Queueing Systems 20 (1995) 37–59

Effective bandwidths: Call admission, traffic policing and filtering for ATM networks

G. de Veciana[a] and J. Walrand[b]

[a] *Department of Electrical and Computer Engineering, University of Texas at Austin, Austin, TX 78712, USA*
[b] *Department of Electrical Engineering and Computer Sciences, University of California at Berkeley, Berkeley, CA 94720, USA*

Received 23 August 1993; revised 19 December 1994

In this paper we review and extend the effective bandwidth results of Kelly [28], and Kesidis, Walrand and Chang [29, 6]. These results provide a framework for call admission schemes which are sensitive to constraints on the mean delay or the tail distribution of the workload in buffered queues. We present results which are valid for a wide variety of traffic streams and discuss their applicability for traffic management in ATM networks. We discuss the impact of traffic policing schemes, such as thresholding and filtering, on the effective bandwidth of sources. Finally we discuss effective bandwidth results for Brownian traffic models for which explicit results reveal the interaction arising in finite buffers.

Keywords: Communication networks, effective bandwidths, large deviations.

1. Introduction

One of the key ideas behind broadband integrated services digital networks (BISDN) using the asynchronous transfer mode (ATM) is the statistical multiplexing of heterogeneous packetized traffic streams and messages via switches and communication links. In order for streams to share resources one must guard against traffic fluctuations by inserting buffers. To ease the task of managing such a network it is desirable to obtain a circuit-switched model for which relatively simple call admission, routing, and network planning algorithms are available. For example, suppose a collection of sources, n_j of type $j \in J$ which require a bandwidth α_j, share a link with capacity c. One can easily check if bandwidth is available by considering whether

$$\sum_{j \in J} n_j \alpha_j \leq c.$$

Unfortunately, the interaction of traffic in networks is typically not linear in the number of sources nor is it usually decoupled across the different types of streams.

There exists, however, a remarkable collection of results for multi-type streams sharing a buffered queue for which an *effective bandwidth* and the accompanying linear constraint can be found such that particular criteria are satisfied. The goal herein is to discuss the structure required and limitations of such results for different criteria (quality of service), such as mean delay and overflow probability, for multi-class queues.

This problem has recently received much attention; in fact as of the writing of this paper much work has appeared. Below we provide a brief review of related work as a guide to the interested reader; a more complete account can be found in Whitt [34]. Among the first studies of effective bandwidths is an analysis of buffer-less systems by Hui [25]. The paper of Kelly [28], discussed herein, reported the first results for buffered systems subject to either mean delay or tail constraints. The work of Guérin, Ahmadi and Naghshineh [23] discussed the manner in which these methods could be incorporated in a framework for resource management.

Effective bandwidth results have been obtained via spectral expansions for Markov fluid traffic models by Gibbens and Hunt [21] and Elwalid and Mitra [17]. These methods have the advantage of providing explicit solutions to multiplexed systems and thus an understanding of the approximate nature of large buffer asymptotics.

An alternative approach has been to investigate large buffer asymptotics via the theory of large deviations. The work of Kesidis et al. [29] and Chang [6], identified under some rather general conditions the existence of effective bandwidths. The work of Glynn and Whitt [34, 22] independently encompasses much of the material herein and provides several additional clean results. Finally a novel result by Duffield and O'Connell [15], considers the case of traffic streams with long range correlations, or self-similar structure, where the scaling and results differ from those discussed in this paper. The work of Doshi [14] showed that by guaranteeing a performance constraint for a heterogeneous multiplexer one really makes weak (if any) promises to individual users. This important point led the further work by de Veciana and Kesidis [12] proving effective bandwidth results for systems using general processor sharing service policies and segregating i.i.d. traffic streams. An extension to feed-forward networks, via a study of the input-output properties of queues, is proposed in de Veciana, Courcoubetis and Walrand [11]. In addition recent numerical as well as analytical studies, e.g., K. Rege [31], G. Choudhury et al. [7] and Botvich and Duffield [4], show that asymptotic effective bandwidth results can be either optimistic or conservative depending on the nature of the arrival streams. These and other studies indicate that caution is warranted in using the effective bandwidth concept.

In this paper we begin by reviewing effective bandwidth results for criteria such as mean packet delay or the probability of large delays first considered by Kelly [28]. We discuss a simple extension of this result to a system with prioritized service and multiple average delay constraints.

Next, in section 3, we extend the approach of Kesidis et al. [29] and Chang [6]

both of whom used large deviations to obtain effective bandwidths where the criterion is the likelihood of a large workload or queue length in a discrete-time queue. We give a direct proof of this result including a large class of stationary ergodic mixing or Markov sources as well as random possibly dependent service times. Some novel examples where these results apply, such as randomized service priority, are presented. In addition we discuss the nature of streams where such results fail.

Packet admission policies which are optimal in the sense of reducing the effective bandwidth of sources are considered in section 4. Among memoryless policies with the same throughput, we show that thresholding is optimal for i.i.d. sources, but not necessarily optimal in general. While filtering the rate of a traffic stream would appear to reduce fluctuations and thus improve performance, a simple example shows that the effective bandwidth of a filtered stream remains unchanged unless a fraction of the traffic is rejected, i.e., the filter does not have unit gain. Further studies of the effective bandwidth at the output of a discrete-time queue and leaky bucket explicitly show how such systems can in fact reduce the effective bandwidth of a traffic stream at the expense of further delays, see de Veciana et al. [11, 10].

Finally in section 5 we discuss related approximations; namely, heavy traffic limits for which explicit solutions can be obtained exhibiting the effect of finite buffers.

2. Classical techniques

We begin by reviewing a result for a multi-class buffered resource of Kelly [28]. He considers a system with independent sources, say n_j streams of type $j \in J$. For sources of type j, bursts of traffic arrive as a Poisson process with rate ν_j; the length of each burst is arbitrary with mean μ_j and variance σ_j^2. The length of a burst is the required service time; the model corresponds to a first-come first-serve $M/GI/1$ queue. Note at the outset that this is not a particularly good model for ATM networks in which the packet size (and hence the service time) is fixed and where arrivals are highly correlated; it does however have some merit in a setting with variable length packets such as Frame Relay networking.

Using the Pollaczek–Khintchine formula one can find the distribution of the workload in the system and in particular the mean delay *before* service $\mathbb{E}D$ of typical customers:

$$\mathbb{E}D = \frac{\sum_{j \in J} n_j \nu_j (\mu_j^2 + \sigma_j^2)}{2\left(1 - \sum_{j \in J} n_j \nu_j \mu_j\right)}.$$

Kelly [28] considers a delay constraint $\mathbb{E}D < d$ which by rearranging terms gives the following linear constraint:

$$\sum_{j \in J} n_j \alpha_j(d) \leq 1, \quad \text{where } \alpha_j(d) = \nu_j \left[\mu_j + \frac{1}{2d} (\mu_j^2 + \sigma_j^2) \right].$$

We call $\alpha_j(d)$ the *effective bandwidth* of a call of class j subject to a bound, d, on the *mean* delay before service.

