EFFECTIVE BANDWIDTHS: CALL ADMISSION, TRAFFIC POLICING & FILTERING FOR ATM NETWORKS

by

G. de Veciana and J. Walrand

Memorandum No. UCB/ERL M93/47

16 June 1993

ELECTRONICS RESEARCH LABORATORY

College of Engineering
University of California, Berkeley
94720

EFFECTIVE BANDWIDTHS: CALL ADMISSION, TRAFFIC POLICING & FILTERING FOR ATM NETWORKS

by

G. de Veciana and J. Walrand

Memorandum No. UCB/ERL M93/47

16 June 1993

EFFECTIVE BANDWIDTHS: CALL ADMISSION, TRAFFIC POLICING & FILTERING FOR ATM NETWORKS

by

G. de Veciana and J. Walrand

Memorandum No. UCB/ERL M93/47

16 June 1993

ELECTRONICS RESEARCH LABORATORY

College of Engineering
University of California, Berkeley
94720

Effective Bandwidths: Call Admission, Traffic Policing & Filtering for ATM Networks

G. de Veciana and J. Walrand
Department of Electrical Engineering and Computer Sciences
University of California at Berkeley
Berkeley CA 94720

June 16, 1993

Abstract

In this paper we review and extend the effective bandwidth results of Kelly[20], and Kesidis, Walrand and Chang[21]. 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 buffers. We present results which are valid for a wide variety of systems and sources, 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.

1 Introduction

One of the key ideas behind BISDN/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. Then one can easily check if bandwidth is available by considering:

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

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 in which an *effective bandwidth* and the accompanying linear constraint can be found such that particular criteria are satisfied. The goal herein is to identify the structure required and limitations of such results for different criteria (quality of service), such as mean delay and cell loss, for multi-class queues.

This problem has been studied via a variety of techniques, each of which reveals additional insight. We begin by reviewing Kelly's results (in §2) for criteria such as mean packet delay or the

probability of large delays [20]. We consider the robustness of his result with respect to the service policy, by studying a prioritized service scheme.

Next we extend the approach of Kesidis, Walrand and Chang [21, 6], using large deviations to obtain effective bandwidths for a variety of systems where the criterion is once again the likelihood of large delays. In §3.1 we give a direct proof of this result, including a large class of stationary ergodic sources. Some novel examples where this result applies are discussed in §3.2, while in §3.3 we comment on the nature of sources for which the result fails. In §3.4 we consider the effect of packet admission policies and filtering on the effective bandwidth required by a source. We conclude with a discussion of related approximations; namely, heavy traffic limits which provide the natural extension of Kelly's effective bandwidth result for constraints on the mean workload.

Our discussion is not comprehensive; there exist a variety of related approaches. Notably, similar results for Markov fluid sources were obtained via spectral expansions, by Gibbens et al. [16], and Elwalid et al. [13]. Most of these results can also be obtained via large deviations, see Kesidis et al. [21] and de Veciana et al. [9].

2 Classical techniques

We begin by reviewing Kelly's result for a multi-class buffered resource [20]. 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 will be the required service time, so 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 (fixed packet size, correlated arrivals); it does however have some merit in a setting with variable length packets such as Frame Relay.

2.1 Effective bandwidths for the mean workload (M/GI/1)

Using the Pollaczek-Khintchine formula one finds the distribution of the workload in the system, as well as mean delay *before* service $\mathsf{E}D$ of typical customers. Following Kelly, we consider the constraint $\mathsf{E}D < d$, where

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

By rearranging terms, a linear constraint is obtained,

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

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 ($\mathsf{E}W$) is not fruitful. For example, in order to guarantee $\mathsf{E}W \leq w$, it suffices to insure $\mathsf{E}D + \mathsf{E}S \leq w$, where $\mathsf{E}S$ denotes the mean service time of customers. Thus a linear constraint is obtained by simply letting $d = w - \mathsf{E}S$ in the formulas above. Note however, that $\mathsf{E}S$ depends explicitly on the proportion of calls of each type, hence, only by assuming this mix is approximately constant (or $w \gg \mathsf{E}S$), 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.

