Effective Bandwidths of Departure Processes from Queues with Time Varying Capacities

Cheng-Shang Chang*
Dept. of Electrical Engineering
National Tsing Hua University
Hsinchu 30043
Taiwan, R.O.C.

Tim Zajic
CIMAT
Apartado Postal 402
36000 Guanajuato, GTO.
México

Abstract

In this paper, we consider a queue with a time varying capacity *and identify the effective bandwidth of the* stationary *departure process from such a queue. Two important observations are made: (i) the effective bandwidths for the* transient *departure process and the* stationary *departure process from such a queue are in general different, and (ii) sometimes it is necessary to build up the queue first in order to have a large excursion of the stationary departure process. The new result on the effective bandwidth of the stationary departure process is applied to intree networks with time varying capacities and priority tandem queues. Algorithms for approximating the tail distributions of queue lengths in such networks are derived.*
Keywords: grade of service, effective bandwidth, large deviations

1 Introduction

Recently, various researchers ([18, 19, 15, 4, 21, 7, 13, 27, 16, 5, 20, 23, 28, 29, 11], among many others) have proposed and developed the theory of effective bandwidth as a tentative solution for various problems that arise in high speed digital networks, in particular ATM networks. This theory provides the asymptotic decay rate of the tail distribution of the queue length at a local node, i.e.,

$$\lim_{x\to\infty} \frac{\log Pr(q \geq x)}{x} = -\theta^*, \tag{1}$$

where q is the stationary queue length and θ^* is the decay rate.

In Chang [3], a general framework for the theory of effective bandwidth was first developed. In this framework, the general form of effective bandwidth was derived in terms of the asymptotic logarithm of the moment generating function. Moreover, a set of rules for network operations was derived: (i) multiplexing, (ii) input-output relation, and (iii) routing. These rules allow one to extend the single queue result to networks. However, the development of this framework was pure algebraic and it concentrated on the upper bound of (1). In Kesidis, Walrand and Chang [21], Chang, Heidelberger, Juneja and Shahabuddin [7], Chang [4] and Whitt [27] , the connection to the theory of large deviations was made. Using the Legendre transform, the large deviation approach provided a probabilistic derivation of the general form of effective bandwidths in [3]. Also, it filled the gap for the lower bound needed for (1) in the single queue case. The lower bounds needed for intree networks with routing were justified by Chang [6] using the sample path large deviation principle and the contraction principle.

With the exception of O'Connell [24], it appears that all previous work on networks has assumed a constant capacity at each local node. In this paper, we consider a queue with a *time varying capacity* and identify the effective bandwidth of the *stationary* departure process from such a queue. For results concerning the *transient* case, see [24]. Two important observations are made in the present paper:

(i) The effective bandwidths for the *transient* departure process (e.g., the departure process from an empty queue at time 0) and the *stationary* departure process from such a queue are in general different. However, the departure process from an empty queue at time 0 and the stationary departure process have the same effective bandwidth when the queue has a constant capacity.

(ii) In order to have a large excursion of the stationary departure process, sometimes it is necessary to build up the queue first at the rate characterized by (1).

Our extension allows us to modify the algorithm in

*Supported in part by the National Science Council, Taiwan, R.O.C., under Contracts NSC 83-0208-M007-091 and NSC 83-0404-E007-052.

Chang [5, 6] for approximating the tail distributions of queue lengths in intree networks with time varying capacities. Since a multiple class queue under a head-of-line priority scheme can be modelled by a queue with a time varying capacity, our extension can also be applied for analyzing the performance of a session in an ATM network with priorities.

The paper is organized as follows: in Section 2, we identify the effective bandwidth of the *stationary* departure process from a queue with a time varying capacity. We then apply this result to intree networks and priority tandem queues in Section 3. The paper is concluded in Section 4.

2 Effective bandwidths of departure processes

Consider a discrete-time queue with one class of customers. Let $a(t)$ and $q(t)$ be the number of arrivals at time t and the number of customers in the queue at time t respectively. Assume that the buffer size is infinite and that the server can serve $c(t)$ customers per unit of time. The quantity $c(t)$ will be referred to as the capacity of the server at time t. Throughout, assume that both $a(t)$ and $c(t)$ are nonnegative. This model is different from Chang [4] in that the capacity is allowed to be time varying. Under a work-conserving policy, i.e., a policy that does not allow idling when there are customers in the queue, the queue is governed by the following Lindley's equation:

$$q(t+1) = (q(t) + a(t+1) - c(t+1))^+ \qquad (2)$$

where $(x)^+ \equiv \max(0, x)$. It is well known (e.g., see Loynes [22], Baccelli and Bremaud [2]) that $q(t)$ converges in distribution to a finite random variable $q(\infty)$ if both $a(t)$ and $c(t)$ are stationary and ergodic, and $Ea(t) < Ec(t)$.

Letting $A(t_1, t_2) = \sum_{t=t_1+1}^{t_2} a(t)$ and $C(t_1, t_2) = \sum_{t=t_1+1}^{t_2} c(t)$, we make the following assumptions.

(A1) $a(t)$ and $c(t)$ are independent.

(A2) For all $\theta \in I\!R$, $\lim_{t\to\infty} \frac{1}{t} \log Ee^{\theta A(0,t)} = \Lambda_A(\theta)$, and $\Lambda_A(\theta)$ is differentiable.

(A3) For all $\theta \in I\!R$, $\lim_{t\to\infty} \frac{1}{t} \log Ee^{\theta C(0,t)} = \Lambda_C(\theta)$, and $\Lambda_C(\theta)$ is differentiable.

(A4) Both $a(t)$ and $c(t)$ are stationary and ergodic, and $Ea(t) < Ec(t)$.

Let $\Lambda_A^*(\alpha)$ be the Legendre transform of $\Lambda_A(\theta)$, i.e,

$$\Lambda_A^*(\alpha) = \sup_\theta [\theta\alpha - \Lambda_A(\theta)]. \qquad (3)$$

Also, let $\Lambda_C^*(\alpha)$ be the Legendre transform of $\Lambda_C(\theta)$. Assumptions (A2-3), known as the Gärtner-Ellis limits [14, 12], imply both $\{A(0,t)/t, t \geq 0\}$ and $\{C(0,t)/t, t \geq 0\}$ satisfy the large deviation principle with good convex rate functions $\Lambda_A^*(\alpha)$ and $\Lambda_C^*(\alpha)$, respectively.

The following theorem shows the decay rate of the tail distribution of the stationary queue length.

