Queueing Systems 27 (1997) 79–97

Admission control of multi-class traffic with service priorities in high-speed networks *

V.G. Kulkarni and N. Gautam

Department of Operations Research, University of North Carolina, Chapel Hill, NC 27599-3180, USA
E-mail: {vg_kulkarni;gautam}@unc.edu

Received 29 November 1995; revised 15 June 1997

We consider a fluid model of a system that handles multiple classes of traffic. The delay and cell-loss requirements of the different classes of traffic are generally widely different and are achieved by assigning different buffers for different classes, and serving them in a strict priority order. We use results from the effective bandwidth of the output processes (see Chang and Thomas (1995)) to derive simple and asymptotically exact call-admission policies for such a system to guarantee the cell-loss requirements for the different classes assuming that each source produces a single class traffic. We compare the admission-control policies developed here with the approximate policy studied by Elwalid and Mitra (1995) for the case of two-class traffic.

Keywords: quality-of-service, fluid-flow models, effective bandwidth, multi-priority traffic, Chernoff dominant eigenvalue

1. Introduction

The concept of effective bandwidth and its use in the admission control for the statistical multiplexing of bursty sources is now well-documented and accepted (see Gibbens and Hunt [11], Kesidis et al. [12], Elwalid and Mitra [8], Choudhury et al. [3], Whitt [18], etc.). In the emerging high-speed networks using asynchronous transfer mode (ATM), each traffic-source is described by its stochastic characteristics, and is assured a quality of service (QoS), as measured by cell-loss probability, delay, delay-jitter, etc. The effective bandwidth is a number associated with a traffic-source such that if the sum of the effective bandwidths of all the sources multiplexed onto a buffer is less than the output rate of that buffer, then the QoS is satisfied for each source.

This method of admission control works quite satisfactorily as long as the QoS requirements of the sources are the same or at least similar. Otherwise buffer-sizing has to be done to assure the most stringent QoS for all the sources. This leads to unnecessarily large buffer sizes.

When the multiplexed traffic has widely differing QoS requirements (as will most

* This work was partially supported by NSF Grant No. NCR-9406823.

certainly be the case since the high-speed network is expected to carry all the traffic: video, voice and data) we need to look for other alternatives. There are two distinct cases that arise in applications.

In the first case (for example, the digitized voice with low-end bit-dropping, or MPEG2 video, or output from a leaky bucket) each source produces multi-class traffic, and although different classes can tolerate different cell-losses, the accepted traffic from a given source must reach the destination in the order in which it was generated. This necessitates first-come-first-served scheduling at each node enroute. This kind of traffic requirement is best handled by buffer-sharing schemes, or space-priority mechanisms (see Çidon et al. [4,5], Elwalid and Mitra [7], and Lin and Sylvester [16]). Kulkarni et al. [15] show that the effective bandwidth concept can be extended to effective bandwidth vectors and used for admission control in this case.

In the second case each source produces a single-class traffic, but different sources have different QoS requirements. For example, real-time traffic has a more stringent delay requirement but can tolerate higher cell-loss; while data traffic can tolerate higher delay but demands much smaller cell losses. In such cases it is feasible to provide a separate buffer for each class and service the real-time traffic buffer with higher priority than the data-traffic buffer. Several service priorities are possible. The simplest scheduling discipline gives full priority to real-time traffic and the channel capacity that is not used by the real-time traffic is made available to the data traffic.

In this paper we concentrate on this second case and consider a fluid model of N distinct classes. There are K_j independent sources ($j = 1, 2, \ldots, N$) producing fluid of type j that is multiplexed into a buffer of size B_j. The fluid is removed from these buffers following a static priority full service (SPFS) policy, i.e., fluid of type j is removed before fluid of type i if $j < i$. The main aim of this paper is to identify the conditions under which the cell-loss probability requirement is satisfied for each class. We do this by using the results on the effective bandwidth of output processes as reported by de Veciana et al. [6], Chang and Thomas [1], and Chang and Zajic [2].

To do this we need to analyze multi-priority fluid models. Some work is already done in this area: see Narayanan and Kulkarni [17] and Zhang [19]. A recent paper by Elwalid and Mitra [9] provides an approximate way of solving the admission control problem for the two priority case. They approximate the busy periods of the high-priority buffers by exponential distributions to obtain tractable solutions, and show that the approximation works quite well. They also incorporate the new results using Chernoff bounds in their analysis. Our results provide a simple admission control criterion that does not use the exponential approximation of Elwalid and Mitra [9]. When there are two types of fluids, the SPFS policy is identical to the generalized processor sharing scheduling discipline analyzed by Zhang et al. [20,21].

The rest of the paper is organized as follows. The next section restates some of the known results from large deviations. It also states an important characterization of the effective bandwidth of the output in terms of that of the input.

The multi-priority fluid model (with K_j sources of type j) is described in section 3 to set the notation. We observe that the multi-priority model is identical to a tandem

fluid model, where the fluid of type j and the output from the buffers $1, 2, \ldots, j-1$ are input to the jth buffer in tandem. This observation is also made in Elwalid and Mitra [9] (for $N = 2$) and is immensely useful in our analysis.

The admission control criterion is formally derived in section 4. The main result is given in theorem 3. An admission criterion that does not use the output analysis is even more simple and is shown to be conservative. Thus, the admission control in practice can be done rather simply, using existing effective bandwidth formulae. The results are illustrated with an example of two classes of exponential on–off sources in section 5.

In section 6 the admission control is fine tuned using Chernoff bounds as stated in Elwalid et al. [9,10]. It takes into account the gain in statistical multiplexing. Three methods are stated to compute the relevant tail probabilities. The results are demonstrated for a two-class exponential on–off source model in section 7. Numerical examples are used to compare the three methods with each other and that in Elwalid and Mitra [9].

2. Preliminaries

Consider a single-buffer fluid model driven by a random environment $\{Z(t),\ t \geqslant 0\}$. The buffer has infinite capacity and is serviced by a channel of constant capacity c. When the environment is in state $Z(t)$, fluid enters the buffer at rate $r(Z(t))$. Let $X(t)$ be the amount of fluid in the buffer at time t. The dynamics of the buffer content process $\{X(t),\ t \geqslant 0\}$ is described by

$$\frac{\mathrm{d}X(t)}{\mathrm{d}t} = \begin{cases} r(Z(t)) - c & \text{if } X(t) > 0, \\ \{r(Z(t)) - c\}^+ & \text{if } X(t) = 0, \end{cases} \tag{1}$$

where $\{x\}^+ = \max(x, 0)$.