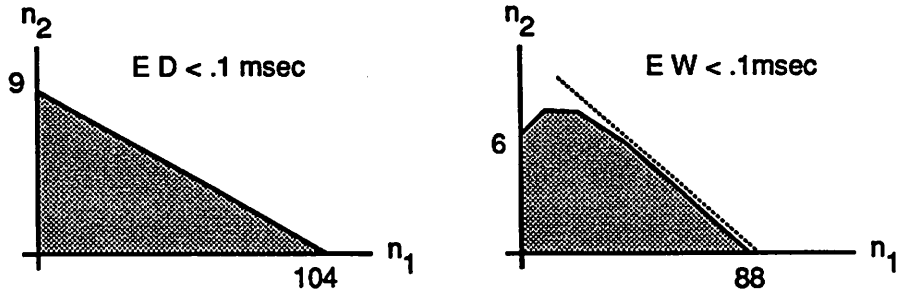

Figure 1: Effective Bandwidths and admissible regions.

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 $\mu_1 = 28.3$ μsec and $\sigma_1^2/\mu_1^2 = 2$. Type 2 sources have a mean traffic rate of 10 Mbps with $\mu_2 = 56.5$ μsec 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 $d = 0.1$ msec. The graph on the right shows the admissible region when the mean sojourn time is constrained to be less than $w = 0.1$ msec. As seen in Figure 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.

In practice one would like to optimize a tradeoff between maximum delay of some sources (voice and video) versus cell loss in others (data). Thus, it is interesting to consider the robustness of effective bandwidths to service policies; one such example follows.

2.1.1 Service priority

For simplicity we consider a 2-class M/GI/1 model with a non-preemptive service policy giving higher priority to Type 1 traffic. Using Little's result one obtains the expected delay before service of the two types of traffic, $\mathsf{E}D_1$ and $\mathsf{E}D_2$, as a function of the traffic statistics and the number of sources of each type(see Ex. 3.9. in Walrand [24]):

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

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

$$\begin{aligned} n_1\alpha_1(d_1) + n_2[\alpha_2(d_1) - \nu_2\mu_2] &\le 1; \\ n_1\alpha_1(\tilde{d_2}) + n_2\alpha_2(\tilde{d_2}) &\le 1 \end{aligned}$$

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

This example exhibits the interaction of traffic with 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. The likelihood of this event

is linear in the number of low priority sources, which explains the first constraint above. 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. This setup permits structured multiplexing of traffic streams subject to various *mean* delay constraints.

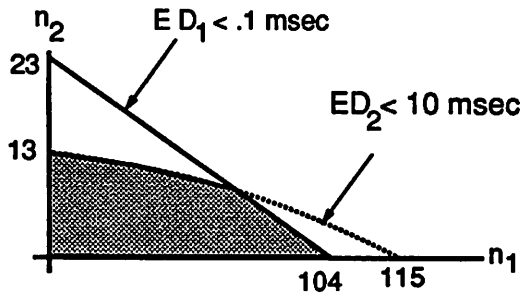

Figure 2: Admissible region for priority service.

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 until service for high priority traffic to be less than $d_1 = 0.1$ msec while delay constraints for Type 2 traffic are relaxed to $d_2 = 10$ msec. The admissible region defined by the above constraints is shown in Figure 2.

In practice it is of interest to consider a delay constraint on high priority traffic while maintaining a packet loss constraint on the low priority traffic. The effective bandwidths for tail distributions of the workload discussed in the next sections can be used to control packet loss for low priority traffic in this model. Intuitively, a bound on the mean delay of high priority traffic coupled with a packet loss constraint for low priority traffic will give rise to two linear inequalities defining the region where one might wish to operate.

2.2 Effective bandwidths for tail probabilities (M/GI/1 and D/GI/1)

We now review Kelly's effective bandwidth result, for M/GI/1 queues where a constraint of the type

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

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.

