Effective Bandwidths with Priorities

Arthur W. Berger and Ward Whitt

IEEE/ACM Transactions on Networking 6 (1998) 447–460

***Abstract*— The notion of effective bandwidths has provided a useful practical framework for connection admission control and capacity planning in high-speed communication networks. The associated admissible set with a single linear boundary makes it possible to apply stochastic-loss-network (generalized-Erlang) models for capacity planning. In this paper we consider the case of network nodes that use a priority-service discipline to support multiple classes of service, and we wish to determine an appropriate notion of effective bandwidths. Just as was done previously for the first-in first-out discipline, we use large-buffer asymptotics (large deviations principles) for workload tail probabilities as a theoretical basis. We let each priority class have its own buffer and its own constraint on the probability of buffer overflow. Unfortunately, however, this leads to a constraint for each priority class. Moreover, the large-buffer asymptotic theory with priority classes does *not* produce an admissible set with linear boundaries, but we show that it nearly does and that a natural bound on the admissible set does have this property. We propose it as an approximation for priority classes. Then there is one linear constraint for each priority class. This linear-admissible-set structure implies a new notion of effective bandwidths, where a given connection is associated with *multiple* effective bandwidths: *one for the priority level of the given connection and one for each lower priority level.* This structure can be used regardless of whether the individual effective bandwidths are determined by large-buffer asymptotics or by some other method.**

1 Introduction and Summary

The desire to provide different quality-of-service (QoS) guarantees to different classes of customers using emerging communication networks is leading to the use of priorities in the allocation of network resources. In particular, asynchronous-transfer-mode (ATM) switches are being built with the capability of supporting multiple priority classes. Also, priority queueing can be used in internet-protocol (IP) routers to support real-time services along with best-effort service. Thus, it is natural to consider admission-control and dimensioning procedures that take account of the priority structure. In this paper we discuss extensions of the effective-bandwidth concept when there are priority classes, allowing any number of priority classes. For accounts of previous work on effective (or equivalent) bandwidth (or capacity) with the first-in first-out (FIFO) discipline, see Chang and Thomas [8], de Veciana, Kesidis and Walrand [17], Guerin, Ahmadi and Naghshineh [23], Kelly [26], Whitt [36] and references therein.

Arthur Berger is with AT&T Labs, 101 Crawfords Corner Rd., Holmdel, NJ 07733-3030; awberger@att.com. Ward Whitt is with AT&T Labs, 180 Park Avenue, Building 103, Florham Park, NJ 07932-0971; wow@research.att.com

Researchers have begun to examine the impact of non-FIFO queueing on bandwidth allocation and admission control in high-speed networks. First, Elwalid and Mitra [18] analyzed a loss-priority model where each ATM connection has some cells designated high priority and others designated low priority, and all cells are buffered in a single FIFO queue, with lower-priority cells being discarded when the queue length exceeds a threshold. A generalization of this model with two or more loss-priorities per connection was analyzed by Kulkarni, Gun and Chimento [28].

Zhang [37] and Elwalid and Mitra [20] considered models with Markov-modulated rate process (MMRP) sources, each belonging to a priority class with its own buffer. In particular, Zhang [37] considered an MMRP model where the state of the underlying Markov chain determines the rate of two or more MMRP sources. Zhang found the exact solution for the joint distribution of the amount of fluid in each queue. Elwalid and Mitra [20] provided an approximate solution to the important special case of Zhang's model in which independent Markov chains determine the arrival rates to two delay-priority queues. With their approximation, they could apply their previous MMRP algorithms to calculate the (approximate) admissible set.

Here we focus on developing an appropriate notion of effective bandwidths for the same model (where each priority class has its own queue and buffer), allowing more general sources (not necessarily MMRPs). Our main conclusion is that the notion of effective bandwidths needs to be modified. With priorities, the admissible set should be determined by a constraint for each priority class, because there is a separate (typically quite different) performance constraint for each priority class. Under appropriate assumptions or approximations, these constraints can be regarded as linear. Then there is one linear constraint for each priority class, which implies a new notion of effective bandwidths: A given connection should have *multiple* effective bandwidths, one for the priority level of the given connection and one for each lower level.

The linear-admissible-set structure is important for applying stochastic-loss-network models to do capacity planning. For capacity planning, it is natural to specify the blocking probabilities for each type of connection request instead of the resource capacities, and solve the loss network model to determine the required network capacities. An example of the loss-network-model approach to capacity planning is the design tool developed for the FIFO discipline by Mitra, Morrison and Ramakrishnan [30]. Our analysis provides a basis for extending such tools to priority classes.

With priorities, the loss network model can be solved using numerical transform inversion, as in Choudhury, Leung and Whitt [10], [11], [12]. The loss network model can be solved repeatedly using a search procedure as in [12] to determine appropriate resource capacities. As indicated in [11], [12], it is also possible to consider alternative sharing schemes besides the complete sharing. Upper limit and guaranteed minimum constraints are proposed as a way to provide multiple grades of service and protect one type of source from overloads from other types of sources. These constraints are appealing because, with them, it is still possible to calculate the blocking probabilities using the numerical transform inversion algorithm [11], [12].

Our contribution is to show how an admissible set with a linear constraint for each priority class can be derived, and to point out the implications for effective bandwidths. To do so, we start by applying large deviations theory to derive the exact admissible set associated with priority classes using large-buffer asymptotics. However, that is only the first step, because unlike for FIFO queues the exact large-buffer asymptotic admissible set with priorities does *not* have linear boundaries. We show that natural approximations and bounds for that admissible set do have linear boundaries. We also examine numerical examples to see how these approximations perform.

We establish the exact large-buffer-asymptotic admissible set for priorities in a companion paper [4]. In support of that work, we have derived new large deviation principles (LDPs) for departure processes from a FIFO queue in [32], extending earlier work by de Veciana, Courcoubetis and Walrand [15], Chang [7], and Chang and Zajic [9]. Those authors found that an LDP for the departure process could be obtained from a stronger sample-path LDP for the input process. However, the sample-path LDP for input processes used applies only to discrete-time processes whose increments have moment generating functions that are finite everywhere, which naturally occurs if the input increments are bounded. This requirement is not too restrictive from an engineering perspective, but it is from the modeling perspective, because many natural models do not have this property. We obtain a more general LDP for departure processes in [32] by requiring that the input process satisfy a sample-path LDP in the function space D of right-continuous real-valued functions with left limits, with an appropriate non-uniform topology, which allows the rate functions to be finite on sample paths with jumps. This extension is needed even to establish an LDP for the departure process from the elementary M/D/1 and M/M/1 fluid queues. The new work shows that the same large-deviation behavior for departure processes originally established in [15] holds more generally. (See (2.12) below.) We show the application to the low-priority workload in Section 2.

Important related work includes LDPs for priorities established by Kulkarni and Gautam [27] and Zhang [38] in papers that appeared after this paper and [4] were submitted. Kulkarni and Gautam [27] obtain the same exact asymptotic admissible set obtained here, but extra conditions as provided in [32] are needed in their supporting LDP for departure processes. Zhang [38] establishes an LDP for the two-queue generalized processor sharing (GPS) discipline, which contains the two-priority model as a special case. Zhang's LDP is based on the same sample-path LDP for departure processes used in [15], [7], [9], and so it too does not apply to the examples considered here. Additional LDPs for GPS are contained in Paschalidis [31] and Bertsimas, Paschalidis and Tsitsiklis [6].

We consider two priority classes, because the key points can be made with just two. However, the results extend immediately to any number of priority classes, because for any priority class under consideration, all higher classes can be lumped together and all lower classes can be ignored. Since the large-buffer asymptotics with priorities typically produces an admissible set with nonlinear boundaries, we develop two approximations for the low-priority steady-state workload. These approximations cause the nonlinear admissible set (i.e., with nonlinear boundaries) to be replaced by a linear admissible set (i.e., with linear boundaries), thereby making it possible to define the new notion of effective bandwidths. The two approximations produce upper and lower bounds on the admissible set. The lower bound is appealing because it is conservative and because it tends to be close to the admissible set based on large-buffer asymptotics. The two approximations both reduce the low-priority steady-state workload to an appropriate FIFO workload. Since the two approximations are general, they also can be used to approximate the full low-priority workload probability distribution and to produce other notions of effective bandwidths besides the notion based on large-buffer asymptotics, e.g. see Kelly [26].

Others have also proposed our lower-bound approximating admissible set with linear boundaries. First, a lower bound for the GPS large-buffer-asymptotic admissible set established by de Veciana and Kesidis [16] reduces to our proposed admissible set when the GPS discipline is specialized to the two-priority case. Kulkarni and Gautam [27] also introduce this approximating admissible set. However, none of these other papers on large-buffer-asymptotics for priorities explore the implications for effective bandwidths. Also none discuss other methods besides large-buffer asymptotics.

Unfortunately, the effective-bandwidth approach based completely on large-buffer asymptotics is often not a very accurate approximation; e.g., see [13]. Moreover, it can be difficult to find tractable source traffic models that accurately fit traffic data. Nevertheless, we believe that our results can be very useful because they identify an appropriate structure for the admissible set. Once we decide to use an admissible set with linear boundaries, which corresponds to multiple effective bandwidths for each priority class, the actual effective bandwidths used can be defined in various ways. We illustrate by specifying alternative measurement-based procedures for obtaining appropriate effective bandwidths, which are based on the linear structure, and which are easy for practitioners to apply. We provide an informal development of effective bandwidths with priorities not based on large-buffer asymptotics in [5]. Related brief informal discussion appears on p. 609 of Ahmadi et al. [3] and in Section 3.7 of Kelly [26]. Additional support for our conclusions is provided by Elwalid and Mitra [20], who found for the case of two priorities that the approximate admissible sets they calculated often had this linear structure.

Here is how the rest of this paper is organized. In Section 2 we develop the large-buffer asymptotics for priorities and present the two bounds on the admissible set, referring to [4] and [32] for technical details. In Section 3 we give the asymptotic-decay-rate functions for standard input processes. There we relate the MMRP to conventional queueing input processes. Choudhury, Mandelbaum, Reiman and Whitt [14] showed that any MMRP can be represented as a limit of Markov-modulated Poisson processes (MMPPs); here we show that this limit applies to the asymptotic-decay-rate functions.

