ELSEVIER

Performance Evaluation 41 (2000) 249–267

PERFORMANCE EVALUATION
An International Journal

www.elsevier.com/locate/peva

Workload bounds in fluid models with priorities

Arthur W. Berger[a], Ward Whitt[b,*]

[a] *Akamai Technologies, 500 Tech. Sq., Cambridge, MA 02139, USA*
[b] *AT&T Labs, Room A117, Shannon Laboratory, 180 Park Avenue, Florham Park, NJ 07932-0971, USA*

Received 30 June 1997; received in revised form 4 January 1999

Abstract

In this paper we establish upper and lower bounds on the steady-state per-class workload distributions in a single-server queue with multiple priority classes. Motivated by communication network applications, the model has constant processing rate and general input processes with stationary increments. The bounds involve corresponding quantities in related models with the first-come first-served discipline. We apply the bounds to support a new notion of effective bandwidths for multi-class systems with priorities. We also apply the lower bound to obtain sufficient conditions for the workload distributions to have heavy tails.

Keywords: Priority queues; Stochastic fluid models; Bounds; Admission control; Effective bandwidths; Large-buffer asymptotics; ATM; Heavy-tailed distributions; Long-tailed distributions

1. Introduction

Motivated by the desire to model asynchronous transfer mode (ATM) switches and internet protocol (IP) routers supporting multiple priority classes, we consider a stochastic fluid queue with unlimited buffer space, constant release rate and m priority classes. We allow the input for each class to arrive in an arbitrary manner. Our main assumptions are that the m single-class cumulative input processes are mutually independent and that each has stationary increments. We want to allow general stationary input processes in order to be able to represent traffic complexity as observed in many recent traffic measurements (e.g., see [6,12,17]).

We assume that the priority service discipline is preemptive resume, which in our fluid context means that the constant output rate available at any instant is applied to the highest-priority work waiting or arriving at that instant. Note that in the application to packet networks the transmission of a packet is not preempted, however, the resulting inaccuracy of assuming preemptive resume is well with the inaccuracy of the overall traffic model, particularly in the region of interest of many packets in queue. Within each priority class, work is served in a first-come first-served (FCFS) order. With this priority discipline, it actually suffices to consider only two priority classes. From the perspective of any

* Corresponding author. Tel.: +1-973-360-8724; fax: +1-973-360-8050.
E-mail addresses: awberger@akamai.com (A.W. Berger), wow@research.att.com (W. Whitt)

0166-5316/00/$ – see front matter
PII: S0166-5316(99)00075-9

class, all lower priority classes can be ignored, while all higher priority classes can be lumped together. Thus, without loss of generality, we consider only two priority classes, with class 1 having priority over class 2. Since class 1 experiences a FCFS system, we are primarily interested in the steady-state workload (buffer content) and waiting time for the low-priority class 2. The waiting time is a virtual waiting time; in particular, the waiting time at time t is the time until a potential additional infinitesimal particle of fluid arriving at time t would be served (processed). The low-priority waiting time at any instant is at least as large as the total workload at that instant divided by the processing rate; it often is strictly larger because subsequent high-priority input has priority over waiting low-priority work.

Our main results for the low-priority workload and waiting time are upper and lower bounds in terms of associated stochastic fluid models with the FCFS service discipline. These bounds allow us to apply results for FCFS systems to bound and approximate the low-priority steady-state workload and waiting time.

Our bounds have many possible applications. We developed them in order to extend the concept of effective bandwidths for admission control to settings with multiple priority classes. That application of the bounds is described in [2], so we will be brief here. The notion of effective bandwidths was originally developed for the FCFS discipline. The general idea is to assign an effective bandwidth e_i to each connection of type i. Then a vector $(n_1, \ldots, n_{|I|})$, where I is the set of connection types, $|I|$ is its cardinality and n_i is the number of connections of type i, is deemed feasible for a system with capacity (i.e., available bandwidth or constant processing rate) c if

$$\sum_{i \in I} n_i e_i \leq c.$$

The associated admissible set with a single linear boundary greatly simplifies engineering, e.g., it makes it possible to apply stochastic-loss-network (generalized-Erlang) models, as in [21], for capacity planning. A theoretical basis for the notion of effective bandwidths and the linear admissible-set structure for the FCFS discipline has been provided by large-buffer asymptotics (large deviations theory), e.g., see [7,9,15,27] for reviews.

As discussed in [2], a corresponding large-buffer asymptotics can be developed for stochastic fluid queues with priorities. The resulting admissible set has a constraint for each priority class. That by itself presents no major problem, but unfortunately these constraints are in general nonlinear. Losing the linearity causes the notion of effective bandwidths to lose much of its appeal. Fortunately, the FCFS bounds introduced here can help. The FCFS bounds produce approximating admissible sets that do have linear boundaries. In particular, the admissible set associated with the upper (lower) bound has linear boundaries and is contained in (contains) the exact admissible set with nonlinear boundaries, i.e., the upper bound on the workload tail probabilities is conservative, leading to a smaller admissible set. In [2] we suggest the conservative upper bound for the steady-state workload and the associated smaller admissible set as the preferred approximation.

Even more important than the candidate approximations for effective bandwidths, we believe, is the proposed structure for the admissible set with multiple priority classes. Regardless of the method used to define effective bandwidths, the analysis suggests that there should be a linear constraint associated with 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. We have made a case for this general admissible-set structure, without referring to large-buffer asymptotics in [3]. This approximating admissible-set was also

suggested by Kulkarni and Gautam [16], but they obtained it by examining the exact admissible set, rather than from general bounds on the steady-state workloads.

Here is how the rest of this paper is organized. In Section 2 we define the stochastic processes of interest in the two-priority model. In Section 3 we show how to construct stationary versions of the stochastic processes defined in Section 2. In Section 4 we apply the stationary versions together with previous large-deviations results in [13,27] to establish the exact large-buffer deviation result.

In Section 5 we establish an exact relation between the low-priority waiting time and the total workload under the assumption that the high-priority input has stationary and independent increments. In Section 6 we establish the lower bound on the low-priority steady-state workload, which we call the reduced service-rate bound. In Section 7 we combine this bound with another lower bound in [8] to obtain general sufficient conditions for the low-priority workload to have a heavy-tailed distribution. In Section 8 we establish the upper bound on the steady-state workload, which we call the empty-buffer bound. Finally, in Section 9 we consider an illustrative numerical example.

We close this introduction by mentioning other related work. Exact analyses of priority models with extra Markovian assumptions have been done by Sugahara et al. [23], Takine and Hasegawa [24] and Zhang [29]. Zhang [29] finds an exact solution for a Markov modulated fluid model with priorities, and Elwalid and Mitra [11] develop an approximation. With admission control, these approaches offer the possibility of calculating the feasible set more accurately, but at the expense of losing the more elementary effective bandwidth approach with linear constraint boundaries.

2. The general fluid model with priorities

In this section we define the basic stochastic processes in the general fluid model with priorities. By "fluid model" we mean that work is processed continuously at a constant rate as if it were a fluid. We let work arrive in an arbitrary manner, i.e., it could arrive continuously or in jumps. There is a single-server (or buffer) with unlimited waiting space. The specific priority discipline is preemptive resume, i.e., higher-priority work immediately preempts lower priority work, and lower priority work resumes service where it stopped when it regains access to the server.