Theorem 2.1 *Under appropriate conditions (cf. Chang [4], Kesidis, Walrand and Chang [21], Whitt [27]), if there exists a unique* $\theta^* > 0$ *such that*

$$\Lambda_A(\theta^*) + \Lambda_C(-\theta^*) = 0, \qquad (4)$$

then

$$\lim_{x\to\infty} \frac{\log Pr(q(\infty) \geq x)}{x} = -\theta^*. \qquad (5)$$

The proof of Theorem 2.1 is composed of two parts: the upper bound and the lower bound. The upper bound can be proved directly by an algebraic approach (see Chang [4], Lemma 3.7). The lower bound utilizes the large deviation principle.

In particular, when the capacity is a fixed constant, i.e., $c(t) = c$ for all t, one has from (4) that

$$\frac{\Lambda_A(\theta^*)}{\theta^*} = c. \qquad (6)$$

Since c is the capacity of the server, $\frac{\Lambda_A(\theta^*)}{\theta^*}$ is called the effective bandwidth of the arrival process subject to the condition that the tail distribution of the queue length has the decay rate θ^*. Let $a^*(\theta) = \frac{\Lambda_A(\theta)}{\theta}$. In view of (6), we shall call the function $a^*(\theta)$ the effective bandwidth function of the arrival process $a(t)$.

To apply Theorem 2.1 to a tandem network, one must obtain the effective bandwidth function (or the Gärtner-Ellis limit) of the *stationary* departure process. For the case with constant capacities, this result was previously obtained in Chang [6]. However, the case without constant capacities is more complicated. One of the main contributions in this paper is to identify the effective bandwidth function of the *stationary* departure process. Since a formal proof for our new result is quite lengthy, we will only provide the derivation with intuition in this conference version. In addition, we remark that the necessary assumptions are admittedly quite technical in this version and their weakening will be a subject of future research.

Let $b(t)$ be the number of departures from the *stationary* queue in (2) at time t. Similarly, define $B(t_1, t_2) = \sum_{t=t_1+1}^{t_2} b(t)$. Note that $\{b(t), t \geq 0\}$ is stationary and ergodic with mean $Eb(t) = Ea(t)$. In order to derive the effective bandwidth function of the departure process, one needs conditions stronger than the Gärtner-Ellis limits. As in Chang [6], we consider the following two conditions.

(A2′) $\{a(t), t = 0, 1, 2, \ldots,\}$ is bounded, nonnegative integer-valued, and adapted to a filtration $\mathcal{F}_t^A$ such that for all $\theta \in I\!R$, $s, t \geq 0$, for some function $0 \leq \Gamma_A(\theta) < \infty$ (independent of s and t),

$$\Lambda_A(\theta)t - \Gamma_A(\theta) \leq \log E(e^{\theta A(s,t+s)} | \mathcal{F}_s^A) \leq \Lambda_A(\theta)t + \Gamma_A(\theta), \quad a.s. \tag{7}$$

where $\Lambda_A(\theta)$ is defined in (A2). Moreover, $\Lambda_A(\theta)$ is differentiable for all θ.

(A3′) $\{c(t), t = 0, 1, 2, \ldots,\}$ is bounded, nonnegative integer-valued, and adapted to a filtration $\mathcal{F}_t^C$ such that for all $\theta \in I\!R$, $s, t \geq 0$, for some function $0 \leq \Gamma_C(\theta) < \infty$ (independent of s and t),

$$\Lambda_C(\theta)t - \Gamma_C(\theta) \leq \log E(e^{\theta C(s,t+s)} | \mathcal{F}_s^C) \leq \Lambda_C(\theta)t + \Gamma_C(\theta), \quad a.s. \tag{8}$$

where $\Lambda_C(\theta)$ is defined in (A3). Moreover, $\Lambda_C(\theta)$ is differentiable for all θ.

These two conditions imply both $a(t)$ and $c(t)$ satisfy the sample path large deviation principle, a stronger concept than the large deviation principle (cf. [6, 10]).

The following condition allows us to deal with the fact that in general, for t fixed, $a(t)$ and $q(0)$ are dependent, as well as $c(t)$ and $q(0)$.

(A5) There exists $0 < \nu < 1$ and a positive integer n such that

$$\sup_j \sup_{A \in \mathcal{F}_{-\infty}^j, B \in \mathcal{F}_{j+n}^\infty} \{ \frac{|P(A \cap B) - P(A)P(B)|}{P(A)P(B)} \} \leq \nu, \tag{9}$$

where $\mathcal{F}_{-\infty}^j = \sigma(a(t),\ -\infty < t \leq j)$ and $\mathcal{F}_{j+n}^\infty = \sigma(a(t),\ j+n \leq t < \infty)$. Similarly (9) remains valid with $a(t)$ replaced by $c(t)$.

We remark that our proof technique remains valid for other mixing conditions (for example, condition (S) of [10]). Indeed, a challenging problem is to find an appropriate such condition which ensures nice mixing properties of the departure process.

Finally, the following assumption is needed to justify applying the large deviation lower bound to closed sets in the proof of Theorem 2.2.

(A6) The following lower bound for the sequence $\{c(t)\}$ holds for all $\alpha \geq \Lambda_A'(0)$,

$$\liminf_{t \to \infty} \frac{1}{t} Pr(C(0,t)/t \geq \alpha) \geq - \inf_{\gamma \geq \alpha} \Lambda_C^*(\gamma) \ . \tag{10}$$

We remark that (A6) holds in case the sequence $\{c(t)\}$ is i.i.d. (see [9, Corollary 2.2.19]). In case (A6) fails, Theorem 2.2 remains valid with the further restriction that $\alpha < \sup\{\beta : \Lambda_C^*(\beta) < \infty\}$.

Theorem 2.2 *Under (A1), (A2′), (A3′), and (A4-6) if there exists a unique $\theta^* > 0$ satisfying (4) and (5), then*

$$\lim_{t \to \infty} \frac{1}{t} \log Pr(B(0,t) \geq \alpha t) = -\Lambda_B^*(\alpha),\ \alpha \geq \Lambda_A'(0), \tag{11}$$

$$\lim_{t \to \infty} \frac{1}{t} \log E e^{\theta B(0,t)} = \Lambda_B(\theta), \quad \theta \geq 0, \tag{12}$$

where

$$\Lambda_B^*(\alpha) = \theta^* \alpha - \sup_{\gamma \leq \alpha} [\theta^* \gamma - \Lambda_A^*(\gamma)] + \inf_{\alpha \leq \gamma} \Lambda_C^*(\gamma), \tag{13}$$

$$\Lambda_B(\theta) = \sup_{\alpha \geq \Lambda'_A(0)} [\theta \alpha - \Lambda_B^*(\alpha)]. \tag{14}$$

Thus, the effective bandwidth function of the stationary departure process, denoted by $b^*(\theta)$, $\theta \geq 0$, has the form $\Lambda_B(\theta)/\theta$.