Note that in the previous section the assumption of Poisson arrivals was necessary in dealing with the multi-class setting; this continues to be the case here although the asymptotics on which the result is based can be obtained for GI/GI/1 and possibly SM/GI/1 (SM: semi-Markov) queues, see Iglehart [18] and Karlin et al. [19]. The main problem in extending Kelly's argument is that a superposition of renewal or semi-Markov traffic streams will not necessarily preserve these properties. Kelly does however consider another case which avoids this problem, a slotted/batch model, corresponding to a D/GI/1 queue with bulk arrivals at given time slots.

We first introduce a general result on the tail distribution of a GI/GI/1 queue. 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

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

Then it can be shown that the distribution of interest becomes exponential, i.e.,

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

where C is a constant that can be computed with some difficulty [18].

Using this result, Kelly obtained effective bandwidths for both M/GI/1 and D/GI/1 queues. 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

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

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

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

and where $\Lambda_j(\delta) = \log\mathbf{E}\exp[\delta S_j]$ 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 service 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 Eq. 4 then becomes

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

A change in 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.

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 buffer occupancy. The basic ideas are drawn from Kesidis, Walrand and Chang [21, 6]. We present a direct proof, some extensions, and discuss some of the practical issues. We use large deviations as a means to obtain estimates for large buffer asymptotics. When estimating tail distributions large deviation bounds are usually more refined than those obtained via central limit theories; the latter are briefly discussed in §4.

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 [10]. A sequence of measures $\{\mu_n\}$, on $\mathbf{R}$, 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 $S_n = n^{-1} \sum_{i=1}^n X_n$, 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 $\Lambda(\theta) \triangleq \lim_{n\to\infty} \frac{1}{n} \log \mathbf{E} \exp[\theta S_n]$ exist (possibly infinite) for all $\theta \in \mathbf{R}$;
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., $\lim_{n\to\infty} |\frac{d\Lambda(\theta_n)}{d\theta}| = \infty$ 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 $\mathbf{E} e^{\theta X_1} < \infty$ for all θ, which corresponds to the original large deviation estimate of Cramér. The result also applies to sequences with weak dependencies.

A more specific characterization of sources for which LDPs hold can be found in [10] and in the appendix. For example, coordinate functions of Markov chains satisfying strong uniformity conditions on the transition kernel and tail will satisfy an LDP, see for example [11]. In this case, the rate function can usually be interpreted in terms of the relative entropy rate of a deviant Markov chain with respect to the original process. For stationary sequences satisfying appropriate mixing and tail conditions similar results hold, see [4].

3.1 General effective bandwidth result

Theorem 3.1 *Let $\{X_n\}$ be a stationary ergodic process with $\mathbf{E}X_n < 0$, which either satisfies an LDP with convex good rate function $I(\cdot)$, such that for all $\theta < \infty$*

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

and $\Lambda^*(\cdot)$ is strictly convex in a neighborhood of $\alpha^* = \arginf_{\alpha>0} \Lambda^*(\alpha)/\alpha$, or satisfies the requirements for the Gärtner-Ellis Theorem[1]. Then the Lindley process

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

[1] Note the Gärtner-Ellis Theorem does not require finite log-moment generating functions.

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

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

Remark Note that if $\Lambda(\theta) < \infty$ then $\lim_{|x|\to\infty} \Lambda^*(x)/|x| = \infty$, so α^* above makes sense (see [10] page 34). Also note that the strict convexity of $\Lambda^*(\cdot)$ is equivalent to the differentiability of $\Lambda(\cdot)$ at some point, see the proof below. Alternatively if the Gärtner-Ellis Theorem is in force, then the steepness and differentiability conditions guarantee the not only that α^* makes sense, but also the strict convexity of $\Lambda^*(\cdot)$ when the random variables are real-valued (see [12] page 224).