In general, there may be m priority classes, but it suffices to consider only two. Hence, let there be two priority classes, with class 1 having preemptive priority over class 2. Let class i have required work arrive according to the stochastic process $\{A_i(t): t \geq 0\}$, i.e., $A_i(t)$ is the input for class i over the interval $[0, t]$. (The process $A_i(t)$ might be the superposition of arrival processes from independent sources.) We assume that $A_i(t)$ has nondecreasing sample paths. Let the work be processed continuously at rate c in order of the priority. Thus, 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

$$V_i(t) = V_i(0) + X_i(t) - \inf_{0 \leq s \leq t} \{\min\{0, V_i(0) + X_i(s)\}\}, \quad t \geq 0, \tag{2.1}$$

where

$$X_i(t) = A_i(t) - S_i(t), \tag{2.2}$$

$$S_1(t) = ct, \tag{2.3}$$

$$S_2(t) = ct - D_1(t), \tag{2.4}$$

$$D_i(t) = A_i(t) + V_i(0) - V_i(t), \tag{2.5}$$

with $A_i(0) = 0$ for all i. The processes $\{S_i(t): t \geq 0\}$ in (2.3) and (2.4) 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, (2.3) 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 the input over $[0, t]$, plus the initial work, minus what is present at time t, as indicated in (2.5). For $i = 2$, the server-availability process can be defined in terms of the departure process of the high-priority class by (2.4). Finally, the process $\{X_i(t): t \geq 0\}$ in (2.2) is the (cumulative) *net input process* for class i, in terms of which the workload process is defined by the usual one-dimensional *reflection map* in (2.1).

It is important to distinguish between the workload process and the waiting time process. The workload $V_i(t)$ is the class i work in the system at time t (e.g., in units of bits), while the (virtual) waiting time $W_i(t)$ is the time required to clear the workload $V_i(t)$ at time t (not counting any class i input after time t). However, the class 2 waiting time at time t depends on the class 1 input after time t. We can define the class i waiting time $W_i(t)$ by

$$W_i(t) = \inf\{u: u \geq 0 \text{ and } S_i(t+u) - S_i(t) \geq V_i(t)\}, \quad t \geq 0. \tag{2.6}$$

Combining (2.3) and (2.6), we see that $W_1(t) = V_1(t)/c$, as it should, but in general we only have

$$W_2(t) \geq \frac{V_2(t)}{c} \quad \text{for all } t. \tag{2.7}$$

Indeed, the low-priority waiting time $W_2(t)$ can be much greater than the scaled low-priority workload $V_2(t)/c$ if the server is frequently occupied with high-priority work.

If customers arrive at random times and bring service requirements, then the processes A_i are pure jump processes, having jumps up equal to the service times and $W_i(t)$ is the virtual waiting time process (the time a potential arrival at time t would have to wait before beginning service). If there are arrivals at time t, then $W_i(t)$ is the time required for all these arrivals to complete service. With Poisson arrivals, the steady-state virtual waiting time coincides with the steady-state actual waiting time (before beginning service) seen by arrivals, by the Poisson Arrivals See Time Averages (PASTA) property, see [28].

If work arrives continuously at a random rate, work can be processed without there being any work in the buffer. This will occur whenever the buffer is empty and the input rate is positive but less than the output rate c.

3. Constructing stationary versions

In Section 2 we indicated how to define the stochastic processes of interest, with general initial conditions. Now we construct stationary versions of these processes, which describe the system in equilibrium or steady-state. For background, see [1,4].

Indeed, so far we have made no stochastic assumptions. Now we assume that the stochastic processes A_i are mutually independent processes, each of which has stationary and ergodic increments with

$$\lim_{t\to\infty} \frac{A_i(t)}{t} = c\rho_i \quad \text{w.p.1 for each } i, \tag{3.1}$$

where $\rho \equiv \rho_1 + \rho_2 < 1$. The stability condition $\rho < 1$ ensures that the average rate that work enters is less than the processing rate c. This condition enables us to construct stationary versions of all the

processes, as we show below. To treat the high-priority workload, we can use standard arguments as in Section 6 of Borovkov [4] and Chapter 2 of Baccelli and Brémaud [1]. However, it is more complicated to obtain a stationary version of the low-priority workload, because the low-priority workload depends on the high-priority departure process, as can be seen from (2.4). Thus, we successively construct stationary versions of the stochastic processes V_1, D_1, V_2, D_2 and W_2. (Since $W_1(t) = V_1(t)/c$, nothing special need be done for W_1.)

For this purpose, let $\Rightarrow$ denote convergence in distribution. First, as is customary, we extend the processes A_i to be over $(-\infty, \infty)$ with stationary increments, but still keep the convention that $A_i(0) = 0$ for $i = 1, 2$, which implies that $X_i(0) = 0$. Then the stationary increments condition on A_1 implies that

$$\hat{V}_1(t) \equiv \sup_{0\le s\le t}\{X_1(t) - X_1(s)\} \stackrel{\mathrm{d}}{=} \sup_{0\le s\le t} X_1(-s), \quad t \ge 0, \tag{3.2}$$

i.e., the stationarity in X_1 allows us to construct the steady-state workload as the simple maximum of the reverse-time net input process, with initial workload 0. Since the final supremum in (3.2) is nondecreasing in t, $V_1(t) \Rightarrow \hat{V}_1$ as $t \to \infty$. Condition (3.1) for $i = 1$ implies that $X_1(t)/t \to c(\rho_1 - 1)$ and $X_1(t) \to -\infty$ as $t \to \infty$ w.p.1. Hence, the steady-state high-priority workload is

$$V_1 \stackrel{\mathrm{d}}{=} \sup_{s\ge 0} X_1(-s) < \infty \quad \text{w.p.1.} \tag{3.3}$$

More generally, there is a stationary version of the stochastic process $\{V_1(t): t \ge 0\}$, which we denote by $\{V_1^*(t): t \ge 0\}$, with

$$V_1^*(t) = \sup_{s\le t}\{X_1(t) - X_1(s)\}, \quad t \ge 0, \tag{3.4}$$

i.e., the random vector $(V_1^*(t_1 + h), \ldots, V_1^*(t_k + h))$ has a distribution independent of h for all k and all k-tuples $(t_1, \ldots, t_k)$, see [4, Chapter 1].

Given the stationary process $\{V_1^*(t): t \ge 0\}$, the associated stationary departure process $\{D^*(t): t \ge 0\}$ defined by (2.5) is

$$D_1^*(t) = A_1(t) + V_1^*(0) - V_1^*(t), \tag{3.5}$$

i.e., it has stationary increments. (However, note that in general variables $A_1(t)$ and $V_1^*(0)$ are not independent.) This in turn makes the associated stochastic process $S_2^*(t)$ and $X_2^*(t)$ have stationary increments. Given that $V_1^*(t)$ has a proper distribution for each t, (3.1) implies that $t^{-1}V_1^*(t) \to 0$ w.p.1 as $t \to \infty$. Since this is an important technical point, we state it as a proposition and prove it.

Proposition 3.1. *Under the assumptions above*

$$\frac{V_1^*(t)}{t} \to 0 \quad w.p.1 \, as \, t \to \infty.$$

Proof. The limit $t^{-1}X(t) \to c(\rho_1 - 1)$ as $t \to \infty$ w.p.1 implies the stronger functional limit $n^{-1}X(nt) \to c(\rho_1 - 1)t$ w.p.1 as $n \to \infty$, with convergence being uniform in t over bounded intervals, see [13, Theorem 4]. Then apply the continuous mapping theorem with the reflection map and general initial condition $V_1^*(0)$ as in Theorem 6.4 (iii) of Whitt [26] to get $n^{-1}V_1(nt) \to 0$ w.p.1 as $n \to \infty$, again uniform in t over bounded intervals, which implies the desired result. □

Given Proposition 3.1, (3.1) and (3.5) imply that

$$\frac{D_1^*(t)}{t} \to c\rho_1 \quad \text{as } t \to \infty \text{ w.p.1}, \tag{3.6}$$

so that

$$\frac{X_2^*(t)}{t} \to c(\rho - 1) \quad \text{as } t \to \infty \text{ w.p.1}. \tag{3.7}$$

Hence, we can repeat the construction above to construct a stationary version $\{V_2^*(t): t \geq 0\}$ of the stochastic process $\{V_2(t): t \geq 0\}$ with $V_2^*(t) < \infty$ w.p.1.

