DECOUPLING BANDWIDTHS FOR NETWORKS:

A Decomposition Approach to Resource Management

by

G. de Veciana, C. Courcoubetis, and J. Walrand

Memorandum No. UCB/ERL M93/50

28 June 1993

DECOUPLING BANDWIDTHS FOR NETWORKS:

A Decomposition Approach to Resource Management

by

G. de Veciana, C. Courcoubetis, and J. Walrand

Memorandum No. UCB/ERL M93/50

28 June 1993

ELECTRONICS RESEARCH LABORATORY

College of Engineering
University of California, Berkeley
94720

DECOUPLING BANDWIDTHS FOR NETWORKS:
A Decomposition Approach to Resource Managment

G. DE VECIANA, C. COURCOUBETIS* AND J. WALRAND

Department of Electrical Engineering and Computer Sciences
University of California at Berkeley
Berkeley CA 94720

**University of Crete, Heraklion, Crete*

June 28 1993

Abstract

We consider large buffer asymptotics for feed-forward *networks* of discrete-time queues with deterministic service rate shared by multiple classes of streams. First we review the concept of *effective bandwidths* for traffic streams subject to a tail constraints on the buffer occupancy. Next we discuss the effective bandwidth of the departure process of such a queue, proving that in fact the effective bandwidth of the output is at worst equal to that of the input, and depending on the service rate, strictly less than that of the input. We then define the notion of a *decoupling bandwidth* guaranteeing that asymptotics within the network are decoupled.

These results provide a framework for call admission schemes which are sensitive to constraints on the tail distribution of the workload in buffers or approximate cell loss probabilities. Our results require relatively weak assumptions on both the traffic streams and service policies. We consider the problem of "optimal" traffic shaping (via buffering) subject to a loss or delay constraint. Finally, we discuss our results in the context of resource management for ATM networks.

1 Introduction

An important open problem in the context of BISDN/ATM is that of designing appropriate resource management schemes comprising call admission, routing and network planning for a heterogeneous collection of users requiring multiple qualities of service. The difficulty of this problem relative to traditional (circuit-switched) telephone networks, lies in the 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 an equivalent circuit-switched model. 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 how much bandwidth is available by considering:

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

When appropriate, such a scheme can be extended to a network, as the counterpart of traditional telephone systems, where a connection is set up if indeed physical resources are available to link the source to the destination. Unfortunately, the interaction of multiple types of traffic and resources 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 *single queue* for which an *effective bandwidth* and the accompanying linear constraint can be found such that an asymptotic constraint on the tail distribution of the buffers' workload is guaranteed. The goal herein is to investigate this idea for a *network* of queues. In this paper we obtain the input/output map for the effective bandwidths of streams sharing a queue. We will show explicitly that the reduction in effective bandwidth due to a buffer depends on the release rate. We then consider the resource management problem for a multi-class feed-forward network via the notion of *decoupling bandwidths*. Our results show that when the queues' service rates are selected appropriately, the asymptotics for queues in the network reduce to that of the single buffer case, and thus by way of effective bandwidths a workable circuit-switched model is obtained.

The prototypical example is that of a Jackson network, for which the steady state distribution is in fact product-form. The queue length distribution for a queue i, is geometric and simply depends on the ratio of the aggregate arrival intensity to the service rate of the queue, i.e., $\rho_i = \lambda_i/\mu_i$. In order to guarantee individual δ-constraints on the tail distribution of the queue lengths

$$\mathbb{P}(Q_i > B) = \rho_i^{B+1} \leq \delta^B,$$

we require that $\lambda_i \leq \mu_i \delta$. Thus to manage resources in this network, it suffices to ensure that the mean rate flowing through each node remains within the interval computed from the constraint. A new traffic stream will not violate the imposed tail constraints if the additional traffic along its route remains within the prescribed intervals.

In this paper we extend this scenario to a feed-forward network with multiple classes of traffic streams, where the queues have deterministic service rate and *arbitrary* work conserving service policies. The intuitive picture for our result is as follows: Consider a large accumulation of customers (packets) for a queue deep in the network. Such an event is necessarily due to an increase in the empirical arrival rate of the traffic streams sharing that queue. When the asymptotics are "decoupled" these deviations from the mean rate are such that other queues shared by these streams are invisible. The likelihood of a traffic streams' deviations can then be computed at the network edge, where the statistics are assumed to be known.

To our knowledge this is the first successful attempt to study the effective bandwidth idea for large buffer asymptotics in networks. We note however, some previous work in this

area using bounding techniques by Kurose [23] and Chang[4]. Also in [10], we presented an alternative point of view, investigating the asymptotics of networks with negligible buffers but where traffic streams were periodically averaged.

We begin by reviewing the notion of effective bandwidths for a single queue in §2 as well as some fundamental concepts in the theory of large deviations. In §3, we consider the large deviations of the departure process of a queue. Having identified the properties of the departure processes, we consider some examples of traffic shaping via buffering in §3.1. In §3.2 we consider queues in tandem. We turn to more general aspects of resource management for networks in §4. A summary of our results and conclusions can be found in §5.

2 Single buffer asymptotics via large deviations

In this section we state the effective bandwidth result for a multi-class discrete-time queue subject to constraints on the tail probability of the buffer occupancy. This discussion is based on an in-depth study in an earlier paper [9], which purported to extend the results of Kelly [20]. The large deviations techniques we use here were inspired by a heuristic of Borovkov, see Walrand and Parekh [26, 25], and the results of Kesidis, Walrand and Chang [22, 4]. See also the recent paper of Whitt [27]. These ideas were furthered in [9], where we explored randomness and dependencies in the arrival and service processes. There exists much related work in this field. Notably, effective bandwidth results for Markov fluid sources were obtained via spectral expansions, by Gibbens and Kelly [17], and Elwalid and Mitra [15]. In addition some early work on this topic can be found in Hui [19] and Guérin, Ahmadi and Naghshineh [18].