The trivial extension to constraints on the expected queue length or the mean sojourn time ($\mathbb{E}W$) is not fruitful. For example, in order to guarantee $\mathbb{E}W \leq w$ it suffices to insure $\mathbb{E}D + \mathbb{E}S \leq w$ where $\mathbb{E}S$ denotes the mean service time of customers. Thus a linear constraint is obtained by simply letting $d = w - \mathbb{E}S$ in the formulas above. Note, however, that $\mathbb{E}S$ depends explicitly on the proportion of calls of each type, hence, only by assuming this mix is approximately constant (or $w \gg \mathbb{E}S$) can we obtain a satisfactory effective bandwidth for the mean sojourn time. From a user's point of view, it suffices for the network to guarantee a mean delay before service since the user can then compute *his own* expected sojourn time. This simple case exemplifies the fact that in obtaining effective bandwidth formulae it is essential to select the criterion carefully.

Example 1

Figure 1 shows approximate admissible regions of operation for two types of sources sharing a 150 Mbps line. Type 1 sources have a mean traffic rate of 1 Mbps; the packets arrive according to a Poisson stream, have a mean service time $\mu_1 = 28.3\,\mu$s and $\sigma_1^2/\mu_1^2 = 2$. Type 2 sources have a mean traffic rate of 10 Mbps with $\mu_2 = 56.5\,\mu$s and $\sigma_1^2/\mu_1^2 = 1$. The graph on the left shows the admissible number of sources when the mean delay before service is less than $d = 0.1$ ms. The graph on the right shows the admissible region when the mean sojourn time is constrained to be less than $w = 0.1$ ms. As seen in fig. 1, a constraint on the mean sojourn time can lead to a nonlinear boundary which is, however, approximately linear for a large range of traffic mixes.

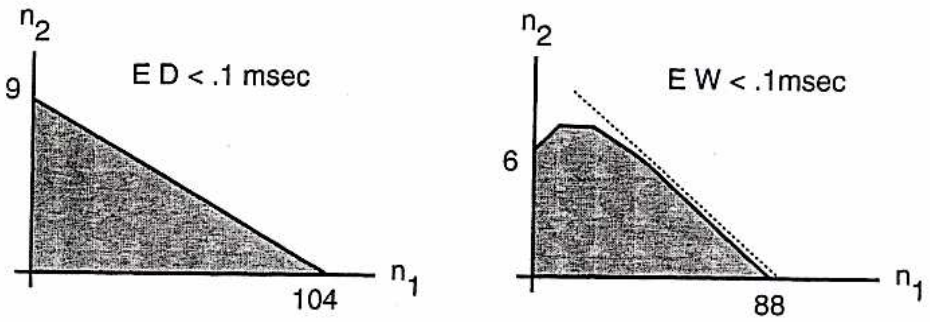

Fig. 1. Effective bandwidths and admissible regions.

In practice one might tradeoff delay characteristics for some types of traffic (voice and video) with that of others (e.g., data). Thus it is interesting to consider effective bandwidths for priority service policies; below we consider one such example.

For simplicity we discuss a 2-class $M/GI/1$ model with a non-preemptive service policy giving high priority to Type 1 traffic. Using Little's result one obtains the expected delay before service of the two types of traffic, $\mathbb{E}D_1$ and $\mathbb{E}D_2$, as a function of the traffic statistics and the number of sources of each type (see Walrand [33, p. 128]):

$$\mathbb{E}D_1 = \frac{\sum_{i=1}^{2} n_i \nu_i (\mu_i^2 + \sigma_i^2)}{2(1 - n_1 \nu_1 \mu_1)}, \quad \mathbb{E}D_2 = \frac{\sum_{i=1}^{2} n_i \nu_i (\mu_i^2 + \sigma_i^2)}{2\left(1 - \sum_{i=1}^{2} n_i \nu_i \mu_i\right)(1 - n_1 \nu_1 \mu_1)}.$$

Now suppose we require that $\mathbb{E}D_1 \leq d_1$ and $\mathbb{E}D_2 \leq d_2$, then the following conditions need to be satisfied:

$$n_1 \alpha_1(d_1) + n_2[\alpha_2(d_1) - \nu_2 \mu_2] \leq 1,$$
$$n_1 \alpha_1(\tilde{d}_2) + n_2 \alpha_2(\tilde{d}_2) \leq 1,$$

where $\tilde{d}_2 = d_2(1 - n_1 \nu_1 \mu_1)$ and $\alpha_j(\cdot)$ is as defined above.

This setup exhibits interaction among traffic streams with different priorities. As might have been expected, the delay constraint on high priority traffic gives rise to a linear constraint where the effective bandwidth of low priority traffic is reduced. Indeed, since Type 1 packets have priority they will only incur extra delays if on arrival a Type 2 packet has begun service. Since the probability of this event is linear in the number of low priority sources, the first constraint above is linear. The delay constraint on low priority traffic also results in a linear relationship, but with a reduced bound $\tilde{d}_2$ which unfortunately depends on the traffic intensity of Type 1 traffic. In principle this permits structured multiplexing of traffic streams subject to various *mean* delay constraints.

Example 2

Consider our previous example, but let the service policy give priority to Type 1 traffic, rather than the first-in-first-out policy assumed above. Suppose we constrain the mean delay before service for high priority traffic to be less than $d_1 = 0.1$ ms while delay constraints for Type 2 traffic are relaxed to $d_2 = 10$ ms. The admissible region defined by the above constraints is shown in fig. 2.

In practice one might further consider imposing a loss (or statistical delay) constraint on high priority traffic while maintaining an average delay constraint

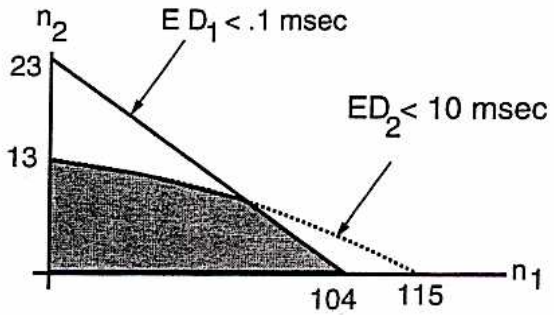

Fig. 2. Admissible region for priority service.

for low priority traffic. The effective bandwidth results for bounds on the tail distributions considered in the sequel can be used to control performance measures related to the tail distributions in a queueing system. Ideally a bound on loss for high priority traffic coupled with an average packet delay constraint for low priority traffic will define a region where we might wish to operate a multi-service system.

We now turn to Kelly's [28] effective bandwidth result for $M/GI/1$ and $D/G/1$ queues where a constraint of the type

$$\lim_{B\to\infty} \frac{1}{B} \log \mathbb{P}(W > B) \leq -\delta \tag{1}$$

is to be satisfied. In this expression B represents a large buffer size under which it is desirable to maintain the workload W and δ represents a statistical constraint on the tail distribution of the workload.

We first introduce a general result on the tail distribution of a $GI/GI/1$ queue, see Feller [18]. Let A denote a random variable distributed as an inter-arrival period, S a random variable distributed as a service time, and suppose there exists a solution k to

$$\mathbb{E} \exp[k(S - A)] = 1. \tag{2}$$

It can be shown that the distribution of interest is asymptotically exponential, i.e.,

$$\lim_{B\to\infty} \mathbb{P}(W > B) \exp[kB] = C, \tag{3}$$

where C is a constant that can be computed with some difficulty, see Iglehart [26].

Kelly used this result to obtain effective bandwidths for both $M/GI/1$ and $D/GI/1$ queues subject to the tail constraint in eq. (1). As above, for each type $j \in J$ let A_j be distributed as an inter-arrival, i.e., either exponential with parameter ν_j or deterministic, and let S_j denote the service time or batch arrivals per slot.