Given that $V_2^*(t)$ is proper, we can apply Proposition 3.1 again to deduce that $t^{-1}V_2^*(t) \to 0$ w.p.1 as $t \to \infty$. Then conditions (2.5) and (3.1) imply that

$$\frac{D_2^*(t)}{t} \to c\rho_2 \quad \text{as } t \to \infty \text{ w.p.1}. \tag{3.8}$$

Finally, we obtain a stationary version W_2^* of $W_2(t)$ defined in terms of (S_2^*, V_2^*) as in (2.6). The supporting theorem is the continuous-time analog of Proposition 6.6 in [5, p. 105]. Let V_i and W_i be random variables with the steady-state distributions of $V_i(t)$ and $W_i(t)$. Henceforth we omit the $*$ notation; we will indicate when stationarity is assumed.

4. The large deviations result

Suppose that we have criteria on the steady-state workload tail probabilities for each priority class that we want satisfied, e.g.

$$P(V_i > b_i) \leq p_i \quad \text{for each priority class } i. \tag{4.1}$$

We think of this tail probability constraint as a surrogate for the constraint that the probability of a buffer overflow from a buffer of size b_i be less than p_i for class i. It is natural to use the workload in (4.1) instead of the waiting time or sojourn time if we are interested in the probabilities of buffer overflows. Then the tail probability $P(V_i > b_i)$ is an approximation for the probability that class i work will overflow a buffer of size b_i when there are separate buffers dedicated to each priority class.

With criteria such as (4.1), we can use the notion of effective bandwidths as in [27] to develop an admission control procedure for sources of each priority class. However, *with priorities, we must proceed recursively over the priority classes.* The possibilities for lower priorities depend on the high-priority sources in service.

Let there be multiple sources of each of the two priority classes. Let the sources be indexed by the pair (i, j), representing source type j of priority class i. Let J_i be the number of source types for priority class i. Let $A_{ij}(t)$ be the arrival process of an (i, j) source. Let $A_{ij}(t)$ be a general input process with nondecreasing sample paths, e.g., $A_{ij}(t)$ represents the number of bits to arrive at a network node from an (i, j) source during the interval $[0, t]$. Let

$$\psi_{A_{ij}}(\theta) = \lim_{t\to\infty} t^{-1} \log E\, e^{\theta A_{ij}(t)}, \tag{4.2}$$

be the *single-source arrival-process asymptotic decay-rate functions* (cumulant generating functions) as in [27, Eq. (1.10)] (without assuming here that A_{ij} has rate 1). We assume that these decay-rate functions

are well defined. Given mutually independent sources, with $n_{ij}(i, j)$-sources, we can form associated aggregate asymptotic decay-rate functions

$$\psi_{A_i}(\theta) = \sum_{j=1}^{J_i} \psi_{A_{ij}}(\theta) n_{ij} \tag{4.3}$$

for priority class i. We give explicit formulas for asymptotic decay-rate functions in [27] and [2, Section IV].

Similarly, we can form the associated asymptotic decay-rate functions for the server-availability processes by letting

$$\psi_{S_i}(\theta) = \lim_{t\to\infty} t^{-1} \log E\, e^{\theta S_i(t)}. \tag{4.4}$$

For the high-priority class, $S_1(t) = ct$, $t \geq 0$, so that

$$\psi_{S_1}(\theta) = c\theta. \tag{4.5}$$

However, the low-priority service-availability process $S_2(t)$ is more complicated, but by (2.4) we can express it in terms of $\psi_{D_1}(\theta)$.

We now show how to use the asymptotic decay-rate functions to define a notion of effective bandwidths for (i, j) sources using criterion (4.1). The analysis of effective bandwidths here is the natural extension of the effective bandwidths for the queue length process in [27], just as in [27, Section 5]. In [27, p. 75], the processes, $A_i(t)$ and $S_i(t)$ are counting processes, and the key equations are (1.12) and (1.17). (In [14,27] three essentially equivalent processes were studied for the standard queuing model: the queue length process, the workload process and the discrete-time waiting time sequence, with each process being essentially a reflection of a net input process, and with each process having its own effective bandwidth equation. Here, with the more general processes $A_i(t)$ and $S_2(t)$, we focus only on the generalization of the queue length process with net input process $X_i(t) = A_i(t) - S_i(t)$, as in [27, Section 5].)

The notion of effective bandwidths is based on an exponential approximation for the workload tail probabilities

$$P(V_i > b_i) \approx e^{-\eta_i b_i}, \tag{4.6}$$

assuming that b_i is relatively large. Given (4.1) and (4.6), we want to choose η_i in (4.6) so that

$$\eta_i \geq \eta_i^* \equiv \frac{-\log p_i}{b_i}. \tag{4.7}$$

The theoretical basis for the exponential approximation (4.6) is an asymptotic result for the workload tail probability $P(V_i > t)$ as $t \to \infty$, Theorem 10 of [27], which is a minor modification of Theorem 4 of Glynn and Whitt [14]. We restate it here in the context of our priority model.

Theorem 4.1. *Consider the general stationary two-priority queuing model in Section 3. If there exists a function ψ_{X_i} and positive constants θ_i^* and ϵ such that*

$$t^{-1} \log E\, e^{\theta[A_i(t) - S_i(t)]} \to \psi_{X_i}(\theta) = \psi_{A_i}(\theta) + \psi_{S_i}(-\theta) \quad \text{as } t \to \infty \text{ for } |\theta - \theta_i^*| < \epsilon, \tag{4.8}$$

with ψ_{X_i} finite in a neighborhood of θ_i^* and differentiable at θ_i^* with

$$\psi_{X_i}(\theta_i^*) \equiv \psi_{A_i}(\theta_i^*) + \psi_{S_i}(-\theta_i^*) = 0, \tag{4.9}$$

and $\psi'_{X_i}(\theta_i^*) > 0$, *then*

$$t^{-1} \log P(V_i > t) \to -\theta_i^* \quad as\, t \to \infty. \tag{4.10}$$

In Theorem 10 of [27] there is a condition that there exists a constant M such that $S_i(\delta) \leq M$ for all sufficiently small δ. That condition is satisfied here because $S_i(t) \leq ct$ for the model in Section 2. As shown in [10], the conditions can be weakened somewhat. The differential of $\psi_{X_i}(\theta)$ can be omitted and, instead of a root to (4.9), it suffices to have

$$\theta_i^* = \sup\{\theta > 0 : \psi_{X_i}(\theta) \leq 0\}, \tag{4.11}$$

but (4.9) is the usual case.

We now apply Theorem 4.1 to develop notions of effective bandwidths and effective capacities for the two priority classes. Let the *effective bandwidth* of an (i, j) source be

$$e_{ij} = \frac{\psi_{A_{ij}}(\eta_i^*)}{\eta_i^*}. \tag{4.12}$$

For class 1, this is the customary procedure. Note that e_{ij} depends only on the source j input process $\{A_{ij}(t) : t \geq 0\}$ of priority i (not on $A_{ik}(t)$ for $k \neq j$).

