Queueing Systems 36, 351–379, 2000
© 2000 Kluwer Academic Publishers. Manufactured in The Netherlands.

Applications of SMP Bounds to Multi-class Traffic in High-speed Networks [*]

N. GAUTAM [**] ngautam@psu.edu
Department of Industrial and Manufacturing Engineering, The Pennsylvania State University,
310 Leonhard Building, University Park, PA 16802, USA

V.G. KULKARNI vg_kulkarni@unc.edu
Department of Operations Research, University of North Carolina, Chapel Hill, NC 27599-3180, USA

Received 3 August 1998; revised 6 April 2000

Abstract. In this paper, we consider the stochastic fluid-flow model of a single node in a high-speed telecommunication network handling multi-class traffic. The node has multiple buffers, one for each class of traffic. The contents of these buffers are multiplexed onto a single output channel using one of the service scheduling policies: the Timed Round Robin Policy or the Static Priority Service Policy. The Quality of Service requirements for each class are based on cell loss probabilities. Using effective bandwidth methodologies and the recently developed bounds for semi-Markov modulated traffic, we solve call admission control problems for the two service scheduling policies at this node. We compare the performance of the effective bandwidth methodologies and the SMP bounds technique. We also numerically compare the performance of the two service scheduling policies.

Keywords: Quality of Service, admission control, multi-class traffic, timed round robin policy, static priority service policy, fluid-flow models, fluid queues

1. Introduction

High-speed telecommunication networks are rapidly becoming a reality. Modeling and analysis of such networks is an important step in their design and operation. In this paper, we mainly concentrate on high-speed networks that use the asynchronous transfer mode (ATM) where information flows in the network in the form of 53-byte packets or cells. These high-speed networks are expected to handle a wide variety of traffic on the same channel. Therefore a cell may carry one of the different types of information: voice, video, data, etc. This creates the need to deal with multi-class traffic. For each class of traffic, a Quality of Service (QoS), measured by cell-loss probability, delay, delay-jitter, etc. needs to be assured. The QoS may be different for each type of traffic. 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-loss.

[*] This work was partially supported by NSF Grant No. NCR-9406823.
[**] N. Gautam's research was partially supported by the IBM Co-operative Dissertation Fellowship Award.

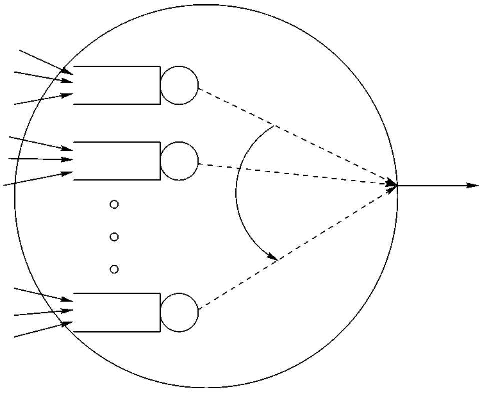


Figure 1. A multi-class node.

The high speed (e.g., 155–622 Mbits/s) of the ATM network implies that it can transmit millions of cells per second. This makes fluid-flow models useful in describing the flow of cells. We analyze the packetized traffic by approximating it by fluids, following the large literature using fluid-flow models for communication systems, (see [1,8], etc.).

Figure 1 shows a schematic representation of a single node designed to handle multi-class traffic. The node has multiple buffers, one for each class. The node follows a given service scheduling policy to transmit the data from these buffers onto the output channel. In [14] a similar scenario using the packetized general processor sharing mechanism and the weighted round robin mechanism are considered for discrete arrival systems.

We study two different service scheduling policies: timed round robin and static priority. Under timed round robin policy (a variation of polling), the scheduler serves the buffers in a fixed cyclical fashion. Takagi [23], Daganzo [5], etc. analyze the different types of policies in polling systems and describe their properties. Some of the common polling policies studied are the full-service exhaustive policy, the gated policy, the weighted round robin mechanism and the timed round-robin policy. We shall concentrate on the timed round-robin policy under which the scheduler serves each buffer for a fixed amount of time in a given cyclic order.

Under static priority service policy each class of fluid has a fixed priority of service. This policy gives full priority to the highest priority traffic and the transmission capacity that is not utilized by highest priority traffic is offered to the next highest priority traffic, etc. Narayanan and Kulkarni [20] analyze multi-class fluid models that use static priority service policy. They develop the marginal buffer-content distributions for each class of fluid. Zhang [25] analyzes the joint distribution of the buffer contents of each class.

For a single node using a single class of traffic the concept of effective bandwidths and its applications to the QoS problem is well established. Gibbens and Hunt [13], Kesidis et al. [15], Elwalid and Mitra [9], Kulkarni [17], Choudhury et al. [4], Whitt [24], etc. discuss the concept of effective bandwidths for single-class traffic.

The effective-bandwidth methodology, although simple to use, is based on an exponential approximation to the tail of the distribution of the buffer content in steady state. This approximation holds only when the buffer sizes are very large, the tail probabilities are small, and under certain assumptions about the input traffic. Several researchers have attempted to redress these shortcomings. For example, Elwalid et al. [7,10] modify the effective-bandwidth methodology and develop the Chernoff Dominant Eigenvalue (CDE) approximation for single-class traffic. To avoid approximations, other approaches have been developed. They include deriving upper and lower bounds for the buffer content process in steady state with a Markov additive input by discretizing time and using extensions of Kingman's exponential bounds for waiting times in the stationary regime in a $G/G/1$ queue (see [2,3,6,16,19,22]). Artiges and Nain [2] obtain exponential bounds for multiplexing multiclass Markovian on–off sources, where the upper bounds are similar to those in [21].

In [18], effective bandwidth approximation and Chernoff dominant eigenvalue approximation are used to solve design and admission control problems under static priority service policy. In this paper we use the bounds obtained for the semi-Markov modulated fluid traffic (see [11]) in the analysis of both the static priority service policy as well as the timed round robin policy.

The paper is organized as follows. In section 2, for a single buffer model, we recapitulate the effective bandwidth approximation and the semi-Markov process bounds. In section 3 we illustrate the multi-class node model. In section 4 we describe the timed round-robin policy and compare the performance of the effective bandwidth methodologies against the SMP bounds technique for QoS problems under this policy. In section 5, we use the SMP bounds technique for admission control problems for the static priority service policy. We compare the admissible regions with those obtained using effective bandwidth techniques and Chernoff bounds. In section 6, we numerically compare the performance of the two service scheduling policies.

2. Preliminary results: effective bandwidths and SMP bounds

Consider a single infinite-sized buffer (with constant output capacity c) that admits traffic from K independent sources, with kth source driven by a random environment process $\{Zk(t),\ t \geqslant 0\}$, $k = 1, 2, \dots, K$ (see figure 2). At time t, source k generates fluid at rate $r_k(Z_k(t))$. Let $X(t)$ be the amount of fluid in the buffer at time t. We are interested in the following probability:

$$\lim_{t\to\infty} P\big\{X(t) > x\big\} = P\{X > x\}. \tag{1}$$

If B is the actual buffer size then $P\{X > B\}$ is taken as the steady state approximation of the buffer overflow probability.

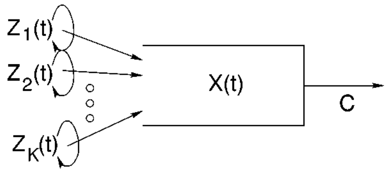


Figure 2. Single buffer fluid model.

2.1. *Effective bandwidths*

Assume that the environment processes $\{Z_k(t),\ t \geqslant 0\}$, $k = 1, 2, \dots, K$, are stationary and ergodic processes satisfying the Gartner–Ellis conditions (see [15]). Then, for a given v $(v > 0)$, the *effective bandwidth* of source k is

$$eb_k(v) = \lim_{t\to\infty} \frac{1}{vt} \log E\left\{\exp\left(v \int_0^t r_k\big(Z_k(t)\big)\,\mathrm{d}t\right)\right\}. \tag{2}$$

When the $\{Z_k(t),\ t \geqslant 0\}$ processes can be modeled as certain special stochastic processes, Kesidis et al. [15], Elwalid and Mitra [9] and Kulkarni [17] illustrate how to compute $eb_k(v)$ in those cases. Let η be the solution to

$$\sum_{k=1}^{K} eb_k(\eta) = c. \tag{3}$$

The effective bandwidth methodology yields the following approximation of the probability in equation (1):

$$P(X > x) \approx \mathrm{e}^{-x\eta}. \tag{4}$$

Using the effective bandwidth approximation, we conclude that the QoS criterion for cell loss probability $P(X > B) < \varepsilon$ is satisfied if $\mathrm{e}^{B\eta} < \varepsilon$, where B is the buffer size. The R.H.S. in (4) is an approximation, not a bound and is valid for large B and small ε.

2.2. *SMP bounds*

Consider the case when $\{Z_k(t),\ t \geqslant 0\}$ $(k = 1, 2, \dots, K)$ are independent semi-Markov processes (SMPs) with state space $\mathcal{S}_k = \{1, 2, \dots, \ell_k\}$ and kernel $G^k(x) = [G^k_{ij}(x)]$. The expected time the kth SMP spends in state i is τ^k_i. The stationary distribution vector of the kth SMP $\{Z_k(t),\ t \geqslant 0\}$ is p^k, where

$$p^k_i = \lim_{t\to\infty} P\big\{Z_k(t) = i\big\}.$$

We describe how to compute $eb_k(v)$ first. Let $\widetilde{G}^k_{ij}(s)$ be the Laplace–Stieltjes transform (LST) of $G^k_{ij}(x)$. For a given $v > 0$, define

$$\begin{aligned}\chi^k_{ij}(v, u) &= \widetilde{G}^k_{ij}\big(-v\big(r_k(i) - u\big)\big),\\ \chi^k(v, u) &= \big[\chi^k_{ij}(v, u)\big].\end{aligned}$$

Then $eb_k(v)$ is given by the smallest positive number such that the Perron–Frobenius eigenvalue of $\chi^k(v, eb_k(v))$ is one. Let η be a solution to equation (3), and denote $\Phi^k(\eta) = \chi^k(\eta, eb_k(\eta))$. Let h^k be the left eigenvector of $\Phi^k(\eta)$ corresponding to the eigenvalue 1, i.e.,

$$h^k = h^k \Phi^k(\eta).$$

Now, let

$$P^k(i,j) = \left[G^k(\infty)\right]_{ij}. \tag{5}$$

We also define

$$H^k = \sum_{i=1}^{\ell_k} \frac{h_i^k}{\eta(r_k(i) - eb_k(\eta))} \left(\sum_{j=1}^{\ell_k} \left(\phi_{ij}^k(\eta)\right) - 1 \right), \tag{6}$$

$$\Psi_{\min}^k(i,j) = \inf_x \left\{ \frac{h_i^k \, \mathrm{e}^{-\eta(r_k(i)-eb_k(\eta))x} \int_x^\infty \mathrm{e}^{\eta(r_k(i)-eb_k(\eta))y} \, \mathrm{d}G_{ij}^k(y)}{(p_i^k/\tau_i^k) \int_x^\infty \mathrm{d}G_{ij}^k(y)} \right\}, \tag{7}$$

and

$$\Psi_{\max}^k(i,j) = \sup_x \left\{ \frac{h_i^k \, \mathrm{e}^{-\eta(r_k(i)-eb_k(\eta))x} \int_x^\infty \mathrm{e}^{\eta(r_k(i)-eb_k(\eta))y} \, \mathrm{d}G_{ij}^k(y)}{(p_i^k/\tau_i^k) \int_x^\infty \mathrm{d}G_{ij}^k(y)} \right\}. \tag{8}$$