Sections 4 and 5 and are devoted to numerical examples evaluating the performance of the effective-bandwidth approach with priorities, using the large-buffer asymptotics to generate the effective bandwidths. In Section 6 we present alternative ways to generate effective bandwidths, not based on large-buffer asymptotics, that exploit the linear-admissible-set structure. In Section 7 we give an example illustrating how the priority structure might be applied in practice. Finally, in Section 8 we draw conclusions based on both the theory and our numerical examples.

2 Large-Buffer Asymptotics

With two priority classes, let class 1 be the high priority class. Our model has input from multiple sources from each of the two priority classes. Let $A_{ij}(t)$ denote the input of work in the interval $[0,t]$ from a type-j source of priority i. Let there be J_i source types of priority i. We let $\{A_{ij}(t) : t \geq 0\}$ be a general stationary process. It could be an MMRP, but it could also be more general. We assume that the processes $A_{ij}(t)$ are mutually independent. We assume that work is processed continuously at a constant rate c whenever work is present. As in previous work on effective bandwidths with the FIFO discipline, e.g., [36], in our mathematical analysis we assume that there is an infinite buffer. The tail probability thus serves as an approximation for the overflow probability.

The notion of effective bandwidth is based on performance criteria on the tail probability of the priority-i steady-state workload V_i for $i = 1, 2$, namely,

$$Prob(V_i > b_i) \leq p_i \quad i = 1, 2 \ , \tag{2.1}$$

and on the exponential approximation:

$$Prob(V_i > b_i) \approx e^{-\eta_i b_i} \quad i = 1, 2 \ , \tag{2.2}$$

which is asymptotically correct as $b_i \to \infty$ under regularity conditions, i.e., the large-buffer asymptotics. (We refer to [4], [36] for more details, including technical conditions needed for the results below.)

Note that, with priorities, the steady-state low-priority delay is different from the steady-state low-priority workload. The low-priority delay can be much bigger than the low-priority workload, because it may be necessary for low-priority work to wait for high-priority work that arrives after the low-priority work. We use workload here to focus on loss criteria. We intend to treat delay criteria in subsequent work.

The effective bandwidth of a type-j source at priority-i is defined to be

$$e_{ij} = \psi_{A_{ij}}(\eta_i^*)/\eta_i^* \ , \quad \text{where } \eta_i^* = -(\log p_i)/b_i \ , \tag{2.3}$$

and $\psi_{A_{ij}}(\theta)$ is the asymptotic-decay-rate function for a type-j source at priority i,

$$\psi_{A_{ij}}(\theta) = \lim_{t\to\infty} t^{-1} \log E e^{\theta A_{ij}(t)} \tag{2.4}$$

and $A_{ij}(t)$ is the input of work of a type-j source at priority-i during the interval $[0,t]$.

Assuming that the system starts with initial workload $V_i(0)$ for class i at time 0, the *workload* for class i at time t can be defined by

$$V_i(t) = V_i(0) + X_i(t) - \inf_{0\leq s\leq t}\{\min\{0, V_i(0) + X_i(s)\}\} \ , \quad t \geq 0 \ , \tag{2.5}$$

where

$$\begin{aligned}
A_i(t) &= A_{i1}(t) + \ldots + A_{iJ_i}(t) \ , && (2.6)\\
X_i(t) &= A_i(t) - S_i(t) \ , && (2.7)\\
S_1(t) &= ct \ , && (2.8)\\
S_2(t) &= ct - D_1(t) \ , && (2.9)\\
D_i(t) &= A_i(t) + V_i(0) - V_i(t) \ , && (2.10)
\end{aligned}$$

with $A_i(0) = 0$ for all i. The processes $\{S_i(t) : t \geq 0\}$ in (2.8) and (2.9) are the *server-availability processes*; i.e., $S_i(t)$ is the total potential processing that can be done for class i in the interval $[0,t]$. The maximum server processing rate is the capacity or available bandwidth c. Clearly, (2.8) holds for the high priority class. The processes $\{D_i(t) : t \geq 0\}$ are the *departure (output) processes*; i.e., the output in completed work during the interval $[0,t]$. The output $D_i(t)$ is clearly the input over $[0,t]$, plus the initial work, minus what is present at time t, as indicated in (2.10). For $i = 2$, the server-availability process can clearly be defined in terms of the departure process of the high priority class by (2.9). Finally, the process $\{X_i(t) : t \geq 0\}$ in (2.7) is the *net input process* for class i, in terms of which the workload process is defined by the usual one-dimensional *reflection map* in (2.5).

In this context, the tail-probability asymptotic decay rate η_i in (2.2) is determined as the root θ of the equation

$$\psi_{A_i}(\theta_i) + \psi_{S_i}(-\theta_i) = 0 \ , \tag{2.11}$$

where ψ_{A_i} and ψ_{S_i} are defined as in (2.4); see Theorem 10 of [36]. Under extra regularity conditions, the high-priority departure-process asymptotic decay rate function is

$$\psi_{D_1}(\theta) = \begin{cases} \psi_{A_1}(\theta) \ , & \theta < \hat{\theta} \\ \psi_{A_1}(\hat{\theta}) + c(\theta - \hat{\theta}) \ , & \theta \geq \hat{\theta} \ , \end{cases} \tag{2.12}$$

where $\hat{\theta}$ is determined by the equation $\psi'_{A_1}(\hat{\theta}) = c$; see [15], [7], [9], and [32].

Using the effective-bandwidth approximation, the set of connections that satisfy the performance criteria (2.1), called the admissible set, is the set of n_{ij} for which

$$\sum_{j=1}^{J_i} e_{ij} n_{ij} \leq C_i \ , \quad i = 1, 2 \ , \tag{2.13}$$

where n_{ij} is the number of type-j priority-i sources and C_i is the "effective capacity," which depends on the server availability process for priority-i. For priority 1, the effective capacity C_1 is simply the speed of the server, c, but for priority 2 the effective capacity depends on the departure process of the aggregate of priority-1 connections.

For the desired simplicity in network planning models, we make further approximations so that the priority-2 constraint in (2.13) is a linear combination of the priority-1 and priority-2 sources. A natural choice would be an admissible set of the form:

$$\sum_{j=1}^{J_1} e_{1j} n_{1j} \leq c \tag{2.14}$$

$$\sum_{j=1}^{J_1} e_{1j} n_{1j} + \sum_{j=1}^{J_2} e_{2j} n_{2j} \leq c \ . \tag{2.15}$$

However we claim that (2.15) is needlessly conservative and significant improvement is obtained if (2.15) is replaced with

$$\sum_{j=1}^{J_1} e^2_{1j} n_{1j} + \sum_{j=1}^{J_2} e_{2j} n_{2j} \leq c \tag{2.16}$$

where e^2_{1j} is the effective bandwidth of a type-j priority-i source as seen by priority 2;

$$e^2_{1j} = \psi_{A_{1j}}(\eta^*_2)/\eta^*_2 \ . \tag{2.17}$$

Equations (2.14), (2.16) and (2.17) immediately generalize to an arbitrary number I of priorities:

$$\sum_{i=1}^{k} \sum_{j=1}^{J_i} e^k_{ij} n_{ij} \leq c \quad \text{for} \quad k = 1, \ldots, I \ , \tag{2.18}$$

where

$$e^k_{ij} = \psi_{A_{ij}}(\eta^*_k)/\eta^*_k \ . \tag{2.19}$$

To obtain the desired linear structure in (2.16), two approximations are suggested: The *empty-buffer approximation* provides a lower bound on the true admissible set, while the *reduced-service-rate* (RSR) *approximation* provides an upper bound. (The bounding properties are intuitive; see [4] for proofs.) The empty-buffer approximation for priority class 2 makes the simplifying assumption that the amount of priority-1 work that queues in the buffer is negligible and thus the priority-1 departure process is approximated by its arrival process. In particular, the empty buffer approximation makes $S_2(t) \approx ct - A_1(t)$, so that $X_2(t) \approx A_1(t) + A_2(t) - ct$ and $V_2(t)$ is approximated by the total workload $V_1(t) + V_2(t)$. In terms of asymptotic-decay-rate functions, the empty-buffer approximation implies that $\psi_{D_1}(\theta) \approx \psi_{A_1}(\theta)$ in (2.12), which is correct if $\theta \leq \hat{\theta}$ (which is often the case), and $\psi_{S_2}(\theta) \approx c\theta - \psi_{A_1}(\theta)$. Applying the effective bandwidth approximation to the empty-buffer system yields the empty-buffer effective-bandwidth (EBEB) admissible set (2.14), (2.16) and (2.17). As mentioned earlier, the EBEB approximation is also developed as an approximation to the exact large-buffer-asymptotic admissible set by Kulkarni and Gautam [27]; see their (17) on p. 87. The lower bound on the GPS large-buffer-asymptotic admissible set derived by de Veciana and Kesidis [16] when specialized to priorities also reduces to the EBEB approximation. Note, however, that our empty-buffer bound applies much more generally (beyond large-buffer asymptotics).

The RSR approximation assumes that the server is continuously available to the lower priority class but at a reduced rate, where the reduction is the long-run average usage of the high priority class. In particular, the RSR approximation makes $S_2(t) \approx (1-\rho_1)ct$ and $\psi_{S_2}(\theta) \approx (1-\rho_1)c\theta$. Applying the effective bandwidth approximation to the RSR system, yields the RSR-effective-bandwidth admissible set (2.14) and (2.16) with

$$e^2_{1j} = c\rho_{1j}, \tag{2.20}$$

where ρ_{1j} is the occupancy of a type-j priority-1 source.

Although both (2.17) and (2.20) are useful to bound the admissible set, for a single estimate of e^2_{1j} we recommend (2.17), because the empty buffer approximation yields the exact effective capacity C_2 (2.13) for a relevant range of η^*_2, and is conservative, while the RSR approximation is not.