2.1 Large deviations

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 [12]. 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 for 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 $\Lambda(\theta) \triangleq \lim_{n\to\infty} \frac{1}{n} \log \mathbb{E} \exp[\theta S_n]$ 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., $\lim_{n\to\infty} |\frac{d\Lambda(\theta_n)}{d\theta}| = \infty$ 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 $\mathbb{E}e^{\theta X_1} < \infty$ for all θ, which corresponds to the original large deviation estimate of Cramér. The result also applies to sequences with weak dependencies.

A more specific characterization of sources for which LDPs hold can be found in [12]. For example, coordinate functions of Markov chains satisfying strong uniformity conditions on the transition kernel and tail will satisfy an LDP, see for example [13]. 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 [3].

2.2 Effective bandwidths for single buffer

Theorem 2.1 [See [9]] *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 $\theta < \infty$*

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

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

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

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

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

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

Remark 2.1 Note that if $\Lambda(\theta) < \infty$ then $\lim_{|x|\to\infty} \Lambda^*(x)/|x| = \infty$, so α^* above makes sense (see [12] page 34). Also note that the strict convexity of $\Lambda^*(\cdot)$ is closely related to the differentiability of $\Lambda(\cdot)$ at some point, i.e., if $\Lambda(\cdot)$ were differentiable then $\Lambda^*(\cdot)$ would be strictly convex. Alternatively if the Gärtner-Ellis Theorem is in force, then the steepness and differentiability conditions guarantee not only that α^* makes sense, but also the strict convexity of $\Lambda^*(\cdot)$ when the random variables are real-valued (see [14] page 224).

By letting $X_n = A_n - c$, where A_n denotes aggregate arrivals for a multi-class queue with service rate c, Theorem 2.1 can be used to establish the following corollary.

Corollary 2.1 [See [9] or [22, 4]] *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 2.1, so that*

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

Suppose they share a deterministic buffer with any work conserving service policy at rate c. Then the following effective bandwidth result holds:

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

and where W denotes the stationary workload.

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 see Kesidis et al. [22]. These include the usual i.i.d. sources, as well as Markov modulated fluids or Poisson processes and Gaussian processes. The extension of our results to continuous-time queues, such as the case of Markov modulated fluids, can be made rigorous via discrete *exponentially good approximations* (see [12] for a definition) in which case the previous arguments will apply. For an investigation of some ideas in quick estimation and policing using effective bandwidths see Kesidis et al. [6, 21].

3 Effective and decoupling bandwidths for departures from a queue

We begin this section by identifying the large deviation rate function of the departure process from a stationary queue. For simplicity we consider a discrete-time queue with constant service rate, though by analogy with Theorem 2.1, extensions to randomized and dependent service times follow through directly.

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

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

and $\Lambda^*(\cdot)$ is strictly convex or satisfies the requirements for the Gärtner-Ellis Theorem with finite log-moment generating function. Then the Lindley process

$$W_{n+1} = [W_n + A_n - c]^+$$

has a stationary distribution, say that of a random variable W, and the associated departure process $\{D_n\}$ satisfies an LDP with with convex good rate function given by $\Lambda^*(\cdot)$ on $[0, c]$ and infinite on $[0, c]^c$.

Remark 3.1 By contrast with Theorem 2.1, here we require that asymptotic log-moment generating functions are bounded. This condition guarantees exponential tightness of the distribution for the paths of normalized partial sums required to use the result of Dembo and Zajic [11] in the proof below.

Proof: The stability condition, $\mathbb{E}A_n < c$, guarantees the existence of a stationary distribution, see Loynes [24] or Walrand [26] Chap. 7. In particular, let $X_n = A_n - c$ and

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

then the distribution of W_0^m converges monotonically to that of W, where $\mathbb{P}(W < \infty) = 1$. Similarly the departure process is monotonically increasing and converging to the stationary distribution.

Since limits $\Lambda(\theta)$ exist and are bounded by Theorem 4.5.10 in [12], or directly from the Gärtner-Ellis Theorem, the rate function for the arrival process can be identified as the convex dual of $\Lambda(\cdot)$, i.e.,

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

Let $S_n^D = \sum_{i=1}^n D_n$, $S_n^A = \sum_{i=1}^n A_n$ denote the partial sums of the departure and arrival processes for $n > 0$ and $S_n^A = \sum_{i=n}^0 A_n$ for $n \le 0$. We begin by considering the departures for a stationary version of this queue. Note that

$$S_n^D \le W + S_n^A, \quad \text{where} \quad W = \max_{i\ge 0} [S_{-i}^A - ic, 0].$$

Using an argument similar to that in Theorem 2.1, or see Chang [4] we have that for $\epsilon > 0$ and large enough n,

$$\begin{aligned}
\mathbb{E}\exp[\theta S_n^D] &\le \mathbb{E}\exp[\ \theta\ (S_n^A + W)\ ] \\
&\le \mathbb{E}\exp[\ \theta\ \max_{i\ge 0}[S_n^A + S_{-i}^A - ic, S_n^A]\ ] \\
&\le \exp[(\Lambda(\theta)+\epsilon)n] + \sum_{i\ge 0} \exp[(\Lambda(\theta)+\epsilon)n + (\Lambda(\theta)+\epsilon-\theta c)i] \\
&\le C\ \exp[(\Lambda(\theta)+\epsilon)n],
\end{aligned}$$

for some finite constant C as long as $\Lambda(\theta) + \epsilon < c\theta$. By Chebychev's bound we then have

$$\mathbb{P}\left(S_n^D \geq n\alpha\right) \leq \exp[-\theta n\alpha]\, C \exp[(\Lambda(\theta) + \epsilon)n]$$

So it follows by letting $n \to \infty$ and $\epsilon \to 0$ that

$$\limsup_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left(\frac{1}{n} S_n^D \geq \alpha\right) \leq -\sup_{\theta}[\theta\alpha - \Lambda(\theta)] = -[\theta^*\alpha - \Lambda(\theta^*)] = -\Lambda^*(\alpha),$$

where we note that we have taken $\alpha < c$ so $\Lambda(\theta^*) < c\theta^*$ guaranteeing the boundedness of the asymptotic log-moment generating function for the departures.