Let the *effective capacity* available for class i be

$$C_i = \frac{-\psi_{S_i}(-\eta_i^*)}{\eta_i^*}. \tag{4.13}$$

We then say that the collection of sources consisting of n_{ij} sources of type j, $1 \leq j \leq J_i$, are feasible, given the aggregate input process for higher-priorities if

$$\sum_{j=1}^{J_i} e_{ij} n_{ij} \leq C_i. \tag{4.14}$$

Note that the admissible set in (4.14) is linear for each i, but the low-priority (class 2) admissible set depends upon the high-priority sources in service via the effective capacity C_2.

The admissibility criterion (4.14) holds if and only if

$$\sum_{j=1}^{J_i} \psi_{A_{ij}}(\eta_i^*) n_{ij} + \psi_{S_i}(-\eta_i^*) \leq 0. \tag{4.15}$$

This is what we want, because then the prevailing class i decay rate θ_i^* will then exceed η_i^* defined in (4.7) by virtue of Theorem 4.1. To see this, note that ψ_{A_i} and ψ_{S_i} are increasing and convex, which implies that $-\psi_{S_i}(-\theta)$ is increasing and concave, so that

$$\psi_{A_i}(\theta) \leq -\psi_{S_i}(-\theta) \quad \text{for } 0 \leq \theta \leq \theta_i^*, \quad \text{and} \quad \psi_{A_i}(\theta) \geq -\psi_{S_i}(-\theta) \quad \text{for } \theta \geq \theta_i^*.$$

Hence, $\theta_i^* \geq \eta_i^*$ as claimed. (This makes $p_i \approx e^{-\eta_i^* b_i} > e^{-\theta_i^* b_i}$.)

Note that the effective capacities for classes 1 and 2 simplify to nice, intuitive expressions. Since

$$\psi_{S_i}(\theta) = c\theta, \quad \psi_{S_2}(\theta) = c\theta + \psi_{D_1}(-\theta), \tag{4.16}$$

$$C_1 = \frac{-\psi_{S_1}(-\eta_1^*)}{\eta_1^*} = \frac{c\eta_1^*}{\eta_1^*} = c, \tag{4.17}$$

$$C_2 = \frac{c\eta_2^* - \psi_{D_1}(\eta_2^*)}{\eta_2^*} = c - \frac{\psi_{D_1}(\eta_2^*)}{\eta_2^*}, \tag{4.18}$$

where $\psi_{D_1}(\theta)$ is given below in (4.19). We call $\psi_{D_1}(\eta_2^*)/\eta_2^*$ in (4.18) the *effective capacity for class* 2 *used up by class* 1.

To proceed further, from (4.18) we see that we need to determine the asymptotic decay-rate function $\psi_{D_1}(\theta)$ for the high-priority departure process, but this is where the nonlinearity comes in. Under regularity conditions, see [2,19] and references cited there,

$$\psi_{D_1}(\theta) = \begin{cases} \psi_{A_1}(\theta), & \theta < \hat{\theta}, \\ \psi_{A_1}(\hat{\theta}) + c(\theta - \hat{\theta}), & \theta > \hat{\theta}, \end{cases} \tag{4.19}$$

with $\hat{\theta}$ determined by the equation

$$\psi'_{A_1}(\hat{\theta}) = c. \tag{4.20}$$

Our two bounds will avoid the nonlinearity in (4.19). For further discussion, see [2].

5. An exact result for a special case

In this section, under an additional assumption, we obtain an exact relation between the low-priority waiting time W_2 and the total workload V. Since V is the same as for the FCFS discipline, this establishes an important connection to FCFS models. This relation can provide the basis for both exact results and approximations for W_2. The extra assumption is that the class 1 input process A_1 has independent as well as stationary increments. Such an assumption might be appropriate for an ATM switch if the high-priority class is predominantly constant bit-rate (CBR) traffic. Due to network jitter and lack of synchronization, it may be reasonable to model the CBR input as a Poisson process.

We exploit the fact that W_2 is the class 1 first passage time to 0 starting from the steady-state workload of both classes. Let $T_{x0}^{(1)}$ denote the class 1 first passage time from x to 0. This first passage time accounts for future random input and the constant output rate c. The independent-increments property makes the future inputs, starting in V independent of V, which we understand to hold when we write $T_{V0}^{(1)}$.

Since we already have assumed that A_1 has stationary increments, the independent-increments assumption makes A_1 a subordinator or, equivalently, a Lévy process with nonnegative nondecreasing sample paths, as in [18, p. 69]. A subordination is characterized by its characteristic Laplace exponent $\phi(s)$, where

$$E\,\mathrm{e}^{-sA(t)} = \mathrm{e}^{-t\phi(s)}, \quad t > 0. \tag{5.1}$$

Theorem 5.1. *With the general stationary model, if in addition the high-priority input process A_1 has independent increments, then*

$$W_2 \stackrel{\mathrm{d}}{=} T_{V0}^{(1)}, \tag{5.2}$$

$$E\,\mathrm{e}^{-sW_2} = E\,\mathrm{e}^{-\eta(s)V}, \tag{5.3}$$

where $\eta(s)$ is the unique continuous solution to the equation

$$\eta\left(\frac{s}{c}\right) = s + \phi\left(\eta\left(\frac{s}{c}\right)\right), \quad \textit{and} \tag{5.4}$$

$$EW_2 = \frac{EV}{c(1-\rho_1)}. \tag{5.5}$$

Proof. As indicated above, W_2 is the first passage time to 0 for class 1 starting with V. The Laplace transform of this first passage time conditional on V is given in [18, p. 79], while η is characterized in p. 74. The constant c in (5.4) occurs because the processing rate here is c instead of 1. By changing the measuring units, we can regard the processing rate as 1:

$$E\,\mathrm{e}^{sA(t)/c} = \mathrm{e}^{t\tilde{\phi}(s)}, \qquad E\,\mathrm{e}^{-sW_2/c} = \mathrm{e}^{-\tilde{\eta}(s)V}, \quad \text{where} \quad \tilde{\eta} = s + \tilde{\phi}(s).$$

Since $\tilde{\phi}(s) = \phi(s/c)$ and $\tilde{\eta}(s) = \eta(s/c)$, we obtain (5.4). Finally, (5.5) holds because $ET_{x0}^{(1)} = x/c(1-\rho_1)$ for each x, see [18]. □

6. The reduced service-rate lower bound

We now drop the extra assumption in Section 5 (unless specifically stated) and consider the distributions of the low-priority steady-state workload V_2 and waiting time W_2. They are hard to determine because the server-availability process S_2 in (2.4) depends on the stochastic fluctuations of the high-priority class. A convenient rough approximation is to assume that the server is continuously available to the low-priority class at a reduced rate, with the reduction accounting for the long-run average usage of the high-priority class. In particular, we call the approximation

$$S_2(t) \approx S_2^{\mathrm{r}}(t) \equiv (1-\rho_1)ct, \quad t \geq 0, \tag{6.1}$$

the *reduced service-rate* (RSR) *approximation*. With the RSR approximation, we can analyze the two priority classes separately, just as in a system without priorities. The RSR approximation decouples the system, making the low-priority class depend upon the high-priority class only through the offered load parameter ρ_1.