3 Computing Decay-Rate Functions

In [36] and other papers, explicit formulas are given for arrival-process asymptotic-decay-rate functions for many specific models, which can be used to generate effective bandwidths based on large-buffer asymptotics, as in (2.3) and (2.17). For example, a very general arrival counting process for queueing models is the *Markovian arrival process* (MAP); e.g., see Lucantoni [29]. For a MAP, the arrival-process asymptotic-decay-rate function is

$$\psi_A(\theta) = pf(D_0 + D_1 e^{\theta}) \ , \tag{3.1}$$

where D_0 and D_1 are square submatrices of the specially structured MAP infinitesimal generator matrix and pf is the Perron-Frobenius eigenvalue; see Section 6 of [36]. The Perron-Frobenius eigenvalue of a square matrix A with this structure is a real eigenvalue greater than the real part of any other eigenvalue; see Theorem 2.6, p. 46, of Seneta [33]. The Perron-Frobenius eigenvalue of the matrix A can be calculated by solving the characteristic equation $det(zI - A) = 0$ and finding the root with maximum real part.

A *Markov modulated Poisson process* (MMPP) is the special cases of a MAP in which $D_0 = M - \Lambda$ and $D_1 = \Lambda$, where M is the infinitesimal generator of the Markovian environment process and Λ is the associated diagonal matrix of Poisson arrival rates in the environment states. Hence, for an MMPP characterized by the pair (M, Λ), (3.1) becomes

$$\psi_A(\theta) = pf(M - \Lambda + \Lambda e^{\theta}) \ . \tag{3.2}$$

Given that we are focusing on workloads, in a queueing model we should consider the total input in required service time. If we consider the total input stemming from an arrival counting process with asymptotic-decay-rate function $\psi_C(\theta)$ bringing i.i.d. service requirements, with a generic service requirement Z, where

$$\psi_Z(\theta) = \log E e^{\theta Z} \ , \tag{3.3}$$

then the overall arrival process is

$$A(t) = \sum_{i=1}^{C(t)} Z_i \ , \quad t \geq 0 \ , \tag{3.4}$$

and the overall arrival-process asymptotic-decay-rate function becomes

$$\psi_A(\theta) = \psi_C(\psi_Z(\theta)) \ ; \tag{3.5}$$

e.g., see Theorem 5 of [36].

Alternatively, it is natural to consider *Markov-modulated rate process* (MMRP) input, as in Elwalid and Mitra [19]. If the infinitesimal generator of the Markovian environment process is M and Λ is the associated diagonal matrix of deterministic arrival rates in the environment states, then the arrival-process asymptotic-decay-rate function is

$$\psi_A(\theta) = pf(M + \Lambda\theta) \ . \tag{3.6}$$

We now show how to relate the MMPP and MMRP decay rate functions in (3.2) and (3.6). Choudhury, Mandelbaum, Reiman and Whitt [14] showed that any MMRP can be represented as a limit of MMPPs. Consider a common Markovian environment process with infinitesimal generator M. Let Λ be the diagonal rate matrix for the MMRP, where all entries are positive (not representing flow out). Then the asymptotic-decay-rate function is (3.6). We can obtain equation (3.6) by considering a limit of MMPP inputs. For each $\epsilon > 0$, we define an MMPP arrival process model. For each ϵ, let the Markovian environment infinitesimal generator be M, the arrival rate matrix be the diagonal matrix Λ/ϵ, the service times be deterministic of size ϵ, and the processing rate be c. Then, by (3.2) and (3.5), the asymptotic-decay-rate function of the MMPP/D/1 input process is

$$\psi_{A,\epsilon}(\theta) = pf\left(M - \frac{\Lambda}{\epsilon} + \frac{\Lambda}{\epsilon} e^{\epsilon\theta}\right) \ . \tag{3.7}$$

If we expand the exponential, letting $e^{\epsilon\theta} = 1 + \epsilon\theta + o(\epsilon^2)$ in (3.7), then we see that (3.7) approaches (3.6) as $\epsilon \to 0$.

4 Examples with M/M/1 Inputs

In this section we consider examples in which the source arrival processes are batch Poisson processes. If $A(t)$ is a compound Poisson process with Poisson rate λ and component i.i.d. jumps having moment generating functions $m_J(\theta)$, then the asymptotic-decay-rate function is

$$\psi_A(\theta) = \lambda(m_J(\theta) - 1) \ . \tag{4.1}$$

If the jumps are size 1, then $m_j(\theta) = e^{\theta}$ and $\psi_A(\theta) = \lambda(e^{\theta} - 1)$. In this section we shall consider the special case in which the jumps are exponential with mean μ^{-1}, as in the M/M/1 workload process, Then $m_J(\theta) = \mu/(\mu - \theta)$ and $\psi_A(\theta) = \lambda\theta/(\mu - \theta)$.

Now suppose that all sources are M/M/1 workload sources with mean service times μ_i^{-1} for class i. Thus, the model is equivalent to the M/M/1 two-priority queue. The goal is to determine the feasible arrival rates λ_1 and λ_2 or, equivalently, the feasible offered loads ρ_1 and ρ_2 where $\rho_i = \lambda_i/(\mu_i c)$. The M/M/1 model nicely illustrates the results since many quantities of interest can be determined analytically. One can view the model as representing (not necessarily accurately) the traffic on an ATM network as Poisson arrivals of bursts of cells.

We compare six cases for the admissible set associated with the M/M/1 two-priority queue:

(1) the exact admissible set,
(2) the admissible set given the RSR approximation,
(3) the admissible set given the empty-buffer approximation,

and, paired with each of the above, the corresponding effective bandwidth approximations:

(4) the admissible set (2.13) with the exact calculation for the effective capacity of C_i,
(5) the admissible set (2.14), (2.16) and (2.20), based on the effective-bandwidth approximation and the RSR approximation,
(6) the admissible set (2.14), (2.16) and (2.17), based on the effective-bandwidth approximation and the empty-buffer approximation.

We computed the exact admissible set by numerically inverting the Laplace transform of the class-2 steady-state workload, using the transform inversion algorithm in [1]. For the M/G/1 priority queue, this transform is given in Kella [25]:

$$E[e^{-sV_2}] = \frac{\hat{z}_1(\hat{h}_2(s))(1 - \rho_1 - \rho_2)s}{\hat{h}_2(s)(s - \hat{z}_1(\hat{h}_2(s))} \ , \tag{4.2}$$

where V_i is the priority-i steady-state workload, $\hat{h}_2(s) = \lambda_2(1 - \hat{g}_2(s))$, and $\hat{g}_i(s)$ is the Laplace-Stieltjes transform of the priority-i service-time distribution (i.e., distribution of work added by a priority-i arrival), $\hat{z}_1(s) = s + \rho_1 - \rho_1\hat{b}_1(s)$, and $\hat{b}_1(s)$ is the Laplace-Stieltjes transform of the priority-1 busy-period distribution, which is given by the Kendall functional equation $\hat{b}_1(s) = \hat{g}_1(s + \rho_1 - \rho_1\hat{b}_1(s))$. Values of the transform $\hat{b}_1(s)$ can be calculated iteratively; see [2].

The corresponding effective-bandwidth approximation has a simple closed form when the batches are exponentially distributed. Then the asymptotic decay rate function for the priority-1 departure process $\psi_{D_1}(\theta)$ is

$$\psi_{D_1}(\theta) = \begin{cases} \dfrac{\lambda_1\theta}{\mu_1-\theta} , & \theta \le \hat{\theta} \\ \lambda_1(\rho_1^{-1/2}-1) + c(\theta-\hat{\theta}) , & \theta > \hat{\theta} , \end{cases} \tag{4.3}$$

where $\hat{\theta} = \mu_1(1-\sqrt{\rho_1})$, by (2.12). Thus the admissible set based on effective bandwidths is:

$$\frac{\rho_1}{1-\mu_1^{-1}\eta_1^*} \le 1 \tag{4.4}$$

$$\frac{\rho_2}{1-\mu_2^{-1}\eta_1^*} \le \begin{cases} 1 - \dfrac{\rho_1}{1-\mu_1^{-1}\eta_2^*} , & \rho_1 \le (1-\mu_1^{-1}\eta_2^*)^2 \\ \mu_1(1-\sqrt{\rho_1})^2/\eta_2^* , & \rho_1 > (1-\mu_1^{-1}\eta_2^*)^2 . \end{cases} \tag{4.5}$$

The empty-buffer approximation (2.16) and (2.17) will coincide with (4.5) when the maximum admissible ρ_1 is less than $(1-\mu_1^{-1}\eta_2^*)^2$, or equivalently $\eta_1^*/\eta_2^* > 2-\mu_1^{-1}\eta_2^*$, which holds when η_1^* is at least twice η_2^*. This typically would be the case, as for example when the priority-2 threshold b_2 is at least twice priority-1's, b_1, and $p_1 \le p_2$.

With the RSR approximation (and M/M/1 input), the class-2 steady-state workload, V_2^a, is distributed as the workload in an associated standard M/M/1 queue, if we rescale time so that the processing rate is 1. For the high-priority class, the workload with parameters (λ_1, μ_1, c) is equivalent to the workload in a model with parameters $(\lambda/c, \mu_1, 1)$, which corresponds to a standard M/M/1 model. (We do not scale the service requirements, which are not time but work.) Thus,

$$P(V_1^a > t) = P(V_1 > t) = \rho_1 e^{-(\mu_1-(\lambda_1/c))t} . \tag{4.6}$$

For the low-priority class, the RSR approximation yields M/M/1 input with parameters $(\lambda_2, \mu_2, c(1-\rho_1))$. The steady-state workload is the same as for the parameter triple $(\lambda_2/c(1-\rho_1), \mu_2, 1)$. Thus,

$$P(V_2^a > t) = \frac{\rho_2}{1-\rho_1} e^{-(\mu_2-(\lambda_2/c(1-\rho_1))t} , \quad t \ge 0 . \tag{4.7}$$

Thus, the class-2 steady-state workload, V_2^1, for the RSR approximation with M/M/1 input, the admissible set is given by

$$\rho_1 \le \rho_1^{\max} \tag{4.8}$$

$$\rho_2 \le \rho_2^{\max}(1-\rho_1) \tag{4.9}$$

where $\rho_i^{\max}$ is the value of z that solves

$$z e^{-\mu_i(1-z)b_i} = p_i \quad i = 1, 2 \tag{4.10}$$

The effective bandwidths with M/M/1 input are

$$\psi_{A_i}(\eta_i^*)/\eta_i^* = \frac{\lambda_i}{\mu_i - \eta_i^*} \quad i = 1, 2 \tag{4.11}$$

and the RSR effective-bandwidth admissible set ((2.14), (2.16) and (2.20)) becomes

$$\frac{\lambda_1}{\mu_1-\eta_1^*} \le c \tag{4.12}$$

$$\frac{\lambda_2}{\mu_2-\eta_2^*} \le c(1-\rho_1) , \tag{4.13}$$

where $\eta_i^* < \mu_i$. Expressed in terms of ρ_1 and ρ_2, the admissible set is

$$\frac{\rho_1}{1-\mu_1^{-1}\eta_1^*} \le 1 \tag{4.14}$$

$$\rho_1 + \frac{\rho_2}{1-\mu_2^{-1}\eta_2^*} \le 1 . \tag{4.15}$$

Even when ρ_1 equals the maximum admissible value, $\rho_1^{\max}$ in (4.8) or $1-\mu_1^{-1}\eta_1^*$ in (4.14), the maximum admissible value of ρ_2, in (4.9) or (4.15) respectively, is still positive. Hence the admissible set is convex.

