On the Performance of Multiplexing Independent Regulated Inputs

Cheng-Shang Chang
Institute of Communications Engineering
National Tsing Hua University
Hsinchu, Taiwan 300, R.O.C.
cschang@ee.nthu.edu.tw

Yuh-ming Chiu
Institute of Communications Engineering
National Tsing Hua University
Hsinchu, Taiwan 300, R.O.C.
ymchiu@gibbs.ee.nthu.edu.tw

Wheyming Tina Song
Department of Industrial Engineering and Engineering Management
National Tsing Hua University
Hsinchu, Taiwan 300, R.O.C.
wheyming@ie.nthu.edu.tw

ABSTRACT

In this paper, we consider the performance analysis problem for a work conserving link with a large number of independent regulated inputs. For such a problem, we derive simple stochastic bounds under a general traffic constraint for the inputs. The bound for queue length is shown to be a stochastic extension of the deterministic worst case bound and it is asymptotically tighter than the bound in Kesidis and Konstantopoulos [23]. We also test the bound by considering periodic inputs with independent starting phases. Based on Sanov's theorem and importance sampling, we propose a fast simulation algorithm that achieves significant variance reduction. The simulations results are compared with our stochastic bound and the bound in [23].

Keywords: Multiplexing, performance bounds, fast simulation

1. INTRODUCTION

As we all know, the current Internet (IPv4) does not provide quality-of-service (QoS) for users. As it is, the current Internet is flooded with "elastic" traffic and congestion occurs randomly in the Internet. Due to the lack of QoS guarantees, real-time traffic, such as voice and video, may not go through congested nodes in time, and the current Internet is not capable of providing satisfactory services for real-time traffic. In view of this, it is a common belief that QoS should be provided in the next generation Internet. A special task force, called the Internet Engineering Task Force (IETF), is assigned the job to standardize the procedure for service guarantees in the next generation Internet.

There are two main IETF proposals for providing QoS in the next generation Internet: the RSVP/Intserv architecture [17, 18, 19, 20, 21, 22] and the Differentiated Services (diff-serv) architecture [13, 14, 15, 16]. In the RSVP/Intserv architecture, each input is regulated before entering the network and allocated a service curve along the path in the network. As a result, QoS is guaranteed for each input. Such an architecture may not be scalable for large high speed networks. To deal with such a scalability problem, the diff-serv architecture groups regulated inputs into several classes and provides QoS guarantees for each class.

Motivated by the diff-serv architecture, we consider the performance analysis problem for a work conserving link with a large number of independent regulated inputs. Such a problem was previously addressed in the literature (see e.g. [1, 10, 23, 24, 27, 29]). Elwalid, Mitra and Wentworth [10] considered such a problem with inputs regulated by a leaky bucket and a peak rate regulator. For stochastic analysis, they further assumed periodic on-off inputs with independent starting phases. Based on the Chernoff bound, they derived admissible set that guarantees a very small loss probability for an equivalent *bufferless* link. The same problem was also considered by LoPresti, Zhang, Kurose and Towsley [27]. They separated the problem into two subproblems: a bufferless link and a buffer with no server. They derived performance bounds based on the bounds for each subproblem. Rajagopal, Reisslein and Ross [29] studied the same problem in [27] with a more general traffic constraint. Kesidis and Konstantopoulos [23, 24] identified the worst-case input traffic when there is *only one* input regulated by a leaky bucket and a peak rate regulator. Based on the worst-case input traffic, they also derived stochastic bounds for multiplexing a large number of independent inputs. Using the central limit theorem and the Chernoff bound, Boorstyn, Burchard, Liebeherr and Oottamakorn [1] came up with new concepts of effective envelopes. The effective envelopes in [1] can then be used for admission control to achieve stochastic guarantees as the existing method for deterministic guarantees.

Our contribution of this paper consists of two parts: in the first part, we derive simple stochastic bounds for *buffered* links under the general traffic constraint in [29]. The bound for queue length is shown to be a stochastic extension of the deterministic worst case bound in [8] and it is asymptotically tighter than the bound in Kesidis and Konstantopoulos [23]. In the second part, we test the bound by considering periodic inputs with independent phases as in [10]. However, the periodic inputs need not be on-off sources. Based on Sanov's theorem and importance sampling, we propose a fast simulation algorithm that achieves significant variance reduction. The simulations results are compared with our stochastic bound and the bound in [23].