Referring to eqs. (2) and (3), note that the tail constraint in eq. (1) will be satisfied if we guarantee that $k \geq \delta$ and hence by monotonicity that

$$\mathbb{E}\exp[\delta(S-A)] \leq 1. \tag{4}$$

For the $M/GI/1$ model, where the aggregate inter-arrival A is exponential with parameter $\nu = \sum_{j\in J} n_j\nu_j$ and S is distributed as S_j with probability $p_j = n_j\nu_j/\nu$, Kelly shows that eq. (4) becomes

$$\sum_{j\in J} n_j\alpha_j(\delta) \leq 1, \quad \text{where} \quad \alpha_j(\delta) = \frac{\nu_j}{\delta}(\exp[\Lambda_j(\delta)] - 1),$$

where $\Lambda_j(\delta) = \log \mathbb{E}\exp[\delta S_j]$ is the log-moment generating function of S_j. For the $D/GI/1$ model, A is a deterministic time slot, say the time to serve one unit of work, and S be distributed as the aggregate work for the sources sharing the queue arriving during a time slot. The constraint in eq. (4) then becomes

$$\sum_{j\in J} n_j\alpha_j(\delta) \leq 1, \quad \text{where} \quad \alpha_j(\delta) = \frac{\Lambda_j(\delta)}{\delta}.$$

Scaling the service rate by a factor of c modifies the above inequality to

$$\sum_{j\in J} n_j\alpha_j(\delta) \leq c,$$

which parallels the bandwidth constraint considered in the introduction.

Note that the assumption of Poisson or slotted arrivals was necessary in dealing with the multiclass setting although the asymptotics on which the result is based can be obtained for $GI/GI/1$ and even $SM/GI/1$ (SM: semi-Markov) queues, see Karlin and Dembo [27]. The main problem in extending Kelly's argument is that a superposition of renewal or semi-Markov traffic streams usually will not preserve these properties.

To summarize, the effective bandwidth characterization gives a simple relationship which might be used for call management schemes which are sensitive to the tail distribution or mean workload in buffers. However, in the present setting, they only hold for a restricted collection of sources. Finally, note that Kelly's $D/GI/1$ model would be a reasonably good model for an output buffer in an ATM switch if dependencies in the arrival processes could be handled; this is one of the goals of the next section.

3. Large deviations

In this section we will establish effective bandwidth results for a wide class of sources subject to constraints on the tail probability of the workload or the buffer

occupancy in a discrete-time queue. The result is drawn from Kesidis et al. [29] and Chang [6]. We present a direct proof via large deviations and discuss some examples of randomized service.

We begin by reviewing the statement and possible requirements for large deviation results to hold. For a complete reference on the subject see Dembo and Zeitouni [13]. A sequence of measures $\{\mu_n\}$ on $\mathbb{R}$ will satisfy a Large Deviation Principle (LDP) with *good rate function*, $I(\cdot)$, if for every closed set F,

$$\limsup_{n\to\infty} \frac{1}{n} \log \mu_n(F) \leq - \inf_{x\in F} I(x),$$

and every open set G,

$$\liminf_{n\to\infty} \frac{1}{n} \log \mu_n(G) \geq - \inf_{x\in G} I(x),$$

and $\{x : I(x) \leq \alpha\}$ is compact for $\alpha < \infty$. We only consider the setting where $\{\mu_n\}$ denote the distributions of the partial sums $n^{-1}S_n = n^{-1}\sum_{i=1}^n X_n$ for a sequence of real-valued random variables $\{X_n\}$. We then say that $\{X_n\}$ satisfies an LDP with good rate function $I(\cdot)$. Below we briefly discuss when such bounds do indeed hold.

The Gärtner–Ellis Theorem establishes the existence of an LDP with convex good rate function for a large class of sources. The requirements are that:

1. The limits $\Lambda(\theta) \triangleq \lim_{n\to\infty}(1/n) \log \mathbb{E} \exp[\theta S_n]$ exist (possibly infinite) for all $\theta \in \mathbb{R}$;
2. The origin is in the interior D_Λ^o of the *effective domain* $D_\Lambda \triangleq \{\theta : \Lambda(\theta) < \infty\}$ of $\Lambda(\cdot)$;
3. $\Lambda(\cdot)$ is differentiable throughout D_Λ^o and *steep*, i.e., $\lim_{n\to\infty} |d\Lambda(\theta_n)/d\theta| = \infty$ whenever $\{\theta_n\}$ is a sequence in D_Λ^o converging to a boundary point of D_Λ^o.

Under conditions 1–3 an LDP holds with the good rate function given by the convex dual $\Lambda^*(\cdot)$ of $\Lambda(\cdot)$:

$$\Lambda^*(x) = \sup_\theta [\theta x - \Lambda(\theta)].$$

This result applies to i.i.d. sequences with $\mathbb{E}e^{\theta X_1} < \infty$ for all θ, which corresponds to the original large deviation estimate of Cramér. The result also applies to sequences with weak dependencies. For example, (random) coordinate functions of Markov chains satisfying strong uniformity conditions on the transition kernel and tails will satisfy an LDP. For stationary sequences satisfying appropriate mixing and tail conditions similar results hold.

THEOREM 3.1 [see Chang [6]]

Let $\{X_n\}$ be a stationary ergodic process with $\mathbb{E}X_n < 0$, which either satisfies an LDP with convex good rate function $I(\cdot)$, such that for all $\theta < \infty$

$$\Lambda(\theta) = \lim_{n\to\infty} \frac{1}{n} \log \mathbb{E} \exp\left[\theta \sum_{i=1}^{n} X_i\right] < \infty,$$

and $\Lambda^*(\cdot)$ is strictly convex or satisfies the requirements for the Gärtner–Ellis Theorem. Then the Lindley process

$$W_{n+1} = [W_n + X_n]^+$$

has a stationary distribution, say that of a random variable W, and for $\delta > 0$,

$$\Lambda(\delta) \leq 0 \Leftrightarrow \lim_{B\to\infty} \frac{1}{B} \log \mathbb{P}(W > B) \leq -\delta.$$

Proof

The stability condition, $\mathbb{E}X_n < 0$, guarantees the existence of a stationary distribution, see Loynes [30]. In particular, let

$$W_n^m = 0, \quad n \leq -m,$$
$$W_{n+1}^m = [W_n^m + X_n]^+, \quad n \geq -m,$$

then the distribution of W_0^m converges monotonically to that of W. Let $S_0 = 0$ and $S_n = \sum_{i=-n}^{-1} X_i$ for $n \geq 1$. Recall that W_0^m is given by

$$W_0^m = \max_{0\leq n\leq m} S_n. \tag{5}$$

Since the sequence $\{X_n\}$ is stationary and ergodic, the limits

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{E} \exp[\theta S_n] = \Lambda(\theta)$$

must exist. Moreover, by theorem 4.5.10 in [13], or directly from the Gärtner–Ellis Theorem, the rate function is in fact the convex dual of $\Lambda(\cdot)$, i.e.,

$$I(\alpha) = \Lambda^*(\alpha) = \sup_{\alpha} [\theta\alpha - \Lambda(\theta)].$$