It has been shown in Kulkarni and Rolski [14] that the buffer content process $\{X(t),\ t \geqslant 0\}$ is stable if

$$E\{r(Z(\infty))\} < c. \tag{2}$$

Typically, in applications we are interested in satisfying a Quality of Service criterion that can be mathematically formulated as follows:

$$\lim_{t\to\infty} P(X(t) > B) \leqslant \varepsilon.$$

One can think of B as the finite buffer size and ε as the upper bound on the overflow probability. It is known that (see Elwalid and Mitra [8], Gibbens and Hunt [11], and Kesidis et al. [12]) in the asymptotic region, i.e., as $B \to \infty$ and $\varepsilon \to 0$, such that $-\log(\varepsilon)/B \to \delta > 0$, there exists a quantity $eb(\delta)$, called the effective bandwidth of the input source, such that the QoS criterion is satisfied if

$$eb(\delta) < c.$$

We recount below an important result by Kesidis et al. [12] that relates $eb(\delta)$ to the input process. Let $A(t)$ be the total fluid input until time t, where

$$A(t) = \int_0^t r\big(Z(u)\big)\,\mathrm{d}u.$$

Define

$$h_A(v) = \lim_{t\to\infty} \frac{1}{t} \log E\big\{ \exp\big(vA(t)\big)\big\}. \tag{3}$$

Kesidis et al. [12] show that the effective bandwidth of the input is given by

$$eb_A(\delta) = \frac{h_A(\delta)}{\delta}. \tag{4}$$

They also state that $h_A(v)$ is an increasing, convex function of v.

The important question is how to compute $h_A(\cdot)$ for a given input process. Elwalid and Mitra [8], and Kesidis et al. [12] show how to do this when $\{Z(t),\ t \geqslant 0\}$ is a Continuous Time Markov Chain (CTMC). Kulkarni [13] shows how to do this when $\{Z(t),\ t \geqslant 0\}$ is a Markov Regenerative Process (MRGP).

Now let $D(t)$ be the total output from the buffer over $[0, t]$. In practice, $D(t)$ may act as an input for a downstream buffer (e.g., in case of *tandem queues*). Hence it is useful to know the effective bandwidth of the output process. Analogous to (3), define

$$h_D(v) = \lim_{t\to\infty} \frac{1}{t} \log E\big\{ \exp\big(vD(t)\big)\big\}. \tag{5}$$

Note that $h_D(v)$ is also a convex, increasing function and $h'_D(v) \leqslant c$, since the peak rate of the output process is bounded above by c, the channel-capacity.

The next theorem (from Chang and Thomas [1], and, Chang and Zajic [2]) establishes the relationship between $h_A(v)$ and $h_D(v)$. ($h'_A(v)$ denotes the derivative of $h_A(v)$ with respect to v.)

Theorem 1. Suppose $\{Z(t),\ t \geqslant 0\}$ is a stationary ergodic process satisfying the Gärtner–Ellis conditions (see Kesidis et al. [12]). Let v^* satisfy

$$h'_A(v^*) = c. \tag{6}$$

Then

$$h_D(v) = \begin{cases} h_A(v) & \text{if } 0 \leqslant v \leqslant v^*, \\ h_A(v^*) - cv^* + cv & \text{if } v > v^*. \end{cases} \tag{7}$$

See figure 1 for an illustration of $h_A(v)$ and $h_D(v)$. Using equation (4) we can relate the effective bandwidth $eb_D(\delta)$ of the output to the effective bandwidth $eb_A(\delta)$ of the input.

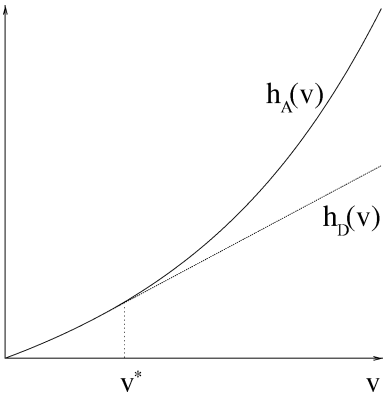

Figure 1. $h_A(v)$ and $h_D(v)$ vs v.

Corollary 2.

$$eb_D(\delta) = \begin{cases} eb_A(\delta) & \text{if } 0 \leqslant \delta \leqslant v^*, \\ c - \dfrac{v^*}{\delta}\left\{c - eb_A(v^*)\right\} & \text{if } \delta > v^*. \end{cases} \tag{8}$$

3. The multi-class single-node model

In this section we consider a single node of a multi-class telecommunications network where N distinct classes of fluid are serviced by a single channel of capacity c. We assume that there are N buffers, and class-j traffic flows into buffer j, $j = 1, 2, \ldots, N$. A scheduler always removes fluid according to a static priority full service policy which is described as follows: assign all the available capacity for the class-1 fluid and assign the leftover channel capacity (if any) that class-1 does not need, to class-2 fluid. Then assign the leftover channel capacity (if any) that class-1 and class-2 do not need, to class-3 fluid, and so on. We later present a mathematical description of this policy.

There are K_j independent and identical sources generating class-j traffic that gets multiplexed onto buffer j. Refer to figure 2 for a schematic representation of the model.

Let the ith source of class j be driven by a random environment process $\{Z_j^i(t),\ t \geqslant 0\}$. We assume that $\{Z_j^i(t),\ t \geqslant 0\}$ $(i = 1, 2, \ldots, K_j)$ are independent and identical stationary and ergodic processes satisfying the Gärtner–Ellis conditions (see Kesidis et al. [12]). At time t, when the environment is in state $Z_j^i(t)$, fluid is generated by the ith source of class j at rate $r_j(Z_j^i(t))$.

Let $A_j^i(t)$ be the total amount of fluid input from the source i of class j into buffer j up to time t, i.e.,

$$A_j^i(t) = \int_0^t r_j\big(Z_j^i(u)\big)\,\mathrm{d}u.$$

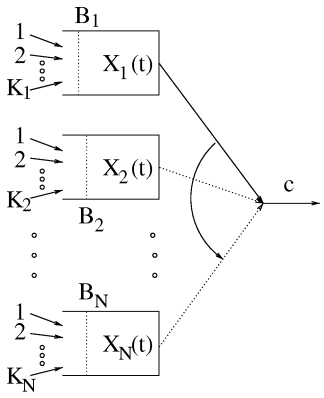

Figure 2. The model.

The corresponding asymptotic log moment generating functions (ALMGF) for each of the K_j sources are identical and equal to

$$h_j(v) = \lim_{t\to\infty} \frac{1}{t} \log E\{ \exp\big(vA_j^i(t)\big)\}. \tag{9}$$

Thus the net ALMGF of all K_j sources inputting class j sources into buffer j is $K_j\, h_j(v)$.