2. STOCHASTIC BOUNDS FOR MULTIPLEXING INDEPENDENT REGULATED INPUTS

We consider a discrete-time work conserving link with n inputs. Time is indexed by $t = 0, 1, 2, \ldots$. The capacity of the link is $n \cdot c$. Let $a_i(t)$ be the number of arrivals at time t for the i^{th} input, $i = 1, \ldots, n$, and $q(t)$ be the queue length of the link at time t. Then the link is governed by the Lindley equation:

$$q(t+1) = (q(t) + \sum_{i=1}^{n} a_i(t+1) - nc)^+, \tag{1}$$

where $x^+ = \max[0, x]$. Assuming $q(0) = 0$, expanding (1) recursively yields

$$q(t) = \max_{0 \le s \le t} [\sum_{i=1}^{n} A_i(t) - A_i(s) - nc(t-s)], \tag{2}$$

where $A_i(t) = \sum_{s=1}^{t} a_i(s)$ is the cumulative number of arrivals by time t for the i^{th} input. As in [1], we make the following assumptions:

(A1) Independence: all the n inputs are independent.

(A2) Stationarity: all the n inputs are stationary, i.e., $\{a_i(t), t = 1, 2, \ldots\}$ forms a stationary sequence.

(A3) Traffic characterization: all the n inputs are f-upper constrained [8] for some deterministic, increasing, and subadditive function f with $f(0) = 0$, i.e.,

$$A_i(t) - A_i(s) \le f(t-s), \quad \forall s \le t. \tag{3}$$

(A4) Rate condition: let $\rho = \lim_{t\to\infty} f(t)/t$ be the envelope rate in [4]. Assume $\rho < c$.

Under (A3) and (A4), one can perform deterministic worst case analysis as follows: using the multiplexing rule in [8], the multiplexed traffic is nf-upper constrained. From the representation in (2), one has the following well-known deterministic worst case bound for the queue length (see e.g., [5] and references therein)

$$\sup_{t \ge 0} q(t) \le n q_{max}, \tag{4}$$

where

$$q_{max} = \sup_{t \ge 0} [f(t) - ct]^+. \tag{5}$$

With the additional assumptions in (A1) and (A2), we show in the next section that the deterministic worst case can be improved by a stochastic bound.

2.1 Queue Length

As discussed in [1, 7, 10, 27, 29], there are two important stochastic properties that can be deduced from (A1-4).

(P1) The means of all the inputs are not greater than ρ. To see this, note from (A3) that $A_i(t) \le f(t)$ for all t. From the stationarity assumption in (A2), it follows that

$$\mathsf{E}[a_i(1)] = \frac{\mathsf{E}[A_i(t)]}{t} \le \frac{f(t)}{t}.$$

Taking limits on both sides yields $\mathsf{E}[a_i(1)] \le \rho$.

(P2) For all $\theta \ge 0$

$$\mathsf{E}e^{\theta(A_i(t) - A_i(t-s))} \le 1 - \frac{s\rho}{f(s)} + \frac{s\rho}{f(s)} e^{\theta f(s)}. \tag{6}$$

To see this, we use (P1) and stationarity to show that $\mathsf{E}[A_i(t) - A_i(t-s)] \le \rho s$ for all s. In conjunction with (A3), we know that $A_i(t) - A_i(t-s)$ is a random variable with mean not greater than ρs and a bounded support $[0, f(s)]$. For such a random variable, it is known from the classical result on convex ordering (also known as variability ordering, see e.g., [30, 35]) that its moment generating function is bounded above by that of the Bernoulli random variable Y with $\mathsf{P}(Y = 0) = 1 - \frac{s\rho}{f(s)}$ and $\mathsf{P}(Y = f(s)) = \frac{s\rho}{f(s)}$.

Now consider the function

$$\Delta(p_1||p_2) = p_1 \log \frac{p_1}{p_2} + (1 - p_1) \log \frac{1-p_1}{1-p_2}$$

for $0 \le p_1, p_2 \le 1$ and $\Delta(p_1||p_2) = \infty$ otherwise. The function $\Delta(p_1||p_2)$ is the two-dimensional special case of the Kullback-Leibler distance (or known as the relative entropy) that measures the distance between two probability vectors. It is easy to verify that $\Delta(p_1||p_2)$ is increasing in p_1 and decreasing in p_2 for $p_1 \ge p_2$. Thus,

$$\Delta(p_1||p_2) \ge \Delta(\tilde{p}_1||\tilde{p}_2)$$

for all $p_1 \ge \tilde{p}_1 \ge \tilde{p}_2 \ge p_2$.