By (2.6) and (6.1), the associated waiting time and workload approximations are related by

$$W_2^{\mathrm{r}}(t) = \frac{V_2^{\mathrm{r}}(t)}{c(1-\rho_1)}, \quad t \geq 0, \tag{6.2}$$

$$W_2^{\mathrm{r}} = \frac{V_2^{\mathrm{r}}}{c(1-\rho_1)}, \tag{6.3}$$

with the steady-state workload being

$$V_2^{\mathrm{r}} = \sup_{t\geq 0}\{A_2(-t) + (1-\rho_1)ct\} = \sup_{t\geq 0}\left\{A_2\left(\frac{-t}{1-\rho_1}\right) + ct\right\}, \tag{6.4}$$

which is the formula for V_1 in (3.2) with $\{A_1(t): t \geq 0\}$ replaced by the scaled process $\{A_2(t/(1-\rho_1)): t \geq 0\}$.

It is intuitively clear that the RSR approximation is typically optimistic, i.e., that we should usually have V_2^r and W_2^r smaller than their counterparts V_2 and W_2. We now present some supporting evidence using stochastic comparison concepts. We say that a random variable U_1 is less than or equal to another U_2 in *increasing convex order* and write $U_1 \leq_{icx} U_2$ if $Ef(U_1) \leq Ef(U_2)$; for all nondecreasing convex real-valued functions f for which the expectations are well defined, see [22] or [1, Chapter 4]. The essential line of reasoning below goes back to Rogozin [20].

Theorem 6.1. *In the general stationary model,* $V_2^r \leq_{icx} V_2$.

Proof. We work with the stationary versions defined in Section 2. Then

$$ES_2(t) = S_2^r(t), \quad t \geq 0,$$

for all t, where $S_2^r(t)$ is defined in (6.1). Hence, the processes $\{S_2(t): t \geq 0\}$ and $\{S_2^r(t): t \geq 0\}$ are ordered by convex stochastic order, i.e., $S_2 \geq_{cx} S_2^r$, by which we mean that

$$Ef(\{S_i(t): t \geq 0\}) \geq Ef(\{S_i^r(t): t \geq 0\}) \tag{6.5}$$

for all real-valued convex functions f on the space of sample paths for which the expectations are well defined (see [1, pp. 198, 220] and [22, Remark 2 in p. 81] for related arguments). By (3.2), V_2 and V_2^r can be written as (nonincreasing) convex real-valued functions of $\{S_2(t): t \geq 0\}$ and $\{S_2^r(t): t \geq 0\}$, respectively. Since nondecreasing convex real-valued functions of arbitrary convex real-valued functions are convex, we have the stated conclusion, i.e.

$$Eg(V_2^r) = Eg \circ f(S_2^r) \leq Eg \circ f(S_2) = Eg(V_2)$$

for all nondecreasing convex g, where f here denotes the convex functions taking S_2^r into V_2^r and S_2^* into V_2^*. □

The $\leq_{icx}$ ordering in Theorem 6.1 implies that $E(V_2^r)^k \leq E(V_2^k)$ for all $k \geq 1$. However, the $\leq_{icx}$ ordering is weaker than ordinary stochastic order $V_2^r \leq_{st} V_2$ which would hold if $Ef(V_2^r) \leq Ef(V_2)$ for all nondecreasing real-valued functions f. We now show that the ordering in Theorem 6.1 cannot be strengthened to stochastic order.

Example 6.1. To see that we need not have $V_2^r \leq_{st} V_2$, we show that it is possible to have $P(V_2^r > 0) > P(V_2 > 0)$. Our example also shows that it is possible to have $P(W_2^r > 0) > P(W_2 > 0)$, so that in general we do not have $W_2^r \leq_{st} W_2$ either. First, if $A_2(t)$ is a pure jump process, then we always have (by Little's law applied to the server)

$$P(V_2^r > 0) = \frac{\rho_2}{(1-\rho_1)c}. \tag{6.6}$$

For our concrete example, let $c = 1$ and initially let $A_2(t) = \rho_2 t$, $t \geq 0$, corresponding to deterministic input. (We will later make $A_2(t)$ a pure jump process.) Let the high-priority input occur in constant lumps of size ρ_1 spaced apart according to i.i.d. random variables distributed as $\rho_1/(1+\rho_2) + X$, where X is exponentially distributed with mean $1 - \rho_1/(1-\rho_2)$. Thus, the mean time between successive class 1 inputs of size ρ_1 is 1. Following each type 1 input of size ρ_1, there is a period of length ρ_1 during which the server works on this input. At the end of this period there is $\rho_1\rho_2$ class 2 work. The server then takes $\rho_1\rho_2/(1-\rho_2)$ time to clear this class 2 work. The remainder of the interval before the next class 1 input,

of length X, the server is processing only the class 2 input. Hence, for this model (using regenerative analysis)

$$P(V_2 > 0) = \frac{\rho_1}{1-\rho_2}, \tag{6.7}$$

so $P(V_2^{\mathrm{r}} > 0) > P(V_2 > 0)$ if and only if $\rho_2(1-\rho_2) > \rho_1(1-\rho_1)$. Since we must have $\rho_1 + \rho_2 < 1$ for stability, this inequality holds whenever $\rho_1 < \rho_2$. For a somewhat extreme case, let $\rho_1 = 0.1$ and $\rho_2 = 0.5$. Then

$$P(V_2^{\mathrm{r}} > 0) = \tfrac{5}{9} > \tfrac{1}{5} = P(V_2 > 0). \tag{6.8}$$

Now we have to make $A_2(t)$ a pure jump process behaving approximately like deterministic input. For this purpose, let $A_2^{(\epsilon)}$ denote the input process having jumps of size $\epsilon\rho_2$ spaced apart by i.i.d. random variables distributed as $\epsilon\rho_2 + \epsilon Y$, where Y is an exponential random variable with mean $1-\rho_2$. Let $V^{(\epsilon)\mathrm{r}}$ denote the RSR approximation associated with $A_2^{(\epsilon)}$. As $\epsilon \to 0$, $A_2^{(\epsilon)}(t)$ approaches deterministic input, so that $P(V_2^{(\epsilon)\mathrm{r}} > 0) \to P(V_2 > 0)$ in (6.7), but (6.6) holds for all ϵ. Hence, the counterexample in (6.8) holds for all sufficiently small ϵ. Finally, this example also serves for the steady-state (virtual) waiting times, because $P(W_2^{\mathrm{r}} > 0) = P(V_2^{\mathrm{r}} > 0)$ and $P(W_2 > 0) = P(V_2 > 0)$ here.

We have yet to establish a result corresponding to Theorem 6.1 for the waiting times. However, we can establish an exact representation for W_2 in terms of V when the high-priority class input has independent increments, as assumed in Section 5. We now show that W_2^{r} is a lower bound for W_2 under this extra condition.

Theorem 6.2. *Under the conditions of Theorem* 5.1,

$$W_2 \geq_{\mathrm{cx}} \frac{V}{c(1-\rho_1)} \geq_{\mathrm{icx}} \frac{V_2^{\mathrm{a}}}{c(1-\rho_1)} = W_2^{\mathrm{r}},$$

so that $W_2 \geq_{\mathrm{icx}} W_2^{\mathrm{a}}$.