Remark 2.3 Note that $\Lambda_B^*(\alpha)$ in (13) consists of the following four cases.

Case 1. $\alpha \leq \Lambda_A'(\theta^*)$ and $\alpha \leq \Lambda_C'(0)$:

In this case, $\sup_{\gamma \leq \alpha}[\theta^* \gamma - \Lambda_A^*(\gamma)] = \theta^* \alpha - \Lambda_A^*(\alpha)$ and $\inf_{\alpha \leq \gamma} \Lambda_C^*(\gamma) = 0$. Thus, $\Lambda_B^*(\alpha) = \Lambda_A^*(\alpha)$.

Case 2. $\alpha \leq \Lambda_A'(\theta^*)$ and $\alpha > \Lambda_C'(0)$:

In this case, $\sup_{\gamma \leq \alpha}[\theta^* \gamma - \Lambda_A^*(\gamma)] = \theta^* \alpha - \Lambda_A^*(\alpha)$ and $\inf_{\alpha \leq \gamma} \Lambda_C^*(\gamma) = \Lambda_C^*(\alpha)$. Thus, $\Lambda_B^*(\alpha) = \Lambda_A^*(\alpha) + \Lambda_C^*(\alpha)$.

Case 3. $\alpha > \Lambda_A'(\theta^*)$ and $\alpha > \Lambda_C'(0)$:

In this case, $\sup_{\gamma \leq \alpha}[\theta^* \gamma - \Lambda_A^*(\gamma)] = \sup_\gamma[\theta^* \gamma - \Lambda_A^*(\gamma)] = \Lambda_A(\theta^*)$ and $\inf_{\alpha \leq \gamma} \Lambda_C^*(\gamma) = \Lambda_C^*(\alpha)$. Thus, $\Lambda_B^*(\alpha) = \theta^* \alpha - \Lambda_A(\theta^*) + \Lambda_C^*(\alpha)$.

Case 4. $\alpha > \Lambda_A'(\theta^*)$ and $\alpha \leq \Lambda_C'(0)$:

In this case, $\sup_{\gamma \leq \alpha}[\theta^* \gamma - \Lambda_A^*(\gamma)] = \Lambda_A(\theta^*)$ and $\inf_{\alpha \leq \gamma} \Lambda_C^*(\gamma) = 0$. Thus, $\Lambda_B^*(\alpha) = \theta^* \alpha - \Lambda_A(\theta^*)$.

Example 2.4 (Constant capacity) In the case $c(t) = c$ for all t, $\Lambda_C(\theta) = c\theta$ and $\Lambda_C^*(\alpha) = 0$ if $\alpha = c$ and $\Lambda_C^*(\alpha) = \infty$ otherwise. Thus, we only have Cases 1-3. In view of $\Lambda_C'(0) = c$,

$$\Lambda_B^*(\alpha) = \begin{cases} \Lambda_A^*(\alpha) & \text{if } \alpha \leq c \\ \infty & \text{otherwise.} \end{cases} \tag{15}$$

Taking the Legendre transform,

$$\Lambda_B(\theta) = \begin{cases} \Lambda_A(\theta) & \text{if } \theta \leq \tilde{\theta} \\ \theta c - \tilde{\theta} c + \Lambda_A(\tilde{\theta}) & \text{otherwise} \end{cases} \tag{16}$$

where $\tilde{\theta}$ is a solution of $\Lambda_A'(\theta) = c$. This was previously obtained in [7, 5, 6].

Example 2.5 (Constant arrival rate) In the case $a(t) = \alpha_0$ for all t, $\Lambda_A(\theta) = \alpha_0\theta$ and $\Lambda_A^*(\alpha) = 0$ if $\alpha = \alpha_0$ and $\Lambda_A^*(\alpha) = \infty$ otherwise. Thus, we only have Cases 3 and 4, i.e.,

$$\Lambda_B^*(\alpha) = \begin{cases} \theta^*(\alpha - \alpha_0) & \text{if } \alpha_0 \le \alpha \le \Lambda_C'(0) \\ \theta^*(\alpha - \alpha_0) + \Lambda_C^*(\alpha) & \text{if } \alpha \ge \Lambda_C'(0) \end{cases} \tag{17}$$

where θ^* is the unique solution of $\alpha_0 = \frac{\Lambda_c(-\theta)}{-\theta}$. Taking the Legendre transform,

$$\Lambda_B(\theta) = \begin{cases} \alpha_0\theta & \text{if } 0 \le \theta \le \theta^* \\ \alpha_0\theta^* + \Lambda_C(\theta - \theta^*) & \text{if } \theta > \theta^*. \end{cases} \tag{18}$$

Though the $M/M/1$ queue in the following example does not satisfy our assumptions, it is interesting to carry out the computation and compare it with the known results. Extensions from our discrete-time queueing models to continuous time models could be possible by the inversion formula in Glynn and Whitt [17].

Example 2.6 ($M/M/1$ queue) Consider an $M/M/1$ queue. Assume the arrival process is a Poisson process with rate λ and the service times are independent and exponentially distributed with mean $1/\mu$. Thus, one has

$$\Lambda_A(\theta) = \lambda(e^\theta - 1) \tag{19}$$

$$\Lambda_A^*(\alpha) = \alpha \log \frac{\alpha}{\lambda} - \alpha + \lambda. \tag{20}$$

Similarly,

$$\Lambda_C(\theta) = \mu(e^\theta - 1) \tag{21}$$

$$\Lambda_C^*(\alpha) = \alpha \log \frac{\alpha}{\mu} - \alpha + \mu. \tag{22}$$

In this example, $\theta^* = \log \frac{\mu}{\lambda}$. Since $\Lambda_A'(\theta^*) = \lambda e^{\theta^*} = \mu = \Lambda_C'(0)$, we only have Cases 1 and 3 in Theorem 2.2.

Case 1. $\lambda \le \alpha < \mu$:

$$\Lambda_B^*(\alpha) = \Lambda_A^*(\alpha) = \alpha log \frac{\alpha}{\lambda} - \alpha + \lambda. \tag{23}$$

Case 3. $\alpha \ge \mu$:

$$\Lambda_B^*(\alpha) = \theta^*\alpha - \Lambda_A(\theta^*) + \Lambda_C^*(\alpha) = \alpha log \frac{\alpha}{\lambda} - \alpha + \lambda. \tag{24}$$

This is consistent with the well known result that the stationary departure process from an $M/M/1$ queue is also a Poisson process.