In the general M/G/1 setting, with the empty-buffer approximation the priority-2 steady-state workload, V_2^e, is distributed as the waiting time in an M/G/1 FIFO queue with arrival rate $(\lambda_1+\lambda_2)/c$ and service distribution $(\lambda_1 G_1 + \lambda_2 G_2)/(\lambda_1+\lambda_2)$, where G_i is the batch-size distribution of priority i. When the distributions are exponential with the same mean, μ^{-1}, for both priorities, the aggregate workload is that in a model with M/M/1 input having parameters $(\lambda_1+\lambda_2, \mu, c)$. The workload is the same as for the parameter triple $((\lambda_1+\lambda_2)/c, \mu, 1)$ and

$$P(V_2^e > t) = \rho e^{-(\mu-(\lambda_1+\lambda_2)/c)t} , \quad t \ge 0 . \tag{4.16}$$

The rate here, $\mu-(\lambda_1+\lambda_2)/c = \mu(1-\rho_1-\rho_2)$ is smaller than the RSR rate $\mu - \lambda_2/c(1-\rho_1) = \mu(1-\rho_2/(1-\rho_1)])$ in (4.7). The admissible set is

$$\rho_1 \le \rho_1^{\max} \tag{4.17}$$

$$\rho_2 \le \rho_2^{\max} - \rho_1 \tag{4.18}$$

where $\rho_1^{\max}$ and $\rho_2^{\max}$ are the same as in the RSR approximation, (4.10). If $b_2 \ge b_1$ and $p_2 \ge p_1$, which is the meaningful case, then $\rho_2^{\max} \ge \rho_1^{\max}$ and the admissible set is convex.

Note that the empty-buffer admissible set, (4.17) and (4.18), is a subset of the RSR admissible set (4.8) and (4.9) and that they approach each other as $\rho_2^{\max} \to 1$. The upper limit $\rho_2^{\max}$ is close to one when the priority-2 performance criterion is loose, i.e. b_2 and p_2 are relatively large and hence η_2^* is small.

If the two batch-size distributions are exponential with different means, μ_1^{-1} and μ_2^{-1}, then the aggregate input is an $M/H_2/1$ queue. The effective bandwidths are given by (4.11) and the effective bandwidth of priority 1 as seen by priority 2, $\psi_{A_1}(\eta_2^*)/\eta_2^*$, is $\lambda_1/(\mu_1-\eta_2^*)$. The admissible set ((2.14), (2.16) and (2.17)) becomes

$$\frac{\lambda_1}{\mu_1-\eta_1^*} \le c \tag{4.19}$$

$$\frac{\lambda_1}{\mu_1-\eta_2^*} + \frac{\lambda_2}{\mu_2-\eta_2^*} \le c , \tag{4.20}$$

where $\eta_1^* < \mu_1$ and $\eta_2^* < \min\{\mu_1, \mu_2\}$. In terms of ρ_1 and ρ_2, the admissible set is:

$$\frac{\rho_1}{1 - \mu_1^{-1}\eta_1^*} \leq 1 \tag{4.21}$$

$$\frac{\rho_1}{1 - \mu_1^{-1}\eta_2^*} + \frac{\rho_2}{1 - \mu_2^{-1}\eta_2^*} \leq 1 \ . \tag{4.22}$$

Note that criterion (4.21) is more stringent than $\rho_1 < 1$, while criterion (4.22) is more stringent than $\rho_1 + \rho_2 < 1$. Also note that the criterion in (4.22) is more stringent than the low-priority constraint with the RSR effective-bandwidth approximation in (4.15),

When $\lambda_1 = c(\mu_1 - \eta_1^*)$ and thus priority 1 is at its maximum admissible load, (4.19), then (4.20) becomes

$$\lambda_2 \leq \left(\frac{\mu_2 - \eta_2^*}{\mu_1 - \eta_2^*}\right)(\eta_1^* - \eta_2^*) \ . \tag{4.23}$$

Provided that $\eta_1^* > \eta_2^*$, which is the usual case, we can always admit some priority-2 work after priority 1 has reached its capacity. Consequently the admissible set is convex.

However, if $\eta_1^* < \eta_2^*$, then the admissible set is *not* convex. If any priority-2 input is present, then the priority-2 constraint is binding, but if no priority-2 input is present, then only the priority-1 constraint is relevant, and more priority-1 input is allowed.

For a numerical example, consider the case where both classes have Poisson arrivals and exponential batch sizes with mean 1. The performance criterion parameters are: $b_1 = 20$, $b_2 = 200$ and $p_1 = p_2 = 10^{-6}$. Hence η_1^* is 0.691 and is ten times bigger than η_2^*. Here $\rho_1^{\max}$ in (4.10) is 0.36, which is 16% bigger than the maximum admissible ρ_1 given the effective-bandwidth approximation: $1 - \mu_1^{-1}\eta_1^* = 0.31$. This illustrates the well-known conservatism that occurs with effective bandwidths, apart from any priorities [13]. However, for class 2 (which has a loose performance criterion), the maximum admissible ρ_2 given the effective-bandwidth approximation matches $\rho_2^{\max}$ to three significant figures, 0.931. Table 1 shows some sample points on the frontier of the six admissible sets. The six cases listed earlier are presented in the order: 6, 4, 5, 3, 1, 2. In Table 1, emp. stands for empty buffer; emp with (without) effective bandwidths corresponds to equations (4.21) and (4.22) ((4.17) and (4.18)). The exact values with (without) effective bandwidths are based on (4.4) and (4.5) ((4.2) with numerical inversion). The RSR values with (without) effective bandwidths are based on (4.14) and (4.15) ((4.8) and (4.9)).

Since η_2^* is less than $\hat{\theta}$ in (4.3) even for ρ_1 equal to $\rho_1^{\max}$, the empty-buffer-effective-bandwidth and the effective bandwidth admissible sets coincide. Although the empty-buffer admissible set does not equal the exact admissible set, they coincide to three significant figures over the range of feasible ρ_1 given the effective-bandwidth approximation for priority 1. However, if we had not used the effective bandwidth of priority 1 *as seen by priority 2*, $\psi_{A_1}(\eta_2^*)/\eta_2^*$, but rather had used the (unadjusted) priority-1 effective bandwidth $\psi_{A_1}(\eta_1^*)/\eta_1^*$ in (4.22), then when $\rho_1 =$ 0.30 we would have thought the maximum admissible ρ_2 would be 0.027 which is 23 times smaller than correct value of 0.631.

maximum admissible ρ_2

	eff. bndwths.			no eff. bndwths.		
ρ_1	emp.	exact	RSR	emp.	exact	RSR
.00	.931	.931	.931	.931	.931	.931
.05	.881	.881	.884	.881	.881	.885
.10	.831	.831	.838	.831	.831	.838
.15	.781	.781	.791	.781	.781	.792
.20	.731	.731	.745	.731	.731	.745
.25	.681	.681	.698	.681	.681	.698
.30	.631	.631	.652	.631	.631	.652
.35	–	–	–	.581	.581	.605

Table 1: M/M/1 example; performance criterion parameters: $b_1 = 20$, $b_2 = 200$, $p_1 = p_2 = 10^{-6}$.

Similar results occur for other parameter values, as long as $\eta_1^* \gg \eta_2^*$, which is the most meaningful case. The condition $\eta_1^* \gg \eta_2^*$ along with the feasibility condition on η_1^*, $\eta_1^* < \mu_1$, ensures that η_2^* is relatively small. This increases the region where the exact effective-bandwidth constraint for priority 2, (4.5), is linear in (ρ_1, ρ_2), and also helps to reduce the conservatism of the empty-buffer effective-bandwidth approximation for the priority-2 constraint (4.22). If η_1^* is also small, then the *priority-1* effective-bandwidth constraint is less conservative; i.e., the right-hand edge of the admissible set based on effective bandwidths is closer to the exact. For instance, suppose b_1 in the previous example is increased from 20 to 150 and thus η_1^* is decreased to $1.33\eta_2^* = 0.092$. Now the maximum admissible ρ_1 given the effective bandwidth approximation matches to three significant figures the true maximum, $\rho_1^{\max}$ in (4.10), which is now 0.908. Note that changing η_1^* changes e_1 and $\rho_1^{\max}$, but for a given feasible ρ_1, the maximum feasible ρ_2 does not change for the exact calculation nor for the approximations including the effective bandwidths.

5 Examples with MMPP Inputs

We now we consider more realistic traffic models. As in Section 4, we consider only two priority classes. Now at least one class has MMPP input. With this more complex input, we no longer calculate the exact admissible set, but we numerically calculate the exact admissible sets associated with the RSR and empty-buffer approximations. To illustrate a range of behavior, we present three examples. In the first example the RSR and empty-buffer approximations and their associated effective bandwidths all yield essentially the same admissible set. Since the approximations serve as upper and lower bounds for the true system, we thus indirectly calculate the exact admissible set. In the second example the RSR and empty-buffer approximations yield common admissible sets, but the associated effective-bandwidth approximations (though comparable to each other) are qualitatively more conservative, as can occur in non-priority FIFO queues [13]. The third example considers a larger number of connections. All three examples are inspired from ATM networks. An arrival represents a cell and adds one unit of work. The processing rate c equals 1 cell/cell-time.

