section 2.2, computing numerical solutions of (3.13) is straightforward since the left hand quantity is monotonic, decreasing for $z \in (-\infty, 0)$. At $z = -\infty$ its value is $(c + K\lambda_2)$ and at $z = 0$ its value is $\{K_1\langle \boldsymbol{w}^{(1)}, \boldsymbol{\lambda}^{(1)}\rangle + K_2\langle \boldsymbol{w}^{(2)}, \boldsymbol{\lambda}^{(2)}\rangle\}$. These quantities correspond to the peak and mean rates to Buffer 2 in the representation of Figure 1.2; the proof of the claim on the mean rate relies on (3.8). The stability condition implies that c exceeds the total mean rate. Hence c is between the values taken by the left hand of (3.13) at $z = -\infty$ and $z = 0$.

The constant $L_0^{(2)}$ is the approximation in (2.22) for

$$\Pr\ \left(R_\Omega(t) + \sum_{i=1}^{K_2} R_{2i}(t) \geq c \right) \ . \tag{3.14}$$

The moment generating functions are $m_\Omega(s) = \sum_{i=1}^{d_\Omega} w_i^{(\Omega)} e^{s\lambda_i^{(\Omega)}}$ and $m_2(s) = \sum_i w_i^{(2)} e^{s\lambda_i^{(2)}}$. Hence,

$$F_2(s) = sc - \log m_\Omega(s) - K_2 \log m_2(s) \ . \tag{3.15}$$

The main step in the numerical procedure for computing $L_0^{(2)}$ is the computation of s^* which solves $F_2'(s) = 0$. Again this is a straightforward procedure since $F_2'(s)$ is monotonic, decreasing for $s \geq 0$ and takes positive and negative values at $s = 0$ and $s = \infty$, respectively.

4 NUMERICAL INVESTIGATIONS

The investigations, which are reported in this section, are designed to shed light on the qualitative properties of our analytic techniques for the prioritized multiplexing system in Figure 1.1, and also to evaluate the effectiveness of the approximations. The main results are in two parts: first, on the distribution of contents of the low priority buffer and, second, on the admissible region $\mathcal{A}(B_1, B_2; L_1, L_2)$.

We developed two tools for this investigation: a computer program which implements for on-off sources all the elements of the theory which has been discussed above, and a fluid simulator for general Markovian sources. Specifically, the former implements the results in Anick et al. [AMS82] for the stationary distribution $\boldsymbol{\pi}^{(\Sigma 1)}(x)$, computes the mean busy period $\overline{b}$ and the state probabilities $\boldsymbol{q}$ at the onset of the empty period for the high priority buffer; next it synthesizes $\boldsymbol{M}^{(\Omega)}$ and $\boldsymbol{\lambda}^{(\Omega)}$; finally, it computes $L_0^{(2)}$ and $z^{(2)}$ for the distribution in the low priority buffer. The fluid simulation yields distribution for both buffers. For the computation of the admissible region both packages were exercised in a search procedure to find the maximum admissible K_2 for each fixed (admissible) K_1.

The parameters are as follows:

class 1 sources: $\alpha_1 = 1/2.5, \quad \beta_1 = 1.0, \quad \lambda_1 = 1.0.$

class 2 sources: $\alpha_2 = 1/2.5, \quad \beta_2 = 1.0, \quad \lambda_2 = 1.2.$

Thus, as in [AMS82], the unit of time is the mean on period and the unit of data is the mean amount in an on period of a high priority source. The channel bandwidth, $c = 12.1$.

Figure 4.1 gives the distributions of the contents in Buffer 2 for $\boldsymbol{K} = (17, 6)$ and $\boldsymbol{K}(17, 7)$ obtained analytically and from simulations. Observe that the analytic results are conservative except, as expected, for quite small buffer content. Table 4.1 gives the numerical values of the two key parameters of the distribution.