Example 2.6 leads us to consider the case when the stationary departure process has the same effective bandwidth function as that of the arrival process. In the following example, we show this holds if the capacity process is a θ^*-conjugate process of the arrival process (For the notion of conjugate process, see p[7] and references therein).

Example 2.7 (Conjugate capacity process) ¿From Example 2.6, we note that if $\Lambda_C(\theta) = \Lambda_A(\theta + \theta^*) - \Lambda_A(\theta^*)$, then $\Lambda_C'(0) = \Lambda_A'(\theta^*)$ and we only have Cases 1 and 3. In this example, one has $\Lambda_C^*(\alpha) = \Lambda_A^*(\alpha) - \theta^*\alpha + \Lambda_A(\theta^*)$. Thus, $\Lambda_B^*(\alpha) = \Lambda_A^*(\alpha)$ for all $\alpha \ge Ea(t)$ and $\Lambda_B(\theta) = \Lambda_A(\theta)$ for all $\theta \ge 0$. A sufficent condition for $\Lambda_C(\theta) = \Lambda_A(\theta+\theta^*) - \Lambda_A(\theta^*)$ is that $\{c(t), t \ge 0\}$ is a θ^*-conjugate process of $\{a(t), t \ge 0\}$.

In the following, we sketch the derivation of the effective bandwidth function of the stationary departure process in Theorem 2.2.

(Upper bound) Note that the number of departures in the interval $(0, t]$ is bounded above by the sum of the number of arrivals in that interval and the number of customers in the queue at time 0. Also, it is bounded above by the capacity. Thus,

$$B(0,t) \le \min[q(0) + A(0,t), C(0,t)]. \tag{25}$$

Since we assume that the queue has reached its steady state, $q(0)$ has the same distribution as $q(\infty)$. In general, both $A(0,t)$ and $C(0,t)$ are not *independent* of $q(0)$. However, in view of (A5)

$$\begin{aligned} &Pr(\min[q(0) + A(0,t), C(0,t)] \ge \alpha t) \qquad (26) \\ \approx\ &Pr(q(0) + A(0,t) \ge \alpha t)Pr(C(0,t) \ge \alpha t) \end{aligned}$$

when t is large. ¿From the large deviation principle for $C(0,t)$, it follows that

$$Pr(C(0,t) \ge \alpha t) \approx \exp(-t \inf_{\alpha \le \gamma} [\Lambda_C^*(\gamma)]). \tag{27}$$

Now write, using (A5)

$$\begin{aligned} &Pr(q(0) + A(0,t) \ge \alpha t) \qquad (28) \\ \approx &\int Pr(q(0) + \gamma t \ge \alpha t) dPr(A(0,t)/t \le \gamma). \end{aligned}$$

Using (5), the large deviation principle for $A(0,t)$ and proceeding as in the proof of the upper bound in Varadhan's integral lemma (see, for example, [9, Section 4.3]), it follows that

$$Pr(q(0) + A(0,t) \ge \alpha t) \approx e^{-\theta^*\alpha + \sup_{\gamma \le \alpha}[\theta^*\gamma - \Lambda_A^*(\gamma)]}. \tag{29}$$

In conjunction with (26) and (27),

$$Pr(\min[q(0)+A(0,t),C(0,t)] \geq \alpha t) \approx \exp(-t\Lambda_B^*(\alpha)), \quad (30)$$

where $\Lambda_B^*(\alpha)$ is defined in (13). In view of (25), the derivation is concluded.

(**Lower bound)** Our proof for the lower bound is based on the sample path large deviation principle in Chang [6]. The basic idea is to identify the path that leads to a large excursion. Let $A^{(n)}(t) = A(0,\lceil nt\rceil)/n$, $B^{(n)}(t) = B(0,\lceil nt\rceil)/n$, and $C^{(n)}(t) = C(0,\lceil nt\rceil)/n$. Under assumptions (A1), (A2′), (A3′) and (A4), $A^{(n)}(t)$ (resp. $C^{(n)}(t)$) satisfies the sample path large deviation principle with the rate function $\int_0^1 \Lambda_A^*(\phi'(t))dt$ (resp. $\int_0^1 \Lambda_C^*(\phi'(t))dt$), where $\phi(0)=0$ and $\phi(t)$ is absolutely continuous in $[0,1]$. Intuitively, one has from the sample path large deviation principle that for small $\epsilon > 0$

$$Pr(\sup_{0\leq t\leq 1} |A^{(n)}(t)-\phi(t)| \geq \epsilon) \approx e^{-n\int_0^1 \Lambda_A^*(\phi'(t))dt}. \quad (31)$$

¿From (2), one can easily show that

$$B(0,t) = \min[q(0) + \inf_{0<s\leq t}[A(0,s)+C(s,t)], C(0,t)]. \quad (32)$$

Let $q^{(n)}(0) = q(0)/n$ and note that

$$\begin{aligned} B^{(n)}(t) &= \min[q^{(n)}(0) \\ &+ \inf_{0<s\leq t}[A^{(n)}(s)+C^{(n)}(t)-C^{(n)}(s)], C^{(n)}(t)]. \end{aligned} \quad (33)$$

We will prove the lower bound by considering the following four cases:

Case 1. $\alpha \leq \Lambda_A'(\theta^*)$ and $\alpha \leq \Lambda_C'(0)$:

Note that from (33)

$$\begin{aligned} &Pr(B^{(n)}(1) \geq \alpha) \\ &\geq Pr(A^{(n)}(s) \geq \alpha s, 0\leq s\leq 1) \\ &\quad Pr(C^{(n)}(1)-C^{(n)}(s) \geq \alpha(1-s), 0\leq s\leq 1). \end{aligned} \quad (34)$$

¿From the sample path large deviation principle for $A^{(n)}(t)$ and $C^{(n)}(t)$, it follows that

$$\begin{aligned} &Pr(A^{(n)}(s) \geq \alpha s, 0\leq s\leq 1) \\ &\approx \exp(-n \inf_{\{\phi(s)\geq\alpha s, 0\leq s\leq 1\}} \int_0^1 \Lambda_A^*(\phi'(t))dt) \end{aligned} \quad (35)$$

and

$$\begin{aligned} &Pr(C^{(n)}(1)-C^{(n)}(s) \geq \alpha(1-s), 0\leq s\leq 1) \approx \\ &\exp(-n \inf_{\{\phi(1)-\phi(s)\geq\alpha(1-s), 0\leq s\leq 1\}} \int_0^1 \Lambda_C^*(\phi'(t))dt). \end{aligned} \quad (36)$$