For the first example, suppose that the priority-1 connections support a constant-bit-rate service and have equally spaced cells. Their superposition is conservatively modeled as a Poisson process. A priority-2 connection represents more bursty traffic and is the popular two-state MMPP where one state is ON while the other state is OFF, and hence the process is equivalent to an interrupted Poisson process (IPP). The MMPP has rate matrix

$$\Lambda = \begin{pmatrix} \lambda_1 & 0 \\ 0 & 0 \end{pmatrix}$$

and infinitesimal generator

$$M = \begin{pmatrix} -r_1 & r_1 \\ r_2 & -r_2 \end{pmatrix} .$$

The parameters λ_1, r_1, r_2 are determined as follows: Let the mean arrival rate of a class-2 connection be $\lambda_1 r_2/(r_1 + r_2) = 0.02$ cells/cell-time, let the fraction of time the connection is in the ON state be $r_2/(r_1 + r_2) = 0.1$, and let the mean number of arrivals during an ON period be $\lambda_1/r_1 = 20$, corresponding to roughly 1 Kbyte. Let the performance-criteria parameter values be $b_1 = 100$, $b_2 = 1,000$, $p_1 = 10^{-9}$ and $p_2 = 10^{-6}$. These performance parameters correspond to the ATM context where the priority-1 queue length should be kept relatively short to satisfy a quality of service commitment on cell delay variation and where the priority-2 connections have access to a non-negligible, "moderate," buffer space. For these parameters, $\eta_1^* = 15\eta_2^* = 0.2072$. The effective bandwidth of the priority-1 Poisson arrival process is:

$$e_1^1 \equiv \psi_{A_1}(\eta_1^*)/\eta_1^* = \rho_1(e^{\eta_1^*} - 1)/\eta_1^* = 1.1111\rho_1 , \tag{5.1}$$

which is only 10% bigger than the class-1 effective bandwidth as seen by priority 2, (2.14),

$$e_1^2 \equiv \psi_{A_1}(\eta_2^*)/\eta_2^* = 1.0069\rho_1 . \tag{5.2}$$

Thus, when priority 1 is at its maximum admissible value, there is relatively little spare capacity for priority-2 connections. From (3.2), the effective bandwidth of a priority-2 connection is

$$e_2 = \psi_{A_2}(\eta_2^*)/\eta_2^* = \left(-\alpha + \sqrt{\alpha^2 + 4\lambda_1 r_2(e^{\eta_2^*} - 1)}\right)/2\eta_2^* , \tag{5.3}$$

where $\alpha = r_1 + r_2 - \lambda_1(e^{\eta_2^*} - 1)$. For the given parameter values, $e_2 = 0.0258$, which is only 30% greater than the mean rate of 0.02. Note that the priority-2 performance criterion is relatively loose: the threshold b_2 is 50 times the mean burst size of 20 cells. This is a regime where effective bandwidths based on (2.2) are more likely to be accurate. The maximum admissible ρ_1 based on the effective bandwidth approximation (5.1) is 0.89996, which almost equals the exact value from the M/D/1 model of 0.90027. Likewise, the maximum admissible number of priority-2 connections, based on the effective-bandwidth approximation (5.3) is $\lfloor c/e_2 \rfloor = \lfloor 1/0.0258 \rfloor = 38$, whereas the exact value is 39, obtained from computing the workload distribution of the $\sum_i$ MMPP$_{\rm i}$/D/1 queue. As in [13], this computation is done by numerically inverting the Laplace-Stieltjes Transform (LST) of the virtual waiting time distribution of the MAP/G/1 queue, given by

$$\hat{w}(s) = s(1-\rho)g[sI + D_0 + D_1\hat{h}(s)]^{-1}e , \tag{5.4}$$

where D_0 and D_1 are the infinitesimal rate matrices of the MAP, $\hat{h}(s)$ is LST of the service-time distribution $H(t)$, e is a column vector of all 1's, ρ is traffic intensity (overall arrival rate times mean service time) and the row vector g is the solution to $gG = g$, $ge = 1$, where the matrix G is given by:

$$G = \int_0^\infty e^{(D_0 + D_1 G)x} dH(x) . \tag{5.5}$$

The priority-2 admissible-set constraint for the RSR effective bandwidth approximation, (2.16) with (2.20), is

$$\rho_1 + 0.0258 n_2 \le 1 , \tag{5.6}$$

which is almost the same as the corresponding empty-buffer effective-bandwidth constraint, (2.16) with (2.17),

$$1.0069\rho_1 + 0.0258 n_2 \le 1 . \tag{5.7}$$

The admissible sets for the (non-effective-bandwidth) RSR and empty-buffer approximations are computed numerically [13]. Table 2 shows the admissible sets of the four cases. The effective bandwidths are slightly conservative compared to the "exact" RSR and EB approximations, but the four regions are essentially the same. Since effective-bandwidth approximations do not always work so well, it is heartening to note that the present example does mirror a relevant case in ATM: We can think of the higher-priority queue carrying constant-bit-rate connections, whose superposition can be conservatively modeled by the relatively nice Poisson or batch-Poisson process, while the second-priority queue has buffer space on the order of 1,000 to 10,000 cells, allowing η_2^* to be relatively small.

In the second example, the connections of both priority classes are two-state ON/OFF MMPPs. A priority-1 connection has a mean arrival rate of 0.02 cells/cell-time, the fraction of time ON is 0.1, and the mean number of arrivals during an ON period is 10. For a priority-2 connection the parameters are respectively 0.04 cells/cell-time, 0.1, and 50. The performance criterion parameter values are $b_1 = 100$, $b_2 = 500$, $p_1 = 10^{-9}$, $p_2 = 10^{-6}$, and hence $\eta_1^* = 7.5\eta_2^* = 0.207$. The effective bandwidth of a priority-1 connection, e_1, is 0.133 and is about 5 times bigger than the effective bandwidth as seen by class 2, $e_1^2 = 0.026$. Thus if one were to use (2.15) for the admissible set, then one would significantly overestimate the class-1 effective bandwidth as seen by class 2 (in the priority-2 constraint).

pr.-1	max. no. priority 2			
	eff. bndwths.		no eff. bndwths.	
ρ_1	emp.	RSR	emp.	RSR
0.0	38	38	39	39
0.1	34	34	35	35
0.2	30	31	31	31
0.3	27	27	27	27
0.4	23	23	23	23
0.5	19	19	19	19
0.6	15	15	15	15
0.7	11	11	11	11
0.8	7	7	7	7
0.9	3	3	3	3

Table 2: First MMPP example: Priority-1 is Poisson, priority-2 connections are two-state MMPPs, $\eta_1^* = 15\eta_2^* = 0.207$.

no.	max. no. priority 2				
	eff. bndwths.		no eff. bndwths.		int.
pr.-1	emp.	RSR	emp.	RSR	eb
0	6	6	8	8	8
1	5	5	8	8	7
2	5	5	8	8	7
3	5	5	8	8	7
4	5	5	7	8	7
5	5	5	7	7	6
6	5	5	7	7	6
7	4	5	7	7	6
8	–	–	6	7	6
9	–	–	6	7	6
10	–	–	6	6	5
11	–	–	6	6	5
12	–	–	5	6	5
13	–	–	5	6	5

Table 3: Second MMPP example: Priority-1 and priority-2 connections are two-state MMPPs, $\eta_1^* = 7.5\eta_2^* = 0.207$.

According to the effective-bandwidth approximation, the maximum number of priority-1 connections is $\lfloor 1/e_1 \rfloor = 7$; while the true value is 13, obtained from numerically computing the workload distribution in the $\sum_i$ MMPP$_i$/D/1 queue. This is an example of the well-known phenomenon that in non-priority FIFO queues effective bandwidths based on (2.2) can be very conservative [13], particularly when the performance parameters (b_1, p_1) are relatively tight for the arrival processes. Our approximations for priorities are still subject to the accuracy of the effective bandwidths in FIFO queues which is evident on the axes of the admissible set where only one class is present. Herein we use effective bandwidths based on (2.2) to have the desired linearity property in (2.16). One could apply our approximations for priorities to alternative concepts of effective bandwidths; see Section 6.

Table 3 shows the admissible sets for this second example. (See the next paragraph for an explanation of the last column "int.eb," which stands for intercept effective bandwidth.) The main point of the table is that the approximations for the priority queueing have added little additional inaccuracy beyond what was inherent from effective bandwidths based on (2.2). That is, the exact admissible set, which has not been computed, is closely bounded by the RSR and empty-buffer approximations. And the admissible set based on effective bandwidths with the exact effective capacity (2.13) (also not computed) is closely bounded by the RSR and empty-buffer effective-bandwidth approximations. For the empty-buffer effective bandwidths, if we had not used the effective bandwidth of class 1 as seen by priority 2, e_1^2, but rather had kept the effective bandwidth as e_1 in the priority-2 constraint, then instead of the vector $(6, 5, 5, 5, 5, 5, 5, 4)$ of admissible priority-2 connections we would have obtained the needlessly conservative vector $(6, 5, 4, 3, 2, 2, 1, 0)$.

This second example also illustrates the usefulness of another approximation, given one wanted to avoid the conservatism of the effective bandwidths based on (2.2) but still retain the linearity of (2.16). Define the effective bandwidth to be the reciprocal of the number of admissible connections given each priority alone, including the case of priority-1 connections subject to the priority-2 performance criterion (the latter corresponds to the empty-buffer priority-2 constraint $Prob(V_1 + V_2 > b_2) < \rho_2$ where only priority-1 connections are present). For the present example $e_1 = 13^{-1}$, $e_2 = 8^{-1}$ and $e_1^2 = 39^{-1}$, where 39 is the maximum number of admissible priority-1 connections given the priority-2 performance criterion. We call this approximation "intercept effective bandwidths," and the last column of Table 3 shows the resulting admissible set. Although for this example the admissible set from the intercept effective bandwidth approximation is a subset of that from the empty-buffer approximation and thus is conservative relative to the true admissible set, this need not hold in general.