From [11,12], we have

$$C_* \, \mathrm{e}^{-\eta x} \leqslant P(X > x) \leqslant C^* \, \mathrm{e}^{-\eta x}, \quad x \geqslant 0, \tag{9}$$

where

$$C^* = \frac{\prod_{k=1}^K H^k}{\min_{\mathcal{A}} \prod_{k=1}^K \Psi_{\min}^k(i_k, j_k)}, \qquad C_* = \frac{\prod_{k=1}^K H^k}{\max_{\mathcal{A}} \prod_{k=1}^K \Psi_{\max}^k(i_k, j_k)},$$
$$\mathcal{A} = \Bigg\{ (i_1, j_1), (i_2, j_2), \ldots, (i_K, j_K)\colon\ i_k, j_k \in \mathcal{S}_k,$$
$$\sum_{k=1}^K r_k(i_k) > c \text{ and } \forall k,\ P^k(i_k, j_k) > 0 \Bigg\}. \tag{10}$$

Using the SMP bounds we conclude that the QoS criterion for cell loss probability $P(X > B) < \varepsilon$ is satisfied if $C^* \mathrm{e}^{-B\eta} < \varepsilon$. Using equation (9), we can describe a bound on $P(X > B)$ that is valid for all B and ε.

3. Multi-class node model

In this section we use the single class model results obtained in section 2 to solve QoS problems in multi-class nodes by making suitable transformations. Consider the model of a multi-class node illustrated in figure 3. The node consists of N input buffers, one

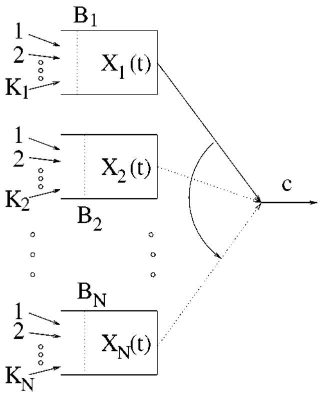


Figure 3. The multi-class node model.

for each class of traffic. The input to buffer j $(j = 1, \ldots, N)$, is from the K_j sources of class j. The ith source of class j is driven by an independent random environment process $Z_{ij} = \{Z_{ij}(t),\ t \geqslant 0\}$ for $i = 1, 2, \ldots, K_j$. At time t, source i of type j generates fluid at rate $r_{ij}(Z_{ij}(t))$. Let $X_j(t)$ be the amount of fluid in buffer j at time t. All the classes of fluids are served by a single channel of constant capacity c, using a specified service scheduling policy (in this paper, we consider timed round robin policy and static priority service policy).

We assume that all N buffers are of infinite capacity. If B_j is the actual size of buffer j $(j = 1, 2, \ldots, N)$, then we take

$$\lim_{t\to\infty} P\big\{X_j(t) > B_j\big\} = P\{X_j > B_j\}$$

as the steady state approximation of the overflow probability from buffer j. Let ε_j be the cell loss probability target for class j traffic $(j = 1, 2, \ldots, N)$. The Quality of Service (QoS) criterion for cell loss that need to be satisfied class j traffic is

$$\lim_{t\to\infty} P\big\{X_j(t) > B_j\big\} = P\{X_j > B_j\} < \varepsilon_j. \tag{11}$$

We first explain the two service scheduling policies, timed round robin policy and static priority service policy. Note that the effective bandwidth and the SMP bounds analysis for the multiclass model is not a trivial extension of that of the single class model. The output channel capacity for each buffer is not a constant in the multiclass node model. Therefore the model requires a careful transformation that results in a constant output channel capacity model for each of the buffers. From the transformed models, we compute $P\{X_j > B_j\}$ using effective bandwidth approximation and SMP bounds techniques for the two policies. We also solve admission control problems for the two policies. Finally, we compare the two policies.

4. Timed round robin policy

Consider the multi-class node model described in section 3 and illustrated in figure 3. All classes of fluids are multiplexed using a *Timed Round Robin* service scheduling policy which is described as follows. The scheduler allocates the entire output capacity c to each of the N buffers in a cyclic fashion. In each cycle, buffer j gets the entire capacity for an interval of length τ_j. Note that during this interval, buffer j could be empty. Hence the scheduler is not work conserving.

Let t_{so} be the total switch-over time during an entire cycle. We assume that t_{so} does not change with time. The *cycle time* T is defined as the amount of time the scheduler takes to complete a cycle, and is given by

$$T = t_{\text{so}} + \sum_{j=1}^{N} \tau_j. \tag{12}$$

First we assume that all buffers are of infinite capacity. The dynamics of the buffer-content process $\{X_j(t),\ t \geqslant 0\}$ is described by

$$\frac{\mathrm{d}X_j(t)}{\mathrm{d}t} = \begin{cases} \displaystyle\sum_{i=1}^{K_j} r_{ij}\bigl(Z_{ij}(t)\bigr) - c & \text{if } X(t) > 0 \text{ and scheduler serving buffer } j, \\ \displaystyle\left\{\sum_{i=1}^{K_j} r_{ij}\bigl(Z_{ij}(t)\bigr) - c\right\}^{+} & \text{if } X(t) = 0 \text{ and scheduler serving buffer } j, \\ \displaystyle\sum_{i=1}^{K_j} r_{ij}\bigl(Z_{ij}(t)\bigr) & \text{if scheduler not serving buffer } j. \end{cases} \tag{13}$$

We assume that the following stability condition is satisfied for buffer j $(j = 1, \dots, N)$:

$$\sum_{i=1}^{K_j} E\bigl\{r_{ij}\bigl(Z_{ij}(\infty)\bigr)\bigr\} < c\,\frac{\tau_j}{T}. \tag{14}$$

4.1. *Effective bandwidth analysis*

If we are given $\tau_1, \tau_2, \dots, \tau_N$ and t_{so}, then the buffer contents of a given buffer (say, j) and its dynamics do not depend on the parameters of any other buffer (say $i \neq j$). Therefore, it is convenient to analyze each buffer separately. Buffer j can be modeled as a single-buffer-fluid model with variable output capacity and input from K_j different sources, such that source i of class j is modulated by an environmental process $\{Z_{ij}(t),\ t \geqslant 0\}$. The output capacity alternates between c (for τ_j units of time) and 0 (for $T - \tau_j$ units of time).

Note that the effective-bandwidth approximation (see section 2.1) and the SMP bounds (see section 2.2) assume that the output channel capacity is a constant. Therefore

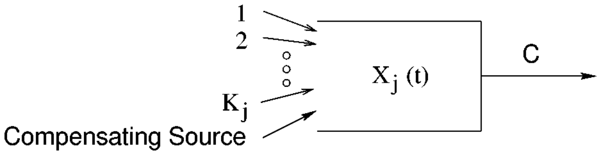


Figure 4. Transformed buffer j model.

to utilize those techniques, we need to first transform our model into an appropriate one with a constant output channel capacity as follows.

Consider a single-buffer-fluid model for buffer j with a constant output channel capacity c whose input is generated by the original K_j sources and a fictitious compensating source. The compensating source is such that it stays on for a deterministic amount of time $T - \tau_j$ and off for a deterministic amount of time τ_j. When the compensating source is on, it generates fluid at rate c and when it is off it generates fluid at rate 0. Note that the compensating source is independent of the original K_j sources. Clearly, the dynamics of the buffer-content process (of buffer j) in equation (13) remain unchanged for this transformed single-buffer-fluid model with $K_j + 1$ input sources (including the compensating source) and constant output capacity c. Refer to figure 4 for an illustration of the transformed model for buffer j.

Using the effective bandwidth computations in [17], we can show that the effective bandwidth of the compensating source described above is given by

$$eb_j^s(v) = \frac{c(T - \tau_j)}{T}. \tag{15}$$

Note that the effective bandwidth of this deterministic source is indeed its mean traffic generation rate. Let the effective bandwidth of source i ($i = 1, 2, \ldots, K_j$) of class j be $eb_{ij}(v)$. Therefore $P(X_j > B_j) \approx \mathrm{e}^{-B_j \eta_j}$, where η_j (using equation (3)) is obtained by solving

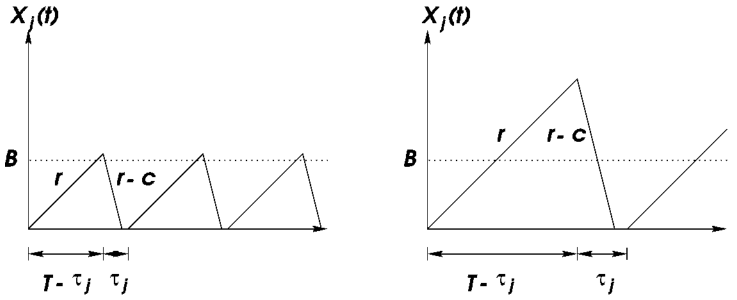
$$\sum_{i=1}^{K_j} eb_{ij}(\eta_j) + c\,\frac{(T - \tau_j)}{T} = c. \tag{16}$$

The QoS criteria for all the classes of traffic are satisfied if for all $j = 1, 2, \ldots, N$,

$$\mathrm{e}^{-B_j \eta_j} < \varepsilon_j. \tag{17}$$

Equations (16) and (17) indicate that the QoS guarantee using the effective-bandwidth approximation technique depends only on the ratio τ_j/T and not on the individual values of τ_j or T. Consider two instances, one with large τ_j and T and the other with small τ_j and T, such that the ratio τ_j/T is the same in both instances. The effective bandwidth approximation implies that the loss probability will be the same in both instances. Intuitively, the probability of buffer overflow should be larger for the longer cycle time T.