Also, let

$$\tau = \inf\{t : f(t)/t \le c\}. \tag{7}$$

Note from (A4) that $\tau < \infty$.

Theorem 1 *Under the assumptions in (A1-4), the queue length process $\{q(t), t \ge 1\}$ (with $q(0) = 0$) reaches its steady state at time τ, i.e., $q(t)$ has the same distribution as $q(\tau)$ for $t \ge \tau$. Moreover, for all $x > 0$*

$$\mathsf{P}(q(\tau) \ge nx) \le \sum_{s=1}^{\tau} e^{-n\Delta(p_1(s,x)||p_2(s))}, \tag{8}$$

where

$$p_1(s,x) = \frac{cs + x}{f(s)}, \tag{9}$$

and

$$p_2(s) = \frac{s\rho}{f(s)}. \tag{10}$$

We argue that Theorem 1 is a stochastic extension of the deterministic worst case bound in (4). For $x > q_{max}$, we have $p_1(s,x) > 1$ for all s and thus $\Delta(p_1(s,x)||p_2(s)) = \infty$. As a result of Theorem 1, $\mathsf{P}(q(\tau) \geq nx) = 0$ for all $x > q_{max}$.

From the subadditive theory, we know that

$$\rho = \lim_{t\to\infty} \frac{f(t)}{t} = \inf_{t\geq 0} \frac{f(t)}{t}.$$

Thus, we always have $0 \leq p_2(s) \leq 1$ for all s. As we assume that $\rho < c$ in (A4), $p_2(s) < p_1(s,x)$ for all $x > 0$. Basically, the Kullback-Leibler distance $\Delta(p_1(s,x)||p_2(s))$ measures how unlikely it is for a Bernoulli random variable with mean $p_2(s)$ to become another Bernoulli random variable with mean $p_1(s,x)$. When n is large, the upper bound in (8) is dominated by the term with the smallest Kullback-Leibler distance, i.e.,

$$\sum_{s=1}^{\tau} e^{-n\Delta(p_1(s,x)||p_2(s))} \approx e^{-n\Delta(p_1(s^*,x)||p_2(s^*))},$$

where s^* is the index of the time that yields the smallest Kullback-Leibler distance. Intuitively, one may view s^* as the *critical time scale* that leads to buffer overflows when n is large (see e.g., [2, 7] for related works on large deviations). More refined approximations along this line can be found in [25].

For the special case that $f(t) = \min[\sigma + \rho t, \pi t]$, Kesidis and Konstantopoulos [23] showed that

$$\mathsf{P}(q(\tau) \geq nx) \leq e^{-n\Delta(\tilde{p}_1(x)||\tilde{p}_2)}, \quad \frac{\rho}{c} q_{max} \leq x \leq q_{max},$$

where $\tilde{p}_1(x) = x/q_{max}$ and $\tilde{p}_2 = \rho/c$. In this case, $q_{max} = (\pi - c)\sigma/(\pi - \rho)$ and $\tau = \sigma/(c-\rho)$. For $1 \leq s \leq \tau$, we have $cs < f(s) \leq cs + q_{max}$ as f is concave in this case. Thus, $p_2(s) \leq \tilde{p}_2$ and $p_1(s,x) \geq \tilde{p}_1(x)$ for all $1 \leq s \leq \tau$. Since $\tilde{p}_1(x) \geq \tilde{p}_2$ for $\frac{\rho}{c} q_{max} \leq x \leq q_{max}$, we have that for all $1 \leq s \leq \tau$ and $\frac{\rho}{c} q_{max} \leq x \leq q_{max}$

$$\Delta(p_1(s,x)||p_2(s)) \geq \Delta(\tilde{p}_1(x)||\tilde{p}_2).$$

This shows that the bound in Theorem 1 is asymptotically tighter than that in [23] when $n \to \infty$. We note that the bound in [23] can be further improved by solving a more complicated optimization problem in [24].

The proof of Theorem 1 is based on the following lemmas.