Proof: The stability condition, $\mathbf{E}X_n < 0$, guarantees the existence of a stationary distribution, see Loynes [22]. In particular, let

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

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

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

must exist. Moreover, by Theorem 4.5.10 in [10], or directly from the Gärtner-Ellis Theorem, the rate function is in fact the convex dual of $\Lambda(\cdot)$, i.e.,

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

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

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

and it follows from Eq. 5 that

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

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

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

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

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

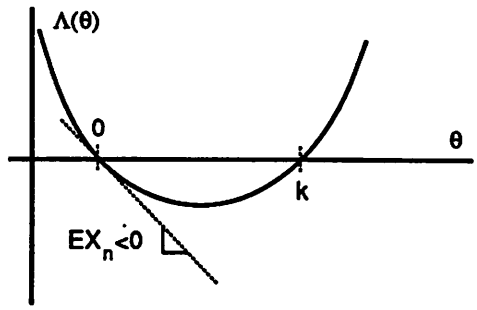

Figure 3: Convexity of log-moment and asymptotic rate.

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

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

Indeed, a short argument shows that $\Lambda(k) = 0$. As mentioned above, the optimizer α^* of Eq. 7 is well defined. The first order optimality conditions require that

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

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 $\frac{d\Lambda^*(\lambda^*)}{d\alpha} = k$. Our convexity requirement and the previous optimiality 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 Figure 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. This in turn proves a result first obtained by Kesidis, Walrand and Chang [21].

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 any work conserving service policy at rate c. Then the following effective bandwidth result holds:*

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

and where W denotes the stationary queue length.

Proof: Since each source satisfies a large deviation principle the limiting log-moment generating functions

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

exist and the rate function for each sources 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 limits

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

exist, and by the contraction principle and convexity of the rate functions, the aggregate satisfies a large deviation principle with good rate function (see [10] page 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,

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

□

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. [21] and Courcoubetis et al. [7]. These include the usual i.i.d. sources, as well as Markov modulated fluids or Poisson processes and Gaussian processes. The extension of these results to continuous-time queues, such as the case of Markov modulated fluids, can be made rigorous via discrete *exponentially good approximations*(see [10] for a definition) in which case the previous arguments will apply.

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 arrivals (rate ν_j) with 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 :*

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

where W denotes the stationary workload or delay before service for 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 interarrival 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 i^{th} arrival which corresponds to a particular stream of type $j \in J$ with probability ν_j/ν. As in the previous corollary, the condition

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

gives the desired result. Since interarrival times are exponential and independent, we find that

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

the log of the Laplace transform for an exponential interarrival with rate ν. The term corresponding to the arriving work can also be simplified to,

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

The condition $\Lambda(\delta)\leq 0$ can then be rewritten as

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

□

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.,

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

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

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

3.2 Other examples

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 device 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 deviations 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. Using this theorem one might attempt to obtain results for queues with dependent arrival and service processes, as we only require that their difference satisfies a large deviation principle.

3.2.1 A noisy server

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:

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

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, so 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 page 22 in [5]). 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 rate, results in a reduced service capacity, which depends in a natural way on the variance of the noise and the autocorrelation between noise samples.

3.2.2 Randomized service

Suppose that in addition to multi-class sources, there exist high and low priority streams queued in segregated buffers. In order to keep tail delays of high priority streams down, we will allow for randomized service 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. This policy is not work conserving as a slot may be assigned to a priority for which no work is available; but we will assume that slots are relatively small and there is a reasonable amount of input traffic. We obtain two effective bandwidth equations for high (J_h) and low (J_l) priority traffic

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

Here, $\alpha_j(\cdot)$ denote the effective bandwidths obtained for sources sharing a deterministic server. Since these constraints are decoupled we can envisage choosing different tail constraints (δ) for the two priorities.

3.3 What happens when the conditions fail?

Given these rather abstract conditions for the existence of effective bandwidths, one might ask what types of sources will not have an effective bandwidth. A particularly insightful account of the phenomena that occur can be found in Anantharam [1]. He considers a GI/GI/1 queue wherein the distribution of $X = S - A$ (difference of the service time and inter-arrivals periods) does not have an exponential tail, for example

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

where $L(x)$ is a slowly varying function, see [1]. For such a system, delays will build up suddenly, i.e., when a *single* customer with a huge excess service time arrives rather than as an accumulation of several deviant service times. Although these conditions are quite unlikely in a networking setup, they point to the radically different behavior of sources with "fat tails".