Proof. Since $ET_{x0}^{(1)} = x/c(1-\rho_1)$ for each x,

$$E(T_{V0}^{(1)}|V) = \frac{V}{c(1-\rho_1)}. \tag{6.9}$$

Thus, for any convex g,

$$E[g(T_{V0}^{(1)})|V] \geq g\left(\frac{V}{c(1-\rho_1)}\right) \quad \text{w.p.1,} \quad \text{and} \quad Eg(T_{V0}^{(1)}) \geq Eg\left(\frac{V}{c(1-\rho_1)}\right),$$

i.e., $T_{V0}^{(1)} \geq_{\mathrm{cx}} V/c(1-\rho_1)$. Hence

$$W_2 \stackrel{\mathrm{d}}{=} T_{V0}^{(1)} \geq_{\mathrm{cx}} \frac{V}{c(1-\rho_1)} > \frac{V_2}{c(1-\rho_1)} \geq_{\mathrm{icx}} \frac{V_2^{\mathrm{r}}}{c(1-\rho_1)} = W_2^{\mathrm{r}}, \tag{6.10}$$

where we have used Theorem 6.1 in the penultimate step. □

The RSR approximation is not only a bound. It also arises as a special case in which class 1 input is a fluid or as a limit in which the class 1 input approaches a fluid input. This implies that the bound is sharp, i.e., is attained in some cases.

We now show that the resulting effective bandwidth approximation is optimistic.

Theorem 6.3. *In the general stationary model*

$$(1-\rho_1)c\theta = \psi_{S_2^r}(\theta) \le \psi_{S_2}(\theta) \quad \text{for all } \theta, \tag{6.11}$$

so that for the workload asymptotic decay-rates in Theorem 3.1 *are ordered by*

$$\theta_2^{*r} \ge \theta_2^*, \tag{6.12}$$

and for any $\eta_2^* > 0$, *the effective capacities are ordered by*

$$C_2^r \equiv \frac{-\psi_{S_2^r}(-\eta_2^*)}{\eta_2^*} \ge \frac{-\psi_{S_2}(-\eta_2^*)}{\eta_2^*} \equiv C_2. \tag{6.13}$$

Proof. The convex order $S_2^r \le_{cx} S_2$ used in the proof of Theorem 6.1 implies that $E\,e^{\theta S_2^r(t)} \le E\,e^{\theta S_2(t)}$ for all θ and t from which (6.11) follows immediately. In turn (6.12) and (6.16) follow easily from (6.11) and (4.13). □

If we use the RSR approximation, then the admission criteria in (4.14) become

$$\sum_{j=1}^{J_1} e_{1j}n_{ij} = \sum_{j=1}^{J_1} \frac{\psi_{A_{1j}}(\eta_1^*)}{\eta_1^*} n_{1j} \le c, \tag{6.14}$$

$$\sum_{j=1}^{J_2} e_{2j}n_{2j} = \sum_{j=1}^{J_2} \frac{\psi_{A_{2j}}(\eta_2^*)}{\eta_2^*} n_{2j} \le c(1-\rho_1), \tag{6.15}$$

where ρ_1 in (6.15) is the utilization of the J_1 class 1 sources, and the target parameters η_i^* are as in (4.7) with the constraints in (4.1) to be met for large b_i, $i = 1, 2$. Since $\rho_1 = \sum_{j=1}^{J_1} \rho_{1j} n_{1j}$, (6.15) can be written as

$$\sum_{j=1}^{J_1} c\rho_{1j}n_{1j} + \sum_{j=1}^{J_2} e_{2j}n_{2j} \le c. \tag{6.16}$$

The pair of constraints (6.14) and (6.16) form a linear feasible set.

7. A further lower bound and heavy tails

In this section we apply [8] to obtain a stochastic lower bound for V_2^r that enables us to obtain a sufficien condition for V_2 to have a heavy-tailed distribution. Following [8], let the low-priority input process b a general stochastic fluid input process determined by a stationary environment process $\{Z_2(t): t \ge 0\}$ We assume that the environment process spends alternating positive times $X_1, Y_1, X_2, Y_2, \ldots$ in state such that the input is above and below a high rate r_2. We assume that $\{(X_n, Y_n)\}$ is a stationary sequenc with $EX_n < \infty$ and $EY_n < \infty$.

Let G be the cumulative distribution function (cdf) of a high-activity period X_n and let $G^c(t) \equiv 1-G(t)$ be the associated complementary cdf (ccdf). Let G_e be the associated stationary excess cdf, defined by

$$G_e(t) = \frac{1}{EX_1}\int_0^t G^c(u)\,du, \quad t \geq 0. \tag{7.1}$$

Theorem 7.1 (from [8]). *Under the assumptions above, if* $r_2 > c(1-\rho_1)$, *then*

$$P(V_2^r > t) \geq F^c(t) \equiv \left(\frac{EX_1}{EX_1 + EY_1}\right) G_e^c\left(\frac{t}{r_2 - c(1-\rho_1)}\right), \tag{7.2}$$

so that

$$\limsup_{t\to\infty} \frac{P(V_2 > t)}{G_e^c(t/(r_2 - c(1-\rho_1)))} \geq \frac{EX_1}{EX_1 + EY_1} > 0. \tag{7.3}$$

Proof. Inequality (7.2) is Theorem 1 of [8]. Since $V_2 \geq_{icx} V_2^r$, we have

$$\int_t^\infty P(V_2 > u)\,du \geq \int_t^\infty P(V_2^r > u)\,du \quad \text{for all } t, \tag{7.4}$$

see [22, p. 8] which implies (7.3).

Property (7.3) can be interpreted as saying that the ccdf of V_2 has a heavier tail than the ccdf G_e^c. For example, if

$$\lim_{t\to\infty} t^\eta G_e^c(t) = \alpha, \tag{7.5}$$

where η and α are positive constants, then Theorem 7.1 implies that

$$\limsup_{t\to\infty} t^\eta P(V_2 > t) > 0. \quad \square \tag{7.6}$$

8. The empty-buffer upper bound

The empty-buffer bound is based on considering what the class 2 departure process would be if there were never any accumulation of class 1 workload, as would occur with continuous deterministic input with $\rho_1 < 1$. If class 1 never had workload, i.e., if $V_1(t) = 0$ for all t, then we would have $D_1(t) = A_1(t)$ and $S_2(t) = ct - A_1(t)$. Thus, we define the *empty-buffer bound* to be

$$S_2(t) \approx S_2^e(t) \equiv ct - A_1(t). \tag{8.1}$$

Suppose that we now consider the departure process starting out empty. In that case $D_1(t) \leq A_1(t)$, $t \geq 0$, so that

$$S_2(t) \geq S_2^e(t), \quad t \geq 0. \tag{8.2}$$

Indeed, by (2.2)–(2.4)

$$X_2^e(t) = A_1(t) + A_2(t) - ct, \quad t \geq 0, \tag{8.3}$$

so that the empty-buffer bound is equivalent to approximating the class 2 workload process by the aggregate workload, i.e.

$$V_2^e(t) = V(t) \equiv V_1(t) + V_2(t), \quad t \geq 0. \tag{8.4}$$

Hence, we have the following elementary comparison result.

Theorem 8.1. *In the general stationary model,* $V_2 \leq_{st} V_2^e = V$.

Proof. Consider the system starting out empty. Clearly the sample paths are ordered: $V_2(t) \leq V(t) = V_2^e(t)$ for all $t \geq 0$. Since stochastic order is preserved under convergence in distribution, the conclusion follows. □

The associated empty-buffer effective bandwidth (EBEB) approximation is also conservative. Paralleling Theorem 6.3, we have the following elementary result.