¿From the convexity of $\Lambda_A^*(\cdot)$,

$$\int_0^1 \Lambda_A^*(\phi'(t))dt \geq \Lambda_A^*(\int_0^1 \phi'(t)dt) = \Lambda_A^*(\phi(1)). \quad (37)$$

Thus, for any ϕ satisfying $\phi(s) \geq \alpha s$ for $0 \leq s \leq 1$, one has

$$\int_0^1 \Lambda_A^*(\phi'(t))dt \geq \inf_{\gamma\geq\alpha} \Lambda_A^*(\gamma). \quad (38)$$

Since $\alpha \geq \Lambda_A'(0)$, $\inf_{\gamma\geq\alpha}\Lambda_A^*(\gamma) = \Lambda_A^*(\alpha)$. Thus, the right hand side of (38) is achieved when $\phi(s) = \alpha s$. This implies

$$\inf_{\{\phi(s)\geq\alpha s, 0\leq s\leq 1\}} \int_0^1 \Lambda_A^*(\phi'(t))dt = \Lambda_A^*(\alpha). \quad (39)$$

and

$$Pr(A^{(n)}(s) \geq \alpha s, 0\leq s\leq 1) \approx e^{-n\Lambda_A^*(\alpha)}. \quad (40)$$

Similarly, under the assumption $\alpha \leq \Lambda_C'(0)$, one has

$$Pr(C^{(n)}(1)-C^{(n)}(s) \geq \alpha(1-s), 0\leq s\leq 1) \approx 1 \quad (41)$$

by identifying the path $\phi(s) = \Lambda_C'(0)s$ for $0 \leq s \leq 1$. Thus,

$$\begin{aligned} Pr(B(0,n) \geq \alpha n) &= Pr(B^{(n)}(1) \geq \alpha) \\ &\geq e^{-n\Lambda_A^*(\alpha)}. \end{aligned} \quad (42)$$

Case 2. $\alpha \leq \Lambda_A'(\theta^*)$ and $\alpha > \Lambda_C'(0)$:

Under the assumption $\alpha > \Lambda_C'(0)$, note that

$$\begin{aligned} &Pr(C^{(n)}(1)-C^{(n)}(s) \geq \alpha(1-s), 0\leq s\leq 1) \\ &\quad\approx e^{-n\Lambda_C^*(\alpha)}. \end{aligned} \quad (43)$$

The most likely path for $C^{(n)}(s)$ is $\phi(s) = \alpha s$ for $0 \leq s \leq 1$. The rest of the proof follows the same argument as in Case 1.

Case 3. $\alpha > \Lambda_A'(\theta^*)$ and $\alpha > \Lambda_C'(0)$:

Note from (33), (A5) and $\alpha > \Lambda_A'(\theta^*)$ that

$$\begin{aligned} &Pr(B^{(n)}(1) \geq \alpha) \\ &\geq Pr(q^{(n)}(0) \geq \alpha - \Lambda_A'(\theta^*), A^{(n)}(s) \geq \Lambda_A'(\theta^*)s, \\ &\quad C^{(n)}(1)-C^{(n)}(s) \geq \alpha(1-s), 0\leq s\leq 1) \\ &\approx Pr(q^{(n)}(0) \geq \alpha - \Lambda_A'(\theta^*)) \\ &\quad Pr(A^{(n)}(s) \geq \Lambda_A'(\theta^*)s, 0\leq s\leq 1) \\ &\quad Pr(C^{(n)}(1)-C^{(n)}(s) \geq \alpha(1-s), 0\leq s\leq 1). \end{aligned} \quad (44)$$

¿From (5),

$$\begin{aligned} &Pr(q^{(n)}(0) \geq \alpha - \Lambda'_A(\theta^*)) \\ &= Pr(q(0) \geq n(\alpha - \Lambda'_A(\theta^*))) \\ &\approx e^{-n\theta^*(\alpha - \Lambda'_A(\theta^*))}. \end{aligned} \tag{45}$$

Similar to the argument in Case 1,

$$\begin{aligned} &Pr(A^{(n)}(s) \geq \Lambda'_A(\theta^*)s, 0 \leq s \leq 1) \\ &\approx e^{-n\Lambda^*_A(\Lambda'_A(\theta^*))}, \end{aligned} \tag{46}$$

$$\begin{aligned} &Pr(C^{(n)}(1) - C^{(n)}(s) \geq \alpha(1-s), 0 \leq s \leq 1) \\ &\approx e^{-n\Lambda^*_C(\alpha)}. \end{aligned} \tag{47}$$

Using the identity $\Lambda_A(\theta^*) = \theta^*\Lambda'_A(\theta^*) - \Lambda^*_A(\Lambda'_A(\theta^*))$, one has

$$\begin{aligned} &Pr(B(0,n) \geq \alpha n) = Pr(B^{(n)}(1) \geq \alpha) \\ &\geq e^{-n(\theta^*\alpha - \Lambda_A(\theta^*) + \Lambda^*_C(\alpha))}. \end{aligned} \tag{48}$$

Case 4. $\alpha > \Lambda'_A(\theta^*)$ and $\alpha \leq \Lambda'_C(0)$:

The proof is similar to that in Case 3 except that

$$Pr(C^{(n)}(1) - C^{(n)}(s) \geq \alpha(1-s), 0 \leq s \leq 1) \approx 1. \tag{49}$$

In view of the upper and lower bounds, one has (11). Our derivation from (11) to (12) basically follows the argument for Varadhan's integral lemma. Note that for $\theta \geq 0$ and $t \geq 0$,

$$\begin{aligned} Ee^{\theta B(0,t)} &= \int_0^\infty Pr(e^{\theta B(0,t)} \geq x)dx \\ &= \int_{-\infty}^{\Lambda'_A(0)} Pr(e^{\theta B(0,t)} \geq e^{\theta\alpha t})\theta t e^{\theta\alpha t}d\alpha \\ &+ \int_{\Lambda'_A(0)}^{\infty} Pr(e^{\theta B(0,t)} \geq e^{\theta\alpha t})\theta t e^{\theta\alpha t}d\alpha. \end{aligned} \tag{50}$$

The first integral in (50) is bounded below by 0 and above by

$$\int_{-\infty}^{\Lambda'_A(0)} \theta t e^{\theta\alpha t}d\alpha = e^{\Lambda'_A(0)\theta t}. \tag{51}$$

Since we assume both $a(t)$ and $c(t)$ are bounded, $b(t)$ is bounded. Thus, there is a constant δ such that

$$\begin{aligned} &\int_{\Lambda'_A(0)}^{\infty} Pr(e^{\theta B(0,t)} \geq e^{\theta\alpha t})\theta t e^{\theta\alpha t}d\alpha \\ &= \int_{\Lambda'_A(0)}^{\delta} Pr(e^{\theta B(0,t)} \geq e^{\theta\alpha t})\theta t e^{\theta\alpha t}d\alpha. \end{aligned} \tag{52}$$