We obtain a lower bound by considering a queue starting from $W_0 = 0$. The cumulative departure process of this queue clearly lower bounds that of the stationary version. Note that

$$\begin{aligned} S_n^D &= S_n^A - W_n \\ &= S_n^A - \max_{1\leq i\leq n}[S_n^A - S_i^A + [n-i]c] \\ &= \min[S_n^A,\ S_{n-1}^A + c,\ \ldots,\ S_1^A + [n-1]c,\ nc] \\ S_n^D - nc &= \min[S_n^A - nc,\ S_{n-1}^A - [n-1]c,\ \ldots,\ S_1^A - c,\ 0]. \end{aligned}$$

Consequently,

$$\begin{aligned} \mathbb{P}\left(S_n^D > n\alpha\right) &= \mathbb{P}\left(S_n^D - nc > n[\alpha - c]\right) \\ &\geq \mathbb{P}\left(S_n^D - nc > n[\alpha - c] \mid S_n^A > n\alpha\right) \times \mathbb{P}(S_n^A > n\alpha) \\ &= \mathbb{P}\left(\min[S_n^A - nc,\ \ldots,\ S_1^A - c,\ 0] > n[\alpha - c] \mid S_n^A - nc > n[\alpha - c]\right) \times \\ &\qquad \mathbb{P}(S_n^A > n\alpha). \end{aligned}$$

Taking the lim inf we find that

$$\begin{aligned} &\liminf_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left(\frac{1}{n} S_n^D > \alpha\right) \geq \\ &\quad \liminf_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left(\min[S_n^A - nc,\ \ldots,\ S_1^A - c,\ 0] > n(\alpha - c) \mid S_n^A - nc > n[\alpha - c]\right) + \\ &\qquad \liminf_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left(S_n^A > n\alpha\right) \geq 0 - \Lambda^*(\alpha). \end{aligned}$$

The bound for the second term follows by a straightforward application of the large deviation principle for the arrival process. The asymptotic probability of the first term can be estimated by way of a result for the conditional distribution of the paths corresponding to the normalized partial sum process. Indeed as exhibited in Figure 1, one can show that the mass of the conditional distribution of paths leading to $S_n^A - nc > n[\alpha - c]$ concentrates on a specific path lying above the endpoint $n(\alpha - c)$, such that the log of the probability that the path's minimum remains above the endpoint goes to zero. This result is essentially a consequence of the rate function's strict convexity. A discussion of this result is beyond the scope of this paper, we refer the reader to the work of Asmussen [2] and Anantharam

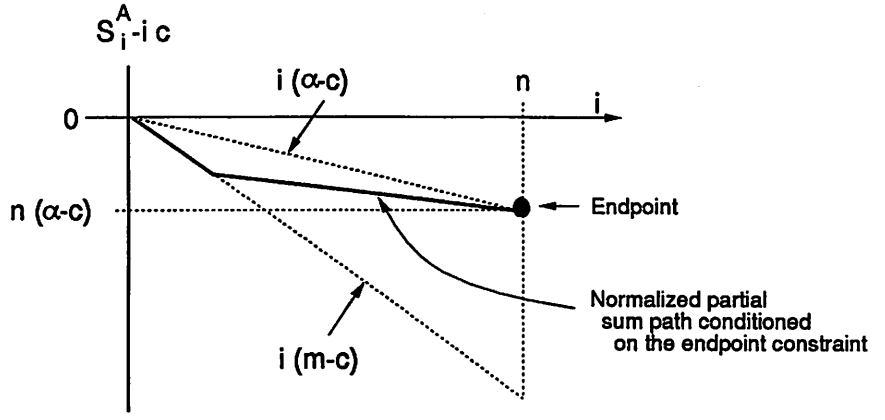

Figure 1: Conditional path subject to constraint on the endpoint.

[1], for an investigation of the normalized partial sum paths of i.i.d. random variables, and Dembo and Zajic [11] for a general result, which we use here, predicated on the existence of an LDP.

Finally, observe that the rate function is clearly infinite on $[0,c]^c$. The remaining steps showing the LDP for open and closed sets are standard, see for example the text of Dembo et al.[12]. □

As a corollary to this theorem consider the scenario in which the arrival process is an aggregate of independent traffic streams.