Also, the effective bandwidth of each of the independent input sources into buffer j is

$$eb_j(\delta) = h_j(\delta)/\delta. \tag{10}$$

The channel-capacity c is used to serve the N buffers according to the following policy. As long as there is fluid in buffer 1, the channel serves it at rate c. When buffer 1 is empty (this can happen only if $\sum_{i=1}^{K_1} r_1(Z_1^i(t)) < c$), the channel serves it at rate

$$\sum_{i=1}^{K_1} r_1\big(Z_1^i(t)\big)$$

and offers the remaining capacity

$$c - \sum_{i=1}^{K_1} r_1\big(Z_1^i(t)\big)$$

to buffer 2. Similarly, as long as there is fluid in buffer 2, (and buffer 1 is empty), the channel serves it at rate

$$\left[c - \sum_{i=1}^{K_1} r_1\big(Z_1^i(t)\big)\right]^+.$$

When buffer 2 is empty it is served at rate

$$\min\left\{\sum_{i=1}^{K_2} r_2\big(Z_2^i(t)\big),\ \left[c-\sum_{i=1}^{K_1} r_1\big(Z_1^i(t)\big)\right]^+\right\}.$$

Similarly, as long as there is fluid in buffer 3, it gets served at rate

$$\left[c-\sum_{i=1}^{K_1} r_1\big(Z_1^i(t)\big)-\sum_{i=1}^{K_2} r_2\big(Z_2^i(t)\big)\right]^+,$$

and so on.

Let $X_j(t)$ be the amount of fluid in the buffer j at time t. We shall analyze the $\{X_j(t),\ t \geqslant 0\}$ process assuming infinite buffers. Due to strict priority rules, we see that $\{X_j(t),\ t \geqslant 0\}$ does not depend upon $K_{j+1}, \ldots, K_N$. We require a modified version of the system stability condition as stated in Kulkarni and Rolski [14] as

$$\lim_{t\to\infty} \sum_{j=1}^{N}\sum_{i=1}^{K_j} E\big\{r_j\big(Z_j^i(t)\big)\big\} < c. \tag{11}$$

Let ε_j be the cell-loss-probability target for class-j traffic. Thus we want to satisfy the following Quality-of-Service criteria for the N classes:

$$G_j(K_1, K_2, \ldots, K_j) = \lim_{t\to\infty} P\big(X_j(t) > B_j\big) \leqslant \varepsilon_j,$$

where B_j is a given number ($j = 1, 2, \ldots, N$).

The main aim of the analysis in the next section is to identify the feasible region

$$\mathcal{K} = \big\{(K_1, K_2, \ldots, K_N)\colon\ G_1(K_1) \leqslant \varepsilon_1, \ldots, G_N(K_1, K_2, \ldots, K_N) \leqslant \varepsilon_N\big\}. \tag{12}$$

4. Analysis

We shall concentrate on the asymptotic region:

$$B_j \to \infty \quad \text{and} \quad \varepsilon_j \to 0, \quad \text{such that} -\log(\varepsilon_j)/B_j \to \delta_j > 0,\ j = 1, 2, \ldots, N.$$

We shall treat each of the priority cases separately.

- **Priority** 1. Since the priority-1 fluid gets uninterrupted service, the call-admission policy is identical to the case where there is no other traffic. The effective bandwidth of K_1 priority-1 fluid sources is given by $K_1\, eb_1(\delta)$ (see (10)). It is known that the QoS criteria $G_1(K_1) \leqslant \varepsilon_1$ is satisfied if

$$K_1\, eb_1(\delta_1) < c. \tag{13}$$

- **Priority** j ($j = 2, 3, \ldots, N$). The capacity available to buffer j is 0 when at least one of the buffers $1, 2, \ldots, j-1$ is non-empty and it is

$$\left[c - \sum_{k=1}^{j-1} \sum_{i=1}^{K_k} r_k\big(Z_k^i(t)\big)\right]^+$$

if all the buffers $1, 2, \ldots, j-1$ are empty. Let $R_{j-1}(t)$ be the sum of the output rates of the buffers $1, 2, \ldots, j-1$ at time t. (Thus $R_0(t) = 0$.) It can be seen that

$$R_{j-1}(t) = \begin{cases} c & \text{if } \sum_{k=1}^{j-1} X_k(t) > 0, \\ \min\left[c, \sum_{i=1}^{K_1} r_1\big(Z_1^i(t)\big)\right] & \text{if } \sum_{k=1}^{j-1} X_k(t) = 0. \end{cases} \tag{14}$$

It is clear that the buffer j gets served with capacity $c - R_{j-1}(t)$ at time t. Thus it is easy to see that the buffer j can be equivalently modeled as one that is served at a constant rate c, but has an additional compensating source producing fluid at rate $R_{j-1}(t)$ at time t. (The compensating source j is independent of the K_j sources of priority j.) Since $R_{j-1}(t)$ is the rate at which fluid is departing from buffers $1, 2, \ldots, j-1$, the effective bandwidth of the compensating source for the jth buffer, $eb_j^s(\delta)$, is equal to the effective bandwidth of the sum of the outputs of buffers $1, 2, \ldots, j-1$. Note that $eb_1^s(\delta) = 0$ for all δ. From corollary 2 of section 2, we have the effective bandwidth of the compensating source for buffer j recursively given by

$$\begin{aligned} &eb_1^s(\delta) = 0 \quad \text{for all } \delta, \\ &eb_j^s(\delta) = \begin{cases} K_{j-1}\, eb_{j-1}(\delta) + eb_{j-1}^s(\delta) & \text{if } 0 \leqslant \delta \leqslant v_j^*, \\ c - \dfrac{v_j^*}{\delta}\big\{c - K_{j-1}\, eb_{j-1}(v_j^*) - eb_{j-1}^s(v_j^*)\big\} & \text{if } \delta > v_j^*, \end{cases} \\ &j \geqslant 2, \end{aligned} \tag{15}$$

where v_j^* is obtained by solving for v in the equation

$$\frac{\mathrm{d}}{\mathrm{d}v}\big[v\big(K_{j-1}\, eb_{j-1}(v) + eb_{j-1}^s(v)\big)\big] = c.$$

Then the QoS criteria is satisfied for buffer j if

$$K_j eb_j(\delta_j) + eb_j^s(\delta_j) < c.$$

Combining the above results we get the following theorem.

Theorem 3. $(K_1, K_2, \ldots, K_N) \in \mathcal{K}$ (see equation (12)), i.e., the QoS criteria

$$G_j(K_1, K_2, \ldots, K_j) \leqslant \varepsilon_j, \quad j = 1, 2, \ldots, N,$$

are satisfied if

$$K_j eb_j(\delta_j) + eb_j^s(\delta_j) < c, \quad \forall j = 1, 2, \ldots, N, \tag{16}$$

where $eb_1^s(\cdot) = 0$, $eb_j^s(\cdot)$ is as in equation (15) and $eb_j(\cdot)$ is as in equation (10).