Lemma 1 *Under (A2),(A3) and (A4), the queue length process $\{q(t), t \geq 1\}$ (with $q(0) = 0$) reaches its steady state at time τ.*

Proof: Our proof is similar to the proofs for bounding busy periods in [8, 26, 4, 11]. For $t > \tau$, it follows from (2) that

$$q(t) = \max\Big[\max_{0\leq s\leq t-\tau-1} [\sum_{i=1}^{n} A_i(t) - A_i(s) - nc(t-s)],$$

$$\max_{t-\tau\leq s\leq t} [\sum_{i=1}^{n} A_i(t) - A_i(s) - nc(t-s)]\Big]. \quad (11)$$

For $0 \leq s \leq t - \tau - 1$, we let $s' = \lfloor \frac{t-s}{\tau} \rfloor \tau + s$. Note that $t - \tau \leq s' \leq t$ and that $s' - s$ is an integer multiple of τ. As all the inputs are f-upper constrained in (A3),

$$A_i(s') - A_i(s) \leq f(s' - s).$$

From the subadditivity of f and the definition of τ, it follows that

$$f(s'-s) \leq \frac{s'-s}{\tau} f(\tau) \leq c(s'-s).$$

Thus,

$$A_i(s') - A_i(s) \leq c(s'-s), \quad i = 1, \ldots, n.$$

For $0 \leq s \leq t - \tau - 1$,

$$\begin{aligned}
&\sum_{i=1}^{n} A_i(t) - A_i(s) - nc(t-s) \\
&= \sum_{i=1}^{n} A_i(t) - A_i(s') + A_i(s') - A_i(s) \\
&\qquad -nc(t-s') - nc(s'-s) \\
&\leq \sum_{i=1}^{n} A_i(t) - A_i(s') - nc(t-s') \\
&\leq \max_{t-\tau\leq s\leq t} [\sum_{i=1}^{n} A_i(t) - A_i(s) - nc(t-s)].
\end{aligned}$$

This implies the first term in (11) is not greater than the second term. Thus, we have

$$q(t) = \max_{t-\tau\leq s\leq t} [\sum_{i=1}^{n} A_i(t) - A_i(s) - nc(t-s)].$$

From the stationarity of the n inputs in (A2) and (2), it then follows that $q(t)$ and $q(\tau)$ have the same distribution. ■

By Chernoff's bound, it is straightforward to show the following lemma from [P2]. The lemma was also previously reported in [1] to introduce the concept of effective envelope that provides admission control criteria for various scheduling algorithms.

Lemma 2 *Under (A1-4),*

$$\begin{aligned}
&\mathsf{P}(\sum_{i=1}^{n} A_i(\tau) - A_i(\tau - s) - ncs \geq nx) \\
&\leq e^{-n\Delta(p_1(s,x)||p_2(s))}.
\end{aligned}$$

Proof: (Theorem 1) In view of Lemma 1, it suffices to prove the upper bound for $q(\tau)$. For $x > 0$, applying the union bound yields

$$\begin{aligned}
&\mathsf{P}(q(\tau) \geq nx) \\
&\leq \sum_{s=0}^{\tau-1} \mathsf{P}(\sum_{i=1}^{n} A_i(\tau) - A_i(s) - nc(\tau - s) \geq nx).
\end{aligned}$$

Theorem 1 then follows directly from Lemma 2. ■

2.2 Extensions

There are two possible extensions of our approach:

(i) Non-identically regulated case: as discussed in [24], one may consider different constraint functions for the n inputs. Assume that A_i is f_i-upper constrained, $i = 1, \dots n$. Let $\rho_i = \lim_{t\to\infty} f_i(t)/t$. Then under the condition that $\sum_{i=1}^n \rho_i < nc$, $\tau = \inf\{t : \sum_{i=1}^n f_i(t) \le nct\}$ is finite and one can use Chernoff's bound (for Lemma 2) and the union bound (for Theorem 1) to show that

$$\mathsf{P}(q(\tau) \ge nx) \le \sum_{s=1}^{\tau} e^{-I_s},$$

where

$$I_s = \sup_{\theta \ge 0}[n(cs + x)\theta - \sum_{i=1}^n \log(1 - \frac{s\rho_i}{f_i(s)} + \frac{s\rho_i}{f_i(s)} e^{\theta f_i(s)})].$$