For example, consider an infinite-sized buffer into which fluid is generated continuously at rate r (deterministic or CBR source). Let this buffer be emptied by a channel



(a) (b)

Figure 5. Buffer content process for different (τ_j, T) values.

whose capacity oscillates between c (for time τ_j) and 0 (for time $T - \tau_j$). For stability, assume that $r < c\tau_j/T$. Then the limiting probability that the buffer content exceeds B is given by

$$P\{X_j > B\} = \begin{cases} 0 & \text{if } T - \tau_j \leqslant \dfrac{B}{r} \\ \left(\dfrac{c}{c-r}\right)\left(1 - \dfrac{\tau_j}{T} - \dfrac{B}{rT}\right) & \text{if } T - \tau_j > \dfrac{B}{r}. \end{cases}$$

Clearly, the probability $P\{X_j > B\}$ increases with T. For this example, the buffer content process $X_j(t)$ is shown in figure 5 for two instances, one with small τ_j and T (figure 5(a)) and the other with large τ_j and T (figure 5(b)), such that the ratio τ_j/T is the same in both instances. From the figure it is easy to see that $P\{X_j > B\}$ is higher for large τ_j and T (figure 5(b)).

Note that the effective bandwidth results are theoretically valid since the effective-bandwidth analysis assumes extremely large buffers ($B \to \infty$). However in practice, this cannot be considered as valid due to finite buffers. Therefore the effective-bandwidth approximation technique fails for moderate to large sized buffers and works only for extremely large sized buffers. The Chernoff dominant eigenvalue approximation (see [10]) also faces the same problem. The SMP bounds below resolve this issue. In section 4.3, the effect of τ_j and T on the performance of the timed round robin policy is explained using numerical examples.

4.2. Semi-Markov process (SMP) bounds analysis

We consider the transformed model of buffer j ($j = 1, 2, \ldots, N$) illustrated in figure 4. We assume that the $\{Z_{ij}(t),\ t \geqslant 0\}$ processes ($i = 1, 2, \ldots, K_j$) are semi-Markov processes. Therefore there are $K_j + 1$ independent sources modulated by SMPs (including the compensating source) that generate traffic into buffer j whose the output capacity is a constant c.

For the SMP bounds analysis for buffer j we follow the single-class traffic analysis in section 2.2 for a buffer with input generated by independent semi-Markovian sources multiplexed together. Let η_j be the smallest positive solution to equation (16).

Using equations (6)–(8), we can obtain H^{ij}, $\Psi^{ij}_{\min}$ and $\Psi^{ij}_{\max}$, respectively, for source i ($i = 1, 2, \dots, K_j$) of class j. The corresponding expressions H^{sj}, $\Psi^{sj}_{\min}$ and $\Psi^{sj}_{\max}$ for the jth compensating source are

$$H^{sj} = \frac{1 - \exp(-\eta_j c((T-\tau_j)/T)\tau_j)}{\eta_j c}\left[\frac{T^2}{(T-\tau_j)\tau_j}\right], \tag{18}$$

$$\Psi^{sj}_{\min} = \begin{bmatrix} 0 & T\exp\left(-\eta_j c\,\dfrac{T-\tau_j}{T}\,\tau_j\right) \\ T\exp\left(-\eta_j c\,\dfrac{T-\tau_j}{T}\,\tau_j\right) & 0 \end{bmatrix}, \tag{19}$$

$$\Psi^{sj}_{\max} = \begin{bmatrix} 0 & T \\ T & 0 \end{bmatrix}. \tag{20}$$

Letting $s = K_j + 1$, we obtain the bounds on the limiting distribution of the buffer content process $\{X_j(t),\ t \geqslant 0\}$ as

$$C_{j*}\,\mathrm{e}^{-\eta_j x} \leqslant P(X_j > x) \leqslant C_j^*\,\mathrm{e}^{-\eta_j x},$$

where η_j is from equation (16),

$$C_j^* = \frac{\prod_{k=1}^{K_j+1} H^{kj}}{\min_{\mathcal{A}^j}\prod_{k=1}^{K_j+1}\Psi^{kj}_{\min}(l_k, m_k)}, \tag{21}$$

$$C_{*j} = \frac{\prod_{k=1}^{K_j+1} H^{kj}}{\max_{\mathcal{A}^j}\prod_{k=1}^{K_j+1}\Psi^{kj}_{\max}(l_k, m_k)}, \tag{22}$$

and

$$\mathcal{A}^j = \Bigg\{(l_1, m_1), (l_2, m_2), \dots, (l_{K_j+1}, m_{K_j+1})\colon\ l_k, m_k \in \mathcal{S}_k, \\ \sum_{k=1}^{K_j+1} r_{kj}(l_k) > c \text{ and } \forall k,\ P^{kj}(l_k, m_k) > 0\Bigg\}. \tag{23}$$

The QoS criteria for all the classes of traffic are satisfied if, for $j = 1, 2, \dots, N$,

$$C_j^*\,\mathrm{e}^{-\eta_j B_j} < \varepsilon_j. \tag{24}$$

From equations (18)–(20), clearly, H^{sj} and $\Psi^{sj}_{\min}$ are functions of τ_j, T and τ_j/T. Hence, C_j^* is a function of both τ_j and T and not simply of the ratio τ_j/T. In the next section we will illustrate some of the differences in the results obtained by using the two techniques, effective-bandwidth approximation and semi-Markov process bounds.

4.3. Effective bandwidth vs SMP bounds

For the sake of simplicity (and getting closed form results), we assume that the input sources are K_j independent and identical alternating on–off sources, that stay on for an exponential amount of time with parameter parameter α_j and off for an exponential amount of time with parameter β_j. When a source is on, it generates traffic continuously at rate r_j into buffer j of size B_j and when it is off, it does not generate any traffic. The scheduler serves buffer j for a deterministic time τ_j at a maximum rate c and stops serving the buffer for a deterministic time $T - \tau_j$.

The effective bandwidth of all the *Kj* sources combined is (see [9,15])

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

Equation (16) reduces to

$$K_j\, eb_j(\eta_j) = \frac{c\tau_j}{T},$$

and solving for η_j, we get

$$\eta_j = \frac{c\tau_j(\alpha_j + \beta_j) - r_j K_j \beta_j T}{(c\tau_j/(K_j T))(r_j T K_j - c\tau_j)}.$$

We can show that (see [12]) equations (21) and (22) reduce to

$$\begin{aligned} C_j^* &= \frac{\big[\frac{r_j T K_j}{\alpha_j c\tau_j}\big]^{K_j}\big(\exp\big(n_j c\,\frac{\tau_j(T-\tau_j)}{T}\big) - 1\big)}{\min_{1\leqslant i\leqslant K_j}\big\{\big(\frac{\alpha_j+\beta_j}{\alpha_j\beta_j}\big)^{K_j}\big(\frac{TK_j\beta_j}{TK_j\beta_j+\eta_j c\tau_j}\big)^{K_j-i}\big\}\eta_j c\tau_j}\left(\frac{T}{T-\tau_j}\right) \\ &= \frac{\big[\frac{r_j T K_j}{\alpha_j c\tau_j}\big]^{K_j}\big(\exp\big(n_j c\,\frac{\tau_j(T-\tau_j)}{T}\big) - 1\big)}{\big(\frac{\alpha_j+\beta_j}{\alpha_j\beta_j}\big)^{K_j}\big(\frac{TK_j\beta_j}{TK_j\beta_j+\eta_j c\tau_j}\big)^{K_j-1}\eta_j c\tau_j}\left(\frac{T}{T-\tau_j}\right), \end{aligned} \tag{25}$$

$$\begin{aligned} C_{*j} &= \frac{\big[\frac{r_j T K_j}{\alpha_j c\tau_j}\big]^{K_j}\big(1 - \exp\big(-n_j c\,\frac{\tau_j(T-\tau_j)}{T}\big)\big)}{\max_{1\leqslant i\leqslant K_j}\big\{\big(\frac{\alpha_j+\beta_j}{\alpha_j\beta_j}\big)^{K_j}\big(\frac{TK_j\beta_j}{TK_j\beta_j+\eta_j c\tau_j}\big)^{K_j-i}\big\}\eta_j c\tau_j}\left(\frac{T}{T-\tau_j}\right) \\ &= \frac{\big[\frac{r_j T K_j}{\alpha_j c\tau_j}\big]^{K_j}\big(1 - \exp\big(-n_j c\,\frac{\tau_j(T-\tau_j)}{T}\big)\big)}{\big(\frac{\alpha_j+\beta_j}{\alpha_j\beta_j}\big)^{K_j}\eta_j c\tau_j}\left(\frac{T}{T-\tau_j}\right). \end{aligned} \tag{26}$$

We now consider three scenarios to compare the performance of the effective-bandwidth approximation and the SMP bounds technique by varying T and τ_j such that τ_j/T remains a constant. For all the numerical examples we use,

$$\alpha_j = 3, \qquad \beta_j = 0.2, \qquad r_j = 3.4, \qquad B_j = 30, \qquad \tau_j/T = 3/13. \tag{27}$$

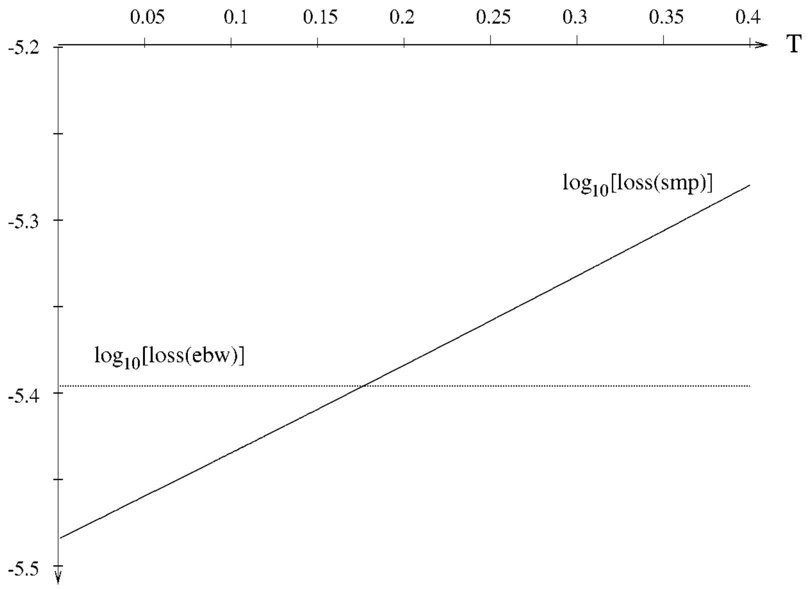


Figure 6. Estimates of the logarithms of loss probability.