Recalling (11), an elementary argument allows to conclude that

$$\begin{aligned} &\int_{\Lambda'_A(0)}^{\infty} Pr(e^{\theta B(0,t)} \geq e^{\theta\alpha t})\theta t e^{\theta\alpha t}d\alpha \\ &\approx e^{t\sup_{\alpha \geq \Lambda'_A(0)}[\theta\alpha - \Lambda^*_B(\alpha)]}. \end{aligned} \tag{53}$$

Note from Case 1 in Remark 2.3 that $\Lambda^*_B(\Lambda'_A(0)) = \Lambda^*_A(\Lambda'_A(0)) = 0$. Thus, the first integral in (50) is dominated by the second integral. In view of (53) and (50), one has (12).

Remark 2.8 **(i)** The effective bandwidths for the *transient* departure process (e.g., the departure process from an empty queue at time 0) and the *stationary* departure process are in general different. As in the proof for both the upper bound and the lower bound, the decay rate of the tail distribution of $q(0)$ plays an important role in the large deviation behavior of $B(0,t)$. However, when the queue has a constant capacity, as discussed in Example 2.4, the decay rate of the tail distribution of $q(0)$ does not contribute to the large deviation behavior of the departure process, and the departure process from an empty queue at time 0 and the stationary departure process have the same effective bandwidth.

(ii) As in Cases 3 and 4 of the proof for the lower bound, sometimes it is necessary to build up the queue in order to have a large excursion of the stationary departure process. For results concerning most likely paths in the buildup of large queue lengths see Anantharam [1], Dembo and Zajic [10] and Chang [6].

3 Applications

3.1 Intree networks with time varying capacities

One of the immediate applications of Theorem 2.2 is to extend the intree network considered in [7, 5, 6] to the case with time varying capacities. Consider a discrete-time intree network with J queues and multiple classes of customers. In such a network, there is a natural partial ordering among the J queues. If part of the departure process of queue i is an input process of queue j, we say queue j is a *successor* of queue i and queue i is a *predecessor* of queue j. Let $Suc(i)$ (resp. $Pre(j)$) be the set of queues that are successors (resp. predecessors) of queue j. We label the J queues in the *postorder* (left subtree, right subtree, root). Since the network is intree, $Suc(i)$ only contains one element. Also, we have from the postorder labelling that $i < j$ for all $i \in Pre(j)$.

We assume that routing depends only on the queue that customers depart from. To model this, we let $\{p_j(n), n \geq 1\}, j = 1\ldots, J-1$, be $J-1$ sequences of routing random variables, i.e., $p_j(n) = 1$ if the n^{th} departure from queue j is routed to the successor of queue j, i.e., queue $Suc(j)$, and $p_j(n) = 0$ if the n^{th}

departure from queue j leaves the intree network. Besides the arrivals from the queues in the predecessors, there are k_j external arrival processes coming to queue j with $\{a_{j,k}(t), t = 0, 1, 2, \ldots\}$, $k = 1, \ldots, k_j$, denoting the number of arrivals at time t.

Denote $q_j(t)$ and $c_j(t)$ the queue length of queue j at time t and its capacity, respectively. We assume that each queue is governed by Lindley's equation, i.e.,

$$q_j(t+1) = (q_j(t) + a_j(t+1) - c_j(t+1))^+, \quad j = 1, \ldots, J, \tag{54}$$

where $a_j(t)$ is the total number of arrivals at queue j at time t.

Now we modify the algorithm in Chang [6] to yield the following heuristic for estimating the stationary queue distributions $q_j(\infty), j = 1, \ldots, J$:

Step 1. Obtain the normalized logarithmic moment generating function for each arrival process, routing process and capacity process (provided that their limits exist).

$$\Lambda_{A,j,k}(\theta) = \lim_{t\to\infty} \frac{1}{t} \log E(e^{\theta A_{j,k}(0,t)}) \tag{55}$$

$$\Lambda_{P,j}(\theta) = \lim_{t\to\infty} \frac{1}{t} \log E(e^{\theta P_j(0,t)}) \tag{56}$$

$$\Lambda_{C,j}(\theta) = \lim_{t\to\infty} \frac{1}{t} \log E(e^{\theta C_j(0,t)}) \tag{57}$$

where $A_{j,k}(t_1, t_2) = \sum_{s=t_1+1}^{t_2} a_{j,k}(s)$, $P_j(t_1, t_2) = \sum_{s=t_1+1}^{t_2} p_j(s)$ and $C_j(t_1, t_2) = \sum_{s=t_1+1}^{t_2} c_j(s)$.

Step 2. For $j = 1, \ldots, J$, define the following functions recursively:

$$\Lambda_j(\theta) = \sum_k \Lambda_{A,j,k}(\theta) + \sum_{k\in Pre(j)} \tilde{\Omega}_k(\theta) \tag{58}$$

$$\Omega_j(\theta) = \sup_{\alpha \geq \Lambda'_j(0)} [\theta\alpha - \Omega_j^*(\alpha)] \tag{59}$$

$$\tilde{\Omega}_j(\theta) = \Omega_j(\Lambda_{P,j}(\theta)) \tag{60}$$

where

$$\Omega_j^*(\alpha) = \theta_j^*\alpha - \sup_{\gamma\leq\alpha}[\theta_j^*\gamma - \Lambda_j^*(\gamma)] + \inf_{\alpha\leq\gamma} \Lambda_{C,j}^*(\gamma) \tag{61}$$

$$\Lambda_j^*(\alpha) = \sup_{\theta\geq 0}[\theta\alpha - \Lambda_j(\theta)] \tag{62}$$

$$\Lambda_{C,j}^*(\alpha) = \sup_{\theta\geq 0}[\theta\alpha - \Lambda_{C,j}(\theta)] \tag{63}$$

and θ_j^* is the solution of $\Lambda_j(\theta) + \Lambda_{C,j}(-\theta) = 0$.

Step 3 . Approximate the steady state distribution $q_j(\infty)$ of queue j as follows:

$$Pr(q_j(\infty) \geq x) \approx e^{-\theta_j^* x}. \tag{64}$$

As discussed in [4, 7, 5, 6], the intuition behind this algorithm is the concept of input-output relations of queueing networks. An intree network consists of the following network operations: (i) multiplexing, (ii) input-output relation through a queue, and (iii) routing. These three operations correspond to the three rules in (58)-(60).