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 according to a work conserving service policy with rate c. Then the aggregate departure process satisfies an LDP with convex good rate function*

$$\Lambda_D^*(\alpha) = \inf_{\sum_{j\in J} n_j\alpha_j = \alpha} \sum_{j\in J} n_j \Lambda_j^*(\alpha_j), \tag{1}$$

on the set $[0,c]$ and infinite elsewhere. Define $\alpha_j^*(\delta)$ by the following duality relationship

$$\Lambda_j(\delta) = \sup_\alpha [\alpha\delta - \Lambda^*(\alpha)] = \alpha_j^*(\delta)\delta - \Lambda_j^*(\alpha_j^*(\delta)).$$

Note that

$$2\alpha_j(\delta) - \alpha_j(0) \;\geq\; \alpha_j^*(\delta) = \left[\frac{d\Lambda_j^*}{d\alpha}\right]^{-1}(\delta) \;\geq\; \alpha_j(\delta). \tag{2}$$

We will call $\alpha_j^*(\delta)$ the **decoupling bandwidth** *of a source of type j (see the remark below for comments). The effective bandwidth of the output traffic stream, $\alpha_D(\delta)$, is given by*

$$\alpha_D(\delta) = \begin{cases} \sum_{j\in J} n_j\alpha_j(\delta) & \text{if} \quad \sum_{j\in J} n_j\alpha_j^*(\delta) < c \\ c - \frac{1}{\delta} \inf_{\sum_{j\in J} n_j\alpha_j = c} \sum_{j\in J} n_j\Lambda_j^*(\alpha_j) & \text{otherwise.} \end{cases}$$

Remark 3.2 We do not claim that the effective bandwidth of *individual* output streams is necessarily reduced as we did in Corollary 3.1 for the *aggregate* departure process. A technical result is missing here to identify the rate function of the test traffic stream over the entire interval $[0, c]$. We conjecture that, individual sources may interact when the decoupling constraint, Eq. 5, is not met. This might happen for example when a bursty source shares a queue with a rather smooth source.

Outline of Proof: We begin by adapting the proof of Theorem 3.1 to the departure process of the test traffic stream. Let $S_n^{A^t}$, and $S_n^{D^t}$ denote the partial sums for the arrivals and departures of our test stream, while $S_n^{A^s}$ and $S_n^{D^s}$ denote the cumulative arrivals and departures for the other streams sharing the queue.

Unfortunately the simple proof for upper bound in the previous theorem does not follow through. This would require a characterization of the tail for the stationary workload of the *test* sequence. Nevertheless we can use the approach taken for the lower bound. Suppose we start with an empty queue $W_0 = 0$, the Loynes' construction permits us to conclude that W_n is monotonically converging to the stationary distribution. Similarly the distribution of the departures for the test sequence converges monotonically to that of the stationary departure process. Thus we propose to bound the cumulative departures for the stationary process, by finding a bound for those of the empty queue, which is monotonically increasing to that of the stationary queue.

First note that for $n > 0$ we have,

$$\begin{aligned} S_n^{D^t} + S_n^{D^s} - nc = S_n^D - nc &= \min[S_n^A - nc, \ldots, i\, S_1^A - c,\, 0] \\ &= \min[S_n^{A^t} + S_n^{A^s} - nc, \ldots,\, S_1^{A^t} + S_1^{A^s} - c,\, 0], \end{aligned}$$

and by flow conservation $S_n^{A^t} \geq S_n^{D^t}$. Fix $\epsilon > 0$ and let $\mathcal{E}_n$ denote the event

$$\mathcal{E}_n = \{S_i^{A^s} \in (im - n\epsilon, im + n\epsilon),\ \text{for } i = 1, \ldots n\},$$

corresponding to the case where the partial sum path of the traffic sharing the queue with the test stream maintains an arrival rate close to its mean $m = \mathsf{E}A_1^s$. This event is of course very likely to occur, in fact we have that

$$\lim_{n\to\infty} \frac{1}{n} \log \mathsf{P}(\mathcal{E}_n) \;=\; 0,$$

by the LDP results in Dembo and Zajic [11].

Observe that the cumulative departures of the test stream from the shared queue, can only increase when the stream sharing this queue is set to zero. Also note that $S_n^{A^t} \geq S_n^{D^t}$. Thus,

$$\begin{aligned} \mathsf{P}(S_n^{D^t} \geq n\alpha) \;\leq\; \mathsf{P}(S_n^{D^t} \geq n\alpha \mid S_n^{A^s} = 0) = \\ \mathsf{P}(S_n^{D^t} - nc \geq n[\alpha - c] \mid S_n^{A^s} = 0,\ S_n^{A^t} \geq n\alpha) \times \mathsf{P}(S_n^{A^t} \geq n\alpha), \end{aligned}$$

and

$$\begin{aligned} \mathsf{P}(S_n^{D^t} - nc \geq n[\alpha - c] \mid S_n^{A^s} = 0,\ S_n^{A^t} \geq n\alpha) \;= \\ \mathsf{P}(\min[S_n^{A^t} - nc, \ldots, S_1^{A^t} - c, 0] \geq n[\alpha - c] \mid S_n^{A^t} - nc \geq n[\alpha - c]). \end{aligned}$$

This conditional distribution converges to one as long as $\alpha \leq c$, in which case by convexity and the contraction principle the partial sum process for the test stream remains above the endpoint $n[\alpha - c]$, so we conclude that,

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}(S_n^{D^t} - nc \geq n[\alpha - c] \mid S_n^{A^s} = 0,\ S_n^{A^t} \geq n\alpha) = 0$$

The second term is bounded using the LDP for the arrival streams. Combining these two results we obtain the following upper bound:

$$\limsup_{n\to\infty} \frac{1}{n} \log \mathbb{P}(S_n^{D^t} \geq n\alpha) \leq 0 + \limsup_{n\to\infty} \frac{1}{n} \log \mathbb{P}(S_n^{A^t} \geq n\alpha) \leq -\Lambda_t^*(\alpha).$$

We prove the lower bound in a similar fashion. By conditioning and using the independence of the streams we obtain,

$$\begin{aligned}\mathbb{P}(S_n^{D^t} > n\alpha) &= \mathbb{P}(S_n^{D^t} - nc > n[\alpha - c]\ ) \geq \\ &\mathbb{P}(\min[S_n^{A^t} + S_n^{A^s} - nc, \ldots, S_1^{A^t} + S_1^{A^s} - c, 0] > n[\alpha - c] + S_n^{D^s} \mid S_n^{A^t} > n[\alpha + 2\epsilon], \mathcal{E}_n) \times \\ &\qquad \mathbb{P}(S_n^{A^t} > n[\alpha + 2\epsilon])\ \times\ \mathbb{P}(\mathcal{E}_n).\end{aligned}$$