Estimate of loss-probability. If we are given α_j, β_j, r_j, K_j, c, B_j, τ_j and T, then the estimate of the cell-loss probability at buffer j using the effective-bandwidth techniques is

$$\text{loss(ebw)} = \mathrm{e}^{-\eta_j B_j}$$

and using the SMP bounds the estimate is

$$\text{loss(smp)} = C_j^* \, \mathrm{e}^{-\eta_j B_j}.$$

Figure 6 shows the results for loss(ebw) and loss(smp) when $K_j = 10$, $c = 15.3$, and T varies from 0.01 to 0.40 while τ_j/T is fixed. Intuitively we expect the loss probability to increase with T since an increase in T would increase the time the server does not serve the buffer. The SMP bounds estimate, loss(smp), increases with T and hence confirms our intuition. The effective-bandwidth estimate, loss(ebw), does not change with T. For small T, since loss(smp) $<$ loss(ebw), we can conclude that the effective-bandwidth technique produces a conservative result. For large T, the estimate of the loss probability is smaller using the effective-bandwidth technique than the SMP bounds technique. This indicates that there may be a risk in using the effective-bandwidth technique as it could result in the QoS criteria not being satisfied.

Estimate of the maximum number of sources. Let ε_j be maximum allowable cell-loss at buffer j. Consider that we are given α_j, β_j, r_j, c, τ_j and T. We are required to find the largest number of class-j sources that can be admitted so that the QoS criterion

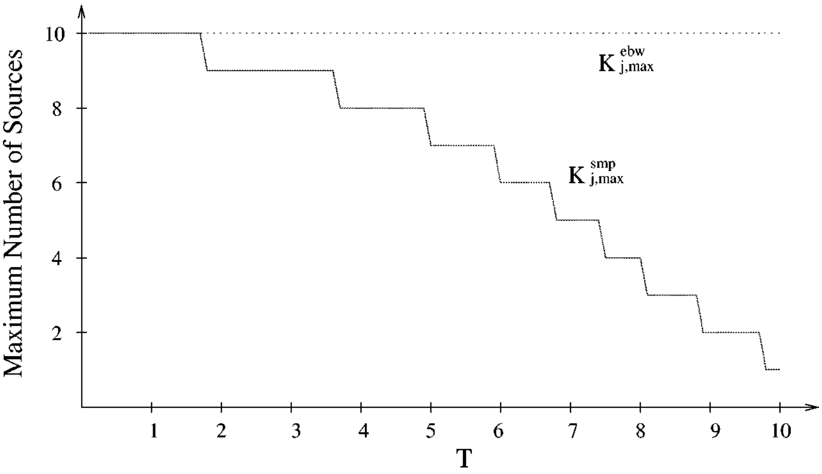


Figure 7. Estimate of the maximum number of sources.

is satisfied. Then, the estimate of the maximum number of sources using the effective-bandwidth technique is

$$K_{j,\max}^{\mathrm{ebw}} = \left\lfloor \frac{1}{eb_j(-\log(\varepsilon_j)/B_j)} \frac{c\tau_j}{T} \right\rfloor.$$

On the other hand, using the upper bound for an SMP we obtain (from equation (25)) C_j^* for a given K_j. Now we choose the largest integer $K_{j,\max}^{\mathrm{smp}}$ that satisfies

$$C_j^* \, \mathrm{e}^{-\eta_j B_j} < \varepsilon_j.$$

Figure 7 shows the results for $K_{j,\max}^{\mathrm{ebw}}$ and $K_{j,\max}^{\mathrm{smp}}$ when $\varepsilon_j = 10^{-5}$, $c = 15.3$, and T varies from 0.01 to 10.00 while τ_j/T is fixed.

As T increases, we expect fewer sources to be allowable into the buffer so that long bursts of traffic can be avoided when the server is not serving. From the figure, $K_{j,\max}^{\mathrm{smp}}$ clearly conforms to our intuition. For large T, we may end up admitting more sources if we used the effective-bandwidth technique and hence the QoS criterion may not be satisfied.

Estimate of the required bandwidth. Consider that we are given the parameters α_j, β_j, r_j, K_j, τ_j and T and we would like to estimate the smallest c value required so that the loss probability is no greater than ε_j. The estimate of the smallest bandwidth required, c, using the effective-bandwidth technique is

$$c_{\min}^{\mathrm{ebw}} = K_j \, eb_j\big(-\log(\varepsilon_j)/B_j\big).$$

The loss probability estimate using the SMP bounds decreases with increase in c. Therefore we perform a search using the bisection method to pick a c between the mean and peak input rates that satisfies

$$C_j^* \, \mathrm{e}^{-\eta_j B_j} = \varepsilon_j,$$

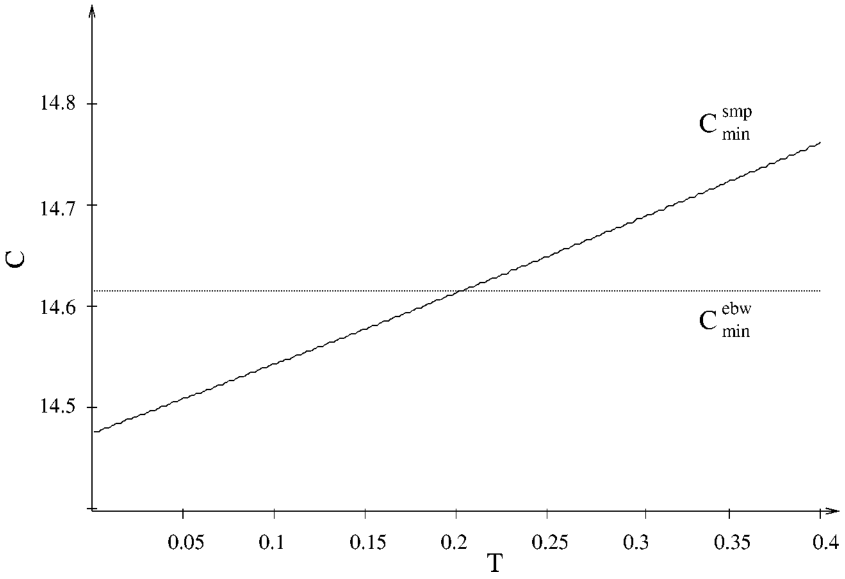


Figure 8. Estimates of the required bandwidth.

where C_j^* is obtained using equation (25). We denote the c value obtained as $c^{\text{smp}}_{\text{min}}$ since it is the smallest output capacity that would result in satisfying the QoS criterion

$$C_j^* \,\mathrm{e}^{-\eta_j B_j} < \varepsilon_j.$$

Figure 8 shows the results for $c^{\text{ebw}}_{\text{min}}$ and $c^{\text{smp}}_{\text{min}}$ when $\varepsilon_j = 10^{-5}$, $K_j = 10$, and T varies from 0.01 to 0.40 while τ_j/T is fixed. Intuitively, the bandwidth required should increase with T so that all the buffer contents are drained out when the server is serving the buffer. The $c^{\text{smp}}_{\text{min}}$ obtained using the SMP bounds technique is consistent with our intuition. On the other hand, $c^{\text{ebw}}_{\text{min}}$ does not vary with T. Therefore on using the effective-bandwidth technique one faces the risk of the QoS criteria not being satisfied.

4.4. Two classes: admission control

Consider the exponential on-off source model described in section 4.3 with two classes of traffic (say, real-time and non-real-time), i.e., $N = 2$. The admission control is performed in the following manner: consider at a given point of time k_1 class 1 sources and k_2 class 2 sources are transmitting. At this time, if a new source arrives into the system, the admission control scheme decides whether or and not to admit this source. A simple admission control scheme is an admissible region such that all points within it denote the number of class 1 and class 2 sources such that their QoS is satisfied. Let the quality of service parameter for buffer j, $j = 1, 2$, under the timed round-robin (trr) discipline be

$$G_j^{\text{trr}}(K_1, K_2) = P(X_j > B_j).$$

The aim is to identify the feasible region

$$\mathcal{K}^{\mathrm{trr}} = \big\{(K_1, K_2)\colon\ G_1^{\mathrm{trr}}(K_1, K_2) < \varepsilon_1,\ G_2^{\mathrm{trr}}(K_1, K_2) < \varepsilon_2\big\}. \tag{28}$$

To begin with, we assume that the cycle time T and the switch-over time t_{so} are fixed known constants. However, the values τ_1 and τ_2 are variable and are appropriately chosen such that $\tau_1 + \tau_2 + t_{\mathrm{so}} = T$. We use the following algorithm to compute the feasible region $\mathcal{K}^{\mathrm{trr}}$ in equation (28). Note that the algorithm can be executed off-line to compute $\mathcal{K}^{\mathrm{trr}}$ and the required $\tau_1,\ \tau_2$. This can be stored and used by table-look-up to implement on-line admission control. The algorithm does not need to be executed at every admission decision, but only when the input parameters change.

Algorithm 1. An algorithm to compute the feasible region:

1. Set $\mathcal{K} = \emptyset$.
2. Let $\tau_1 = T$ and $\tau_2 = 0$. (The scheduler always serves only buffer 1, hence there are no switch-over times and no compensating source.)
3. Obtain the maximum number of admissible class-1 sources $K_1^{\max}$ as the maximum value of K_1 such that

   $$C_1^* \,\mathrm{e}^{-\eta_1 B_1} < \varepsilon_1,$$

   where

   $$C_1^* = \frac{[r_1 K_1/(\alpha_1 c)]^{K_1}}{((\alpha_1 + \beta_1)/(\alpha_1\beta_1))^{K_1}(K_1\beta_1/(K_1\beta_1 + \eta_1 c))^{K_1 - 1}} \tag{29}$$

   and

   $$\eta_1 = \frac{c(\alpha_1 + \beta_1) - r_1 K_1 \beta_1}{(c/K_1)(r_1 K_1 - c)}.$$
4. $\mathcal{K} = \mathcal{K} \cup \{(0, 0), (1, 0), \ldots, (K_1^{\max}, 0)\}$.
5. Let $\tau_2 = T$ and $\tau_1 = 0$. (The scheduler always serves only buffer 2, hence there are no switch-over times and no compensating source.)
6. Obtain the maximum number of admissible class-2 sources $K_2^{\max}$ as the maximum value of K_2 such that

   $$C_2^* \,\mathrm{e}^{-\eta_2 B_2} < \varepsilon_2,$$

   where

   $$C_2^* = \frac{[r_2 K_2/(\alpha_2 c)]^{K_2}}{((\alpha_2 + \beta_2)/(\alpha_2\beta_2))^{K_2}(K_2\beta_2/(K_2\beta_2 + \eta_2 c))^{K_2 - 1}} \tag{30}$$

   and

   $$\eta_2 = \frac{c(\alpha_2 + \beta_2) - r_2 K_2 \beta_2}{(c/K_2)(r_2 K_2 - c)}.$$