Sources without sufficient randomness are also excluded from our framework; consider for example a traffic source for which realizations are deterministic square waves. By randomizing the initial phase, the source becomes stationary and ergodic but will not satisfy a large deviation principle; however, the effective bandwidth is degenerate and equal to the mean arrival rate, regardless of the tail constraint. Chang [6] develops an interesting point of view unifying stochastic and deterministic sources via the notion of envelope processes.

3.4 On cell admission policies 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. Thus, suppose that at time n, A_n packets arrive. 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 reinsurance of policies (see page 287 in Asmussen [2]).

Proposition 3.1 *Suppose $\{A_n\}$ is an i.i.d. sequence satisfying a large deviations principle. Consider all memoryless rejection policies, $h(\cdot)$, with the same throughput μ, i.e., such that $\mathsf{E}h(A_n) = \mu \leq \mathsf{E}A_n$. Let $h^*(a) = \min[a, T]$, where T is determined by $\mathsf{E}h^*(A_n) = \mu$. 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 $\Lambda_h(\theta) = \log \mathsf{E} \exp[\theta h(A_0)]$ and $\Lambda_{h^*}(\theta) = \log \mathsf{E} \exp[\theta h^*(A_0)]$ 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 $\mathsf{E}e^{\delta h(A_0)} \geq \mathsf{E}e^{\delta h^*(A_0)}$ since $\mathsf{E}[h(A_0) - h^*(A_0)] = 0$, 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. Intuitively, when there are dependencies among the arrivals, the optimal $h(\cdot)$ may reflect the dynamics of the process. Before considering an example of such a source, we 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.,

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

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^*,

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

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. In the sequel we consider whether this can be done by filtering.

The Markov source shown in Figure 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.,

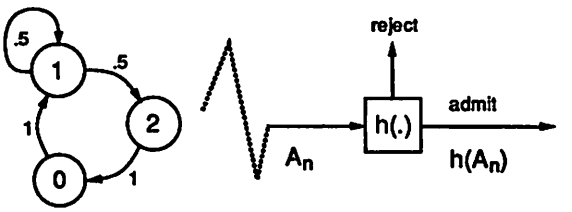

Figure 4: A source for which thresholds are not optimal.

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 was shown in [21] to be

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

where $\mathrm{sp}[\phi(\delta)P]$ 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 Figure 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 from 0.5. This explains why the effective bandwidth remains constant for all constraints δ.

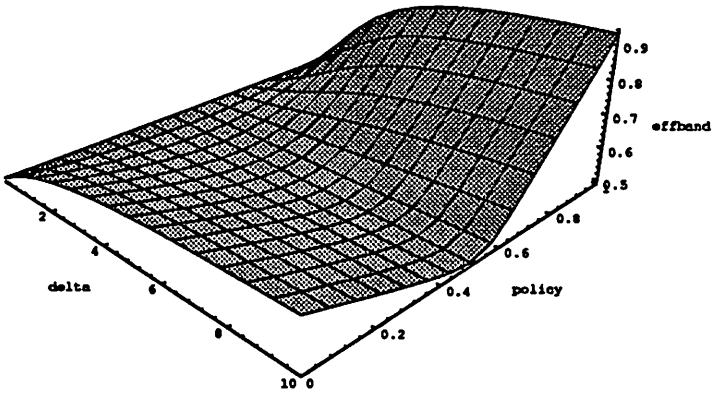

Figure 5: Effective bandwidth vs tail constraint and admission policy.

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, some sources (e.g., packetized voice and video) can tolerate loss, however consecutive losses will lead to a degradation in the quality of service. Thus even an optimal memoryless policy is imperfect in a practical sense. 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.