Using differentiation, one can easily show that the above supremum is achieved when θ satisfies

$$\sum_{i=1}^n \frac{s\rho_i \exp(\theta f_i(s))}{1 - \frac{s\rho_i}{f_i(s)} + \frac{s\rho_i}{f_i(s)} \exp(\theta f_i(s))} = n(cs + x).$$

(ii) Causal non-reentry networks with bounded delays: though we only consider a single node in this paper, our approach can be generalized to causal non-reentry networks with bounded delay as in [26]. In a non-reentry network, an input visits a node at most once. Let $A_i(t)$ be the cumulative number of external arrivals by time t for the i^{th} input, and $A_{i,j}(t)$ be the cumulative number of arrivals at node j by time t for the i^{th} input. Assume that A_i is f_i-upper constrained. As we assume the network guarantees a bounded delay for each input, there is a finite constant $d_{i,j}$ such that

$$A_{i,j}(t) \ge A_i((t - d_{i,j})^+). \tag{12}$$

Also, as the network is causal, departures cannot be earlier than arrivals. This implies

$$A_i(t) \ge A_{i,j}(t). \tag{13}$$

From (12) and (13), it follows that

$$A_{i,j}(t) - A_{i,j}(s) \le A_i(t) - A_i((s - d_{i,j})^+) \le f_i(t - s + d_{i,j}).$$

Thus, as long as all the external arrivals are independent, the inputs at a node in such a network can be bounded above by *independent* random variables. The problem for bounding the queue length distribution at a particular node is then reduced to the non-identically regulated case.

3. FAST SIMULATION FOR MULTIPLEXING INDEPENDENT PERIODIC INPUTS

To see whether the upper bound derived in the previous section is tight, in this section we run simulations for multiplexing independent *periodic* inputs that satisfy the constraints in (A1)-(A4). To be specific, we consider the work conserving link with capacity $n \cdot c$ as in the previous section. Let $\{a(t), t \ge 1\}$ be a deterministic periodic sequence with period τ, i.e., $a(t + \tau) = a(t)$ for all t. Generate the i^{th} input by assigning $a_i(t) = a(t + U_i)$, where U_i is uniformly distributed over $\{0, 1 \dots, \tau - 1\}$ and independent of everything else. Clearly, the n inputs generated this way are stationary and independent and thus satisfy the assumptions (A1) and (A2). One only needs to choose $a(t)$ appropriately so that (A3) and (A4) can be satisfied.

As $\{a_i(t), t \ge 1\}$, $i = 1, \dots, n$, are periodic, the multiplexed process $\{\sum_{i=1}^n a_i(t), t \ge 1\}$ to the work conserving link is not *ergodic*. As a result, the queue length process $\{q(t), t \ge 1\}$ is also not ergodic. This implies that the performance of such a work conserving link cannot be measured by taking a single long run of simulation. Instead, one has to run several *independent* runs to obtain an accurate estimate of the performance. Fortunately, the following lemma shows that $q(t)$ reaches its steady state at time τ, and each run only needs to be carried out up to time τ.

Lemma 3 *Let $\rho = \frac{1}{\tau}\sum_{t=1}^{\tau} a(t)$ be the average rate. If $\rho \le c$, then the queue length process $\{q(t), t \ge 1\}$ (with $q(0) = 0$) reaches its steady state at time τ, i.e., $q(t)$ has the same distribution as $q(\tau)$ for $t \ge \tau$.*

The proof is similar to that in Lemma 1 and thus omitted.

To estimate the buffer overflow probability $\mathsf{P}(q(\tau) \ge nx)$, a standard approach is to run N independent copies of $q(\tau)$, denoted by $q_j(\tau), j = 1, \dots, N$, and form the estimate

$$\frac{1}{N}\sum_{j=1}^{N} I(q_j(\tau) \ge nx),$$

where $I(q_j(\tau) \ge x)$'s are indicator functions. The problem of such an approach is that the buffer overflow probability could be very small when the number of multiplexed inputs n is very large. In that case, the event for buffer overflow is *rare* and the number of independent copies N required for an accurate estimate is huge. This is not efficient and we shall seek a fast simulation approach as described in the next section.

3.1 Importance Sampling

Our approach to speed up our simulation is based on *importance sampling* that accelerates the occurrence of the rare event of interest. In importance sampling, the system is simulated by using a new set of probability distributions (change of measures), and an unbiased and consistent estimate is obtained by multiplying the simulation output by a likelihood ratio.