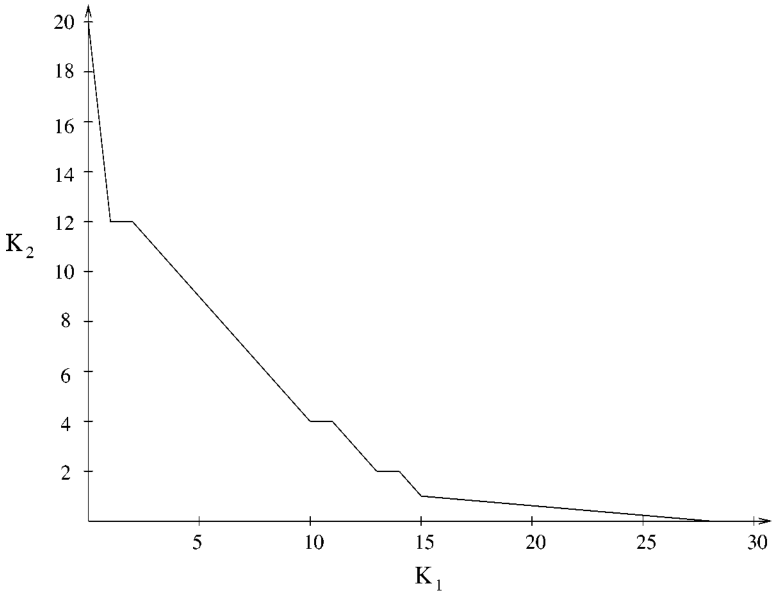


Figure 9. Admissible region $\mathcal{K}^{\text{trr}}$.

```
7. $\mathcal{K} = \mathcal{K} \cup \{(0,1), (0,2), \ldots, (0, K_2^{\max})\}$.
8. Set $K_1 = 1$.
9. While $K_1 < K_1^{\max}$:
    (i) Compute the minimum required $\tau_1$ ($\leqslant T - t_{\text{so}}$) such that the loss probability is less than $\varepsilon_1$.
    (ii) Compute the available $\tau_2$ ($= T - t_{\text{so}} - \tau_1$).
    (iii) Given $\tau_2$, compute the maximum possible $K_2$ value by minimizing over the set $\mathcal{A}^2$ (see equation (23)) for $K_2 + 1$ sources.
    (iv) $\mathcal{K} = \mathcal{K} \cup \{(K_1, 1), (K_1, 2), \ldots, (K_1, K_2)\}$.
    (v) $K_1 = K_1 + 1$.
10. Return $\mathcal{K}^{\text{trr}} = \mathcal{K}$.
```

Using algorithm 1, we plot the admissible region $\mathcal{K}^{\text{trr}}$ in figure 9 using the following numerical values:

$$\begin{aligned} &\alpha_1 = 1, \quad \beta_1 = 0.3, \quad r_1 = 1.0, \quad \varepsilon_1 = 10^{-6}, \quad B_1 = 8, \quad T = 1.22, \quad t_{\text{so}} = 0.02, \\ &\alpha_2 = 1, \quad \beta_2 = 0.2, \quad r_2 = 1.23, \quad \varepsilon_2 = 10^{-9}, \quad B_2 = 10 \quad \text{and} \quad c = 13.22. \end{aligned} \tag{31}$$

Note that there is a steep fall in the admissible region from $(0, 20)$ to $(1, 12)$. This is due to using equation (30) for the K_2 sources and using equation (25) for the $K_2 + 1$

Table 1
Values of (τ_1, τ_2) to be used for given K_1 and K_2. A '–' in the table denotes that the corresponding combination K_1 and K_2 if not admissible.

$K_1 \setminus K_2$	0	1	2	3	4	5	6	7	8	9
0	(0, 1.22)	(0, 1.22)	(0, 1.22)	(0, 1.22)	(0, 1.22)	(0, 1.22)	(0, 1.22)	(0, 1.22)	(0, 1.22)	(0, 1.22)
1	(0.0510, 1.1490)	(0.0510, 1.1490)	(0.0510, 1.1490)	(0.0510, 1.1490)	(0.0510, 1.1490)	(0.0510, 1.1490)	(0.0510, 1.1490)	(0.0510, 1.1490)	(0.0510, 1.1490)	(0.0510, 1.1490)
2	(0.1072, 1.0928)	(0.1072, 1.0928)	(0.1072, 1.0928)	(0.1072, 1.0928)	(0.1072, 1.0928)	(0.1072, 1.0928)	(0.1072, 1.0928)	(0.1072, 1.0928)	(0.1072, 1.0928)	(0.1072, 1.0928)
3	(0.1703, 1.0297)	(0.1703, 1.0297)	(0.1703, 1.0297)	(0.1703, 1.0297)	(0.1703, 1.0297)	(0.1703, 1.0297)	(0.1703, 1.0297)	(0.1703, 1.0297)	(0.1703, 1.0297)	(0.1703, 1.0297)
4	(0.2425, 0.9575)	(0.2425, 0.9575)	(0.2425, 0.9575)	(0.2425, 0.9575)	(0.2425, 0.9575)	(0.2425, 0.9575)	(0.2425, 0.9575)	(0.2425, 0.9575)	(0.2425, 0.9575)	(0.2425, 0.9575)
5	(0.3246, 0.8754)	(0.3246, 0.8754)	(0.3246, 0.8754)	(0.3246, 0.8754)	(0.3246, 0.8754)	(0.3246, 0.8754)	(0.3246, 0.8754)	(0.3246, 0.8754)	(0.3246, 0.8754)	(0.3246, 0.8754)
6	(0.4145, 0.7855)	(0.4145, 0.7855)	(0.4145, 0.7855)	(0.4145, 0.7855)	(0.4145, 0.7855)	(0.4145, 0.7855)	(0.4145, 0.7855)	(0.4145, 0.7855)	(0.4145, 0.7855)	–
7	(0.5071, 0.6929)	(0.5071, 0.6929)	(0.5071, 0.6929)	(0.5071, 0.6929)	(0.5071, 0.6929)	(0.5071, 0.6929)	(0.5071, 0.6929)	(0.5071, 0.6929)	–	–
8	(0.5975, 0.6025)	(0.5975, 0.6025)	(0.5975, 0.6025)	(0.5975, 0.6025)	(0.5975, 0.6025)	(0.5975, 0.6025)	(0.5975, 0.6025)	–	–	–
9	(0.6833, 0.5167)	(0.6833, 0.5167)	(0.6833, 0.5167)	(0.6833, 0.5167)	(0.6833, 0.5167)	(0.6833, 0.5167)	–	–	–	–

sources (including the compensating source) respectively for the two points (0, 20) and (1, 12). Note that the correct choice of τ_1 and τ_2 depend upon K_1 and $K_2 \in \mathcal{K}^{\text{trr}}$. Table 1 gives values of (τ_1, τ_2) for selected values of $(K_1, K_2) \in \mathcal{K}^{\text{trr}}$. Note that from step 9 in algorithm 1, τ_1 (and hence τ_2) depends only on K_1. However, there could be other choices of (τ_1, τ_2) for a given (K_1, K_2).

Next we discuss the effect of varying T, the cycle time. Intuitively, for $T \gg t_{\text{so}}$, an increase in T would result in a smaller admissible region and a decrease in T would result in a bigger admissible region. We confirm our intuition by observing the results in figure 10 (using the numerical values in (31)) for the cases $T = 1.22$ and $T = 12.02$. When T is approximately of the same order of magnitude as t_{so}, a significant fraction of the server off-time is the switch-over time. Hence it is not clear how the admissible region would change with T in this case. From figure 10 (using the numerical values in (31)) we can see that for the cases $T = 0.14$ and $T = 1.22$, one region is not the subset of the other. Hence we conclude that it is not straightforward to obtain an optimal value of T such that the feasible region is maximized. Note that if $t_{\text{so}} = 0$, then the optimal T is such that $T \to 0$.

Before moving on to the next service scheduling policy, namely, static priority policy, we briefly explain how the admission control can be carried out in reality by

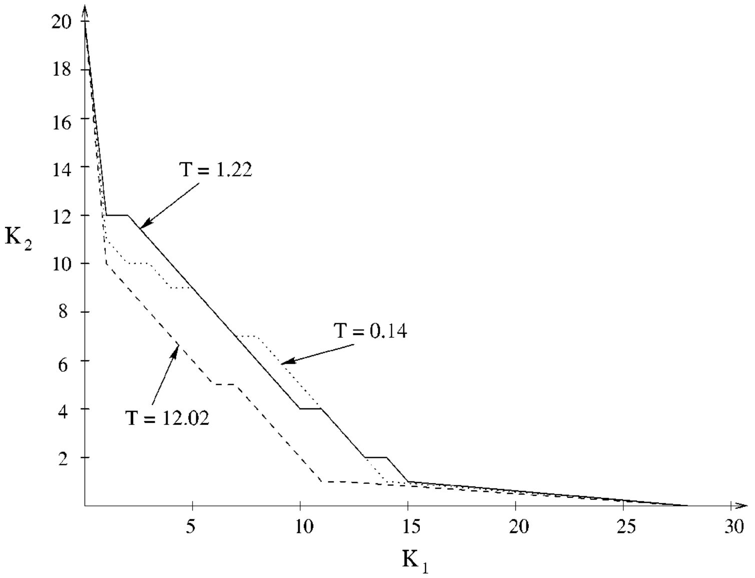


Figure 10. $\mathcal{K}^{\mathrm{trr}}$ as a function of T.

discussing a few implementation issues here. For the two classes of traffic, admission control is implemented using the following algorithm:

Algorithm 2. An algorithm to implement admission control:

```
1. Obtain the input parameters (viz. $\alpha_1$, $\alpha_2$, $\beta_1$, $\beta_2$, $r_1$, $r_2$, $\varepsilon_1$, $\varepsilon_2$, $B_1$, $B_2$, $T$, $t_{\mathrm{so}}$
   and $c$).
2. Use algorithm 1 to generate (off-line) a table similar to table 1 and store the table.
3. When a source arrives, look up the table.
   If the $(\tau_1, \tau_2)$ entry for the new $(K_1, K_2)$ value is missing
      then reject the source
   else
      accept the source and use the new $(\tau_1, \tau_2)$ values.
4. Wait until another new source arrives or the input parameters change.
   If a new source arrives, go to step 3 else go to step 1.
```

Note from table 1 that the (τ_1, τ_2) values need not be modified when a source departs. Hence we do not consider source departing events in algorithm 2. Also this means that the (τ_1, τ_2) values need not be changed very often.

5. Static priority service policy

In this section we analyze the *static priority service policy* (for the model in section 3 and illustrated in figure 3) to multiplex the multi-class traffic. Under this service policy, traffic of class j has higher service priority over traffic of class i, if $i > j$. The scheduler serves the traffic of class j only if there is no fluid of higher priority in the buffers. Thus all the available channel capacity (a maximum of c) is assigned for the class-1 fluid and the leftover channel capacity (if any) that class-1 does not need, to class-2 fluid. Any leftover channel capacity that class-1 and class-2 do not need, is assigned to class-3 fluid, and so on.