Thus for $\epsilon > 0$ there is an n_ϵ such that

$$\forall n > n_\epsilon, \ \mathbb{E} \exp[\theta S_n] \leq \exp[(\Lambda(\theta) + \epsilon)n],$$

and it follows from eq. (5) that

$$\mathbb{E}\exp[\theta W_0^m] \le \sum_{n=0}^{m} E\exp[\theta S_n] \le \sum_{n=0}^{n_\epsilon} E\exp[\theta S_n] + \sum_{n>n_\epsilon} \exp[(\Lambda(\theta)+\epsilon)n].$$

Now, since the first sum is bounded, if $\Lambda(\theta) < -\epsilon$, we have that $\mathbb{E}\exp[\theta W_0^m] = C < \infty$, and it follows by the Chebyshev inequality that $\mathbb{P}(W_0^n > B) \le C\exp[-\theta B]$ so in fact

$$\limsup_{B\to\infty} \frac{1}{B}\mathbb{P}(W > B) \le -\theta \quad \text{as long as} \quad \Lambda(\theta) < 0. \tag{6}$$

On the other hand note that $\mathbb{P}(W > B) \ge \mathbb{P}(S_n > B)$, so by letting $n = \lfloor B/\alpha \rfloor$ for $\alpha > 0$ we find

$$\liminf_{B\to\infty} \frac{1}{B}\log\mathbb{P}(W > B) \ge \frac{1}{\alpha}\liminf_{n\to\infty} \frac{1}{n}\log\mathbb{P}\left(\frac{S_n}{n} > \alpha\right) \ge -\frac{\Lambda^*(\alpha)}{\alpha},$$

where the last inequality corresponds to the large deviations lower bound. We may select α giving the tightest bound

$$\liminf_{B\to\infty} \frac{1}{B}\log\mathbb{P}(W > B) \ge -\inf_{\alpha>0} \frac{\Lambda^*(\alpha)}{\alpha} = -k \quad \text{where in fact} \quad \Lambda(k) = 0. \tag{7}$$

Before arguing that $\Lambda(k) = 0$, we note that the optimizer α^* of eq. (7) is well defined. Indeed if $\Lambda(\theta) < \infty$ then $\lim_{|x|\to\infty} \Lambda^*(x)/|x| = \infty$, so α^* above makes sense (see Dembo and Zeitouni [13, p. 34]). Also note that the strict convexity of $\Lambda^*(\cdot)$ at α^* is equivalent to the differentiability of $\Lambda(\cdot)$. Alternatively if the Gärtner–Ellis theorem is in force, then the steepness and differentiability conditions guarantee

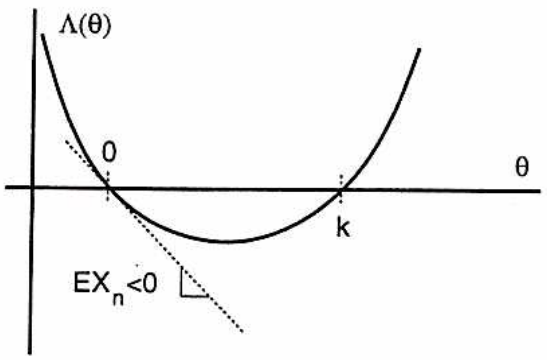

Fig. 3. Convexity of log-moment and asymptotic rate.

not only that α^* makes sense, but also the strict convexity of $\Lambda^*(\cdot)$ when the random variables are real-valued (see Ellis [16, p. 224]).

The first order optimality conditions require that

$$\frac{d\Lambda^*(\alpha^*)}{d\alpha}\alpha^* = \Lambda^*(\alpha^*), \quad \text{so} \quad k = \frac{\Lambda^*(\alpha^*)}{\alpha^*} = \frac{d\Lambda^*(\alpha^*)}{d\alpha}.$$

Recall that $\Lambda(\cdot)$ and $\Lambda^*(\cdot)$ are convex duals, and consider $\Lambda(k) = \sup_\lambda[\lambda k - \Lambda^*(\lambda)]$. Once again by differentiating we find that the supremum is attained at some λ^* such that $d\Lambda^*(\lambda^*)/d\alpha = k$. Our convexity requirement and the previous optimality criterion imply that $\lambda^* = \alpha^*$. Putting these results together we find that $\Lambda(k) = \alpha^* k - \Lambda^*(\alpha^*) = 0$.

Finally note that if $\delta > 0$ and $\Lambda(\delta) \leq 0$ then by convexity it follows that $\delta \leq k$, see fig. 3, so the result follows from the upper and lower bounds, eqs. (7), (6). □

Given this result it is now clear that an effective bandwidth result will hold in a multi-class setup as soon as $\Lambda(\delta) \leq 0$ is linear across the number of sources.

COROLLARY 3.1

Consider a collection of independent sources, n_j of each type $j \in J$, with slotted arrival processes $\{A_n^j\}$, each satisfying the conditions in theorem 3.1. Suppose they share a deterministic buffer with rate c according to a work conserving service policy. Then the following effective bandwidth result holds:

$$\sum_{j\in J} n_j \alpha_j(\delta) \leq c \quad \text{where} \quad \alpha_j(\delta) = \frac{\Lambda_j(\delta)}{\delta} \Leftrightarrow \lim_{B\to\infty} \frac{1}{B} \log \mathbb{P}(W \geq B) \leq -\delta,$$

and where W denotes the stationary workload.

Proof

Each source satisfies a large deviation principle where the limiting log-moment generating functions

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{E} \exp\left[\theta \sum_{i=1}^{n} A_i^j\right] = \Lambda_j(\theta),$$

exist and the rate function for each source is $\Lambda_j^*(\alpha_j) = \sup_\theta[\theta\alpha_j - \Lambda_j(\theta)]$. Let A_n denote the aggregate arrivals at time n and $X_n = A_n - c$ the net arrivals at this slot. Using the independence of the sources we find that the limit

$$\lim_{n\to\infty}\frac{1}{n}\log\mathbb{E}\exp\left[\theta\sum_{i=1}^{n}X_i\right]=\sum_{j\in J}n_j\Lambda_j(\theta)-c\theta=\Lambda(\theta)$$

exists, and by the contraction principle and convexity of the rate functions, the aggregate satisfies a large deviation principle with good rate function given by (see Dembo and Zeitouni [13, p. 110]):

$$I(\alpha)=\inf_{\sum_{j\in J}n_j\alpha_j=\alpha+c}\sum_{j\in J}n_j\Lambda_j^*(\alpha_j).$$

The corollary follows from the previous theorem and the independence of the sources,

$$\Lambda(\delta)\le 0\Leftrightarrow\sum_{j\in J}n_j\frac{\Lambda_j(\delta)}{\delta}\le c. \qquad \square$$

The usefulness of this result is predicated on being able to compute or estimate (possibly on-line) the effective bandwidth of a source. For a summary of some analytical formulae that are available, see Kesidis et al. [29] and Courcoubetis and Weber [8]. These include the usual i.i.d. sources, as well as Markov modulated fluids or Poisson processes and Gaussian processes.

One can also extend Kelly's $M/GI/1$ model to sources with possibly dependent service times.