	$\log_{10} G(0)$	asymptotic slope of $\log_{10} G(x)$
simulations	-1.59	-0.54
analysis	-1.74	-0.44

(a) $(K_1, K_2) = (17, 6)$

	$\log_{10} G(0)$	asymptotic slope of $\log_{10} G(x)$
simulations	-1.40	-0.47
analysis	-1.54	-0.39

(b) $(K_1, K_2) = (17, 7)$

Table 4.1: Key parameters of data in Figure 4.1 for Low Priority Buffer.

In Table 4.2 note that the utilization of the channel by the combined sources is moderately high.

	class 1	class 2	total
$\boldsymbol{K} = (17, 6)$	.40	.17	.57
$\boldsymbol{K} = (17, 7)$	.40	.20	.60

Table 4.2: Utilization of the channel by sources of the two classes.

In Table 4.3 we give the analytically computed mean and busy periods of the high priority buffer for various numbers of the sources of class 1. The overwhelming dominance of the empty period over the busy period is the main motivation for the fine modelling of activity during the empty period in our characterization of the output process.

K_1	15	17	20
mean busy period	.22	.24	.29
mean empty period	18606	1712.4	158.85

Table 4.3: Analytically computed mean and busy periods of the high priority buffer.

Proceeding to the second part of the investigations on admission control, we consider equal sized buffers, $B_1 = B_2 = 5$, and bounds on loss probabilities which are more stringent for the high priority class, $L_1 = 10^{-7}$ and $L_2 = 10^{-5}$. Figure 4.2 gives the admissible sets of sources, as computed by our analytic approximations, and by a combination of simulations and exact analysis. The maximum number of class 1 sources which may be admitted while meeting the quality of service guarantee for the class is 21. Note that for $K_1 \leq 12$, the total rate from all high priority sources is less than the channel capacity and, consequently, in equilibrium the high priority buffer is always empty. Hence, (see Figure 1.2) for $K_1 \leq 12$ the second buffer is oblivious of the presence of the first buffer and the analysis of the priority system is identical to that of a non-prioritized multiplexing system with two source classes, for which exact results are known [KOS84].

The central observation on Figure 4.2 is that the admissible set is reasonably well approximated and conservatively bounded by the set obtained by using our analytic approximations. It should also be pointed out that all the data in Figure 4.2 from the analytic approximations take seconds to compute on a conventional workstation, while a single simulation takes hours.

In Figure 4.3 we compute the admissible set by our analytic approximations for a system which is larger than the one considered in Figure 4.2. The parameters of the system are given on the figure. Note that the channel capacity and the buffers are twice as large, the class 2 sources have a smaller peak rate and the requirements on the loss probabilities are several orders of magnitude more stringent and realistic.

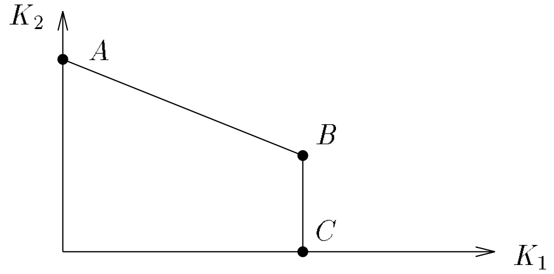
The two corner points of the admissible set are $(K_1, K_2) = (0, 129)$ and $(53, 20)$. That is, to satisfy the service requirements of class 1 at most 53 class 1 sources may be admitted; 20 class 2 sources may additionally be admitted while satisfying the requirements on class 2. The channel utilizations by class 1 and 2 sources are 0.60 and 0.11 for a total of 0.71 at the latter corner point, and at the former the total utilization is 0.73. These numbers for the total channel utilization are representative for all points on the boundary of the admissible set, and somewhat higher than the corresponding numbers for the system in Figure 4.2. Also, at the latter corner point the mean busy and empty periods are 0.128 and 58.8 at Buffer 1; for a comparison the reader is referred to the last column of Table 4.3.