Proceeding as above, we take the log and the normalized lim inf to find a lower bound. First note that

$$\liminf_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left(\frac{1}{n} S_n^{A^t} > \alpha + 2\epsilon\right) + \lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}(\mathcal{E}_n)\ \geq\ -\Lambda_t^*(\alpha + 2\epsilon) + 0.$$

Second we note that $S_n^{A^s} \geq S_n^{D^s}$ and that we have conditioned on $\mathcal{E}_n$, so the conditional probability is lower bounded by

$$\begin{aligned}&\mathbb{P}(\min[S_n^{A^t} + S_n^{A^s} - nc, \ldots, S_1^{A^t} + S_1^{A^s} - c, 0] > n[\alpha - c] + S_n^{A^s} \mid S_n^{A^t} > n[\alpha + 2\epsilon],\ \mathcal{E}_n)\ \geq \\ &\quad \mathbb{P}(\min[S_n^{A^t} + n[m - \epsilon] - nc, \ldots, S_1^{A^t} + m - n\epsilon - c, 0]\ > \\ &\qquad\qquad n[\alpha - c] + n[m + \epsilon] \mid S_n^{A^t} > n[\alpha + 2\epsilon])\ \geq \\ &\quad \mathbb{P}(\min[S_n^{A^t} + n[m - c], \ldots, S_1^{A^t} + m - c, 0] > n[\alpha + 2\epsilon + m - c] \mid S_n^{A^t} > n[\alpha + 2\epsilon]).\end{aligned}$$

Once more taking limits we have that

$$\begin{aligned}\lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}(\min[S_n^{A^t} + n[m - c], \ldots, S_1^{A^t} + m - c, 0]\ > \\ n[\alpha + 2\epsilon + m - c] \mid S_n^{A^t} > n[\alpha + 2\epsilon])\ =\ 0.\end{aligned}$$

As in Theorem 3.1 the conditional distribution converges to one as long as $\alpha + 2\epsilon \leq c - m$, where $m = \mathbb{E}A_1^s$, in which case by convexity and the contraction principle the partial sum process for the test stream remains above the endpoint $n[\alpha + 2\epsilon + m - c]$. Finally we let $\epsilon \to 0$ to obtain

$$\liminf_{n\to\infty} \frac{1}{n} \log \mathbb{P}(S_n^{D^t} > n\alpha) \geq -\liminf_{\epsilon\to 0} \Lambda_t^*(\alpha + 2\epsilon) = -\Lambda_t^*(\alpha)$$

Assuming Eq. 5 is in effect an argument identical to that in Corollary 3.1 shows that the effective bandwidth of the departures for our test stream is equal to that of the input:

$$\begin{aligned}\alpha_{D^t}(\delta) &= \frac{\Lambda_{D^t}(\delta)}{\delta} = \frac{1}{\delta}\sup_{0<\alpha<c}[\alpha\delta - \Lambda^*_{D^t}(\alpha)] \\ &= \frac{1}{\delta}\sup_{0<\alpha<c-\mathbf{E}A_1^t}[\alpha\delta - \Lambda^*_t(\alpha)] \\ &= \frac{1}{\delta}[\alpha^*_t(\delta)\delta - \Lambda_{D^t}(\alpha^*_t(\delta)] = \frac{\Lambda_t(\delta)}{\delta} = \alpha_t(\delta).\end{aligned}$$

□

3.1 Examples: Departure processes and traffic shaping via buffering

In [9], we considered the impact on the effective bandwidth of memoryless rejection (or marking) policies as well as rate filtering. We proved therein that among all memoryless rejection policies with the same throughput a threshold function is optimal in the sense that it minimizes the effective bandwidth at the output, but *only* for i.i.d. arrival streams. An examination of the impact of linear filtering with unit gain, in the case of Gaussian sources, suggested that the effective bandwidth is invariant to filtering. Here we consider the effect of traffic shaping via buffering.

The result in Corollary 3.1 suggests the impact of a buffered traffic shaping device. Suppose we are given the task of shaping a traffic stream such that the effective bandwidth is minimized. Moreover we are given a δ-constraint on the tail of the overflow probability, corresponding to a rough loss constraint. Alternatively a constraint on the delay before entering the network could be considered. The goal is to select the optimal deterministic release rate c such that the queue length satisfies the constraint, and the effective bandwidth of the departure process $\alpha_D(\delta)$ is minimized, i.e.,

$$\begin{aligned}&\min_{c>0} && \alpha_D(\delta) \\ &\text{such that} && \lim_{B\to\infty}\tfrac{1}{B}\log\mathbf{P}(W>B) \le -\delta\end{aligned} \tag{6}$$

The results in the corollary prove that it is optimal to serve the customers at precisely the effective bandwidth of the input, in the sense that this will minimize the effective bandwidth of the output while satisfying the tail constraint. In retrospect this result is quite intuitive: When the release rate is large, the stream is oblivious to the buffer and its effective bandwidth remains unchanged. When the service rate is reduced to the minimum acceptable level, i.e., the effective bandwidth of the input, then queueing in combination with deterministic release, work to our advantage by smoothing the traffic stream entering the network. Below we present two simple examples which should make these observations concrete.