COROLLARY 3.2

Consider a collection of independent sources, n_j of each type $j\in J$, such that a source of type j has Poisson packet arrivals (rate ν_j) with possibly dependent associated service times $\{S_n^j\}$ satisfying a large deviation principle. Suppose they share a buffer with any work conserving policy. Then the following effective bandwidth result holds:

$$\sum_{j\in J}n_j\alpha_j(\delta)\le 1 \text{ with } \alpha_j(\delta)=\frac{\nu_j}{\delta}(\exp[\Lambda_j(\delta)]-1)\Leftrightarrow\lim_{B\to\infty}\frac{1}{B}\log\mathbb{P}(W\ge B)\le-\delta,$$

where W denotes the stationary workload ahead of a typical packet.

Proof

Once again we use our main theorem where $X_i=S_i-A_i$, i.e., A_i denotes the aggregate inter-arrival time, so it is Poisson with rate $\nu=\sum_{j\in J}n_j\nu_j$ and S_i is the work corresponding to the ith arrival which corresponds to a particular stream of type $j\in J$ with probability ν_j/ν. As in the previous corollary, the condition

$$\Lambda(\delta) = \lim_{n\to\infty} \frac{1}{n} \log \mathbb{E} \exp\left[\delta \sum_{i=1}^{n} (S_i - A_i)\right] \leq 0,$$

gives the desired result. Since inter-arrival times are exponential and independent, we find that

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{E} \exp\left[-\delta \sum_{i=1}^{n} A_i\right] = \log\left[\frac{\nu}{\nu+\delta}\right],$$

i.e., the log of the Laplace transform for an exponential inter-arrival with rate ν. After some work the limit corresponding to the arriving work can also be simplified to,

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{E} \exp\left[\delta \sum_{i=1}^{n} S_i\right] = \log\left[\sum_{j\in J} n_j \frac{\nu_j}{\nu} \exp\left[\Lambda_j(\delta)\right]\right],$$

where $\Lambda_j(\delta) = \lim_{n\to\infty}(1/n)\mathbb{E}\exp\left[\delta \sum_{i=1}^{n} S_i^j\right]$. The condition $\Lambda(\delta) \leq 0$ can then be rewritten as

$$\sum_{j\in J} n_j \frac{\nu_j}{\delta} (\exp\left[\Lambda_j(\delta)\right] - 1) \leq 1. \qquad \square$$

Note that the two corollaries are essentially the same. Indeed the asymptotic log-moment generating function of incoming work per unit time for a stream of type j is that of a compound Poisson process, i.e.,

$$\Lambda_j^c(\delta) = \log\left[e^{\nu_j(\exp\left[\Lambda_j(\delta)\right]-1)}\right].$$

Thus assuming we serve at unit rate the effective bandwidth result in corollary 3.1 applies with

$$\alpha_j(\delta) = \frac{\Lambda_j^c(\delta)}{\delta} = \frac{\nu_j}{\delta} (\exp\left[\Lambda_j(\delta)\right] - 1).$$

Until now we have focused on modeling the variability in sources while assuming deterministic service processes. The generality of theorem 3.1 allows us to consider randomness in the service process and thus to obtain constraints which are sensitive not only to source fluctuations, but also to fluctuations at the server. For example, corollary 3.1 is easily extended to the case where the service process is independent of the arrivals and satisfies a large deviation principle. In this case we find the same effective bandwidths obtained previously, but the capacity c is

modified to reflect the randomness in the server as well as the tail constraint. We present two simple examples of servers with slotted arrivals which should elucidate this and other applications.

Consider a multi-class slotted model where the service rate is no longer deterministic. Suppose for example, that due to interference with concurrent processes the output bandwidth is modeled by an auto-regressive Gaussian process centered at c:

$$C_{n+1} = aC_n + N_{n+1}, \quad \text{where } |a| < 1,$$

and N_n is a white Gaussian process with power σ^2. It follows from the Gärtner–Ellis theorem that C_n satisfies a large deviation principle. In fact the asymptotic log-moment generating function of the service process $\{c + C_n\}$ is

$$\Lambda_c(\theta) = c\theta + \frac{\theta^2\sigma^2}{2(1-a)^2}$$

(see Bucklew [5, p. 22]). In order to satisfy a δ constraint on the tail we need only require (see theorem 3.1):

$$\sum_{j\in J} n_j \alpha_j(\delta) \leq c - \delta \frac{\sigma^2}{2(1-a)^2}.$$

The risk associated with fluctuations in the service results in a reduced service capacity which depends in a natural way on the variance of the noise and the autocorrelation between noise samples.

Suppose that in addition to multi-class sources we specify high and low priority traffic types, J_h and J_l respectively, which are queued in segregated buffers. In order to reduce large delays for high priority streams, we choose a randomized service policy which is biased towards high priority packets with probability $p_h > 0.5 > p_l$. Thus at each time slot the server flips a biased coin selecting the priority type to be processed at rate c. Note that this policy is not work conserving, i.e., service may be assigned to a priority with no work to be done. We obtain effective bandwidth constraints for high and low priority traffic:

$$\sum_{j\in J_h} n_j \alpha_j(\delta) \leq -\frac{\log(p_l + p_h \exp[-\delta c])}{\delta},$$

$$\sum_{j\in J_l} n_j \alpha_j(\delta) \leq -\frac{\log(p_h + p_l \exp[-\delta c])}{\delta},$$

where $\alpha_j(\cdot)$ denote the effective bandwidths obtained for sources obtained in

corollary 3.1. Since these constraints are decoupled we can envisage choosing different tail constraints (δ) for the two priorities.

Given the rather abstract conditions for the existence of effective bandwidths presented above, one might ask which types of sources will not have an effective bandwidth. This question is closely related to the manner in which overflows occur in queues, see Anantharam [1]. For a $GI/GI/1$ queue in which the distribution of $X = S - A$ (difference of the service time and inter-arrivals periods) has an exponential tail, one can show that overflows in asymptotically large buffers will occur as an accumulation of traffic over a large period of time, i.e., a large deviation in the empirical net input rate. If however X does not have an exponential tail, for example

$$\mathbb{E}X^2 < \infty \text{ and there is a } q > 0 \text{ s.t. } \mathbb{P}(X > x) = x^{-q}L(x),$$

where $L(x)$ is a slowly varying function, delays will build up suddenly, e.g., when a *single* customer with a huge excess service time arrives rather than as long term accumulation. This type of behavior does not fall in the traditional large deviations framework. Similarly the long range dependencies in self-similar traffic models are such that overflow asymptotics need to be viewed on a different time scale, leading to modified tail behavior, see Duffield and O'Connell [15]. Traffic streams without sufficient randomness are excluded from our framework, however, Chang [6] has developed an interesting point of view unifying stochastic and deterministic sources via the notion of envelope processes.

4. On cell admission and filtering

It is reasonable to ask how packet admission policies might decrease the effective bandwidth of a source. Consider a single arrival process $\{A_n\}$ and *memoryless* policies $h(\cdot)$, which reject (or set to low priority) some fraction of the arrivals. If A_n packets arrive at time n, we allow $h(A_n)$ to go through unchanged and reject or lower the priority of the remaining $A_n - h(A_n)$. Intuitively it is plausible that a threshold function $h^*(a) = \min[a, T]$, for some T, may be optimal among some collection of policies. In fact we will show that this is true if we consider all such policies with the *same* throughput μ and if arrivals are i.i.d. but may not hold otherwise. The following result was inspired by a problem concerning optimal re-insurance of policies, see Asmussen [2, p. 287].