5.1. Two classes: admission control

We concentrate on the case of two-class traffic, although the analysis can be extended to more than 2 classes. The K_j class-j sources, $j = 1, 2$, are independent and identical on-off sources with exponential on and off times, on-time parameter α_j, off-time parameter β_j and on-time rate r_j (similar to the model in section 4.2). Let the Quality of Service parameter for buffer j, $j = 1, 2$, under the static priority policy (spp) regime be

$$G_j^{\text{spp}}(K_1, K_2) = \lim_{t\to\infty} P\bigl(X, (t) > B_j\bigr) = P(X_j > B_j).$$

The aim is to identify the feasible region

$$\mathcal{K}^{\text{spp}} = \bigl\{(K_1, K_2)\colon\ G_1^{\text{spp}}(K_1, K_2) < \varepsilon_1,\ G_2^{\text{spp}}(K_1, K_2) < \varepsilon_2\bigr\}. \tag{32}$$

Note that even though $X_1(t)$ and $X_2(t)$ are dependent, for the QoS performance analysis, we require only the marginal distributions of $X_1(t)$ and $X_2(t)$. Since the marginal distributions (that take the dependence of $X_1(t)$ and $X_2(t)$ into account) can be easily computed, we do not present the joint distribution of $X_1(t)$ and $X_2(t)$ here.

Unlike the timed round-robin policy where we obtained the feasible admissible region $\mathcal{K}^{\text{trr}}$ using only the SMP bounds, for the static priority service policy, we obtain feasible admissible region $\mathcal{K}^{\text{spp}}$ using three different methods, namely, effective bandwidth approximation, Chernoff dominant eigenvalue (CDE) approximation and the SMP bounds. Each method produces a different feasible region $\mathcal{K}^{\text{spp}}$.

In the next section we illustrate how to compute the feasible region $\mathcal{K}_{\text{smp}}$ using SMP bounds. We compare $\mathcal{K}_{\text{smp}}$ with the feasible regions obtained using effective bandwidth approximation ($\mathcal{K}_{\text{ebw}}$ and its relaxation $\mathcal{N}$) and Chernoff dominant eigenvalue approximation ($K_{\text{cde}}^{(1)}$ and its relaxation $\mathcal{K}_{\text{cde}}^{(2)}$) that are explained and computed in [18].

5.2. SMP bounds

Consider the transformed model depicted in figure 11. The sample paths of the buffer content processes $\{X_1(t),\ t \geqslant 0\}$ and $\{X_2(t),\ t \geqslant 0\}$ in this model are identical to those in the original model in figure 3 (for $N = 2$). This observation is made in [10] and is immensely useful in our analysis. Note that the output from buffer 1 can be modeled as an SMP. Hence, the input to buffers 1 and 2 can be modeled as ones with multiplexing independent SMP sources. Therefore, we can use the SMP bounds in section 2.2 to obtain a feasible admissible region that is guaranteed to be conservative.

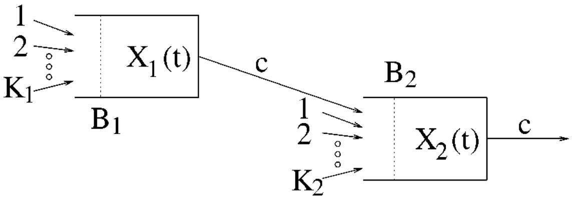


Figure 11. The transformed model.

Buffer 1. If $K_1 \leqslant c/r_1$, then $P\{X_1 > B_1\} = 0$, since buffer 1 will always be empty. Now for the case $K_1 > c/r_1$, the steady-state distribution of the buffer-content process is bounded as

$$C_{*1}\,\mathrm{e}^{-\eta_1 B_1} \leqslant P\{X_1 > B_1\} \leqslant C_1^*\,\mathrm{e}^{-\eta_1 B_1},$$

where

$$\eta_1 = \frac{K_1(c\alpha_1 + c\beta_1 - K_1\beta_1 r_1)}{c(K_1 r_1 - c)}, \tag{33}$$

$$C_1^* = \frac{\left(\frac{K_1 r_1}{K_1 r_1 - c}\,\frac{\alpha_1}{(\alpha_1+\beta_1)}\right)^{K_1}}{\left(\frac{c\alpha_1}{\beta_1(K_1 r_1 - c)}\right)^{\lceil c/r_1\rceil}},$$

and

$$C_{*1} = \left(\frac{K_1 r_1 \beta_1}{c(\alpha_1 + \beta_1)}\right)^{K_1}.$$

Buffer 2. We first model the K_2 exponential on-off sources as a single (K_1+1)-state SMP with the states denoting the number of priority-2 sources that are on and then derive expressions for H^1, $\Psi^1_{\max}(i,j)$ and $\Psi^1_{\min}(i,j)$ defined in equations (6)–(8) (see appendix). In [18] it is shown that the output process from buffer 1 can be modeled as an SMP. In appendix, we derive the corresponding expressions H^2, $\Psi^2_{\max}(i,j)$ and $\Psi^2_{\min}(i,j)$ for the SMP model of the output from buffer 1. Therefore we can analyze the input to buffer 2 as traffic from two sources (output from buffer 1 and the (K_2+1)-state SMP), each modulated by an SMP.

We begin by obtaining η_2. Using the effective bandwidth of the output from a buffer described in [18], we can show that η_2 solves either

$$K_1\,eb_1(\eta_2) + K_2\,eb_2(\eta_2) = c \quad \text{and} \quad \eta_2 \leqslant v^* \tag{34}$$

or

$$\frac{v^*}{\eta_2} K_1\,eb_1\big(v^*\big) + K_2\,eb_2(\eta_2) = \frac{cv^*}{\eta_2} \quad \text{and} \quad \eta_2 > v^*, \tag{35}$$

where

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

and for $j = 1, 2$

$$eb_j(v) = \frac{r_j v - \alpha_j - \beta_j + \sqrt{(r_j v - \alpha_j - \beta_j)^2 + 4\beta_j r_j v}}{2v} \tag{36}$$

Therefore using the expressions for H^1, $\Psi^1_{\max}(i, j)$, $\Psi^1_{\min}(i, j)$, H^2, $\Psi^2_{\max}(i, j)$ and $\Psi^2_{\min}(i, j)$ from appendix, we have

$$C_2^* = \frac{H^1 H^2}{\min\limits_{(i_1,j_1),(i_2,j_2):\ \min\{i_1 r_1, c\}+i_2 r_2 > c,\, p_{i_1 j_1}>0,\, p_{i_2 j_2}>0} \Psi^1_{\min}(i_1, j_1)\Psi^2_{\min}(i_2, j_2)}$$

and

$$C_{*2} = \frac{H^1 H^2}{\max\limits_{(i_1,j_1),(i_2,j_2):\ \min\{i_1 r_1, c\}+i_2 r_2 > c,\, p_{i_1 j_1}>0,\, p_{i_2 j_2}>0} \Psi^1_{\max}(i_1, j_1)\Psi^2_{\max}(i_2, j_2)}.$$

We obtain the feasible admissible region K_{smp} as the set of all values (K_1, K_2) that satisfy

$$C_1^* \,\mathrm{e}^{-\eta_1 B_1} < \varepsilon_1, \qquad C_2^* \,\mathrm{e}^{-\eta_2 B_2} < \varepsilon_2. \tag{37}$$

5.3. *Comparisons*

We compare the region $\mathcal{K}_{\text{smp}}$ with the regions obtained using the CDE approximation, $\mathcal{K}^{(1)}_{\text{cde}}$, and $\mathcal{K}^{(2)}_{\text{cde}}$, and the regions obtained by effective-bandwidth approximation $\mathcal{K}_{\text{ebw}}$ and $\mathcal{N}$ (using numerical results from [18]). We represent the regions under consideration in figure 12 using the following numerical values:

$$\begin{array}{llllll} \alpha_1 = 1.0, & \beta_1 = 0.2, & r_1 = 1.0, & \varepsilon_1 = 10^{-9}, & B_1 = 10, & \\ \alpha_2 = 1.0, & \beta_2 = 0.2, & r_2 = 1.23, & \varepsilon_2 = 10^{-6}, & B_2 = 10 & \text{and} \quad c = 13.2. \end{array} \tag{38}$$

The region obtained by the SMP bounds, $\mathcal{K}_{\text{smp}}$, is conservative. Therefore if an admissible region has points in $\mathcal{K}_{\text{smp}}$, then those points are guaranteed to satisfy the QoS criteria. Our numerical investigation suggests that

$$\mathcal{N} \subset \mathcal{K}_{\text{ebw}} \subset \mathcal{K}_{\text{smp}},$$

although we do not have a proof of it. This means that the effective-bandwidth approximation produces overly conservative results for these parameter values. Usually the effective bandwidth produces conservative results but it is not guaranteed to be conservative, unlike the results from SMP bounds. On one hand, the effective-bandwidth approximation is computationally easy, on the other hand, it could either be too conservative (and hence leading to under-utilization of resources) or be unconservative (and hence uncertain about meeting the QoS criteria).

The CDE approximation, although computationally slower than the effective-bandwidth approximation, is typically faster than the SMP bounds technique. However, there are examples where we can show that the CDE approximation produces regions

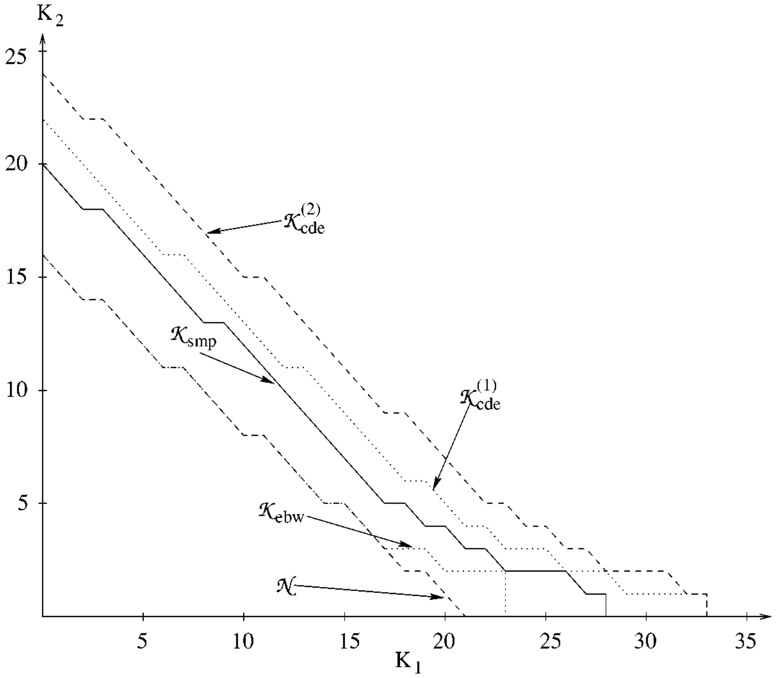


Figure 12. Regions $\mathcal{N}$, $\mathcal{K}_{\text{ebw}}$, $\mathcal{K}_{\text{smp}}$, $\mathcal{K}^{(1)}_{\text{cde}}$, $\mathcal{K}^{(2)}_{\text{cde}}$.