Next we describe the approximation to $\mathcal{K}$ that eliminates the need to compute v_j^* and $eb_j^s(\cdot)$. Let $\mathcal{N}$ be the set of points $(K_1, K_2, \ldots, K_N)$ such that

$$\mathcal{N} = \left\{ (K_1, K_2, \ldots, K_N): \sum_{k=1}^{j} K_k eb_k(\delta_k) < c, \quad \text{for } j = 1, 2, \ldots, N \right\}. \tag{17}$$

Using (15) and the fact that $h_j(v)$ is an increasing, convex function, one can prove that $eb_j^s(v) < \sum_{i=1}^{j-1} K_i eb_i(v)$. Hence, we can easily prove the following theorem.

Theorem 4. $\mathcal{N} \subset \mathcal{K}$.

Thus the admission-control policy based on the simpler set of inequalities (17), rather than (16), is more conservative. We illustrate the results by means of an example in the next section.

5. Exponential on–off sources

Consider the multiclass node (in section 3) with two classes of traffic. Class-1 is the real-time traffic while class-2 is the non-real-time traffic. Class-1 traffic is given higher service priority over class-2 traffic. We assume that there are two buffers, with class-j traffic coming into buffer j, $j = 1, 2$. A scheduler removes (capacity c) fluid from the buffers according to a static priority full service policy. Each of the K_j class-j sources, $j = 1, 2$, are independent and identical on–off sources with exp(α_j) on-times and exp(β_j) off-times. When a class-j source is on, it generates fluid at rate r_j and when it is off, it generates fluid at rate 0.

From theorem 3 in section 4, we can derive the following results. $(K_1, K_2) \in \mathcal{K}$ (see equation (12)), i.e., the QoS criteria

$$G_1(K_1) \leqslant \varepsilon_1, \qquad G_2(K_1, K_2) \leqslant \varepsilon_2$$

are satisfied if

$$\begin{aligned} &\text{(i)} \quad K_1 eb_1(\delta_1) < c, \quad \text{and} \\ &\text{(ii)} \quad K_1 eb_1(\delta_2) + K_2\, eb_2(\delta_2) < c \qquad \text{if } \delta_2 \leqslant v^*, \\ &\qquad \frac{v^*}{\delta_2} K_1 eb_1(v^*) + K_2 eb_2(\delta_2) < \frac{cv^*}{\delta_2} \quad \text{if } \delta_2 > v^*, \end{aligned} \tag{18}$$

where

$$v^* = \frac{\beta_1}{r_1}\left(\sqrt{\frac{c\alpha_1}{\beta_1(K_1 r_1 - c)}} - 1\right) + \frac{\alpha_1}{r_1}\left(1 - \sqrt{\frac{\beta_1(K_1 r_1 - c)}{c\alpha_1}}\right)$$

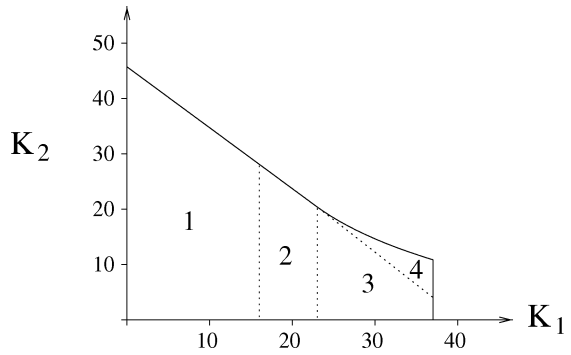

Figure 3. Feasible values of (K_1, K_2).

and (from Elwalid and Mitra [8] and Kesidis et al. [12])

$$eb_j(\delta_j) = \frac{r_j\delta_j - \alpha_j - \beta_j + \sqrt{(r_j\delta_j - \alpha_j - \beta_j)^2 + 4\beta_j r_j \delta_j}}{2\delta_j}.$$

The acceptance region for the following numerical problem is shown in figure 3.

$$\alpha_1 = 2.4, \quad \beta_1 = 0.4, \quad r_1 = 2.0, \quad \varepsilon_1 = 10^{-7}, \quad B_1 = 10,$$
$$\alpha_2 = 1.0, \quad \beta_2 = 0.4, \quad r_2 = 1.2, \quad \varepsilon_2 = 10^{-5}, \quad B_2 = 8 \quad \text{and} \quad c = 32.1.$$

Refer to figure 3 and inequalities (18). In region 1, $K_1 r_1 < c$ and hence $v^* = \infty$. Therefore (i) is trivially satisfied, and (ii) reduces to $K_1\, eb_1(\delta_2) + K_2\, eb_2(\delta_2) < c$. In region 2 $K_1 r_1 > c$ and $\delta_2 \leqslant v^*$. Therefore we still use $K_1 eb_1(\delta_2) + K_2 eb_2(\delta_2) < c$. Now if $\delta_2 > v^*$ and we continue to use $K_1 eb_1(\delta_2) + K_2 eb_2(\delta_2) < c$ we get region 3 (bounded above by the dotted line). Instead, if we use

$$\frac{v^*}{\delta_2} K_1 eb_1(v^*) + K_2 eb_2(\delta_2) < \frac{cv^*}{\delta_2},$$

we get an extra set of feasible values of (K_1, K_2) that constitute region 4. Thus the region $\mathcal{K}$ defined by (16) is the union of regions $1, 2, 3$ and 4, while the approximate region $\mathcal{N}$, defined by (17) is the union of regions $1, 2$ and 3. Note that $\mathcal{N}$ can be significantly smaller than $\mathcal{K}$.

6. Chernoff bounds

It can be shown that the region $\mathcal{K}$ (and hence of course $\mathcal{N}$) from the previous sections, is conservative, mainly because the statistical multiplexing gains are not taken advantage of. In this section we show how we can use the Chernoff Dominant Eigenvalue (CDE) approximation (see Elwalid et al. [9,10]) to further fine tune the call admission control problem analysis. The CDE approximation for the tail probability (for $j = 1, 2, \ldots, N$) is given by

$$\lim_{t\to\infty} P\big(X_j(t) > B_j\big) \approx L_j \mathrm{e}^{-\zeta_j B_j},$$

where L_j is the fraction of the class j fluid that would be lost if there was no buffer.