Note that the intercept-effective-bandwidth overcomes the conservatism of the effective-bandwidths based on (2.2) but at the expense of greater computational complexity. Unlike the RSR and empty-buffer effective bandwidths, the computational complexity of the intercept-effective bandwidths increases with the number of connections, but not as quickly as the (exact) empty-buffer approximation. For example, the number of states in the Markov process for the superposition of N_i homogeneous 2-state MMPPs, $i = 1, 2$, is $(N_1 + 1)(N_2 + 1)$, [13]. Thus, if one can numerically solve systems of 100 states, then the empty-buffer approximation can be used for systems with about 10 connections in each priority, whereas the intercept-effective-bandwidth approximation can be used in systems with a 100 connections since only one priority is considered at a time. In the special case where the per-connection arrival process of work (ATM cells) is the same for both priorities, the numerical computation of the empty-buffer approximation simplifies and the approximation becomes identical to the intercept-effective-bandwidth approximation.

In the third example we consider a larger number of connections. Suppose that the connections of both priorities are ON/OFF MMPPs with the same parameter values: mean rate 0.01, fraction of time ON 0.1, and mean burst size 20. Let $b_1 = 500$, $b_2 = 5,000$, and $p_1 = p_2 = 10^{-6}$. Here $\eta_1^* = 10^* \eta_2^* = 0.028$, and e_1 is 1.7 times the mean rate while e_1^2 and e_2 are only 5% greater than the mean rate. The admissible sets are given in Figure 1. The effective bandwidth approximations again give a low estimate for the number of admissible priority-1 connections, 57, whereas the correct value is 66. Otherwise the approximations are rather tight. Also shown is the case of effective bandwidths where FIFO service is used and all connections are subject to the class-1 criterion; here since the connections of each class are stochastically equivalent, the upper edge of the admissible set is a line with slope minus one. Last, Figure 1 shows the case where the unadjusted priority-1 effective bandwidth, e_1, is used in the priority-2 constraint, as in equation (2.15), as opposed to e_1^2, (2.16). Note that half of the potential gain (measured in terms of area of the admissible set) from using priorities instead of FIFO is not realized if equation (2.15) is used.

6 Alternative Approaches

The results in this paper can be used without invoking large-buffer asymptotics. First, any notion of effective bandwidth previously developed for the FIFO discipline can be directly applied together with the empty-buffer or RSR approximation, because those approximations reduce the problem to the FIFO case. The most important conclusion in this paper is the appropriate new structure for effective bandwidths with priorities, i.e., that there should be multiple effective bandwidths associated with a connection of a given priority class. We now specify two practical engineering methods for the FIFO discipline and show how they can be extended to priorities, exploiting the linearity of the admissible set.

6.1 A Boundary-Point Method

The first method is based on measurements at boundary points of the admissible set. In particular, suppose that the FIFO-service method is based on determining the maximum number of admissible connections of a given type when no other connection types are present, using measurements from a simulation, testbed network or an actual network. To determine e_{ij}^i, consider only priority-i type-j connections for one fixed j. Find the upper limit $\bar{n}_{ij}$ for each connection type alone to obtain parameter specification.

$$e_{ij}^i = c/\bar{n}_{ij} \ , \tag{6.1}$$

which corresponds to the constraint

$$e_{ij}^i n_{ij} \le c \ . \tag{6.2}$$

(In using (6.1) we ignore integrality constraints, i.e., the requirement that the number of connections must be some integer. Assuming that the capacity c is relatively large, this effect should be minor.)

So far we have determined the effective bandwidths e_{ij}^k for $k = i$. Now we determine e_{ij}^k for $k > i$. First fix i and k with $k > i$. We consider a feasible number of priority-i type-j connections established on the link, say n_{ij}^o. This number might be the maximum number admissible given the priority-i criterion, $\bar{n}_{ij}$, or it might be a lower value that corresponds to a designed engineering point. Given n_{ij}^o, we then see how many priority-k type-ℓ connections can be admitted for any fixed ℓ, considering the priority-k performance criterion. Suppose that this number is $m_{k\ell}^o$. We then let

$$e_{ij}^k = (c - e_{k\ell}^k m_{k\ell}^o)/n_{ij}^o \ . \tag{6.3}$$

Equation (6.3) corresponds to the constraint

$$e_{ij}^k n_{ij} + e_{k\ell}^k n_{k\ell} \le c \ . \tag{6.4}$$

In (6.4), we first determine a value for n_{ij}, n_{ij}^o. Then, with that value n_{ij}^o in place, we determine the upper limit on $n_{k\ell}$, $m_{k\ell}^o$. Since the inequality (6.4) should be an equality at the upper limit (again ignoring integrality problems) and since $e_{k\ell}^k$ has previously been determined, we can solve for the single missing parameter e_{ij}^k, obtaining the equation (6.3).

In the case where n_{ij}^o is chosen to be the maximum number admissible, $\bar{n}_{ij}$, then $m_{k\ell}^o$ is a natural measure of the benefit from using per-priority effective bandwidths, since $m_{k\ell}^o$ would be zero with effective bandwidths based on FIFO service. Moreover, when n_{ij}^o equals $\bar{n}_{ij}$, equation (6.3) can be expressed as:

$$e_{ij}^k = e_{ij}^i \left(1 - \frac{e_{k\ell}^k m_{k\ell}^o}{c}\right) \ . \tag{6.5}$$

In (6.5) e_{ij}^k equals e_{ij}^i times a factor that is between zero and one. The larger the value of $m_{k\ell}^o$, the smaller is the value of e_{ij}^k relative to e_{ij}^i. Thus, another measure of the benefit of per-priority effective bandwidths is how much smaller e_{ij}^k is relative to e_{ij}^i. In cases when e_{ij}^k is close to e_{ij}^i, the complexity of using distinct effective-bandwidths probably outweighs the potential efficiency gains.

From equations (6.1) and (6.3), we obtain all the effective-bandwidth parameters e_{ij}^k with $i \le k$. We have obtained these parameters by exploiting the linearity of the constraint set. Given this linearity, it suffices to consider only priority-i type-j connections when we determine the effective-bandwidth parameters e_{ij}^i via (6.1). Similarly, for $i < k$, it suffices to consider only priority-i type-j connections and priority-k type-ℓ connections for any ℓ when we determine the effective-bandwidth parameters e_{ij}^k via (6.3). A significant point is that we need consider only two connection types in this calculation. To determine e_{ij}^k, we consider priority-i type-j connections and priority-k type-ℓ connections for some (any) ℓ.

Since the linear admissible set is only an approximation, we might not actually want to fit the parameters by considering connections at their upper and lower limits. Instead, we might want to exploit knowledge of the typical operating region and determine a linear approximation to a more accurate admissible set by constructing a linear hyperplane tangent to the boundary for each priority class. This observation applies to the determination of both e_{ij}^i and e_{ij}^k for $k > i$. For example, the more accurate admissible set might be determined by simulation, perhaps using source traces, or by system measurements.

6.2 A Traffic-Descriptor Method

A second practical measurement-based approach can be based on a standardized traffic descriptor. Consider Variable-Bit-Rate (VBR) ATM connections for which the Sustainable-Cell-Rate (SCR) traffic descriptor is specified [21]. The SCR constitutes an upper bound on the mean rate of the connection. Suppose that in the FIFO context the effective bandwidth for these connections is chosen to be some factor times the connection's SCR. Thus, for this subsection let e_n represent the effective bandwidth of the n^{th} connection established on the link, and let SCR_n denote the SCR for this connection. Then

$$e_n = \alpha \cdot SCR_n \tag{6.6}$$

and, given that there are N connections established on the link, the FIFO admissible set would now have the form:

$$\sum_{n=1}^{N} e_n = \sum_{n=1}^{N} \alpha \cdot SCR_n \le c\ . \tag{6.7}$$

Note that in this case connections are not grouped in "types."

A theoretical determination of the parameter α in (6.6) and (6.7) would depend on many factors including additional characteristics of individual connections besides the SCR. However, a more heuristic approach could be based on historical measurements of realized connections and network performance. In fact, for some years now, various network operators have been using this approach for frame relay networks, wherein the Committed Information Rate (CIR) is analogous to the SCR. A conservative value for α might be picked initially, and then subsequently reduced as long as the performance commitment for the connections continues to be met. From a conservative, worst-case perspective, since the realized mean rate could be as big as the SCR (or CIR), α would need to be greater than one. However, in frame relay networks, measurements have shown the mean rate to be significantly less than the CIR, and values of α of $1/2$ or even $1/4$ have been used. Of course, the service provider should gather measurements on an ongoing basis to track changes in overall load and connection characteristics during network busy periods. For example, as frame relay networks include switched connections as well as semi-permanent ones, and as the applications using the frame relay networks begin to include voice and video, we expect that the value of α will need to be increased.

To extend the traffic-descriptor method to account for priorities, multiple factors α are determined. Again for this subsection, let e_{in}^k represent the effective bandwidth of the n^{th} connection established at priority i, as seen by priority k, where

$$e_{in}^k = \alpha_i^k \cdot SCR_n\ , \tag{6.8}$$

for chosen factors α_i^k, with $\alpha_i^{k+1} > \alpha_i^k$. Likewise, given N_i connections are established on the link at priority i, the admissible set has the form:

$$\sum_{i=1}^{k}\sum_{n=1}^{N_i} e_{in}^k = \sum_{i=1}^{k}\sum_{n=1}^{N_i} \alpha_i^k \cdot SCR_n \le c, \quad k = 1, \ldots, I\ . \tag{6.9}$$

It is more difficult to select the multiple correction factors α_i^k needed in (6.8) and (6.9) than it is to select the single correction factor α needed in (6.7). The large-buffer asymptotics and the EBEB approximation could be used to generate candidate relative values $\bar{\alpha}_i^k$; then we can set $\alpha_i^k = \beta\bar{\alpha}_i^k$ for a single parameter β, and adjust the single parameter β based on experience. It is important to recognize that more work needs to be done to develop a complete engineering solution, but we have identified an appropriate framework with (6.8) and (6.9).