Example 1: We begin by considering a simple example of buffering a Gaussian process

$\{A_n\}$ with mean μ and finite asymptotic variability

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

The log-moment generating function and its dual are given by:

$$\Lambda(\delta) = \mu\delta + \frac{\delta^2\sigma^2}{2}, \quad \Lambda^*(\alpha) = \frac{(\alpha-\mu)^2}{2\sigma^2}.$$

The effective bandwidth of the arrival process is $\alpha(\delta) = \mu + \frac{\delta\sigma^2}{2}$, while the decoupling bandwidth is $\alpha^*(\delta) = \mu + \delta\sigma^2$. The effective bandwidth of the departure process from a deterministic server at rate c is

$$\alpha_D(\delta) = \begin{cases} \alpha(\delta) & \text{if } \alpha^*(\delta) \le c \\ c - \frac{(c-\mu)^2}{2\sigma^2\delta} & \text{otherwise.} \end{cases}$$

Figure 2 exhibits the quantities we are considering as a function of δ. The effective bandwidth of the departure process is identical that of the arrival process for $\delta \le (c-\mu)/\sigma^2$, after which it is reduced and converges hyperbolically to the logical maximum of c, the service rate of the queue.

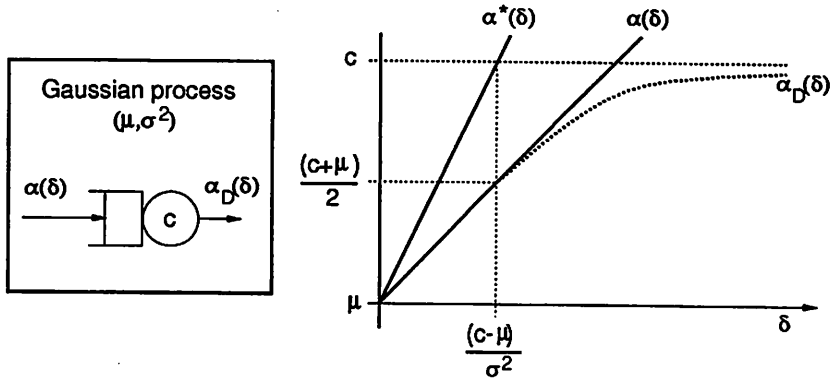

Figure 2: Input/Output effective bandwidths for a buffered for Gaussian source.

Note that although the notation is not suggestive, the effective bandwidth of the output *depends* on the service rate of the buffer. For given δ-constraint on buffer overflows, the effective bandwidth of the departure process is minimized when we serve at rate $c = \alpha(\delta)$ in which case we have

$$\alpha_D(\delta) = \mu + \frac{\delta\sigma^2}{2}[1 - \frac{\delta}{8}] \le \alpha(\delta) = \mu + \frac{\delta\sigma^2}{2}.$$

This expression exhibits explicitly the benefits of optimal traffic shaping via buffering in the sense of Eq. 7. A similar, but more involved calculation, for a buffer with a δ-constraint on the distributions of delays could be carried out.

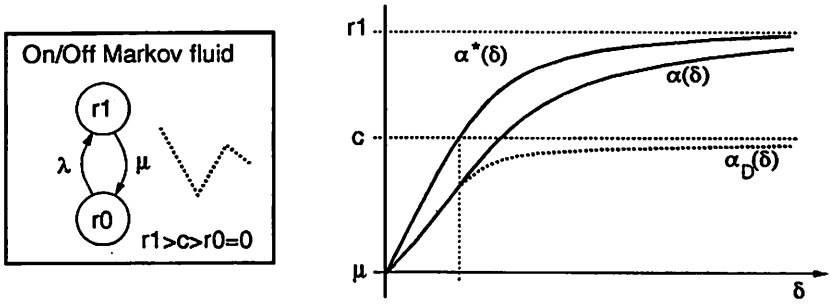

Figure 3: Input/Output effective bandwidths for a buffered On/Off Markov fluid source.

Example 2: A second example of interest is that of an On/Off Markov fluid. The source has the following dynamics: it turns off (on) with intensity μ (λ), and generates traffic at rate r_0 (r_1) when off (on). We will suppose $r_1 > c > r_0 = 0$. The log-moment generating function and its convex dual are given by (see [22, 8]):

$$\Lambda(\delta) = \frac{\delta r_1 - \mu - \lambda + \sqrt{(\delta r_1 - \mu + \lambda)^2 + 4\lambda\mu}}{2}, \quad \Lambda^*(\alpha) = \frac{\left[\sqrt{(\alpha\mu)} - \sqrt{(r_1 - \alpha)\lambda}\right]^2}{r_1}$$

where $m = \frac{\lambda r_1}{\lambda+\mu}$ is the mean arrival rate. After some algebra we find $\alpha(\delta) = \frac{\Lambda(\delta)}{\delta}$ and

$$\alpha^*(\delta) = \begin{cases} \frac{r_1}{2}\left(1 - \sqrt{\frac{(\delta r_1 - \mu + \lambda)^2}{4\lambda\mu + (\delta r_1 - \mu + \lambda)^2}}\right) & \text{if } \delta r_1 - \mu + \lambda < 0 \\ \frac{r_1}{2}\left(1 + \sqrt{\frac{(\delta r_1 - \mu + \lambda)^2}{4\lambda\mu + (\delta r_1 - \mu + \lambda)^2}}\right) & \text{otherwise} \end{cases}$$

The effective bandwidth of the departure process is then given by

$$\alpha_D(\delta) = \begin{cases} \alpha(\delta) & \text{if } \alpha^*(\delta) \leq c \\ c - \frac{[\sqrt{(c\mu)} - \sqrt{(r_1 - c)\lambda}]^2}{r_1\delta} & \text{otherwise.} \end{cases}$$

Figure 3 shows a typical graph of the input, output and decoupling bandwidths associated with an On/Off Markov fluid. The effective bandwidth of the input stream increases to the peak rate with δ while that of the output process converges to the service rate c of the buffer. Note if $c > r_1$ the buffer will not affect the stream at all.