Theorem 8.2. *In the general stationary model*

$$\psi_{S_2^e}(\theta) \geq \psi_{S_2}(\theta) \quad \text{for all } \theta < 0,$$

so that the workload asymptotic decay-rates in Theorem 3.1 *are ordered by*

$$\theta_2^{*e} \leq \theta_2^*,$$

and for all $\eta_2^* > 0$, *the effective capacities are ordered by*

$$C_2^e \equiv \frac{-\psi_{S_2^e}(-\eta_2^*)}{\eta_2^*} = c - \frac{\psi_{A_1}(\eta_2^*)}{\eta_2^*} \leq c - \frac{\psi_{D_1}(\eta_2^*)}{\eta_2^*} = -\frac{\psi_{S_2}(-\eta_2^*)}{\eta_2^*} \equiv C_2. \tag{8.5}$$

At first glance, the empty-buffer bound may seem very crude, but it can be surprisingly accurate. It often happens that the bulk of the workload is low-priority work. Indeed, in support of the empty-buffer approximation, we point out that it is asymptotically exact as $\rho_2 \to 1 - \rho_1$ for any ρ_1 (in heavy traffic), see [25]. In that limit, the total workload is growing, being of order $O(1/(1-\rho))$, where $\rho = \rho_1 + \rho_2 \to 1$, while the class 1 workload remains unchanged. Hence there definitely are scenarios where the empty-buffer bound provides an excellent approximation.

Paralleling (6.14) and (6.15), the admission criteria with the empty-buffer approximation are (6.14) and

$$\sum_{j=1}^{J_2} e_{2j} n_{2j} = \sum_{j=1}^{J_2} \frac{\psi_{A_{2j}}(\eta_2^*)}{\eta_2^*} n_{2j} \leq c - \frac{\psi_{A_1}(\eta_2^*)}{\eta_2^*} = C_2^e. \tag{8.6}$$

Since

$$\frac{\psi_{A_1}(\eta_2^*)}{\eta_2^*} = \sum_{j=1}^{J_1} \frac{\psi_{A_{1j}}(\eta_2^*) n_{1j}}{\eta_2^*}, \tag{8.7}$$

the two constraints (6.14) and (8.6) are fully linear. Note that $\psi_{A_{1j}}(\eta_2^*)/\eta_2^*$ in (8.7) is similar to the effective bandwidth of a class 1 source of type j, except η_2^* is present as opposed to η_1^*. We call $\psi_{A_{1j}}(\eta_2^*)/\eta_2^*$ the

effective bandwidth of a $(1, j)$ source as seen by class 2, and denote it as e_{1j}^2. Thus, the admission criteria for the effective bandwidth empty-buffer approximation can be written as

$$\sum_{j=1}^{J_1} e_{1j} n_{1j} \le c, \tag{8.8}$$

$$\sum_{j=1}^{J_1} e_{1j}^2 n_{1j} + \sum_{j=1}^{J_2} e_{2j} n_{2j} \le c. \tag{8.9}$$

9. An illustrative example

The reduced service rate (RSR) and empty-buffer (EB) approximations provide upper and lower bounds, respectively, for the priority-2 effective capacity, C_2 (4.13). In particular, from (6.11), (6.13) and (8.5)

$$C_2^{\mathrm{r}} = (1 - \rho_1)c \ge C_2 \ge c - \frac{\psi_{A_1}(\eta_2^*)}{\eta_2^*} = C_2^{\mathrm{e}}. \tag{9.1}$$

Thus, the difference $C_2^{\mathrm{r}} - C_2^{\mathrm{e}}$ is an upper bound on the error if either C_2^{r} or C_2^{e} is used as an approximation for C_2. For a perspective on the size of this error, it is natural to normalize by the aggregate capacity c obtaining the normalized error bound denoted as

$$E \equiv \frac{(C_2^{\mathrm{r}} - C_2^{\mathrm{e}})}{c}. \tag{9.2}$$

From (9.1), E can be expressed as

$$E = \frac{\psi_{A_1}(\eta_2^*)}{c\eta_2^*} - \rho_1, \tag{9.3}$$

or equivalently from (3.1)

$$E = \frac{1}{c}\left[\frac{\psi_{A_1}(\eta_2^*)}{\eta_2^*} - \lim_{t\to\infty}\frac{A_1(t)}{t}\right]. \tag{9.4}$$

Note that the normalized error bound depends on the aggregate high-priority arrival process, $A_1(t)$, and the low-priority performance parameters represented by $\eta_2^* = -\log(p_2)/b_2$ (4.7). Also note that in the boundary case where the priority-1 arrival process is a constant rate fluid, E equals zero.

For the application to packet communication networks, one would like the normalized error bound to be less than the noise in the traffic model, $A_i(t)$. Often the traffic models deviate from reality by more than 10%, particularly if a forecast is involved. Thus, if E is less than 10% then the error from the RSR or EB approximations is within the noise of the model.

As a first example, consider the case of an ATM network where the high-priority class supports constant bit-rate (CBR) connections. As mentioned in Section 5, due to network jitter and the lack of synchronization across the connections, the superposition of the jittered CBR streams can be modeled, often conservatively, as a Poisson process. If $A_1(t)$ is a compound Poisson process with Poisson rate $c\rho_1$ and

component unit-size jumps (where a jump represents the arrival of an ATM cell, which has a constant size), then $\psi_{A_1}(\theta) = c\rho_1(e^\theta - 1)$ and

$$E = \rho_1 \left[\frac{e^{\eta_2^*} - 1}{\eta_2^*} - 1 \right] = \frac{\rho_1 \eta_2^*}{2} + O(\eta_2^{*2}). \tag{9.5}$$

For a particular example, if the priority-2 buffer threshold, b_2, is 500 and the probability that the work in system exceeds b_2 should be no more than $p_2 = 10^{-3}$, then η_2^* is 0.0138. If ρ_1 is 0.50, then E is 0.003, which is well within the noise of the traffic models.

As a second example, suppose the aggregate priority-1 arrival process is a two-state Markov modulated Poisson process (MMPP) where one state is on while the other is off, and hence the process is also equivalent to an interrupted Poisson process. The MMPP has rate matrix

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

and infinitesimal generator

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

and where each arrival adds one unit of work. The asymptotic decay-rate function can be expressed in closed form:

$$\psi_{A_1}(\theta) = \tfrac{1}{2}(-\alpha + \sqrt{\alpha^2 + 4\lambda_1 r_2(e^\theta - 1)}), \tag{9.6}$$

where $\alpha = r_1 + r_2 - \lambda_1(e^\theta - 1)$.

Table 1
Normalized error bound E, given priority-1 arrival process is an ON/OFF MMPP with mean rate 0.01, and various fraction of ON times and mean burst sizes, and given priority-2 performance parameters $p_2 = 10^{-3}$ and various buffer thresholds b_2

Fraction of time ON	Mean burst size	Buffer threshold, b_2	Normalized error bound, E
0.1	10	100	0.011
0.1	10	500	0.0013
0.1	10	1000	0.00062
0.1	100	100	0.079
0.1	100	500	0.030
0.1	100	1000	0.010
0.05	10	100	0.016
0.05	10	500	0.0015
0.05	10	1000	0.00070
0.05	100	100	0.17
0.05	100	1000	0.014
0.05	100	500	0.061
0.01	10	100	0.023
0.01	10	500	0.0017
0.01	10	1000	0.00076
0.01	100	100	0.88
0.01	100	500	0.29
0.01	100	1000	0.020

For a particular example, suppose that λ_1, r_1, and r_2 are specified by the mean rate of $A_1(t)$, $\lambda_1 r_2/(r_1+r_2)$, equaling 0.01, and the fraction of time on, $r_2/(r_1+r_2)$, equaling 0.1, 0.05, or 0.01, and the mean number of arrivals during an on period (mean burst size), λ_1/r_1 equaling 10 or 100, and the capacity $c = 1$. For this arrival process and for priority-2 performance parameters of $p_2 = 10^{-3}$ and $b_2 \in \{100, 500, 1000\}$, Table 1 reports the resulting normalized error bound, E. The parameter values were chosen to show where the RSR and EB approximations begin to perform poorly. When the mean burst size is as big as the buffer threshold, as when they both are 100, E is relatively large, particularly for the bursty case where the fraction of time on is only 1%. However, for low-priority traffic in packet networks, where significant queuing can be expected, the buffer should be an order of magnitude bigger than the mean burst size. For these cases, the normalized error bound is less than 10%, which is within the noise of typical traffic models.