Mathematically L_j, $j = 1, 2, \ldots, N$, can be written as

$$L_j = \lim_{t\to\infty} \frac{\int_0^t \{R_{j-1}(t) + [\sum_{i=1}^{K_j} r_j(Z_j^i(t))] - c\}^+ \,\mathrm{d}t}{\int_0^t \{R_{j-1}(t) + \sum_{i=1}^{K_j} r_j(Z_j^i(t))\}\,\mathrm{d}t}, \tag{19}$$

where $R_{j-1}(t)$ is the sum of the rates at which fluid is output from buffers $1, 2, \ldots, j-1$ at time t as defined in (14) with $R_0(t) = 0$ for all t. Note that L_j is a function of c, $K_1, \ldots, K_j$, and the parameters of the sources of class j or less. Typically it may not be computationally simple to calculate L_j exactly. Hence Elwalid et al. [10] suggest a method of estimating L_j by using Chernoff's theorem. We explain it briefly below.

We characterize the input sources of class-j by a function $m_j(w)$, which is similar to $h_j(v)$ function, and is defined as

$$m_j(w) = \lim_{t\to\infty} \log E\{\exp(wr_j(Z_j^i(t)))\}. \tag{20}$$

Note that $m_j(w)$ does not depend on i, for $i = 1, 2, \ldots, K_j$, since the K_j sources are identical. Let

$$u_j^* = \sup_{w\geqslant 0} \left\{ cw - \sum_{j=1}^{N} K_j m_j(w) \right\}.$$

and w_j^* be obtained by solving

$$\sum_{j=1}^{N} K_j\, m_j'(w_j^*) = c.$$

Then the Chernoff estimate of L_j as given in Elwalid et al. [9,10] is

$$L_j \approx \frac{\exp(-u_j^*)}{w_j^* \sigma(w_j^*)\sqrt{2\pi}}, \tag{21}$$

where

$$\sigma^2(w_j^*) = \sum_{j=1}^{N} K_j\, m_j''(w_j^*).$$

The main problem in the above analysis is computing $m_j(w)$. If $\{Z_j^i(t),\ t \geqslant 0\}$ can be modeled as a stationary and ergodic process with state space $\mathcal{S}_j$ and stationary probability vector, π_j, we have

$$m_j(w) = \log \left\{ \sum_{k\in\mathcal{S}_j} \pi_j^k \mathrm{e}^{wr_j(k)} \right\}. \tag{22}$$

Then, we identify the new feasible region $\overline{\mathcal{K}}$ using the following theorem.

Theorem 5. $(K_1, K_2, \ldots, K_N) \in \overline{\mathcal{K}}$, i.e., the QoS criteria

$$G_j(K_1, K_2, \ldots, K_j) = L_j \mathrm{e}^{-\zeta_j B_j} \leqslant \varepsilon_j, \quad j = 1, 2, \ldots, N,$$

are satisfied if

$$K_j eb_j(\zeta_j) + eb_j^s(\zeta_j) < c, \tag{23}$$

where $\zeta_j,\ j = 1, 2, \ldots, N$, is given by

$$\zeta_j = -\frac{\log(\varepsilon_j / L_j)}{B_j},$$

$eb_1^s(\cdot) = 0,\ eb_j^s(\cdot)$ is as in equation (15) and $eb_j(\cdot)$ is as in equation (10) with $eb_j^s(v) = eb_j(v) = 0$ for $v < 0$.

We illustrate the above results with the following example.

7. CDE method for exponential on–off sources

In this section we consider the model of section 5 and show how the acceptance region changes with the choice of the method used to compute L_j. We also compare the results obtained with those in Elwalid and Mitra [9]. Using the results from theorem 5 in section 6, we get the following feasible region $\overline{\mathcal{K}}$.

$(K_1, K_2) \in \overline{\mathcal{K}}$, i.e., the QoS criteria

$$G_1(K_1) = L_1 \mathrm{e}^{-\zeta_1 B_1} \leqslant \varepsilon_1, \qquad G_2(K_1, K_2) = L_2 \mathrm{e}^{-\zeta_2 B_2} \leqslant \varepsilon_2, \tag{24}$$

are satisfied if

$$K_1 eb_1(\zeta_1) < c, \qquad K_2 eb_2(\zeta_2) + eb_2^s(\zeta_2) < c, \tag{25}$$

where $\zeta_j,\ j = 1, 2$, is given by

$$\zeta_j = -\frac{\log(\varepsilon_j / L_j)}{B_j},$$

$$eb_2^s(v) = \begin{cases} K_1\, eb_1(v) & \text{if } 0 \leqslant v \leqslant v^*, \\ c - \dfrac{v^*}{v}\big\{c - K_1\, eb_1(v^*)\big\} & \text{if } v > v^*, \\ 0 & \text{if } v < 0 \end{cases} \tag{26}$$

and

$$eb_j(v) = \begin{cases} \dfrac{r_j v - \alpha_j - \beta_j + \sqrt{(r_j v - \alpha_j - \beta_j)^2 + 4\beta_j r_j v}}{2v} & \text{if } v \geqslant 0, \\ 0 & \text{if } v < 0. \end{cases}$$

Note that ζ_j would be negative when $L_j < \varepsilon_j$. We need to explicitly account for this possibility.

The major effort in the above analysis lies in computing L_1 and L_2. We discuss three methods. The estimate of L_i ($i = 1, 2$) obtained by method j ($j = 1, 2, 3$) is denoted by $L_i^{(j)}$. The feasible regions obtained by using $L_1^{(j)}$ and $L_2^{(j)}$ is denoted by

$$\overline{\mathcal{K}}^{(j)} = \left\{(K_1, K_2)\text{: } L_1^{(j)}\mathrm{e}^{-\zeta_1 B_1} \leqslant \varepsilon_1,\ L_2^{(j)}\mathrm{e}^{-\zeta_2 B_2} \leqslant \varepsilon_2\right\}.$$

Method 1. $L_1^{(1)} = 1$ and $L_2^{(1)} = 1$. We know that $L_1 \leqslant 1$ and $L_2 \leqslant 1$, hence this is a conservative estimate. Then the admissible region $\overline{\mathcal{K}}^{(1)}$ is the same as $\mathcal{K}$ in section 5 and is shown in figure 3.