The complete computation of buffer distributions for any (K_1, K_2) takes only a few seconds on a workstation even for this larger system.

5 CONCLUDING DISCUSSION

We have given a simple, fast and robust analytic technique for determining the behavior of a prioritized multiservice system in conditions of bursty offered traffic. We have, as a first step, considered only two services, "Guaranteed Service" and "Engineered Best Effort Service" with very different quality of service specifications and obtained admissible sets of traffic sources which meet the specifications. Some future directions of work are briefly discussed below.

Real-time administration of admission control calls for even speedier procedures for delineating the boundary of the admissible set. This suggests further approximation of the set by linear boundaries:



The points A and C are obtained by analyzing conventional, non-prioritized multiplexing systems. The point B may be obtained by making use of the results and techniques developed in this paper. The linear segment AB may be characterized thus, $\sum K_j e_j = c$, in which case $\{e_j\}$ may be called "effective bandwidths" of sources of the different classes.

The ability to handle highly complex traffic sources, such as video sources, for the priority scheme of this paper calls for simpler techniques which approximate the mean busy period $\bar{b}$ and the distribution $\boldsymbol{q}$ of states at the onset of the empty period of the high priority buffer. Various competing techniques are currently being researched.

Another topic under study is the replacement of the priority mechanism of access to the trunk by weighted round robin. Although this departs from the notion of plain Best Effort service, it offers attractive features. It is believed that the analytic methodology of this paper can be extended to handle the discipline.

ACKNOWLEDGEMENT

We gratefully acknowledge the support of our colleagues Flavio Bonomi and Alan Weiss. From Flavio we derived benefits from discussions ranging from the systems and architectural aspects to the characteristics of busy periods of buffers. From Alan we got the benefit of his insights into large deviations.

References

[AMS82] D. Anick, D. Mitra, and M. M. Sondhi, "Stochastic theory of a data handling system with multiple sources", *Bell Syst. Tech. J.*, **61**, pp. 1871–1894, 1982.

[BMM94] F. Bonomi, J. Meyer, S. Montagna, and R. Paglino, "Minimal on/off source models for ATM traffic", *Proc. ITC14*, Eds. J. Labetoulle and J. W. Roberts, Elsevier, pp. 387–400, 1994.

[CHJ93] C. S. Chang, P. Heidelberger, S. Juneja, and P. Shahabuddin, "Effective bandwidth and fast simulation of ATM intree networks", *Proc. Performance 93*, Elsevier, pp. 41–58, 1993.

[CIK91] E. G. Coffman, B. M. Igelnik, and Y. A. Kogan, "Controlled stochastic model of a communication system with multiple sources", *IEEE Trans. Inf. Theory*, **37**(5), pp. 1379–1387, 1991.

[CLW94] G. L. Choudhury, D. M. Lucantoni, and W. Whitt, "On the effectiveness for admission control in ATM networks", *Proc. ITC14*, Eds. J. Labetoulle and J. W. Roberts, Elsevier, pp. 411–420.

[CSE93] N. R. Chaganty and J. Sethuraman, "Strong large deviation and local limit theorems", *Ann. Prob.* **21**(3), pp. 1671–1690, 1993.

[CSZ92] D. Clark, S. Shenker, and L. Zhang, "Supporting real-time applications in an integrated services packet network: architecture and mechanism", *Proc. ACM SIGCOMM '93*, pp. 14–26, 1992.

[DUF93] N. G. Duffield, "Exponential bounds for queues with Markovian arrivals", preprint 1993.

[EMI91] A. I. Elwalid and D. Mitra, "Analysis and design of rate-based congestion control of high speed networks, I: stochastic fluid models, access regulation", *Queueing Systems* **9**, pp. 29–64, 1991.

[EMI93] A. I. Elwalid and D. Mitra, "Effective bandwidth of general Markovian traffic sources and admission control of high speed networks", *IEEE/ACM Trans. Networking* **1**(3), pp. 329–343, 1993.