The general solution of the optimal buffering problem in Eq. 7 is given by:

$$\alpha_D(\delta) = \alpha(\delta) - \frac{\Lambda^*(\alpha(\delta))}{\delta},$$

which in the case of the On/Off Markov fluid is

$$\alpha_D(\delta) = \alpha(\delta) - \frac{\left[\sqrt{(\alpha(\delta)\mu)} - \sqrt{(r_1 - \alpha(\delta))\lambda}\right]^2}{r_1\delta}.$$

3.2 Example: Asymptotics for tandem queues

In this section we consider the large buffer asymptotics for stable queues in tandem. As in the single queue example there exists a stationary distribution, see Walrand [26] page 245.

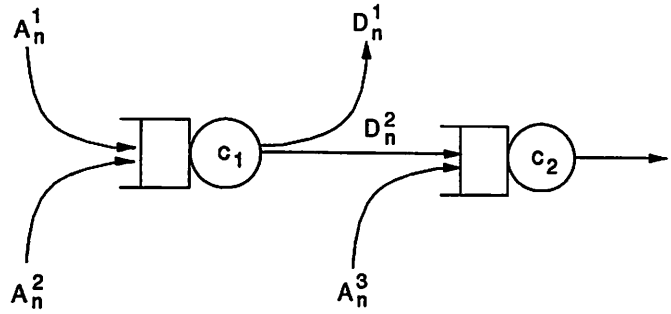

Figure 4: Tandem queues.

Using our characterization for the stationary output process we discuss resource management for a simple scenario shown in Figure 4. The figure shows two queues, shared by three traffic streams, we will assume the effective and decoupling bandwidths of the three streams are given by $\alpha_i(\delta)$, $\alpha_i^*(\delta)$, $i = 1,2,3$. We suppose the goal is to guarantee that

$$\lim_{B\to\infty} \frac{1}{B} \log \mathbb{P}(W_i > B) \leq -\delta \text{ for } i = 1,2.$$

Using the result in Corollary 2.1, the above constraint will be satisfied for the first queue if

$$\alpha_1(\delta) + \alpha_2(\delta) \leq c_1.$$

Now consider the second queue. If the service rate of the first queue is such that

$$\alpha_1(0) + \alpha_2^*(\delta) \leq c_1.$$

then by Corollary 3.2 the effective bandwidth of the stream $\{D_n^2\}$ entering the second queue is the same as the arrival process $\{A_n^2\}$ i.e., equals $\alpha_2(\delta)$. Thus the constraint on the second queue is guaranteed if

$$\alpha_2(\delta) + \alpha_3(\delta) \leq c_2.$$

4 Resource management for networks

In this section we reconsider the example in §3.2 and extend our results to a generalized networking scenario for which we discuss some of the practical aspects of resource management.

We begin by making two observations on the tandem queue example. First, note that in order to guarantee decoupling we only need constraints on the aggregate traffic flow

along routes sharing multiple queues (e.g., $\alpha_2^*(\delta) < c_1 - \alpha_1(0)$). Thus resource management based on effective bandwidths should be organized on a per route basis. Second, note that the decoupling constraint may or may not be subsumed by the effective bandwidth constraint (i.e., $\alpha_1(0) + \alpha_2^*(\delta) \stackrel{?}{\leq} \alpha_1(\delta) + \alpha_2(\delta)$). Clearly it is advantageous to have a unique criterion for resource management. In Remark 3.1 we discussed a tight upper bound for the decoupling bandwidth showing that for a single source this constraint will not be outrageous. In practice, for a queue shared by multiple non-interacting routes, (i.e., multiple streams sharing only this queue) the decoupling constraint will often be in effect once the effective bandwidth requirements are satisfied. So guaranteeing decoupling will typically not be more conservative than the effective bandwidth requirement. However the smoothing effect of queues (suggested by Corollary 3.1) is expected to make an effective bandwidth scheme based on the statistics of traffic at the network edge somewhat conservative for a queue deep in the network. A final point to address is the conservative nature of effective bandwidths in and of themselves. An enlightening discussion of this issue can be found in Choudhury et al. [5]; they have shown regimes for which the tail asymptotics considered herein are not very precise, while others where they are seen to work quite well. Undoubtedly one should be careful in interpreting these asymptotic results, however subject to verification these approximations provide a reasonably simple integrated approach to resource management.

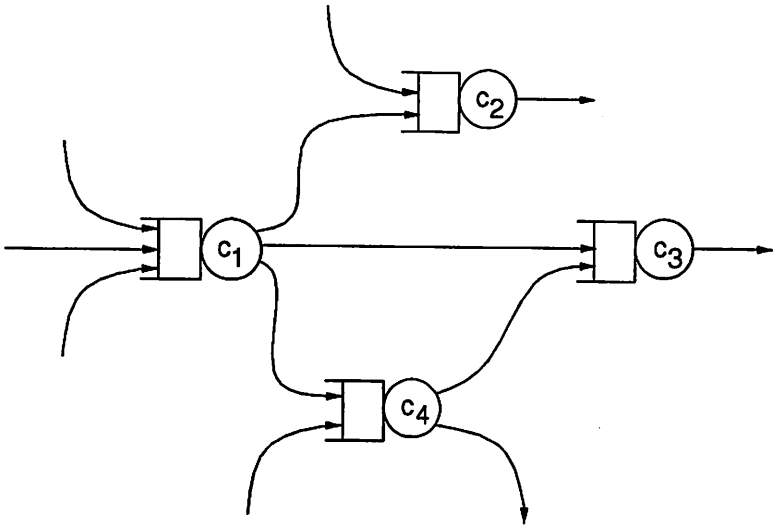

Figure 5: Resource management for networks.