Method 2. Using the independent on–off nature of the inputs of class-1, we can obtain an exact expression for L_1 of equation (19) as

$$L_1^{(2)} = \sum_{i=\lceil c/r_1\rceil}^{K_1} \left(1 - \frac{c}{ir_1}\right) \frac{K_1!}{i!\,(K_1 - i)!} \frac{(\beta_1)^i\,(\alpha_1)^{K_1 - i}}{(\alpha_1 + \beta_1)^{K_1}}. \tag{27}$$

To compute $L_2^{(2)}$, we use (21). First note that we analyze the buffer content process of the second buffer by assuming that the output rate is always c and the input is from $K_2 + 1$ sources, viz., K_2-exponential on–off sources of type 2 and one compensating source producing fluid at rate $R_1(t)$ at time t. Using equation (22) the m-function for the K_2 exponential on–off sources is seen to be

$$m_2(w) = \log\left\{\frac{\alpha_2}{\alpha_2 + \beta_2} + \frac{\beta_2}{\alpha_2 + \beta_2}\mathrm{e}^{w\,r_2}\right\}. \tag{28}$$

We show in appendix 1 that the m-function for the compensating source (see equations (34) and (39)) is given by

$$m_1(w) = \begin{cases} \log\left\{\sum_{k=0}^{M} \pi_1^k \mathrm{e}^{wkr_1}\right\} & \text{if } K_1 \leqslant \left\lfloor \dfrac{c}{r_1}\right\rfloor, \\ \log\left\{\pi_1^M\,\mathrm{e}^{wc} + \sum_{k=0}^{M-1} \pi_1^k\,\mathrm{e}^{wkr_1}\right\} & \text{if } K_1 > \left\lfloor \dfrac{c}{r_1}\right\rfloor, \end{cases} \tag{29}$$

where M and the probabilities $\pi_1^0, \pi_1^1, \ldots, \pi_1^M$ are derived in appendix 1 (equations (32), (33) and (38)). Using $m_1(w)$ and $m_2(w)$, compute

$$u_2^* = \sup_{w \geqslant 0} \left\{cw - m_1(w) - K_2 m_2(w)\right\}$$

and obtain w_2^* by solving

$$m_1'(w_2^*) + K_2 m_2'(w_2^*) = c.$$

Then compute $L_2^{(2)}$ using equation (21) and obtain the feasible region $\overline{\mathcal{K}}^{(2)}$.

Method 3. Compute $L_1^{(3)} = L_1^{(2)}$. Instead of using Chernoff theorem to compute L_2, we directly compute $L_2^{(3)}$ by the following.

If $K_1 \leqslant \lfloor c/r_1 \rfloor$,

$$L_2^{(3)} = \sum_{k=\lceil \frac{c-ir_1}{r_2} \rceil}^{K_2} \sum_{i=0}^{M} \pi_1^i \left(\frac{\alpha^{K_2-k}\beta^k}{(\alpha+\beta)^{K_2}} \right) \left(\frac{ir_1 + kr_2 - c}{ir_1 + kr_2} \right) \frac{K_2!}{k!\,(K_2-k)!}$$

and if $K_1 > \lfloor c/r_1 \rfloor$,

$$L_2^{(3)} = \sum_{k=\lceil \frac{c-ir_1}{r_2} \rceil}^{K_2} \left(\frac{\alpha^{K_2-k}\beta^k}{(\alpha+\beta)^{K_2}} \right) \frac{K_2!}{k!\,(K_2-k)!} \sum_{i=0}^{M} \pi_1^i \left(1 - \frac{c}{\min(c, ir_1) + kr_2} \right),$$

where M and the probabilities $\pi_1^0, \pi_1^1, \ldots, \pi_1^M$ are derived in the appendix (see equations (32), (33) and (38)). The admissible region obtained by this method is $\overline{\mathcal{K}}^{(3)}$. Utilizing the fact that

$$L_1^{(3)} = L_1^{(2)} \leqslant L_1^{(1)} = 1 \quad \text{and} \quad L_2^{(3)} \leqslant L_2^{(2)} \leqslant L_2^{(1)} = 1,$$

we can easily prove the following theorem that summarizes the ordering of the regions obtained in methods 1, 2 and 3.

Theorem 6. $\mathcal{N} \subset \overline{\mathcal{K}}^{(1)} \subset \overline{\mathcal{K}}^{(2)} \subset \overline{\mathcal{K}}^{(3)}$.

We now illustrate the regions $\mathcal{N}$, $\overline{\mathcal{K}}^{(1)}$, $\overline{\mathcal{K}}^{(2)}$ and $\overline{\mathcal{K}}^{(3)}$ with a numerical example and we also compare them with the region (denoted by $\overline{\mathcal{K}}^{(EM)}$), obtained using the procedure in Elwalid and Mitra [9]. The following numerical values are used to create the plots in figures 4 and 5:

$$\begin{array}{llllll} \alpha_1 = 1.0, & \beta_1 = 0.2, & r_1 = 1.0, & \varepsilon_1 = 10^{-9}, & B_1 = 10, & \\ \alpha_2 = 1.0, & \beta_2 = 0.2, & r_2 = 1.23, & \varepsilon_2 = 10^{-6}, & B_2 = 10 & \text{and} \quad c = 13.2. \end{array} \tag{30}$$

Using the numerical values in (30), we illustrate theorem 6 in figure 4. Note that $\overline{\mathcal{K}}^{(2)}$, obtained using method 2 is a much larger feasible region than $\overline{\mathcal{K}}^{(1)}$ obtained using method 1. But $\overline{\mathcal{K}}^{(2)}$ requires a lot more computation time to obtain than $\overline{\mathcal{K}}^{(1)}$. Also note that the region $\overline{\mathcal{K}}^{(3)}$ obtained by method 3 is much larger that $\overline{\mathcal{K}}^{(1)}$ or $\overline{\mathcal{K}}^{(2)}$ but takes a lot of computational time to obtain.

In figure 5 we compare method 2 (feasible region $\overline{\mathcal{K}}^{(2)}$) and method 3 (feasible region $\overline{\mathcal{K}}^{(3)}$) with the approximation method in Elwalid and Mitra [9] (feasible region $\overline{\mathcal{K}}^{(EM)}$) for the numerical values in (30). Consider the feasible K_2 values obtained for $13 \leqslant K_1 \leqslant 20$. Clearly method 2 has a larger feasible set mainly because it uses the exact effective bandwidth of the output of buffer 1 which is much smaller than that of the input when $\zeta_2 \geqslant v^*$. But consider the feasible K_2 values obtained for

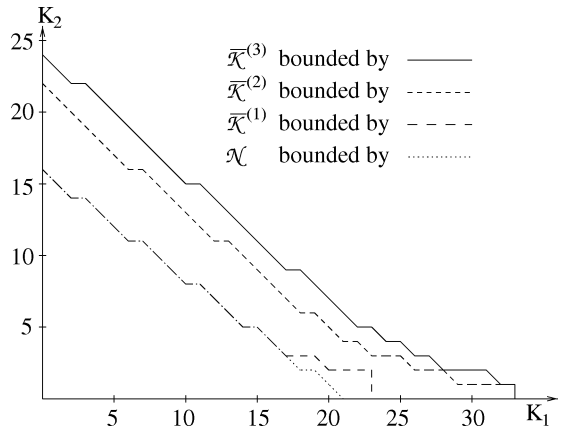

Figure 4. Regions $\mathcal{N}$, $\overline{\mathcal{K}}^{(1)}$, $\overline{\mathcal{K}}^{(2)}$ and $\overline{\mathcal{K}}^{(3)}$.