7 Example Engineering Application

In this section we apply the general principles of the previous sections to a particular engineering example. Consider an ATM network that is being designed as an infrastructure to support the services of a network operator. Suppose the network nodes have four priorities. In the highest priority are placed the ATM connections that support circuit emulation. In the second priority are connections carrying Variable-Bit-Rate (VBR) speech, i.e., speech where silence has been eliminated. In the third priority are VBR non-real-time connections supporting Frame Relay, and in the lowest priority are Unspecified-Bit-Rate (UBR) connections, with no minimum rate guarantees, supporting best effort, elastic, data traffic. As part of the network design, suppose the network operator is using effective bandwidths for dimensioning link capacities via loss-network models. (When effective bandwidths are used for dimensioning, as opposed to connection admission control, rougher approximations are appropriate, consistent with the uncertainty of the future traffic demands.)

We wish to specialize equation (2.18) to the present example of four priorities, $I = 4$. For brevity in this section, the phrase "circuit-emulation connection" means an ATM connection that supports circuit emulation, and likewise for the other services. Since circuit-emulation connections encounter cell-scale congestion within the highest priority queue, one should use an effective bandwidth somewhat higher than the peak cell rate in order to meet a tight cell-delay-variation commitment. This then gives some potential to use a lower effective bandwidth for these connections in the lower priority constraints. However, as illustrated in the first example of Section 5, the potential gain is modest, and thus, for simplicity, we will use the same effective bandwidth in all of the priority constraints. As a consequence, the priority-1 constraint is subsumed within the other constraints and thus can be ignored. Also for simplicity, we use peak cell rate without a multiplicative factor as the effective bandwidth. We can group the circuit-emulation connections into types, based on the rate, which is appropriate if we wish to apply loss-network models. Thus, for priority 1, we have e_{1j}^k equal to the peak cell rate for connections of bandwidth category j, independent of $k = 1, \ldots, 4$.

For the VBR speech connections, we assume the traffic entering the operator's network is non-ATM, pulse-code-modulation 64Kbps circuits, and the operator has chosen an encoding algorithm that eliminates the silences, and packages the resulting bit stream into ATM cells. We can use a stochastic model for the characteristics of this VBR flow and estimate the number of such connections that can be supported for a given bandwidth; e.g., see Sriram and Whitt [35] and Heffes and Lucantoni [24]. Thus, we can use the boundary-point method of Section 6.1 to obtain effective bandwidths for these connections at priority 2. Note that here we have a dependency, that frequently occurs, where the effective bandwidth depends on the capacity which in turn is the object to be determined. Thus in principle, an iteration is necessary, though, for given a range of interest, the effective bandwidth for the VBR speech connections may vary only slightly and to further simplify the calculation a constant value could be used. For example, if the speech encoding algorithm is 32 Kbps adaptive differential pulse code modulation with silence elimination, if ATM adaptation layer 2 is used, and if the bandwidth for the non-circuit-emulation connections will be at least the speed of a $T1$, 1.5 Mbps, then from Sriram, Lyons and Wang [34] one can assume an effective bandwidth of 21.9 Kbps.

For the effective bandwidths of the VBR speech connections as seen by the third-priority Frame-Relay connections, the large-buffer asymptotic-decay-rate approximation of Section 2 should be appropriate, given a relatively large buffer threshold in the performance criteria for the Frame-Relay connections. Likewise, for the effective bandwidth of the VBR speech connections as seen by the fourth-priority UBR connections, where there is a very loose, if any, performance criterion, the large-buffer asymptotic approximation would again be appropriate, and moreover, the asymptotic approximation may be close enough to the mean rate, that the mean rate itself can be used as the effective bandwidth; see the third example in Section 5. In summary, for the priority-2 speech connections there is just one category, thus no dependence on subscript j, and $e_2 > e_2^3 \geq e_2^4$, where e_2^4 or both e_2^4 and e_2^3 could be the mean rate.

For the effective bandwidth of the Frame-Relay connections, it would be reasonable to use the standardized-traffic-descriptor method of Section 6.2, as this is already being used in existing Frame-Relay networks. As with the rates for circuit-emulation connections, the sustainable-cell-rates (SCRs) can be grouped into categories. To reduce complexity, the multiplicative factor α (see Section 6.2) could be picked to be the same for all SCRs; alternatively, it could be different for different rate categories, or be dependent on whether the connection is switched or semi-permanent, as one could expect the former to have a higher occupancy than the latter. (More work should be done to determine good choices.) As for the effective bandwidth of the Frame-Relay connections as seen by the fourth-priority UBR connections, one could again use the standardized traffic descriptors with a reduced value of α, or, as we will choose here, one might simply use the historical, measured mean rate. Thus, for the Frame-Relay connections, e_{3j} would be greater than e_{3j}^4, where the type j could depend on a rate category and/or switched versus semi-permanent connection, and where e_{3j} is given by the traffic descriptor method, and e_{3j}^4, for simplicity, is the mean rate.

For the fourth priority, UBR connections with no minimum rate guarantee, the simplest policy would be to assign an effective bandwidth of zero, in which case the fourth-priority constraint would be subsumed into the third. However, given that the network operator wishes to engineer some capacity for the fourth priority connections, a simple policy would be to dimension a given amount of bandwidth, or a given fraction, say β, of the total capacity, for the aggregate of all fourth priority connections.

In summary, for this example of dimensioning an ATM infrastructure to support four different services, equation (2.18) would specialize to:

$$\sum_{j=1}^{J_1} e_{1j} n_{1j} + e_2 n_2 \leq c \tag{7.1}$$

$$\sum_{j=1}^{J_1} e_{1j} n_{1j} + e_2^3 n_2 + \sum_{j=1}^{J_3} e_{3j} n_{3j} \leq c \tag{7.2}$$

$$\sum_{j=1}^{J_1} e_{1j} n_{1j} + e_2^4 n_2 + \sum_{j=1}^{J_3} e_{3j}^4 n_{3j} + \beta c \le c \ , \qquad (7.3)$$

where the effective bandwidths e_{ij}^k are determined by the methods discussed above.

8 Conclusions

We have considered the problems of connection admission control and dimensioning when there are priority classes and the performance criteria are expressed in terms of buffer overflow probabilities, which translate into steady-state workload tail probabilities in an infinite-buffer model, i.e., $P(V_i > b_i) \le p_i$ as in (2.1). For all priority classes except the highest, the steady-state workload is different from the steady-state delay. The steady-state workload is the appropriate quantity when the concern is buffer overflow. We have focused on the special case of two priority classes, but our analysis extends directly to any number of priority classes.

An important general observation about effective bandwidths with priorities is that there needs to be a constraint in the admissible set for each priority class. In the constraint for priority class i, all higher priority classes play a role (but no lower priority classes). This implies that a connection at priority level i is associated with *multiple* effective bandwidths, one for the priority level of the given connection and one for each lower priority level. For two priority classes, this means that there are two effective bandwidths for priority class 1: one as seen by priority 1 and another as seen by priority 2, i.e., e_{1j}^1 and e_{1j}^2. It is important that e_{1j}^2 be used in the constraint associated with priority class 2 instead of e_{1j}^1. Often e_{1j}^1 is significantly larger than e_{1j}^2, so that using e_{1j}^1 instead of e_{1j}^2 produces a serious error, significantly underestimating the capacity available to priority 2.

If large-buffer asymptotics are used to compute individual effective bandwidths, then we propose the empty-buffer effective-bandwidth (EBEB) approximation, (2.17), (2.19) – where the effective bandwidth of a type-j source of priority i seen by class k where $i < k$, is $e_{ij}^k = \psi_{A_{ij}}(\eta_k^*)/\eta_k^*$ – as a relatively simple approximation for the admissible set with priorities. The effective-bandwidth approximations considered in Sections 2–5 are based on large-buffer asymptotics, under assumptions yielding exponential tail probabilities. The complex structure found in many traffic measurements on existing communication networks indicates that the suitability of these assumptions needs to be carefully checked in applications. Thus alternative heuristic ways to define effective bandwidths were described in Section 6. They make strong use of the linear-admissible-set structure.

With large-buffer asymptotics, we propose the EBEB approximation instead of the exact effective-bandwidth approximation primarily because the EBEB approximation produces a linear admissible set, i.e., an admissible set constructed from linear constraints, whereas the exact effective-bandwidth approximation does not. The reduced-service-rate (RSR) effective bandwidth approximation, (2.14), (2.16), (2.20) also produces a linear admissible set, but it tends to be less accurate as well as not conservative. A linear admissible set greatly helps for doing capacity planning using loss network models.

We have noted that the EBEB approximation is a conservative approximation to the exact effective-bandwidth approximation. Hence, when the exact effective-bandwidth approximation is conservative (which is usually, but not always the case [13]), the overall EBEB approximation is itself conservative. Our experience with numerical examples is that the EBEB approximation is often very close to the exact effective-bandwidth approximation. The most likely source of error, if there is significant error, is the gap between the exact effective-bandwidth approximation based on large-buffer asymptotics, and the exact admissible set.

The effective-bandwidth analysis also dramatically shows when it makes good engineering sense to introduce priority classes. (There can also be other reasons to introduce priorities.) Consistent with intuition, it is appropriate to make class 2 a low-priority class when the performance requirements for class 2 are much less stringent than the performance requirements for class 1. The effective-bandwidth analysis characterizes the performance requirements via the effective performance criteria $\eta_i^* \equiv -\log p_i/b_i$ in (2.3), associated with the constraint $P(V_i > b_i) \le p_i$. Hence, when $\eta_1^* \gg \eta_2^*$, it makes sense for class 2 to be the low-priority class.

The class-i effective performance criterion $\eta_i^* \equiv \log p_i/b_i$ is clearly more strongly affected by the buffer size b_i than the overflow probability p_i because of the logarithm. Thus, the strong ordering $\eta_1^* \gg \eta_2^*$ is likely to occur when $b_1 \ll b_2$. In turn, the ordering $b_1 \ll b_2$ is likely to occur, not so much because buffer sizes are different, but because b_1 is kept small in order to meet more stringent delay requirements for the high-priority class. (The steady-state delay and workload are identical for the high priority class.) For parameter values that we deemed reasonable, it often seemed appropriate to have $\eta_1^*/\eta_2^* \approx 10$. For ratios in this range, we found the two-priority structure to be effective.