$\mathcal{K}^{(1)}_{\text{cde}}$ and $\mathcal{K}^{(2)}_{\text{cde}}$, with points (K_1, K_2) that would actually result in a higher cell loss than what is allowable! Therefore while using the CDE approximation, we run the risk of the QoS criteria not being satisfied.

Using SMP bounds is computationally intensive. However, similar to the implementation for the timed round robin policy in section 4.4, here too the computation can be done off line and the feasible region can be stored in a table. The computation needs to be repeated only when the input parameters change.

6. Comparisons: timed round robin vs. static priority

In this section we compare the timed round-robin policy against the static priority service policy. From section 4, it is clear that the admissible region obtained by the timed round-robin policy is dependent on the value of the cycle time T chosen. In the following comparison, we choose T to be equal to $c(B_1 + B_2) + t_{\text{so}}$. All other numerical values are as in (31). In figure 13 we compare the two policies, timed round-robin and static priority by viewing their respective admissible regions (using SMP bounds) for 2-class exponential on-off sources with parameters in (31).

From the figure we see the timed round-robin policy results in a smaller admissible region. This is because the timed round-robin policy is not a work conserving service discipline unlike the static priority service policy. Clearly, static priority service pol-

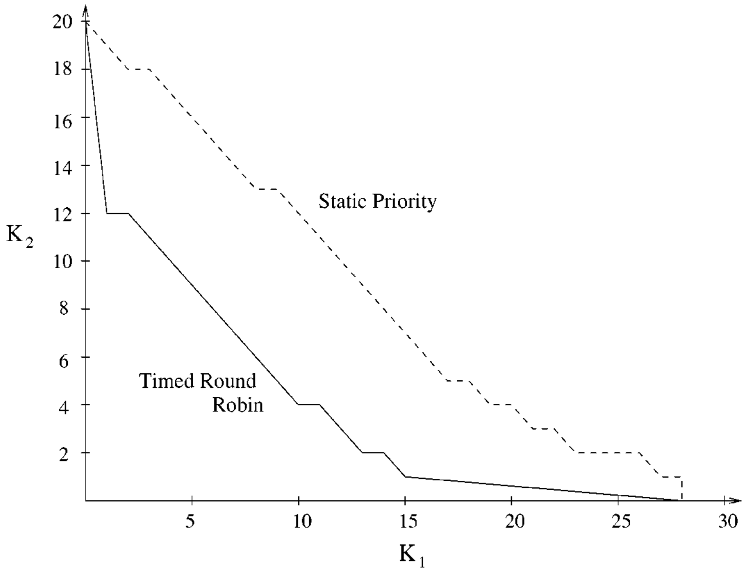


Figure 13. Timed round-robin vs. static priority.

icy does not achieve fairness among the classes of traffic. Therefore it may not be an appropriate policy to use at all times.

7. Conclusions and extensions

In this paper we analyze a single multi-class node of a high speed network with N buffers, one for each class of traffic, and each buffer served according to a given service scheduling policy. We study the overflow probability computations for the N buffers served according to the timed round robin policy and the static priority service policy.

For the timed round robin policy, using a transformed model, we apply the standard effective-bandwidth approximation to produce estimates for the overflow probability in the asymptotic case when the buffers are of extremely large sizes. We also obtain estimates of the overflow probability using the SMP bounds. On comparing the performance of the two techniques we see that the SMP bounds conform to our intuition whereas the effective-bandwidth approximation produces counter-intuitive results. Hence we use the SMP bounds to obtain feasible admissible regions to solve call admission control problems. We also show that the switch-over time plays an important role especially while chosing an optimal cycle time T.

For the static priority service policy, we use the SMP bounds technique to obtain a conservative admissible region to solve call admission control problems. Using admissible regions, we compare the SMP bounds method with other existing approximate methods and conclude that depending on what kind of trade-off one would like to do between computational time, conservativeness and size of the admissible region, an appropriate method can be used. On comparing the admissible region obtained using the SMP bounds for the timed round-robin policy and the static priority service policy, we see that due to the non-work-conserving nature of the timed round-robin scheduling policy, we obtain a smaller admissible region.

We also like to point out that for both the service scheduling policies we can use SMP bounds to solve the admission control problem in the following manner: precompute all the required parameters and store the points in the admissible region in a table. When a source arrives, perform a table-look-up and check whether or not adding this new source would result in a point in the admissible region. Also, for the timed round robin policy, the table can also store the new τ_1 and τ_2 values.

Note that it is possible to extend the computational results to $N > 2$ classes. Other extensions that will be considered in the future include analyzing different polling policies, e.g., exhaustive policy and gated policy. We have so far only dealt with a single node in a network. It is possible to extend the analysis to a series network of nodes, each using a given service scheduling policy. The main concern for using effective-bandwidth-based approaches is that the output from the different buffers will be correlated due to the service policy followed.

Appendix

The input traffic to buffer 2 can be modeled as two streams of traffic produced by two independent sources modulated by SMPs. The first is modulated by a $(K_2 + 1)$-state SMP which is the aggregate source formed using the K_2 exponential on-off sources that input traffic into buffer 2. In fact, the $(K_2 + 1)$-state SMP is a continuous time Markov chain (CTMC). The second source is the output from buffer 1 which can be modeled as an SMP. We begin by analyzing the $(K_2 + 1)$-state SMP and obtain expressions for H^1, $\Psi^1_{\max}(i, j)$ and $\Psi^1_{\min}(i, j)$. Then we analyze the SMP model of the output from buffer 1 and obtain expressions for H^2, $\Psi^2_{\max}(i, j)$ and $\Psi^2_{\min}(i, j)$. We follow the analysis and notations in section 2.2.

Using the expression for η_2 in equations (34) and (35), and for $eb_j(v)$ in equation (36), define

$$c^1 = K_2\, eb_2(\eta_2)$$

and

$$c^2 = c - K_2\, eb_2(\eta_2).$$

The $(K_2 + 1)$-state SMP. For $i = 0, 1, \ldots, K_2$ and $j = 0, 1, \ldots, K_2$, we define the following:

$$G^1_{i,j}(x) = \begin{cases} \dfrac{i\alpha_2}{i\alpha_2 + (K_2 - i)\beta_2}\big(1 - \exp\big\{-\big(i\alpha_2 + (K_2 - i)\beta_2\big)x\big\}\big) & \text{if } j = i - 1, \\ \dfrac{(K_2 - i)\beta_2}{i\alpha_2 + (K_2 - i)\beta_2}\big(1 - \exp\big\{-\big(i\alpha_2 + (K_2 - i)\beta_2\big)x\big\}\big) & \text{if } j = i + 1, \\ 0 & \text{otherwise,} \end{cases}$$

$$G_i^1(x) = 1 - \exp\bigl\{-\bigl(i\alpha_2 + (K_2 - i)\beta_2\bigr)x\bigr\},$$
$$\tau_i^1 = \frac{1}{i\alpha_2 + (K_2 - i)\beta_2},$$
$$P_{ij}^1 = G_{i,j}^1(\infty),$$

and

$$p_i^1 = \frac{a_i^1 \tau_i^1}{\sum_{m=0}^{K_2} a_m^1 \tau_m^1} = \frac{K_2!}{i!(K_2 - i)!} \frac{\alpha_2^{K_2 - i} \beta_2^i}{(\alpha_2 + \beta_2)^{K_2}}.$$

Then $\Psi^1(\eta_2)$ is given by

$$\phi_{i,j}^1(\eta_2) = \begin{cases} \dfrac{i\alpha_2}{i\alpha_2 + (K_2 - i)\beta_2 - (ir_2 - c^1)\eta_2} & \text{if } j = i - 1, \\ \dfrac{(K_2 - i)\beta_2}{i\alpha_2 + (K_2 - i)\beta_2 - (ir_2 - c^1)\eta_2} & \text{if } j = i + 1, \\ 0 & \text{otherwise.} \end{cases}$$

Also, the eigenvectors are obtained by solving

$$h^1 = h^1 \Phi^1(\eta_2).$$

Therefore

$$H^1 = \sum_{i=0}^{K_2} \frac{h_i^1}{\eta_2(ir_2 - c^1)} \left(\sum_{j=0}^{K_2} \bigl(\phi_{ij}^1(\eta_2)\bigr) - 1 \right) \quad \text{and}$$
$$\Psi_{\max}^1(i, j) = \Psi_{\min}^1(i, j) = \frac{h_i^1 \, \mathrm{e}^{-\eta_2(ir_2 - c^1)x} \int_x^\infty \mathrm{e}^{\eta_2(ir_2 - c^1)y} \, \mathrm{d}G_{ij}^1(y)}{(p_i^1/\tau_i^1) \int_x^\infty \mathrm{d}G_{ij}^1(y)}$$
$$= \frac{h_i^1}{p_i^1} \frac{1}{i\alpha_2 + (K_2 - i)\beta_2 - \eta_2(ir_2 - c^1)}.$$

The output from buffer 1. To model the output from buffer 1 as an SMP we need to consider two scenarios. In the first we let $K_1 \leqslant c/r_1$ and thus buffer 1 would always be empty and output from buffer 1 can be modeled as a continuous time Markov chain. In the second when $K_1 > c/r_1$ we need to model the output as an SMP. We treat the two cases separately.