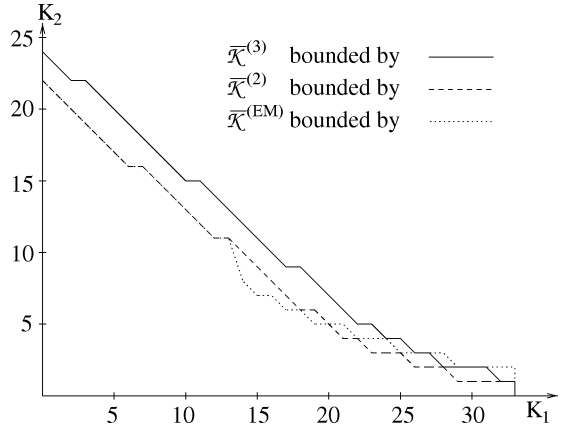

Figure 5. Feasible regions $\overline{\mathcal{K}}^{(2)}$ and $\overline{\mathcal{K}}^{(3)}$ vs $\overline{\mathcal{K}}^{(EM)}$.

$20 \leqslant K_1 \leqslant 33$. Method 2 gives a smaller set of feasible values because in Elwalid and Mitra [9] the output from buffer 1 is approximated by a CTMC with a much smaller state space and hence the effective bandwidth of the approximated output underestimates the actual effective bandwidth. Moreover the approximation method illustrated in Elwalid and Mitra [9] is not necessarily faster than method 2. The region $\overline{\mathcal{K}}^{(EM)}$ could turn out to be not conservative. This can be seen by comparing with the feasible region $\overline{\mathcal{K}}^{(3)}$ obtained by method 3. The feasible K_2 values for $K_1 = 28,\ 33$ and 34 in the region $\overline{\mathcal{K}}^{(EM)}$ obtained using the approximation method in Elwalid and Mitra [9] could result in the QoS criteria not being satisfied.

8. Conclusions

In this paper we have derived simple and asymptotically exact admission control policies for the N-priority system with general sources. These policies use the standard effective bandwidth formulae along with critical numbers $v_j^*,\ j = 2, 3, \ldots, N$, for implementation. A further simplification can be done that eliminates the use of v_j^*, and results in a more conservative policy. We have then used Chernoff bounds to fine tune the policies to obtain larger admissible regions. We have compared the different approximations and simplifications to the admissible region. Depending on what kind of trade-off one would like to do between computational time and size of the feasible region, an appropriate method can be used. We have also illustrated how the approximate policies reported by Elwalid and Mitra [9] compare with ours.

Note that it is possible to extend the computational results in section 7 to $N > 2$ priorities. We consider the example of a 2-priority node firstly because it is easy to illustrate the results using a 2-dimensional graph, and secondly because comparable known results are done for 2-priority cases only.

Appendix: Output process from buffer 1

To compute the m-function of the compensating source, we study the output process from buffer 1. There are K_1 independent and identical exponential on–off sources (with parameters $\alpha_1,\ \beta_1$ and r_1 as defined in section 5) that generate class 1 fluid into buffer 1 and a channel serves the buffer at a maximum capacity c. Let $N_1(t)$ be the number of class 1 sources on at time t and $X_1(t)$ be the amount of class-1 fluid in the buffer 1 at time t. Define

$$Y(t) = \begin{cases} N_1(t) & \text{if } X_1(t) = 0, \\ \left\lceil \dfrac{c}{r_1} \right\rceil & \text{if } X_1(t) > 0. \end{cases} \tag{31}$$

Define

$$M = K_1 \quad \text{if } K_1 \leqslant \left\lfloor \frac{c}{r_1} \right\rfloor \quad \text{and} \quad M = \left\lceil \frac{c}{r_1} \right\rceil \quad \text{if } K_1 > \left\lfloor \frac{c}{r_1} \right\rfloor. \tag{32}$$

Let $R_1(t)$ be the output rate from buffer 1 at time t. We consider two cases:

Case (i) $K_1 r_1 \leqslant c$. In this case $M = K_1$, the buffer 1 is always empty, i.e., $X_1(t) = 0$ for all t. The process $\{Y(t),\ t \geqslant 0\}$ is a CTMC on $\{0, 1, \ldots, K_1\}$ and $R_1(t) = r_1\, Y(t)$ for all t. Therefore using (31), for $i = 0, 1, 2, \ldots, K_1$,

$$\pi_1^i = \lim_{t\to\infty} P\big\{R_1(t) = i\, r_1\big\} = \lim_{t\to\infty} P\big\{N_1(t) = i\big\} = \frac{K_1!}{i!(K_1 - i)!} \frac{(\beta_1)^i (\alpha_1)^{K_1 - i}}{(\alpha_1 + \beta_1)^{K_1}}. \tag{33}$$

Then using equation (22),

$$m_1(w) = \log\left\{\sum_{k=0}^{M} \pi_1^k \mathrm{e}^{wkr_1}\right\}. \tag{34}$$

Case (ii) $K_1 r_1 > c$. In this case $M = \lceil c/r_1 \rceil$. We can see that the $\{Y(t),\ t \geqslant 0\}$ process (see (31)) is a Semi-Markov Process (SMP) on state space $\{0, 1, \ldots, M\}$ with kernel

$$G(t) = \big[G_{i,j}(t)\big].$$

For $i = 0, 1, \ldots, M-1$ and $j = 0, 1, \ldots, M$, let

$$G_{i,j}(t) = \begin{cases} \dfrac{i\alpha_1}{i\alpha_1 + (K_1 - i)\beta_1}\big(1 - \exp\big\{-(i\alpha_1 + \big(K_1 - i)\beta_1\big)t\big\}\big) & \text{if } j = i-1, \\ \dfrac{(K_1 - i)\beta_1}{i\alpha_1 + (K_1 - i)\beta_1}\big(1 - \exp\big\{-(i\alpha_1 + \big(K_1 - i)\beta_1\big)t\big\}\big) & \text{if } j = i+1, \\ 0 & \text{otherwise.} \end{cases}$$

To describe $G_{M,j}(t)$, we need to define the first passage time in $\{X_1(t),\ t \geqslant 0\}$ process as described below:

$$T = \min\big\{t > 0:\ X_1(t) = 0\big\}.$$

Then for $j = 0, 1, \ldots, M-1$, we have

$$G_{M,j}(t) = P\big\{T \leqslant t,\ N_1(T) = j | X_1(0) = 0,\ N_1(0) = M\big\}.$$

(Note that $G_{M,M}(t) = 0$.)