[EHL94] A. I. Elwalid, D. Heyman, T. V. Lakshman, D. Mitra, and A. Weiss, "Bounds for time-reversible, Markovian buffered multiplexing systems and applications to the admission control of video teleconferencing sources", to be submitted to the special issue of *IEEE JSAC* on "Advances in the Fundamentals of Networking", 1994.

[GAN91] R. Guerin, H. Ahmadi, and M. Naghshineh, "Equivalent capacity and its application to bandwidth allocation in high-speed networks", *IEEE JSAC* **9**, pp. 968–981, 1991.

[GHU91] R. J. Gibbens and P. J. Hunt, "Effective bandwidths for the multi-type UAS channel", *Queueing System* **9**, pp. 17–28, 1991.

[HOW71] R. A. Howard, *Dynamic Probabilistic Systems, vol. 1: Markov Models* (Wiley, New York, 1971).

[HUI90] J. Y. Hui, *Switching and Traffic Theory for Integrated Broadband Networks* (Kluwer, Boston, 1990).

[KEL91] F. P. Kelly, "Effective bandwidths at multi-type queues", *Queueing Syst.* **9**, pp. 5–15, 1991.

[KOS84] L. Kosten, "Stochastic theory of data-handling systems with groups of multiple sources" in *Performance of Computer Communication Systems*, Eds. H. Rudin and W. Bux, Elsevier, pp. 321–331, 1984.

[KOS86] L. Kosten, "Liquid models for a type of information buffer problem", *Delft Prog. Report 11*, pp. 71–86, 1986.

[LI89] S.-Q. Li, "Study of information loss in packet voice systems", *IEEE Trans. Commun.*, **COM-37**, pp. 1192–1202, 1989.

[MIT88] D. Mitra, "Stochastic theory of a fluid model of producers and consumers coupled by a buffer", *Adv. Appl. Prob.* **20**, pp. 646–676, 1988.

[NEU81] M. F. Neuts, *Matrix Geometric Solutions in Stochastic Models* (John Hopkins University Press, Baltimore, 1981).

[NRS91] 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**, pp. 378–387, 1991.

[PET65] V. V. Petrov, "On the probabilities of large deviations for sums of independent random variables", *Theory of Prob. and its Applications* **X**(2), pp. 287–298, 1965.

[ROB92] J. W. Roberts, "Performance evaluation and design of multiservice networks", Final Report of the COST 224 Project, Commission of the European Communities, 1992.

[SEL91] T. E. Stern and A. I. Elwalid, "Analysis of a separable Markov-modulated rate model for information-handling systems", *Adv. Appl. Prob.* **23**, pp. 105–139, 1991.

[SGU94] A. Simonian and J. Guibert, "Large deviations approximation for fluid queues fed by a large number of on/off sources", *Proc. ITC14*, Eds. J. Labetoulle and J. W. Roberts, Elsevier, pp. 1013–1022, 1994.

[VCW94] G. de Veciana, C. Courcoubetis, and J. Walrand, "Decoupling bandwidths for networks: a decomposition approach to resource management", *Proc. IEEE INFOCOM*, 1994.

[WHI93] W. Whitt, "Tail probabilities with statistical multiplexing and effective bandwidths for multi-class queues", *Telecommun. Syst.* **2**, pp. 71–107, 1993.

[WIL65] J. H. Wilkinson, *The Algebraic Eigenvalue Problem* (Clarendon, Oxford, UK, 1965).

[ZHA93] J. Zhang, "Performance study of Markov modulated fluid flow models with priority traffic", *Proc. IEEE INFOCOM*, pp. 10–17, 1993.

Figure 4.1: Distribution of contents of low priority Buffer 2 obtained by analytic approximations and simulations. The system parameters are in the text and on Figure 4.2.

Figure 4.2: The admissible set of high and low priority sources which satisfy requirements on loss probabilities for given buffer sizes.

Figure 4.3: The admissible set obtained by analytic approximations.