Case (i). If $K_1 \leqslant c/r_1$ then buffer 1 is always empty and the output from buffer 1 is a $K_1 + 1$ state SMP/CTMC identical to that of the input to buffer 1. Hence we have, for

$i = 0, 1, \dots, K_1$ and $j = 0, 1, \dots, K_1$,

$$
\begin{aligned}
G^2_{i,j}(x) &= \begin{cases} \dfrac{i\alpha_1}{i\alpha_1 + (K_1 - i)\beta_1}\bigl(1 - \exp\bigl\{-\bigl(i\alpha_1 + (K_1 - i)\beta_1\bigr)x\bigr\}\bigr) & \text{if } j = i - 1, \\ \dfrac{(K_1 - i)\beta_1}{i\alpha_1 + (K_1 - i)\beta_1}\bigl(1 - \exp\bigl\{-\bigl(i\alpha_1 + (K_1 - i)\beta_1\bigr)x\bigr\}\bigr) & \text{if } j = i + 1, \\ 0 & \text{otherwise,} \end{cases} \\
G^2_i(x) &= 1 - \exp\bigl\{-\bigl(i\alpha_1 + (K_1 - i)\beta_1\bigr)x\bigr\}, \\
\tau^2_i &= \frac{1}{i\alpha_1 + (K_1 - i)\beta_1}, \\
P^2_{ij} &= G^2_{i,j}(\infty),
\end{aligned}
$$

and it is easy to derive

$$
p^2_i = \frac{a^2_i \tau^2_i}{\sum_{m=0}^{K_1} a^2_m \tau^2_m} = \frac{K_1!}{i!(K_1 - i)!}\,\frac{\alpha_1^{K_1 - i}\beta_1^i}{(\alpha_1 + \beta_1)^K_1}.
$$

Then $\Phi^2(\eta_2)$ is given by

$$
\phi^2_{i,j}(\eta_2) = \begin{cases} \dfrac{i\alpha_1}{i\alpha_1 + (K_1 - i)\beta_1 - (ir_1 - c^2)\eta_2} & \text{if } j = i - 1, \\ \dfrac{(K_1 - i)\beta_1}{i\alpha_1 + (K_1 - i)\beta_1 - (ir_1 - c^1)\eta_2} & \text{if } j = i + 1, \\ 0 & \text{otherwise.} \end{cases}
$$

Also, the eigenvectors are obtained by solving

$$
h^2 = h^2 \Phi^2(\eta_2).
$$

Hence we have

$$
\begin{aligned}
H^2 &= \sum_{i=0}^{K_1} \frac{h^2_i}{\eta_2(ir_1 - c^2)} \left(\sum_{j=0}^{K_1} \bigl(\phi^2_{ij}(\eta_2)\bigr) - 1 \right) \quad \text{and} \\
\Psi^2_{\max}(i, j) = \Psi^2_{\min}(i, j) &= \frac{h^2_i\, \mathrm{e}^{-\eta_2(ir_1 - c^2)x} \int_x^\infty \mathrm{e}^{\eta_2(ir_1 - c^2)y}\, \mathrm{d}G^2_{ij}(y)}{(p^2_i/\tau^2_i) \int_x^\infty \mathrm{d}G^2_{ij}(y)} \\
&= \frac{h^2_i}{p^2_i}\, \frac{1}{i\alpha_1 + (K_2 - i)\beta_1 - \eta_2(ir_1 - c^2)}.
\end{aligned}
$$

Case (ii). If $K_1 > c/r_1$, we do the following analysis. Let $M = \lceil c/r_1 \rceil$. Then the output from buffer 1 can be modeled as an SMP on state space $\{0, 1, 2, \dots, M\}$ (see [18]). For $i = 0, 1, \dots, M-1$ and $j = 0, 1, \dots, M$, let

$$G^2_{i,j}(t) = \begin{cases} \dfrac{i\alpha_1}{i\alpha_1 + (K_1 - i)\beta_1}\bigl(1 - \exp\bigl\{-\bigl(i\alpha_1 + (K_1 - i)\beta_1\bigr)t\bigr\}\bigr) & \text{if } j = i - 1, \\ \dfrac{(K_1 - i)\beta_1}{i\alpha_1 + (K_1 - i)\beta_1}\bigl(1 - \exp\bigl\{-\bigl(i\alpha_1 + (K_1 - i)\beta_1\bigr)t\bigr\}\bigr) & \text{if } j = i + 1, \\ 0 & \text{otherwise.} \end{cases}$$

Let

$$T = \min\bigl\{t > 0\colon\ X_1(t) = 0\bigr\}.$$

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

$$G^2_{M,j}(t) = P\bigl\{T \leqslant t,\ \overline{N}(T) = j \mid X_1(0) = 0,\ \overline{N}(0) = M\bigr\},$$

where $\overline{N}(t)$ denotes the number of priority 1 sources on at time t. (Note that $G^2_{M,M}(t) = 0$.) We need $G^2(\infty) = [G^2_{i,j}(\infty)]$ in our analysis. We have for $i = 0, 1, \dots, M-1$ and $j = 0, 1, \dots, M$,

$$G^2_{i,j}(\infty) = \begin{cases} \dfrac{i\alpha_1}{i\alpha_1 + (K_1 - i)\beta_1} & \text{if } j = i - 1, \\ \dfrac{(K_1 - i)\beta_1}{i\alpha_1 + (K_1 - i)\beta_1} & \text{if } j = i + 1, \\ 0 & \text{otherwise,} \end{cases} \tag{39}$$

$$G^2_{M,j}(\infty) = \widetilde{G}^2_{M,j}(0),$$

where $\widetilde{G}^2_{M,j}(s)$ is the Laplace–Stieltjes transform (LST) of $G^2_{M,j}(t)$, and can be computed using the analysis in [20].

We also need the expression for the sojourn time τ^2_i in state i, for $i = 0, 1, \dots, M$. We have

$$\tau^2_i = \begin{cases} \dfrac{1}{i\alpha_1 + (K_1 - i)\beta_1} & \text{if } i = 0, 1, \dots, M-1, \\ \displaystyle\sum_{j=0}^{M-1} \widetilde{G}^{2\prime}_{M,j}(0) & \text{if } i = M. \end{cases}$$

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

$$p^2_i = \frac{a^2_i \tau^2_i}{\sum_{k=0}^{M} a^2_k \tau^2_k}, \tag{40}$$

where

$$a^2 = a^2 G^2(\infty).$$

Following the analysis in [12] define

$$\bar{\phi}^2_{ij}(\eta_2, m) = \begin{cases} \widetilde{G}^2_{ij}\big(-\eta_2\big(ir_1 - c^2\big)\big) & \text{if } 0 \leqslant i \leqslant M-1, \\ m\widetilde{G}^2_{ij}\big(-\eta_2\big(c - c^2\big)\big) & \text{if } i = M. \end{cases}$$

Solve for m such that the Perron–Frobenius eigenvalue of $\overline{\Phi}^2(\eta_2, m)$ is 1. Hence we obtain h^2 from

$$h^2\overline{\Phi}^2(\eta_2, m) = h^2.$$

It can be shown that random variables with distribution $G^2_{Mj}(x)/G^2_{Mj}(\infty)$ have a decreasing failure rate. Hence $\Psi^2_{\min}(M, j)$ and $\Psi^2_{\max}(M, j)$ occur at $x = \infty$ and $x = 0$ respectively. Thus we have for $(i, j) \in \{0, 1, \ldots, M\}$,

$$H^2 = \sum_{i=0}^{M} \frac{h^2_i}{\eta_2(ir_1 - c^2)} \left(\sum_{j=0}^{M} \big(\bar{\phi}^2_{ij}(\eta_2, m)\big) - 1 \right),$$

$$\Psi^2_{\min}(i, j) = \inf_x \left\{ \frac{h^2_i \,\mathrm{e}^{-\eta_2(ir_1 - c^2)x} \int_x^\infty \mathrm{e}^{\eta_2(ir_1 - c^2)y}\, \mathrm{d}G^2_{ij}(y)}{p^2_i/(i\alpha_1 + (K_2 - i)\beta_1 - \eta_2(ir_1 - c^2)) \int_x^\infty \mathrm{d}G^2_{ij}(y)} \right\},$$

and

$$\Psi^2_{\max}(i, j) = \sup_x \left\{ \frac{h^2_i \,\mathrm{e}^{-\eta_2(ir_1 - c^2)x} \int_x^\infty \mathrm{e}^{-\eta_2(ir_1 - c^2)y}\, \mathrm{d}G^2_{ij}(y)}{p^2_i/(i\alpha_1 + (K_2 - i)\beta_1 - \eta_2(ir_1 - c^2)) \int_x^\infty \mathrm{d}G^2_{ij}(y)} \right\}.$$

Acknowledgements

The authors thank the associate editor and the referee for suggestions that led to considerable improvements in the content and presentation of this paper.

References

[1] D. Anick, D. Mitra and M.M. Sondhi, Stochastic theory of a data handling system with multiple sources, Bell System Tech. J. 61 (1982) 1871–1894.

[2] D. Artiges and P. Nain, Upper and lower bounds for the multiplexing of multiclass Markovian on/off sources, Performance Evaluation 27/28 (1996) 673–698.

[3] C.S. Chang and J. Cheng, Computable exponential bounds for intree networks with routing, in: *Proc. of IEEE INFOCOM*, 1995, pp. 197–204.

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

[5] C.F. Daganzo, Some properties of polling systems, Queueing Systems 6 (1990) 137–154.

[6] N.G. Duffield, Exponential bounds for queues with Markovian arrivals, Queueing Systems 17 (1994) 413–430.

[7] 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 J. Selected Areas Commun. 13(6) (1995) 1004–1016.

[8] A.I. Elwalid and D. Mitra, Analysis and design of rate-based congestion control of high speed networks, part I: Stochastic fluid models, access regulation, Queueing Systems 9 (1991) 29–64.

[9] 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) (1993) 329–343.

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

[11] N. Gautam, V.G. Kulkarni, Z. Palmowski and T. Rolski, Bounds for fluid models driven by semi-Markov inputs, Probab. Engrg. Inform. Sci. 13 (1999) 429–475.

[12] N. Gautam, Quality of service for multi-class traffic in high-speed networks, Doctoral dissertation, Department of Operations Research, University of North Carolina, Chapel Hill, NC, 1997.

[13] R.J. Gibbens and P.J. Hunt, Effective bandwidths for the multi-type UAS Channel, Queueing Systems 9 (1991) 17–28.

[14] G. Kesidis, *ATM Network Performance* (Kluwer Academic, Dordrecht, 1996).

[15] G. Kesidis, J. Walrand and C.S. Chang, Effective bandwidths for multiclass Markov fluids and other ATM sources, IEEE/ACM Trans. Networking 1(4) (1993) 424–428.

[16] J.F.C. Kingman, Inequalities in the theory of queues, Cambridge Phil. Soc. 59 (1964) 359–361.

[17] V.G. Kulkami, Effective bandwidths for Markov regenerative sources, Queueing Systems 24 (1997) 137–153.

[18] V.G. Kulkarni and N. Gautam, Admission control of multi-class traffic with service priorities in high-speed networks, Queueing Systems 27 (1997) 79–97.

[19] Z. Liu, P. Nain and D. Towsley, Exponential bounds with applications to call admission, J. Assoc. Comput. Mach. 44 (1997) 366–394.

[20] A. Narayanan and V.G. Kulkarni, First passage times in fluid models with an application to two-priority fluid systems, in: *Proc. of the IEEE Internat. Computer Performance and Dependability Symposium*, 1996.

[21] Z. Palmowski and T. Rolski, A note on martingale inequalities for fluid models, Statist. Probab. Lett. 31(1) (1996) 13–21.

[22] S.M. Ross, Bounds on the delay distribution in $G/G/1$ queues, J. Appl. Probab. 11 (1974) 417–421.

[23] H. Takagi, *Analysis of Polling Systems* (MIT Press, Cambridge, MA, 1986).

[24] W. Whitt, Tail probabilities with statistical multiplexing and effective bandwidths for multiclass queues, Telecommun. Systems 2 (1993) 71–107.

[25] J. Zhang, Performance study of Markov-modulated fluid flow models with priority traffic, in: *INFOCOM'93*, 1993, pp. 10–17.