PROPOSITION 4.1

Suppose $\{A_n\}$ is an i.i.d. sequence satisfying a large deviation principle. Consider all memoryless rejection policies, $h(\cdot)$, with the same throughput μ, i.e., such that $\mathbb{E}h(A_n) = \mu \leq \mathbb{E}A_n$. Let $h^*(a) = \min[a, T]$, where T is determined by $\mathbb{E}h^*(A_n) = \mu$. Among these policies, the one which results in the smallest effective bandwidth is h^*.

Proof

Note that $\{h(A_n)\}$ and $\{h^*(A_n)\}$ also satisfy large deviation principles where $\Lambda_h(\theta) = \log \mathbb{E} \exp[\theta h(A_0)]$ and $\Lambda_{h^*}(\theta) = \log \mathbb{E} \exp[\theta h^*(A_0)]$ are the corresponding log-moment generating functions. As seen in corollary 3.1, the effective bandwidth of these sources will be $\alpha_h(\delta) = \Lambda_h(\delta)/\delta$ and $\alpha_{h^*}(\delta) = \Lambda_{h^*}(\delta)/\delta$, respectively. We wish to show that $\alpha_h(\delta) \geq \alpha_{h^*}(\delta)$, so it suffices to show $\Lambda_h(\delta) \geq \Lambda_{h^*}(\delta)$. Since $e^z \geq 1 + z$, by letting $z = \delta[h(A_0) - h^*(A_0)]$ we have that

$$e^{\delta h(A_0)} \geq e^{\delta h^*(A_0)} + \delta e^{\delta h^*(A_0)}[h(A_0) - h^*(A_0)] \geq e^{\delta h^*(A_0)} + \delta e^{\delta T}[h(A_0) - h^*(A_0)],$$

where we use the fact that if $h(A_0) \geq h^*(A_0)$ then $h^*(A_0) = T$. Now taking expectations on both sides we have that $\mathbb{E}e^{\delta h(A_0)} \geq \mathbb{E}e^{\delta h^*(A_0)}$ since $\mathbb{E}[h(A_0) - h^*(A_0)] = 0$, and it follows that $\Lambda_h(\delta) \geq \Lambda_{h^*}(\delta)$. □

This result is perhaps not as surprising as the observation that it will not hold for arbitrary sources. When there are dependencies in the arrival process the optimal $h(\cdot)$ may reflect the dynamics of the process. Before considering an example of such a source, let us roughly examine where the previous argument fails.

Consider once again h and h^*, with the same throughput μ and an arbitrary source $\{A_i\}$ satisfying a large deviations principle. As seen above, it would suffice to show that in fact $\Lambda_h(\delta) \geq \Lambda_{h^*}(\delta)$, where these are now the asymptotic log-moment generating functions, e.g.,

$$\Lambda_h(\delta) = \lim_{n \to \infty} \frac{1}{n} \log \mathbb{E} \exp\left[\delta \sum_{i=1}^{n} h(A_i)\right].$$

To roughly understand the behavior of this limit, suppose we could show a central limit result for the given h:

$$\frac{\sum_{i=1}^{n} h(A_i) - n\mu}{\sqrt{n}} \to N(0, \sigma_h^2).$$

Thus $\sum_{i=1}^{n} h(A_i)$ is approximately normally distributed, say $N(n\mu, n\sigma_h^2)$. Taking the limit and log-moment generating function of this distribution, we obtain

$$\Lambda_h(\delta) \approx \delta\mu + \frac{\sigma_h^2 \delta^2}{2},$$

and of course the counterpart for h^*,

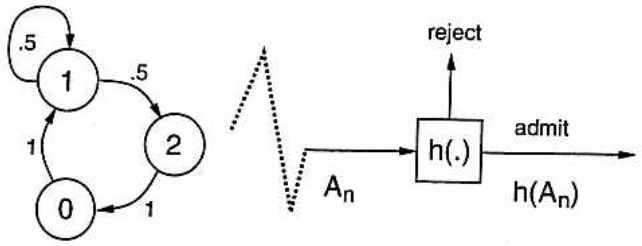

Fig. 4. A source for which thresholds are not optimal.

$$\Lambda_{h^*}(\delta) \approx \delta\mu + \frac{\sigma_{h^*}^2 \delta^2}{2}.$$

Thus h^* would be optimal if for all other h we had $\sigma_h \geq \sigma_{h^*}$. The problem is that σ_h is a function of both $h(\cdot)$ and the dependencies in the source. The goal of an optimal policy would be to reduce the asymptotic variance.

The Markov fluid source shown in fig. 4 is an example of a traffic stream for which the threshold policy is not optimal. The amount of work arriving in each slot will be the label of the state, i.e., 0, 1 or 2. The steady state distribution of this chain is $(\frac{1}{4}, \frac{1}{2}, \frac{1}{4})$, so the mean arrival rate is 1. We will consider memoryless rejection policies $h(\cdot)$ with a throughput of $\frac{1}{2}$, so that $\frac{1}{2}h(1) + \frac{1}{4}h(2) = \frac{1}{2}$. Among these there exists one threshold policy which we denote by $h^*(a) = \min[a, \frac{2}{3}]$. The effective bandwidth of this source can be computed to be

$$\alpha_h(\delta) = \frac{\log(\mathrm{sp}[\phi(\delta)P])}{\delta},$$

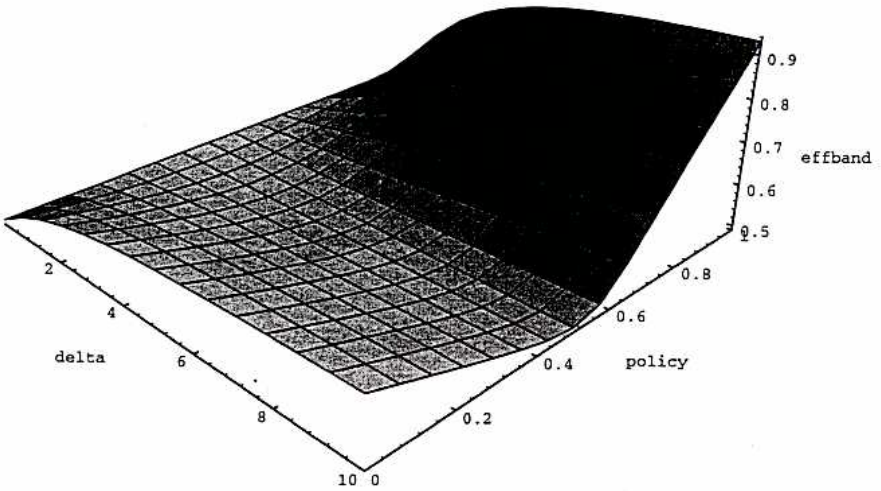

Fig. 5. Effective bandwidth versus tail constraint and admission policy.