We note that a formal statement like

$$\lim_{x\to\infty} \frac{\log Pr(q_j(\infty) \geq x)}{x} = -\theta_j^* \tag{65}$$

is possible. However, our proof for the upper bound and the lower bound cannot be applied inductively as in [6], where sample path large deviations are preserved. A formal proof for (65), at least for the lower bound, may require identifying the most likely path for each arrival process, service process and routing process in the intree network. This seems quite complicated, but it can be done as in the backward phase in [7] for determining appropriate conjugate processes.

3.2 Priority tandem queues with cross traffic

One of the simplest ways to implement an ATM network with multiple grades of services (GOS) is to assign services with priorities. To model a session in such a network, we consider a tandem queue with cross traffic. Assume the session passes through the network from queue 1 to queue J, where the capacity of queue j is c_j, $j = 1, \ldots, J$. Denote $a_{1,j}(t)$, $j = 1, \ldots, J$, as the number of *stationary* arrivals of this session at queue j at time t. Thus, $a_{1,1}(t)$ is the number of external arrivals at time t and $a_{1,j}(t)$, $j = 2, \ldots, J$, is the number of *stationary* departures from queue $j-1$ at time t. The cross traffic at queue j, $j = 1, \ldots, J$, is modelled by a sequence of nonnegative random variables $\{a_{2,j}(t), t \geq 0\}$, $j = 1, \ldots, J$. Assume the cross traffic at each queue has higher priority than the session $\{a_{1,j}(t), t \geq 0\}$. Also, the cross traffic at queue j, $j = 1, \ldots, J$, leaves the tandem queue immediately after departing from queue j. Thus, the bandwidth for the session at each queue is the remaining bandwidth after accommodating the local cross traffic. Since the cross traffic is time varying, the bandwidth for the session at each queue is also time varying.

To analyze such a network, we transform the problem into a network with time varying capacities so that the algorithm developed in the previous section can be applied. Let $b_{2,j}(t)$ be the number of departures from queue j due to the cross traffic at time t, and $c_j(t)$ be the bandwidth available for this session at queue j at time t. As argued in the previous paragraph, we have $c_j(t) = c_j - b_{2,j}(t)$. In order to apply the algorithm in the previous section, we need to compute

$$\Lambda_{C,j}(\theta) = \theta c_j + \lim_{t\to\infty} \frac{1}{t} \log E(e^{-\theta \sum_{s=1}^{t} b_{2,j}(s)}) \ \ \forall\, \theta \in I\!R. \tag{66}$$

Note that $\{b_{2,j}(t), t \geq 0\}$, $j = 1, \ldots, J$, are *stationary* departure processes from queues with *constant* capacities. As shown in Chang [6], Theorem 4.7, for $\theta \geq 0$,

$$\lim_{t\to\infty} \frac{1}{t} \log E(e^{\theta \sum_{s=1}^{t} b_{2,j}(s)}) = \begin{cases} \Lambda_{2,j}(\theta) & \text{if } \theta \leq \tilde{\theta}_j \\ \theta c_j - \tilde{\theta}_j c_j + \Lambda_{2,j}(\tilde{\theta}_j) & \text{otherwise} \end{cases} \tag{67}$$

where

$$\Lambda_{2,j}(\theta) = \lim_{t\to\infty} \frac{1}{t} \log E(e^{\theta \sum_{s=1}^{t} a_{2,j}(s)}), \tag{68}$$

and $\tilde{\theta}_j$ is the solution of $\Lambda'_{2,j}(\theta) = c_j$. We claim that (67) also holds for $\theta < 0$. The proof will be given at the end of this section.

Using (67) and (66), we can apply the algorithm in the previous section as follows:

Step 1. Obtain the normalized logarithmic moment generating function for the external arrival process and each cross traffic (provided that their limits exist).

$$\Lambda_{1,1}(\theta) = \lim_{t\to\infty} \frac{1}{t} \log E(e^{\theta \sum_{s=1}^{t} a_{1,1}(s)}), \tag{69}$$

$$\Lambda_{2,j}(\theta) = \lim_{t\to\infty} \frac{1}{t} \log E(e^{\theta \sum_{s=1}^{t} a_{2,j}(s)}), \tag{70}$$

$$j = 1, \ldots, J.$$

Step 2. For $j = 1, \ldots, J$, define the following functions recursively:

$$\Lambda_{C,j}(-\theta) = \begin{cases} -\theta c_j + \Lambda_{2,j}(\theta) & \text{if } \theta \leq \tilde{\theta}_j \\ -\tilde{\theta}_j c_j + \Lambda_{2,j}(\tilde{\theta}_j) & \text{otherwise} \end{cases} \tag{71}$$

$$\Lambda_{1,j+1}(\theta) = \sup_{\alpha \geq \Lambda'_{1,j}(0)} [\theta\alpha - \Lambda^*_{1,j+1}(\alpha)] \tag{72}$$

where

$$\Lambda^*_{1,j+1}(\alpha) = \theta^*_j \alpha - \sup_{\gamma \leq \alpha} [\theta^*_j \gamma - \Lambda^*_{1,j}(\gamma)] + \inf_{\alpha \leq \gamma} \Lambda^*_{C,j}(\gamma) \tag{73}$$

$$\Lambda^*_{C,j}(\alpha) = \sup_{\theta \geq 0} [\theta\alpha - \Lambda_{C,j}(\theta)] \tag{74}$$

$$\Lambda^*_{1,1}(\alpha) = \sup_{\theta \geq 0} [\theta\alpha - \Lambda_{1,1}(\theta)] \tag{75}$$

and θ^*_j is the solution of $\Lambda_{1,j}(\theta) + \Lambda_{C,j}(-\theta) = 0$.