Consider a set of queues Q. Let R denote a collection of routes, were each route is determined as an ordered subset of Q. Let $R(i)$, denote the set of routes sharing queue i truncated before queue i, i.e., we need only consider the downstream paths. In order to guarantee decoupling we must ensure that the traffic flows on routes, $R(i)$, are decoupled with respect to downstream queues. If the routes in $R(i)$ are such that when $r_1, r_2 \in R(i)$ then $r_1 \cap r_2 = \{i\}$ then the decoupling constraints at queue i would be given by

$$\alpha_r^*(\delta) + \sum_{j \in R(i)-r} \alpha_j(0) < c_i.$$

for all $r \in R(i) - \{i\}$, where $\alpha_r(\delta), \alpha_r^*(\delta)$ denote the effective and decoupling bandwidths of the traffic flowing along route r. However, in order to have flexible routing which is failure proof and distributes loads, networks are likely to have alternative routes leading to the same destinations, such is the case in Figure 5. In this case the aggregate flows sharing downstream queues, must satisfy a decoupling constraint at upstream queues. The traffic flows in any collection of routes in $R(i)$ sharing downstream queues will need to satisfy a decoupling constraint which is in fact stronger than those above. For example the network in Figure 5 has seven possible routes $\{\{1,2\},\{1,3\},\{1,4\},\{1,4,3\},\{2\},\{4\},\{4,3\}\}$. Due to the downstream interaction of routes $\{1,3\}$ and $\{1,4,3\}$ we require that in addition to the individual decoupling constraints above we have

$$\alpha^*_{\{1,3\}}(\delta) + \alpha^*_{\{1,4,3\}}(\delta) + \sum_{j \in R(1) - \{\{1,3\},\{1,4,3\}\}} \alpha_j(0) < c_1.$$

Resource management based on such a scheme, guarantees loss constraints for *aggregate* traffic in the network, while complex service policies at each node may tradeoff performance among individual users requiring different qualities of service.

5 Summary

In this paper we have obtained a series of novel results on the large buffer asymptotics for multi-class traffic streams sharing a feed-forward network of queues. We began by identifying the large deviation rate function for the output of a multi-class queue for aggregate and individual streams. This result showed explicitly the smoothing properties of queues in terms of the reduction of the effective bandwidth of streams. We introduced the notion of *decoupling bandwidths*, to ensure that the asymptotics for every queue in a feed-forward network are essentially decoupled, and hence that resource management can be carried out by considering queues individually using the effective bandwidths for streams as specified by users at the network edge. These results are only valid for networks with "large" shared buffers, but allow arbitrary work conserving service policies. Further research will focus on the interplay between service policies which can successfully guarantee specific qualities of service to certain users and the asymptotics of large but finite buffers.

References

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

[2] S. Asmussen. Conditional limit theorems relating the random walk to its associate, with appplications to risk processes and the GI/G/1 queue. *Advances in Applied Probability*, 14:143–170, 1982.

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

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

[5] G. Choudhury, D. Lucantoni, and W. Whitt. Squeezing the most out of ATM. *Preprint*, 1993.

[6] C. Courcoubetis, G. Kesidis, A. Ridder, J. Walrand, and R. Weber. Call acceptance and routing using inferences from measured buffer occupancy. *to appear in IEEE Trans. Comm.*, 1993.

[7] G. de Veciana. *Design Issues in ATM Networks: Traffic Shaping and Congestion Control.* PhD thesis, Dept. of EECS, University of California, Berkeley, 1993.

[8] 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.

[9] G. de Veciana and J. Walrand. Effective bandwidths: Call admission, traffic policing and filtering for ATM networks. *Submitted to Queueing Systems*, 1993.

[10] G. de Veciana and J. Walrand. Traffic shaping for ATM networks: Asymptotic analysis and simulations. *Submitted to IEEE/ACM Trans. Networking*, 1993.

[11] A. Dembo and T. Zajic. Large deviations for sample path of partial sums. *Preprint*, 1992.

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

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

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

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

[16] M. R. Frater. *Estimation of the Statistics of Rare Events in Data Communications Systems.* PhD thesis, Dept. of Systems Engineering Research School of Physical Sciences, The Australian National University, 1990.

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

[18] R. Guérin, H. Ahmadi, and M. Naghshineh. Equivalent capcity and its application to bandwidth allocation in high-speed networks. *IEEE J. Select. Areas Commun.*, 9:968–981, 1991.

[19] J. Y. Hui. Resource allocation for broadband networks. *IEEE J. Select. Areas Commun.*, 6:1598–1608, 1988.

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

[21] G. Kesidis and J. Walrand. Traffic policing in ATM networks. *Submitted to IEEE/ACM Trans. Networking*, 1993.

[22] G. Kesidis, J. Walrand, and C.S. Chang. Effective bandwidths for multiclass Markov fluids and other ATM sources. *IEEE/ACM Trans. Networking.*, Aug. 1993.

[23] J.F. Kurose. On computing per-session performance bounds in high-speed multi-hop computer networks. In *Proc. ACM SIGMETERICS and PERFORMANCE '92*, pages 128–139, Newport, Rhode Island, June 1992.

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

[25] S. Parekh and J. Walrand. A quick simulation of excessive backlogs in networks of queues. *IEEE Trans. Automatic Control*, 34:54–66, January 1989.

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

[27] W. Whitt. Tail probabilities with statistical multiplexing and effective bandwidths in multi-class queues. *To appear in Telecommunication Systems*, 1993.