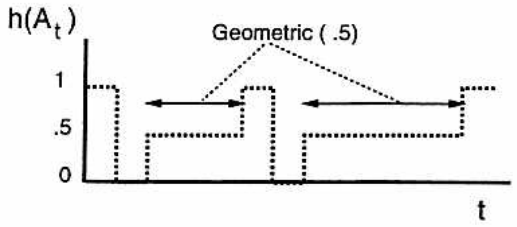

Fig. 6. Sample paths for optimal policy.

where sp $[\phi(\delta)P]$ denotes the spectrum of the product of the transition matrix, P, and a diagonal matrix $\phi(\delta)$ with components $(1, \exp[\delta h(1)], \exp[\delta h(2)])$. Figure 5 shows the effective bandwidth for a range of tail constraints δ over all memoryless policies with a throughput $\frac{1}{2}$; they are parametrized by the value of $h(1)$, where $0 \leq h(1) \leq 1$ and $h(2) = 2[1 - h(1)]$. Clearly $h(1) = 0.5(h(2) = 1)$ is the optimal admission policy since the effective bandwidth is minimal and equal to the throughput 0.5. This somewhat surprising result becomes obvious when one considers the sample paths of the source when this policy is used, see fig. 6. Indeed, the arrivals alternate almost deterministically between the levels 0, 0.5, 1, staying at levels 0 and 1 for a single time slot and at 0.5 for a geometrically distributed number of slots. The deviant behavior for this source may modify the amount of time spent at state 0.5, but this will not significantly affect the average traffic rate of the stream. This explains why the effective bandwidth remains constant for all constraints δ.

Although the notion of optimality, in the sense of minimizing the effective bandwidth for a given throughput, is reasonable, in practice one would further like to reduce the number of correlated losses. Indeed, while some sources (e.g., packetized voice and video) can tolerate loss, consecutive losses can lead to a degradation in the quality of service. Thus even an optimal memoryless policy may be imperfect in practice. The proper formulation is to minimize the effective bandwidth subject to a quality of service constraint, which might reflect the sensitivity of the source to losses. For example, recent detailed studies for variable bit rate video traffic consider the dynamics of loss and traffic policing schemes, see Reininger and Raychaudhuri [32]. In particular, a coder may adapt the level of quantization when the traffic rate exceeds a threshold, and thereby improve the overall performance, while maintaining the traffic within negotiated rate constraints.

We complete this discussion of admission policies with an insightful example suggested by Courcoubetis and Weber [9]. Consider a stationary Gaussian arrival process, $\{A_n\}$ with mean μ, and finite asymptotic variability

$$\sigma^2 = \lim_{n\to\infty} \frac{1}{n} \operatorname{Var}\left(\sum_{i=1}^{n} A_i\right) < \infty.$$

In this case one can show that the effective bandwidth of the source is given by

$$\alpha_A(\delta) = \mu + \frac{\delta\sigma^2}{2}.$$

We will denote the spectral density of the arrival process by $A(f) = \sum_{n=-\infty}^{\infty} e^{in2\pi f} R(n)$ where $R(n) = \operatorname{Cov}(A_i, A_{i+n})$ is the covariance function, and note that in fact $A(0) = \sigma^2$. It is reasonable to consider filtering the source in order to reduce loss. In fact, we will consider all filters $H(f)$ with the same dc gain, $H(0) = G \leq 1$, so that the throughput $G\mu$ is a fraction of the mean arrival rate. The spectrum of the output process will be $D(f) = |H(f)|^2 A(f)$ and an asymptotic variability $D(0) = |H(0)|^2 A(0) = G^2\sigma^2$. For a *fixed* dc gain, the effective bandwidth of the output process,

$$\alpha_D(\delta) = G\mu + G^2 \frac{\delta\sigma^2}{2},$$

is independent of the filter. Intuitively, large buffer asymptotics correspond to averaging over long periods of time, which in turn supersede the smoothing effect of the filter. Note however, that by choosing to reject a fraction of the input traffic, in some cases a significant (almost quadratic) reduction of the effective bandwidth can be obtained. One would expect these conclusions to be approximately true for non-Gaussian sources. Berger and Whitt [3] come to similar conclusions for the popular leaky bucket scheme, showing that asymptotic variability of the output traffic stream depends only on the total capacity of the system, i.e., what is lost, rather than the relative size of the job versus the token buffer. A further study by de Veciana et al. [11, 10] characterizes the effective bandwidth at the output process from such devices showing that though this second order property is invariant the effective bandwidth as a whole is in fact modified.

5. Heavy traffic approximations

Heavy traffic approximations provide further approximate effective bandwidth results for the mean workload and tail distribution as well as some additional insight.

In their study of dependencies in packet queues Fendick et al. [19, 20] consider superpositions of Poisson streams with batch arrivals. In particular suppose that a traffic stream in class $j \in J$ consists of batch arrivals with mean m_j and squared coefficient of variation c_{bj}^2 at rate ν_j; the packet service times are i.i.d.

with mean μ_j and squared coefficient of variation c_{sj}^2. Service is provided by a single server with a first-come first-serve discipline. In this case one can show the mean workload in the system is

$$\mathbb{E}D = \frac{\sum_{j\in J} n_j \nu_j (m_j \mu_j^2 c_{sj}^2 + m_j^2 \mu_j^2 (c_{bj}^2 + 1))}{2\left(1 - \sum_{j\in J} n_j \nu_j m_j \mu_j\right)}.$$

As in section 2, by rearranging terms in the constraint $\mathbb{E}D < d$, the effective bandwidth of a batch arrival stream subject to a mean delay before service less than d can be defined:

$$\sum_{j\in J} n_j \alpha_j(d) \le 1, \quad \text{where} \quad \alpha_j(d) = \nu_j \left[m_j \mu_j + \frac{1}{2d} (m_j \mu_j^2 c_{sj}^2 + m_j^2 \mu_j^2 (c_{bj}^2 + 1)) \right].$$

One might ask if this result generalizes when the arrivals are not Poisson but renewal, the batches are not instantaneous but spaced, or if the inter-arrival spacing and batches are dependent. These cases have been analyzed in the heavy traffic regime by Fendick and Whitt [19]. Their results give approximate effective bandwidths for superpositions of such streams subject to a mean delay constraint. A more complete discussion of heavy traffic approximations in this context can be found in Whitt [34].

For illustrative purposes let us further consider the heavy traffic approximation to a discrete-time multi-class deterministic queue with service rate c. Suppose packets in a stream of a given class $j \in J$, have stationary arrivals $\{A_n^j\}$, with mean μ_j^{-1}. Let $A_{1,t}^j$ denote the cumulative arrivals up to time t, and suppose the arrival process satisfies a central limit theorem such that $n^{-1/2}[A_{1,nt}^j - \mu_j nt] \to N(0, \sigma_j^2)$. Consider scaling the net input X_t, as $X_{nt}/\sqrt{n}$, such that $\sum_{j\in J} n_j \mu_j - c = \alpha/\sqrt{n}$. In the limit, as $n \to \infty$, the scaled workload converges weakly to a regulated Brownian motion with mean drift α and variance $\sigma^2 = \sum_{j\in J} n_j \sigma_j^2$, so that

$$\frac{X_{nt}}{\sqrt{n}} \xrightarrow{w} \sigma B_t + \alpha t.$$

In this regime Harrison's [24] results for regulated Brownian flows apply. When $\alpha < 0$, the steady state distribution of the workload, denoted by the random variable W, is exponential with mean

$$\frac{1}{\lambda} = \mathbb{E}W = \frac{\sum_{j\in J} n_j \sigma_j^2}{2|\alpha|}.$$

Thus, when the system operates in heavy traffic, using the fact that $\sqrt{n} = \alpha / [\sum_{j\in J} n_j \mu_j - c]$ we can unravel our scaling to find that

$$X_t \approx \sigma B_t + \left[\sum_{j\in J} n_j \mu_j - c\right] t.$$

By imposing a tail constraint on the exponentially distributed workload for the unscaled process, $\mathbb{P}(W > B) \leq \exp[-\delta B]$, we obtain the following approximate requirement:

$$\sum_{j\in J} n_j \left[\mu_j + \frac{\delta \sigma_j^2}{2}\right] \leq c.$$

This expression corresponds to a second order version of our original effective bandwidth result for tail constraints, see corollary 3.1. Indeed, if the effective bandwidths are differentiable, as will be the case if the arrival rates are bounded, then

$$\sum_{j=1}^{J} n_j \alpha_j(\delta) \approx \sum_{j=1}^{J} n_j \left[\mu_j + \frac{\delta \sigma_j^2}{2}\right] + o(\delta^2),$$

where $\mu_j = \mathbb{E}A_0^j$, and $\sigma_j^2 = \lim_{n\to\infty} t^{-1} \operatorname{Var}(A_{1,t}^j)$ are the mean and asymptotic variability of the arrival streams. This result is of course exact for Gaussian processes. It is tempting to use simple second order approximations if the errors introduced are insignificant. This issue must however be addressed via simulation. As in previous cases, the precision of this bound will depend on the types of sources and the load on the system.