The RSR and EB approximations for the effective capacity C_2 can be used to approximate the admissible sets for the number of priority-1 and priority-2 connections that can be admitted while satisfying the performance parameters. We use the RSR and EB approximations derived herein to examine these admissible sets in detail in [2]. As the RSR approximation gives an upper bound on C_2, it yields an optimistic approximation for the admissible set, and likewise since the EB approximation gives a lower bound on C_2, it yields a conservative approximation. When the priority-2 performance parameters are significantly looser than those for priority-1 (η_2^* an order of magnitude smaller than η_1^*), then for a given number of priority-1 connections, the RSR and EB estimates for the number of admissible priority-2 connections are often close — equaling the same integer value, or differing by 1 or 2.

References

[1] F. Baccelli, P. Brémaud, Elements of Queueing Theory, Springer, New York, 1994.
[2] A.W. Berger, W. Whitt, Effective bandwidths with priorities, IEEE/ACM Trans. Networking 6 (1998) 447–460.
[3] A.W. Berger, W. Whitt, Extending the effective bandwidth concept to networks with priority classes, IEEE Commun. Magazine 36 (1998) 78–83.
[4] A.A. Borovkov, Stochastic Processes in Queueing Theory, Springer, New York, 1976.
[5] L. Breiman, Probability, Addison-Wesley, Reading, MA, 1968.
[6] R. Cáceres, P.G. Danzig, S. Jamin, D.J. Mitzel, Characteristics of wide-area TCP/IP conversations, Comput. Commun. Rev. 21 (1991) 101–112.
[7] C.S. Chang, J.A. Thomas, Effective bandwidths in high-speed digital networks, IEEE J. Select. Areas Commun. 13 (1995) 1091–1100.
[8] G.L. Choudhury, W. Whitt, Long-tail buffer-content distributions in broadband networks, Perform. Eval. 30 (1997) 177–190.
[9] G. de Veciana, G. Kesidis, J. Walrand, Resource management in wide-area ATM networks using effective bandwidths, IEEE J. Select. Areas Commun. 13 (1995) 1081–1090.
[10] N.G. Duffield, N. O'Connell, Large deviations and overflow probabilities for the general single-server queue with applications, Math. Proc. Cambridge Philos. Soc. 118 (1995) 363–374.
[11] A.I. Elwalid, D. Mitra, Analysis, approximations and admission control of a multi-service multiplexing system with priorities, Proceedings of the IEEE Infocom '95, 1995, pp. 463–472.
[12] A. Feldmann, A.C. Gilbert, W. Willinger, T.G. Kurtz, The changing nature of network traffic: scaling phenomena, Comput. Commun. Rev. 28 (1998) 5–29.
[13] P.W. Glynn, W. Whitt, Ordinary CLT and WLLN versions of $L = \lambda W$, Math. Oper. Res. 13 (1988) 674–692.
[14] P.W. Glynn, W. Whitt, Logarithmic asymptotics for steady-state tail probabilities in a single-server queue, in: J. Gani, J. Galambos (Eds.), Studies in Applied Probability, Essays in Honour of Lajos Takács, Applied Probability Trust, Sheffield, 1994, pp. 131–156.
[15] F. Kelly, Notes on effective bandwidths, in: Stochastic Networks, Clarendon Press, Oxford, 1996, pp. 141–168.
[16] V.G. Kulkarni, N. Gautam, Admission control of multi-class traffic with service priorities in high-speed networks, Queueing Syst. 27 (1997) 79–97.

[17] W.E. Leland, M.S. Taqqu, W. Willinger, D.V. Wilson, On the self-similar nature of Ethernet traffic, IEEE/ACM Trans. Networking 2 (1994) 1–15.
[18] N.U. Prabhu, Stochastic Storage Processes, Springer, New York, 1980.
[19] A.A. Puhalskii, W. Whitt, Functional large deviation principles for waiting and departure processes, Probab. Eng. Inform. Sci. 12 (1998) 479–507.
[20] B.A. Rogozin, Some extremal properties in the theory of mass service, Theoret. Probab. Appl. 11 (1966) 44–151.
[21] K.W. Ross, Multiservice Loss Models for Broadband Telecommunication Networks, Prentice-Hall, Englewood Cliffs, NJ, 1995.
[22] D. Stoyan, Comparison Methods for Queues and Other Stochastic Models, Wiley, New York, 1983.
[23] A. Sugahara, T. Takine, Y. Takahashi, T. Hasegawa, Analysis of a nonpreemptive priority queue with SPP arrivals of high class, Perform. Eval. 21 (1995) 215–238.
[24] T. Takine, T. Hasegawa, The workload in the MAP/G/1 queue with state-dependent services: its application to a queue with preemptive resume priority, Stochastic Models 10 (1994) 183–204.
[25] W. Whitt, Weak convergence theorems for priority queues: preemptive resume discipline, J. Appl. Probab. 8 (1971) 74–94.
[26] W. Whitt, Some useful functions for functional limit theorems, Math. Oper. Res. 5 (1980) 67–85.
[27] W. Whitt, Tail probabilities with statistical multiplexing and effective bandwidths in multi-class queues, Telecommun. Syst. 2 (1993) 71–107.
[28] R.W. Wolff, Poisson arrivals see time averages, Oper. Res. 30 (1982) 223–231.
[29] J. Zhang, Performance study of Markov modulated fluid flow models with priority traffic, Proceedings of the IEEE Infocom '93, 1993, pp. 10–17.

Arthur W. Berger received the Ph.D. degree in applied mathematics from Harvard University, Cambridge, MA, in 1983. He joined AT&T Bell Laboratories and subsequently AT&T Labs and Lucent/Bell Labs. He is currently a Senior Research Scientist at Akamai Technologies, Cambridge, MA. This work was done while he was at AT&T Labs. He has worked in the areas of network planning, performance analysis of telecommunication switching systems, and congestion controls and traffic engineering for B-ISDN/ATM networks and for high-speed IP routers and networks. He has been active in ITU Study Groups 2 and 13, the US T1S1 Committee, and the ATM Forum. His research interests are in applied probability, and traffic controls and traffic engineering for communication networks. Dr. Berger is a member of the IEEE Communication Society and ACM SIGCOMM.

Ward Whitt received the Ph.D. degree in operations research from Cornell University, Ithaca, NY, in 1969. He was on the faculty of Stanford University in 1968–1969 and Yale University in 1969–1977. He joined AT&T Bell Laboratories in 1977 and then AT&T Laboratories in 1996. He is currently a member of the IP Network Management and Performance Department, Internet and Networking Systems Research Laboratory, AT&T Labs, Florham Park, NJ. His research has primarily been in queuing theory and its applications to telecommunication systems. Dr. Whitt is a member of the Institute for Operations Research and Management Sciences.