We need $G(\infty) = [G_{i,j}(\infty)]$ in our analysis. We have for $i = 0, 1, \ldots, M-1$ and $j = 0, 1, \ldots, M$,

$$\begin{aligned} G_{i,j}(\infty) \ &= \begin{cases} \dfrac{i\alpha_1}{i\alpha_1 + (K_1 - i)\beta_1} & \text{if } j = i-1, \\ \dfrac{(K_1 - i)\beta_1}{i\alpha_1 + (K_1 - i)\beta_1} & \text{if } j = i+1, \\ 0 & \text{otherwise,} \end{cases} \\ G_{M,j}(\infty) &= \widetilde{G}_{M,j}(0), \end{aligned} \tag{35}$$

where $\widetilde{G}_{M,j}(s)$ is the Laplace–Stieltjes transform (LST) of $G_{M,j}(t)$, and can be computed using the analysis in Narayanan and Kulkarni [17].

We also need the expression for the sojourn time μ_i in state i, for $i = 0, 1, \ldots, M$. We have

$$\mu_i = \begin{cases} \dfrac{1}{i\alpha_1 + (K_1 - i)\beta_1} & \text{if } i = 0, 1, \ldots, M-1, \\ \sum_{j=1}^{M-1} \widetilde{G}'_{M,j}(0) & \text{if } i = M. \end{cases}$$

Then we have for $i = 0, 1, \ldots, M$

$$\pi_1^i = \lim_{t\to\infty} P\{Y(t) = i\} = \frac{p_i\mu_i}{\sum_{k=0}^{M} p_k\mu_k}, \tag{36}$$

where

$$p = p\,G(\infty).$$

It is easy to see that

$$R_1(t) = \begin{cases} r_1\,Y(t) & \text{if } Y(t) < M, \\ c & \text{if } Y(t) = M. \end{cases} \tag{37}$$

Therefore using (36) and (37), we have for $i = 0, 1, \ldots, M$

$$\begin{aligned} \lim_{t\to\infty} P\{R_1(t) = i\} = \lim_{t\to\infty} P\{Y(t) = i\} &= \frac{p_i\mu_i}{\sum_{k=0}^{M} p_k\mu_k} \quad \text{if } i < M, \\ \lim_{t\to\infty} P\{R_1(t) = c\} = \lim_{t\to\infty} P\{Y(t) = M\} &= \frac{p_M\mu_M}{\sum_{k=0}^{M} p_k\mu_k} \quad \text{if } i = M. \end{aligned} \tag{38}$$

Using the equations (22), (38) and (36), we see that

$$m_1(w) = \log\left\{\pi_1^M \mathrm{e}^{wc} + \sum_{k=0}^{M-1} \pi_1^k \mathrm{e}^{wkr_1}\right\}. \tag{39}$$

References

[1] C.S. Chang and J.A. Thomas, Effective bandwidth in high-speed digital networks, IEEE Journal on Selected Areas in Communications 13(6) (1995) 1091–1100.

[2] C.S. Chang and T. Zajic, Effective bandwidths of departure processes from queues with time varying capacities, in: *INFOCOM '95* (1995) pp. 1001–1009.

[3] G.L. Choudhury, D.M. Lucantoni and W. Whitt, On the effectiveness of effective bandwidths for admission control in ATM networks, in: *Proceedings of ITC-14* (Elsevier Science, B.V., 1994) pp. 411–420.

[4] I. Çidon, R. Guérin and A. Khamisy, On protective buffer policies, Technical Report RC 18113, IBM Research Division (1992).

[5] I. Çidon, L. Georgiadis, R. Guérin and A. Khamisy, Optimal buffer sharing, IEEE Journal on Selected Areas in Communications 13(7) (1995) 1229–1240.

[6] G. de Veciana, C. Courcoubetis and J. Walrand, Decoupling bandwidths for networks: A decomposition approach to resource management, in: *INFOCOM '94* (1994) pp. 466–473.

[7] A.I. Elwalid and D. Mitra, Fluid models for the analysis and design of statistical multiplexing with loss priorities on multiple classes of bursty traffic, IEEE Trans. Communications 42(11) (1992) 2989–3002.

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

[9] A.I. Elwalid and D. Mitra, Analysis, approximations and admission control of a multi-service multiplexing system with priorities, in: *INFOCOM '95* (1995) pp. 463–472.

[10] A.I. Elwalid, D. Heyman, T.V. Lakshman, D. Mitra and A. Weiss, Fundamental bounds and approximations for ATM multiplexers with applications to video teleconferencing, IEEE Journal on Selected Areas in Communications 13(6) (1995) 1004–1016.
[11] R.J. Gibbens and P.J. Hunt, Effective bandwidths for the multi-type UAS channel, Queueing Systems 9 (1991) 17–28.
[12] G. Kesidis, J. Walrand and C.S. Chang, Effective bandwidths for multiclass Markov fluids and other ATM sources, IEEE/ACM Trans. on Networking 1(4) (1993) 424–428.
[13] V.G. Kulkarni, Effective bandwidths for Markov regenerative sources, Queueing Systems 24 (1996) 137–154.
[14] V.G. Kulkarni and T. Rolski, Fluid model driven by an Ornstein–Ühlenbeck Process, Probab. Engrg. Inform. Sci. 8 (1994) 403–417.
[15] V.G. Kulkarni, L. Gün and P.F. Chimento, Effective bandwidth vectors for multiclass traffic multiplexed in a partitioned buffer, IEEE Journal on Selected Areas in Communications 13(6) (1995) 1039–1047.
[16] A.Y.-M. Lin and J.A. Silvester, Priority queueing strategies and buffer allocation protocols for traffic control at an ATM integrated broadband switching system, IEEE Journal on Selected Areas in Communications 9(9) (1991) 1524–1536.
[17] A. Narayanan and V.G. Kulkarni, First passage times in fluid models with an application to two-priority fluid systems, in: *Proceedings of IPDS '96* (1996).
[18] W. Whitt, Tail probabilities with statistical multiplexing and effective bandwidths for multiclass queues, Telecommun. Syst. 2 (1993) 71–107.
[19] J. Zhang, Performance study of Markov-modulated fluid flow models with priority traffic, in: *INFOCOM '93* (1993) pp. 10–17.
[20] Z.L. Zhang, D. Towsley and J. Kurose, Statistical analysis of generalized processor sharing scheduling discipline, IEEE Journal on Selected Areas in Communications 13(6) (1995) 1071-80.
[21] Z.L. Zhang, D. Towsley and J. Kurose, Call admission control scheme under the generalized processor sharing scheduling discipline, Technical Report UM-CS-95-10, University of Massachusetts (1995).