Here we give a brief review of how importance sampling works. For further reading, we refer to the survey paper by Heidelberger [12] and reference therein. Let X be a discrete random variable that has a probability mass function $p(x)$. Consider estimating the probability, γ, that X is in some set A. The standard approach, as discussed in the previous section, is to draw N samples, $X_1, \cdots, X_n$ from $p(x)$ and form the estimate

$$\frac{1}{N}\sum_{i=1}^{N} I(X_i \in A).$$

An alternative approach to estimate γ is to draw N samples, $X'_1, \cdots, X'_n$, from a different probability mass function $p'(x)$

and form the estimate

$$\frac{1}{N}\sum_{i=0}^{N} I(X_i' \in A)L(X_i'),$$

where $L(x) = p(x)/p'(x)$ is the likelihood ratio. Such an estimate is consistent and unbiased as

$$\lim_{N\to\infty}\frac{1}{N}\sum_{i=0}^{N} I(X_i' \in A)L(X_i')$$
$$= \mathsf{E}[I(X_1' \in A)L(X_1')], \quad a.s.,$$

and

$$\mathsf{E}[I(X_1' \in A)L(X_1')] = \sum_{x\in A}\frac{p(x)}{p'(x)}p'(x)$$
$$= \sum_{x\in A} p(x) = \mathsf{E}[I(X \in A)] = \gamma.$$

The question is then how one finds the change of measure $p'(x)$ to speed up the occurrence of the event. As discussed in [12], the large deviation principle is quite effective in finding the appropriate change of measure and this is the approach we are going to use in the next section.

3.2 Fast Simulation Algorithm

In this section, we derive the change of measure in our simulations via the large deviation principle. The derivation consists of two steps: in the first step, we use the Crämer theorem to find the critical time scale that leads to buffer overflow. In the second step, we use the Sanov theorem to find the appropriate change of measure under the critical time scale. Note that the only randomness in our simulation is the starting phase of each input, which is uniformly distributed over $\{0, 1 \ldots, \tau - 1\}$. As we shall see later, the change of measure that accelerates the occurrence of buffer overflow is non-uniform and tends to have synchronized starting phases.

We first give a brief review of the Crämer theorem that will be used later. For further reading on the Crämer theorem (and the large deviation principle), we refer to Bucklew [3], Dembo and Zeitouni [9] and Shwartz and Weiss [33]. The Crämer theorem is the large deviation principle (LDP) for i.i.d. sequences of random variables. Let $\{X_i, i = 1, \ldots, n\}$ be a sequence of independent identically distributed (i.i.d) random variables with the same distribution as X. Let $m = \mathsf{E}[X]$ and $\phi(\theta) = \mathsf{E}[e^{\theta X}]$ be the mean and moment generating function of X respectively. Define the functions

$$\Lambda(\theta) = \log\phi(\theta), \tag{14}$$

and its Legendre transform

$$\Lambda^*(\gamma) = \sup_\theta\left[\theta\gamma - \Lambda(\theta)\right]. \tag{15}$$

Theorem 2 *(Crämer [6]) : Let $S_n = \frac{1}{n}\sum_{i=1}^{n} X_i$. Assume $\phi(\theta) < \infty$ for all $\theta \in R$.*

(i) *(Upper bound) For every closed set $F \subset R$,*

$$\limsup_{n\to\infty}\frac{1}{n}\log\mathsf{P}\left(S_n \in F\right) \le -\inf_{\gamma\in F}\Lambda^*(\gamma). \tag{16}$$

(ii) *(Lower bound) For every open set $G \subset R$,*

$$\liminf_{n\to\infty}\frac{1}{n}\log\mathsf{P}\left(S_n \in G\right) \ge -\inf_{\gamma\in G}\Lambda^*(\gamma). \tag{17}$$

The function $\Lambda^*(\gamma)$ is called the rate function.