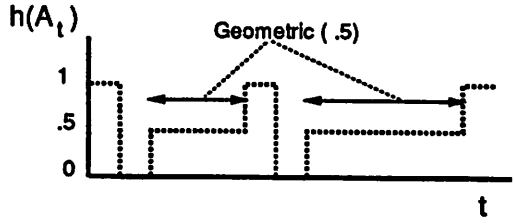

Figure 6: Sample paths for optimal policy.

For example, recent detailed studies for variable bit rate video traffic consider the dynamics of loss and traffic policing schemes [23]. 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. The dynamics of loss have been studied via simulation showing that the loss process can be approximated to first order by an On/Off Markov fluid whose parameters depend on the buffer utilization. With this knowledge in hand, we eventually decide whether by the use of adaptive coding and thresholding obtain significant performance improvements.

We complete this discussion of admission policies, with an insightful example suggested by Courcoubetis and Weber [8]. Consider a stationary Gaussian arrival process, $\{A_n\}$ with mean μ, and finite asymptotic variability

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

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 $R(n) = \mathrm{Cov}(A_i,\ A_{i+n})$ 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 as well as non-linear filters. For example, in the case of the popular leaky bucket scheme a similar result was observed by Berger et al. [3]. In the next chapter we present an alternative view point which elucidates the benefits of traffic shaping.

4 Heavy traffic approximations

As mentioned in the introduction, the large deviation asymptotics obtained herein complement related central limit or heavy traffic approximations. We briefly discuss this topic, as it it provides

further generalizations of Kelly's effective bandwidth results for constraints on the mean workload, and relates back to the large deviation results for tail distributions. We base our discussion on the work of Fendick, Saksena, and Whitt [14, 15] and Harrison [17].

In their study of dependencies in packet queues Fendick et al. begin by considering a superposition of Poisson streams with batch arrivals. For example, suppose 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-in first-out discipline. In this case one can show the mean workload in the system is

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

As in §2.1, by rearranging terms in the constraint $\mathsf{E}D < d$, the effective bandwidth of a batch arrival stream subject to a mean delay before service less than d can be defined:

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

One might ask if this result generalizes when the arrivals are not Poisson but renewal, the batches are not instantaneous but spaced, or the interarrival spacing and batches are dependent. Fendick et al. analyze such systems under heavy traffic, i.e., as the traffic intensity $\rho_n \to 1$. Using their results, effective bandwidths for such traffic streams can be defined, which make sense when the queue is heavily loaded, see [14].

For illustrative purposes consider a multiclass deterministic queue with service rate c. Suppose packets in a stream of type $j \in J$, have interarrivals $\{a_n^j\}$, with mean μ_j^{-1} and variance v_j^2. Let A_t^j denote the cumulative arrivals up to time t, and suppose the process satisfies a central limit theorem such that $n^{-1/2}[A_{[nt]}^j - \mu_j nt] \to N(0,\sigma_j^2)$, e.g., for renewal arrival processes $\sigma_j^2 = \mu_j^3v_j^2$ [24]. Consider scaling the queue length 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 queue length 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 results for regulated Brownian flows apply. When $\alpha < 0$, the steady state distribution of the queue length, denoted by the random variable X, is exponential with mean

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

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

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

By imposing a tail constraint on the exponentially distributed queue length of the unscaled process, $\mathsf{P}(X > B) \le \exp[-\delta B]$, we obtain the following approximate requirement:

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

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

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

where $\mu_j = \mathsf{E}A_0^j$, and $\sigma_j^2 = \lim_{n\to\infty} t^{-1}\mathrm{Var}(\sum_{i=1}^{t} A_i^j)$ 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.

Harrison's 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 netput 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 $\mathsf{E}X$ is given by

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

(see [17] page 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 queue length constraint of the form $\mathsf{E}X < d$. Specifically,

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

which is analogous to the results discussed in §2.1.