In addition to providing the EBEB approximation, we have provided ways to check its accuracy. The empty-buffer and RSR approximations produce convenient upper and lower bounds on both the exact effective-bandwidth approximation and the full low-priority steady-state workload tail probabilities. When the admissible sets based on the empty-buffer and RSR effective-bandwidth approximations are close, we can conclude that the EBEB approximation must be close to the exact effective-bandwidth approximation. It is our experience in examples with $\eta_1^* \gg \eta_2^*$ that the EBEB approximation could essentially be identified with the exact effective-bandwidth approximation.

The most difficult challenge is validating the effective-bandwidth approximation or producing a more accurate admissible set, if necessary. The empty-buffer and RSR approximations reduce the exact calculation of the steady-state low-priority workload tail probability to an exact calculation with the FIFO discipline. When the four approximations – all combinations of empty-buffer and RSR, effective bandwidth and exact – are close, then we know that the EBEB approximation is accurate. Indeed, since the admissible sets tend to be ordered, it suffices to compare only the EBEB approximation to the RSR exact calculation.

A partial check on the accuracy of the effective-bandwidth approximation is provided by considering each priority class alone and computing the maximum admissible number of connections for the FIFO case. This determines the disparity on the axes of the admissible set in the priority case. Sometimes one can get a useful bound without doing this computation: the maximum admissible number of connections is upper bounded by the link rate divided by the mean rate of a connection, and sometimes the empty-buffer effective bandwidth is not much bigger than the mean rate. Indeed, for some relevant cases the effective bandwidths e_{1j}^2 and e_{2j} were within 20% of the mean rate, and that the effective-bandwidth constraint for the lower-priority class could be shown to be very accurate. As a consequence, the effective-bandwidth approximations tended to work better for the lower-priority class than for the higher-priority class. A natural approach, then, is to use the EBEB approximation only for the lower-priority constraint, and use a more refined analysis to generate the higher-priority constraint(s).

To numerically calculate empty-buffer and RSR exact admissible sets with MMPP input, we used the numerical transform inversion algorithm in [13]. As indicate in [13], that algorithm is only effective for up to about 100 environment states. That means we can treat 100 identical sources of one type but only 10 sources each of two types. Thus, the inversion algorithm is far from being able to treat examples with many sources of many types.

A promising approximation strategy – intercept effective bandwidth – is to characterize the admissible set only on the axes and connect the points by lines. On each axis only one of the two classes is present, so that the model is simplified. Our examples based on large-buffer asymptotics indicate that the empty-buffer exact and overall exact admissible sets tend to be slightly concave, so that the intercept-effective-bandwidth approximation is not necessarily conservative, but it is often quite close to the full empty-buffer exact admissible set.

Finally, the RSR and empty-buffer approximations for the low-priority workload process open the way for other approaches to determine an appropriate admissible set, including approaches that are not based on large-buffer exponential asymptotics. Promising practical measurement-based approaches that exploit the linear admissible-set structure were presented in Section 6.

Acknowledgment

We thank Gagan Choudhury for his generous assistance with the numerical calculations. We thank Roch Guerin for encouraging us to be clear about the engineering relevance and the relationship to prior work.

References

[1] J. Abate and W. Whitt, "The Fourier-series method for inverting transforms of probability distributions," *Queueing systems*, vol. 10, pp. 5–88, 1992.

[2] J. Abate and W. Whitt, "Solving probability transform probability transform functional equations for numerical inversion," *Opns. Res. Letters*, vol. 12, pp. 275–281, 1992.

[3] H. Ahmadi, P. F. Chimento, R. A. Guerin, L. Gün, B. Lin, R. O. Onvural and T. E. Tedijanto, "NBBS traffic management overview," *IBM Systems J.*, vol. 34, pp. 604–628, 1995.

[4] A. W. Berger and W. Whitt, "A general framework for effective bandwidths with priorities and loss criteria," AT&T Labs, 1997.

[5] A. W. Berger and W. Whitt, "Extending the effective-bandwidth concept to networks with priority classes," *IEEE Communications Magazine*, 1998, to appear.

[6] D. Bertsimas, I. Ch. Paschalidis and J. N. Tsitsiklis, "Asymptotic buffer overflow probabilities in multiclass multiplexers: an optimal control approach," *IEEE Trans. Aut. Control*, vol. 43, pp. 315–337, 1998.

[7] C. S. Chang, "Sample path large deviations and intree networks," *Queueing Systems*, vol. 20, pp. 7–36, 1995.

[8] C. S. Chang and J. A. Thomas, "Effective bandwidths in high-speed digital networks," *IEEE J. Sel. Areas Commun.*, vol. 13, pp. 1091–1100, 1995.

[9] C. S. Chang and T. Zajic, "Effective bandwidths of departure processes from queues with time varying capacities," *Proc. IEEE Infocom '95*, pp. 1101–1009.

[10] G. L. Choudhury, K. K. Leung and W. Whitt, "An algorithm to compute blocking probabilities in multi-rate multi-class multi-resource loss models," *Adv. Appl. Prob.*, vol. 27, pp. 1104-1143, 1995.

[11] G. L. Choudhury, K. K. Leung and W. Whitt, "An inversion algorithm to compute blocking probabilities in loss networks with state-dependent rates," *IEEE/ACM Trans. Networking*, vol. 3, pp. 585–601, 1995.

[12] G. L. Choudhury, K. K. Leung and W. Whitt, "Efficiently providing multiple grades of service with protection against overloads in shared resources," *AT&T Technical Journal*, vol. 74, pp. 50–63, 1995.

[13] G. L. Choudhury, D. M. Lucantoni and W. Whitt, "Squeezing the most out of ATM," *IEEE Trans. Commun.*, vol. 44, pp. 203–217, 1996.

[14] G. L. Choudhury, A. Mandelbaum, M. I. Reiman and W. Whitt, "Fluid and diffusion limits for queues in slowly changing environments," *Stochastic Models*, vol. 13, pp. 121–146, 1997.

[15] G. de Veciana, C. Courcoubetis and J. Walrand, "Decoupling bandwidths for networks, a decomposition approach for networks," *Proc. IEEE Infocom '94*, vol. 2, pp. 466–474, 1994.

[16] G. de Veciana and G. Kesidis, "Bandwidth allocation for multiple qualities of service using generalized processor sharing," *IEEE Trans. on Information Theory*, vol. 42, pp. 268–272, 1996.

[17] G. de Veciana, G. Kesidis and J. Walrand, "Resource management in wide-area ATM networks using effective bandwidths," *IEEE J. Sel. Areas Commun.*, vol. 13, pp. 1081–1090, 1995.

[18] A. I. Elwalid and D. Mitra, "Fluid models for the analysis and design of statistical multiplexing with loss priorities on multiple classes of brusty traffic," *Proc. IEEE Infocom '92*, pp. 415–425, 1992.

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

[20] A. I. Elwalid and D. Mitra, "Analysis, approximations and admission control of a multi-service multiplexing system with priorities," *Proc. IEEE Infocom '95*, pp. 463–472, 1995.

[21] International Telecommunications Union (ITU), "Traffic control and congestion control in B-ISDN," ITU-T Recommendation I.371, Geneva, May, 1996.

[22] K. W. Fendick, V. R. Saksena and W. Whitt, "Dependence in packet queues," *IEEE Trans. Commun.*, vol. 37, pp. 1173–1183, 1989.

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

[24] H. Heffes and D. M. Lucantoni, "A Markov modulated characterization of packetized voice and data traffic and related statistical multiplexer performance," *IEEE J. Sel. Areas Commun.*, vol. SAC-4, pp. 856–868, 1986.

[25] O. Kella, "Parallel and tandem fluid networks with dependent and Lévy inputs," *Ann. Appl. Prob.*, vol. 3, pp. 682–695, 1993.

[26] F. Kelly, "Notes on effective bandwidths," in *Stochastic Networks*, Clarendon Press, Oxford, pp. 141–168, 1996.

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

[28] V. G. Kulkarni, L. Gun and P. F. Chimento, "Effective bandwidth vectors for multiclass traffic multiplexed in a partitioned buffer," *IEEE J. SEl. Areas Commun.*, vol. 13, pp. 1039–1047, 1995.

[29] D. M. Lucantoni, "The BMAP/G/1 queue: a tutorial" in *Models and Techniques for Performance Evaluation of Computer and Communication Systems*, eds. L. Danatiello and R. Nelson, Springer, New York, pp. 330–358, 1993.

[30] D. Mitra, J. A. Morrison and K. G. Ramakrishnan, "ATM network design and optimization: A multirate loss network framework," *IEEE/ACM Trans. Networking* vol. 4, pp. 531–543, 1996.

[31] I. Ch. Paschalidis, *Large deviations in high speed communications networks*, Ph.D. thesis, M.I.T., 1996.

[32] A. A. Puhalskii and W. Whitt, "Functional large deviation principles for waiting and departure processes," AT&T Labs, 1996.

[33] E. Seneta, *Non-negative Matrices and Markov Chains*, second edition, Springer-Verlag, New York, 1981.

[34] K. Sriram, T. G. Lyons, Y. T. Wang, "Anomalies due to delay and loss in AAL2 packet voice systems: performance models and methods of mitigation," *IEEE J. on Selected Areas of Comm.*, to appear.

[35] K. Sriram and W. Whitt, "Characterizing superposition arrival processes in packet multiplexers for voice and data," *IEEE J. Sel. Areas Commun.*, vol. SAC-4, pp. 833–846, 1986.

[36] W. Whitt, "Tail probabilities with statistical multiplexing and effective bandwidths in multi-class queues," *Telecommunication Systems*, vol. 2, pp. 71–107, 1993.

[37] J. Zhang, "Performance study of Markov modulated fluid flow models with priority traffic," *Proc. IEEE Infocom '93*, pp. 10–17, 1993.

[38] Z.-L. Zhang, "Large deviations and the general processor sharing scheduling for a two-queue system," *Queueing Systems*, vol. 26, pp. 229–254, 1997.