The explicit results for buffered Brownian flows give us a unique opportunity to investigate the effective bandwidth concept for finite storage systems. As above, we suppose the net input can be modeled as a Brownian flow with drift $\mu = \sum_{j\in J} n_j \mu_j$ and variance $\sigma^2 = \sum_{j\in J} n_j \sigma_j^2$. In this case the mean workload $\mathbb{E}W$ is given by

$$\mathbb{E}W = -\frac{\sigma^2}{2\alpha} + \frac{B}{1 - \exp[-2\alpha B/\sigma^2]},$$

(see Harrison [24, p. 90]). Although α and σ^2 are linear in the number of sources, the presence of an exponential nonlinearity couples the traffic streams for finite buffers.

As $B \to \infty$, for $\alpha < 0$ we find an effective bandwidth result for a mean workload constraint of the form $\mathbb{E}W < d$. Specifically,

$$\sum_{j \in J} \alpha_j(d) \leq 1, \quad \text{where} \quad \alpha_j(d) = \mu_j + \frac{\sigma_j^2}{2d},$$

which is analogous to the results discussed in section 2.

References

[1] V. Anantharam, How large delays build up in a GI/G/1 queue, Queueing Systems 5 (1988) 345–368.

[2] S. Asmussen, *Applied Probability and Queues* (Wiley, 1987).

[3] A.W. Berger and W. Whitt, The impact of a job buffer in a token-bank rate-control throttle, Stoch. Models 8 (1992) 685–717.

[4] D.D. Botvich and N.G. Duffield, Large deviations, the shape of the loss curve, and economies of scale in large multiplexers, Technical Report DIAS-APG-94-12, Dublin Institute for Advanced Studies (1994).

[5] J.A. Bucklew, *Large Deviation Techniques in Decision, Simulation and Estimation* (Wiley, New York, 1990).

[6] C.S. Chang, Stability, queue length and delay of deterministic and stochastic queueing networks, IEEE Aut. Contr. 39 (1994) 913–931.

[7] G. Choudhury, D. Lucantoni and W. Whitt, Squeezing the most out of ATM, submitted (1993).

[8] C. Courcoubetis and J. Walrand, Note on effective bandwidth of ATM traffic, preprint (1991).

[9] C. Courcoubetis and R. Weber, Effective bandwidths for stationary sources, preprint (1992).

[10] G. de Veciana, Leaky buckets and optimal self-tuning rate control, *Globecom '94 Proc.*, San Francisco (1994) pp. 1207–1211.

[11] G. de Veciana, C. Courcoubetis and J. Walrand, Decoupling bandwidths for networks: A decomposition approach to resource management for networks, *IEEE Infocom Proc. '94*, also submitted to IEEE/ACM Trans. Networking (1993).

[12] G. de Veciana and G. Kesidis, Bandwidth allocation for multiple qualities of service using generalized processor sharing, Technical Report SCC-94-01, U.T. Austin, ECE Department, also in *Globecom '94 Proc.*, San Francisco (1994) pp. 1550–1554.

[13] A. Dembo and O. Zeitouni, *Large Deviations Techniques and Applications* (Jones and Bartlett, Boston, 1992).

[14] B.T. Doshi, Deterministic rule based traffic descriptors for broadband ISDN: Worst case behavior and concept equivalent bandwidth, *Globecom* (1993) pp. 1759–1764.

[15] N.G. Duffield and N. O'Connell, Large deviations and overflow probabilities for the general single-server queue, with applications, Technical Report DIAS-APG-93-30, Dublin Institute for Advanced Studies (1993).

[16] R.S. Ellis, *Entropy, Large Deviations and Statistical Mechanics* (Springer, 1985).

[17] A.I. Elwalid and D. Mitra, Effective bandwidth of general Markovian traffic sources and admission control of high speed networks, IEEE/ACM Trans. Networking 1(4) (1993).

[18] W. Feller, *An Introduction to Probability and Its Applications*, Vols. 1, 2 (Wiley, 1971).

[19] K. Fendick, V. Saksena and W. Whitt, Dependence in packet queues, IEEE Trans. Commun. 37 (1989).

[20] K. Fendick and W. Whitt, Measurements and approximations to describe the offered traffic and predict the average workload in a single-server queue, Proc. IEEE 77 (1989).

[21] R.J. Gibbens and P.J. Hunt, Effective bandwidths for the multi-type UAS channel, Queueing Systems 9 (1991) 17–28.
[22] P. Glynn and W. Whitt, Large deviations behavior of counting processes and their inverses, Queueing Systems 17 (1994) 107–128.
[23] R. Guérin, H. Ahmadi and M. Naghshineh, Equivalent capacity and its application to bandwidth allocation in high-speed networks, IEEE JSAC 9 (1991) 968–981.
[24] M. Harrison, *Brownian Motion and Stochastic Flow Systems* (Krieger, 1990).
[25] J.Y. Hui, Resource allocation for broadband networks, IEEE JSAC 6 (1988) 1598–1608.
[26] D.L. Iglehart, Extreme values in the GI/G/1 queue, Ann. Math. Stat. 43 (1972) 627–635.
[27] S. Karlin and A. Dembo, Limit distributions for maximal segmental score among Markov-dependent partial sums, Ann. Prob. 24 (1992) 113–140.
[28] F.P. Kelly, Effective bandwidths at multi-class queues, Queueing Systems 9 (1991) 5–16.
[29] G. Kesidis, J. Walrand and C.S. Chang, Effective bandwidths for multiclass Markov fluids and other ATM sources, IEEE/ACM Trans. Networking 1(4) (1993).
[30] R.M. Loynes, The stability of a queue with non-independent inter-arrivals and service times. Proc. Camb. Phil. Soc. 58 (1962) 497–520.
[31] K. Rege, A methodology for designing admission criteria and engineering rules for atm systems, *ITC Sponsored Seminar on Teletraffic Analysis Methods For Current And Future Telecom Networks*, Banglore, India (1993).
[32] D. Reininger and D. Raychaudhuri, Bit-rate characteristics of a VBR MPEG video encoder for ATM networks, *IEEE Infocom* (1993).
[33] J. Walrand, *An Introduction to Queueing Networks* (Prentice-Hall, 1988).
[34] W. Whitt, Tail probabilities with statistical multiplexing and effective bandwidths in multiclass queues, Telecom. Syst. 2 (1993) 71–107.