5 Summary

We have discussed a variety of effective bandwidth results which can be used to simplify network management for shared buffers. In particular we found desirable linear constraints depending on the number and types of input sources such that a statistical constraint on the asymptotic probability of overflow or the mean workload were satisfied. These results extend those of Kelly to a wide class of sources, via central limit and large deviation asymptotics. Interesting extensions to systems with random servers or prioritized service were also considered. Unfortunately few useful results are available identifying the effective bandwidth of individual sources which have shared a buffer, hence the circuit-switched model of a network of interacting streams is only approximate.

In addition an attempt was made at identifying admission policies which are optimal in the sense of reducing the effective bandwidth of sources. For memoryless policies, we found that although thresholding will be optimal for i.i.d. sources, this will not be true in general. Furthermore, we considered filtering the rate of a traffic source to reduce fluctuations in the hope of improving performance. However, we found that the effective bandwidth of filtered sources depends purely on the fraction of traffic rejected. These two observations suggest that in so far as large buffer asymptotics are concerned, the only method available to reduce a source's contribution to the occupancy of a buffer is rejecting some of the traffic, moreover the optimal way to do this is not easy to compute.

References

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

[2] S. Asmussen. *Applied Probability and Queues.* John Wiley & Sons, 1985.

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

[4] W. Bryc and A. Dembo. Large deviations and strong mixing. *Preprint*, 1993.

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

[6] C.S. Chang. Stability, queue length and delay of deterministic and stochastic queueing networks. *submitted to IEEE AC*, 1992.

[7] C. Courcoubetis and J. Walrand. Note on effective bandwidth of ATM traffic. *Preprint*, 1991.

[8] C. Courcoubetis and R. Weber. Effective bandwidths for stationary sources. *Preprint*, 1992.

[9] G. de Veciana, C. Olivier, and J. Walrand. Large deviations of birth death Markov fluids. *Probability in the Engineering and Informational Sciences*, 7:237–255, 1993.

[10] A. Dembo and O. Zeitouni. *Large Deviations Techniques and Applications.* Jones & Bartlett, Boston, 1992.

[11] J.D. Deuschel and D.W. Stroock. *Large Deviations.* Academic Press, Boston, 1989.

[12] R.S. Ellis. *Entropy, Large Deviations and Statistical Mechanics.* Springer-Verlag, 1985.

[13] A. I. Elwalid and D. Mitra. Effective bandwidth of general Markovian traffic sources and admission control of high speed networks. *Submitted IEEE Networks*, 1992.

[14] K. Fendick, V. Saksena, and W. Whitt. Dependence in packet queues. *IEEE Trans. Comm.*, 37, 1989.

[15] K. Fendick and W. Whitt. Measurements and approximations to describe the offered traffic and predict the average workload in a single-server queue. *Proceedings of the IEEE*, 77, 1989.

[16] R.J. Gibbens and P.J. Hunt. Effective bandwidths for the multi-type UAS channel. *Queueing Systems*, 9:17–28, 1991.

[17] M. Harrison. *Brownian Motion and Stochastic Flow Systems.* Krieger Publishing, 1990.

[18] D. L. Iglehart. Extreme values in the GI/G/1 queue. *Ann. Math. Statist.*, 43:627–635, 1972.

[19] S. Karlin and A. Dembo. Limit distributions for maximal segmental score among Markov-dependent partial sums. *Ann. Probab.*, 24:113–140, 1992.

[20] F.P. Kelly. Effective bandwidths at multi-class queues. *Queueing Systems*, 9:5–16, 1991.

[21] G. Kesidis, J. Walrand, and C.S. Chang. Effective bandwidths for multiclass Markov fluids and other ATM sources. *To appear in IEEE Trans. Comm.*, 1992.

[22] R.M. Loynes. The stability of a queue with non-independent inter-arrivals and service times. *Proc. Camb. Phil. Soc.*, 58:497–520, 1962.

[23] D. Reininger and D. Raychaudhuri. Bit-rate characteristics of a VBR MPEG video encoder for ATM networks. In *INFOCOM Proceedings*, 1993.

[24] J. Walrand. *An Introduction to Queueing Networks.* Prentice-Hall, 1988.