Step 3. Approximate the steady state distribution $q_j(\infty)$ of queue j as follows:

$$Pr(q_j(\infty) \geq x) \approx e^{-\theta^*_j x}. \tag{76}$$

Now we show that (67) also holds for $\theta < 0$, i.e,

$$\lim_{t\to\infty} \frac{1}{t} \log E(e^{\theta \sum_{s=1}^{t} b_{2,j}(s)}) = \Lambda_{2,j}(\theta), \ \ \theta < 0. \tag{77}$$

Let $\{b^0_{2,j}(t), t \geq 0\}$ be the number of departures from queue j due to the cross traffic at time t if the tandem queue is empty at time 0. As shown in Chang [6], Theorem 4.7, for all $\theta \in I\!R$,

$$\lim_{t\to\infty} \frac{1}{t} \log E(e^{\theta \sum_{s=1}^{t} b^0_{2,j}(s)}) = \begin{cases} \Lambda_{2,j}(\theta) & \text{if } \theta \leq \tilde{\theta}_j \\ \theta c_j - \tilde{\theta}_j c_j + \Lambda_{2,j}(\tilde{\theta}_j) & \text{otherwise.} \end{cases} \tag{78}$$

¿From a standard sample path argument (see e.g. [25, 26]), one has the stochastic ordering result,

$$\sum_{s=1}^{t} b^0_{2,j}(s) \leq_{st} \sum_{s=1}^{t} b_{2,j}(s), \ \ \forall\, t. \tag{79}$$

Thus, for $\theta < 0$, one has from (78) that

$$\lim_{t\to\infty} \frac{1}{t} \log E(e^{\theta \sum_{s=1}^{t} b_{2,j}(s)}) \leq \Lambda_{2,j}(\theta). \tag{80}$$

Also, note that $\sum_{s=1}^{t} b_{2,j}(s) \leq \sum_{s=1}^{t} a_{2,j}(s) + q_j(0)$. Thus, for $\theta < 0$,

$$Ee^{\theta \sum_{s=1}^{t} b_{2,j}(s)} \geq Ee^{\theta \sum_{s=1}^{t} a_{2,j}(s)} e^{\theta q(0)}. \tag{81}$$

In view of (A2$'$) and (A5),

$$\log Ee^{\theta \sum_{s=1}^{t} b_{2,j}(s)} \geq \Lambda_{2,j}(\theta)t - \Gamma_{2,j}(\theta) + \log Ee^{\theta q(0)}, \tag{82}$$

for some $\Gamma_{2,j}(\theta) < \infty$. Dividing both sides by t and letting $t \to \infty$ yields

$$\lim_{t\to\infty} \frac{1}{t} \log E(e^{\theta \sum_{s=1}^{t} b_{2,j}(s)}) \geq \Lambda_{2,j}(\theta), \ \ \theta < 0. \tag{83}$$

In conjunction with (80), one has (77).

4 Conclusions

In this paper, we identified the effective bandwidth of the departure process from a queue with a time varying capacity. This new result was applied to in-tree networks with time varying capacities and priority tandem queues with cross traffic. Algorithms for approximating the tail distributions of the queue lengths in such networks were derived.

There are two possible extensions of our approach. First, one could extend our result from the discrete-time model to the continuous time model. This could be done by the inversion formula in Glynn and Whitt [16] and a stochastic dominance result in Chang, Thomas and Kiang [8]. The other extension is for fast simulations of loss probabilities in intree networks with time varying capacities. As in Chang, Heidelberger, Juneja and Shahabuddin [7], this requires an additional backward phase to determine the appropriate conjugate processes for arrival processes, service processes and routing processes. Extensions along these lines are currently under investigation.

References

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

[2] F. Baccelli and P. Bremaud, *Palm Probabilities and Stationary Queues.* New York: Springer-Verlag, 1987.

[3] C.S. Chang, "Stability, queue length and delay, Part II: Stochastic queueing networks," *IEEE Conference on Decision and Control*, Vol 1, pp. 1005-1010, 1992.

[4] C.S. Chang, "Stability, queue length and delay of deterministic and stochastic queueing networks," *IEEE Transactions on Automatic Control*, Vol.39, pp. 913-931, 1994.

[5] C.S. Chang, "Approximations of ATM networks: effective bandwidths and traffic descriptors," *IBM RC 18954*, 1993.

[6] C.S. Chang, "Sample path large deviations and intree networks," to appear in *Queueing Systems.*

[7] C.S. Chang, P. Heidelberger, S. Juneja and P. Shahabuddin, "Effective Bandwidth and Fast Simulation of ATM Intree Networks," *Performance Evaluation*, Vol. 20, pp. 45-66, 1994.

[8] C.S Chang, J.A. Thomas, and S.-H. Kiang, "On the stability of open networks: an unified approach by stochastic dominance," *Queueing Systems*, Vol. 15, pp. 239-260, 1994.

[9] A. Dembo and O. Zeitouni. *Large Deviations Techniques and Applications.* Boston: Jones and Barlett Publishers, 1992.

[10] A. Dembo and T. Zajic, "Large deviations from empirical mean and measure to partial sums processes," *preprint*, 1993.

[11] N.G. Duffield, "Exponential bounds for queues with Markovian arrivals," *preprint*, 1993.

[12] R. Ellis, "Large deviations for a general class of random vectors, " *Ann. Probab.*, Vol. 12, pp. 1-12, 1984.

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

[14] J. Gärtner, "On large deviations from invariant measure," *Theory Probab. Appl.*, Vol. 22, pp. 24-39, 1977.

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

[16] P.W. Glynn and W. Whitt, "Logarithmic asymptotics for steady-state tail probabilities in a single-server queue," *J. Appl. Prob*, Vol. 31A, pp. 131-156, 1994.

[17] P.W. Glynn and W. Whitt, "Large deviations behavior of counting processes and their inverses," *Queueing Systems*, Vol. 17, pp. 107-128, 1994.

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

[19] F.P. Kelly, "Effective bandwidths at multi-class queues," *Queueing Systems*, Vol. 9, pp. 5-16, 1991.

[20] G. Kesidis and J. Walrand, "Traffic policing and enforcement of effective bandwidth constraints in ATM networks," *preprint.*

[21] G. Kesidis, J. Walrand and C.S. Chang, "Effective bandwidths for multiclass Markov fluids and other ATM sources," *IEEE/ACM Trans. Networking*, Vol. 1, pp. 424-428, 1993.

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

[23] R. Nagarajan, J. Kurose and D. Towsley, "Local allocation of end-to-end quality-of-service in high speed networks," *preprint.*

[24] N. O'Connell, "Large deviations in queueing networks," *DIAS Technical Report No. DIAS-APG-9413*, 1994.

[25] S.M. Ross, *Stochastic Processes.* New York: J. Wiley & Sons, 1983.

[26] D. Stoyan, *Comparison Methods for Queues and Other Stochastic Models*, Berlin: J. Wiley & Sons, 1983.

[27] W. Whitt, "Tail probability with statistical multiplexing and effective bandwidths in multi-class queues," *Telecommunication Systems*, Vol. 2, pp. 71-107, 1993.

[28] G. De. Veciana and J. Walrand, "Effective Bandwidths: call admission, traffic policing & filtering for ATM networks," Submitted to *IEEE/ACM Trans. Networking*, 1993.

[29] G. De. Veciana, C. Courcoubetis and J. Walrand, "Decoupling bandwidths for networks: a decomposition approach to resource management," Submitted to *IEEE/ACM Trans. Networking*, 1993.