The Crämer theorem gives both an upper bound and a lower bound on how unlikely it is for the partial sum of a sequence of i.i.d. random variables to deviate from its mean. Intuitively, for $\gamma > m$ (and γ is a continuous point of $\Lambda^*(\cdot)$), one has the following approximation

$$\mathsf{P}(S_n \ge \gamma) \approx e^{-n\Lambda^*(\gamma)}. \tag{18}$$

To apply the Crämer theorem in our setting, note from the union bound that for $x > 0$

$$\max_{1\le s\le\tau}\mathsf{P}\Big(\sum_{i=1}^{n} A_i(\tau) - A_i(s) \ge n\cdot(c(\tau - s) + x)\Big)$$
$$\le \mathsf{P}(q(\tau) \ge nx)$$
$$\le \sum_{s=1}^{\tau}\mathsf{P}\Big(\sum_{i=1}^{n} A_i(\tau) - A_i(s) \ge n\cdot(c(\tau - s) + x)\Big).$$

From (18), it follows that (see e.g., [2, 34, 7] for previous works on a more general setting)

$$\mathsf{P}(q(\tau) \ge nx) \approx \exp\Big(-n\min_{1\le s\le\tau}\Lambda_s^*(cs + x)\Big), \tag{19}$$

where $\Lambda_s^*(\gamma) = \sup_\theta[\theta\gamma - \Lambda_s(\theta)]$ is the Legendre transform of $\Lambda_s(\theta) = \log\mathsf{E}e^{\theta A_1(s)}$. As discussed in Section 2, one may view s^*, the index of time that yields the minimum in (19), as the critical time scale that leads to buffer overflow (the critical time scale may not be unique). Finding the critical time scale s^* can be used for accelerating the occurrence of buffer overflow needed in our simulation. To compute the critical time scale s^*, we need to find $\Lambda_s^*(cs + x)$ from the Legendre transform of $\Lambda_s(\theta)$. This can be done by solving $\theta(s)$ from

$$cs + x = \Lambda_s'(\theta(s)), \tag{20}$$

and then letting

$$\Lambda_s^*(cs + x) = \theta(s)(cs + x) - \Lambda_s(\theta(s)). \tag{21}$$

Since $a_1(t) = a(t + U_1)$ for some uniformly distributed random variable U_1, we have

$$\mathsf{E}e^{\theta A_1(s)} = \frac{1}{\tau}\sum_{k=0}^{\tau-1} e^{\theta A(k,k+s)}, \tag{22}$$

and

$$\Lambda_s(\theta) = \log\Big(\frac{1}{\tau}\sum_{k=0}^{\tau-1} e^{\theta A(k,k+s)}\Big), \tag{23}$$

where $A(s, t) = \sum_{\ell=s+1}^{t} a(\ell)$. The equation in (20) can then be rewritten as

$$cs + x = \frac{\sum_{k=0}^{\tau-1} A(k, k + s)e^{\theta(s)A(k,k+s)}}{\sum_{k=0}^{\tau-1} e^{\theta(s)A(k,k+s)}}. \tag{24}$$

In the second step, we need to find the change of measure for our simulation under the critical time scale. For this, we need to introduce the Sanov theorem. Consider a random variable U on a finite alphabet $\Sigma = \{0, 1, 2, \dots, \tau - 1\}$, with the probability mass function $\underline{r} = (r_0, \dots, r_{\tau-1})$, i.e.

$$\mathsf{P}(U = k) = r_k, k = 0, \dots, \tau - 1.$$

Given the first n independent samples $U_1, \dots, U_n$, one might estimate the probability mass function $\underline{r}$ by forming the *empirical distribution*,

$$\underline{L}_n = (L_{n,0}, \dots, L_{n,\tau-1}),$$

where

$$L_{n,k} \equiv \frac{1}{n} \sum_{j=1}^{n} I (U_i = k),$$

and $I(U_i = k)$ is the indicator function, i.e., $I(U_i = k) = 1$ if $U_i = k$ and $I(U_i = k) = 0$ otherwise. The Sanov theorem below gives both an upper bound and a lower bound on how unlikely it is for the empirical distribution to deviate from the true probability mass function.

Theorem 3 *(Sanov [31]) Let $M_1(\Sigma)$ be the space of all probability mass functions on the alphabet Σ.*

(i) *(Upper bound) For every closed set $F \in M_1(\Sigma)$,*

$$\limsup_{n\to\infty} \frac{1}{n} \log \mathsf{P} (\underline{L}_n \in F) = - \inf_{\underline{p}\in F} \Delta(\underline{p}||\underline{r}),$$

where

$$\Delta(\underline{p}||\underline{r}) = \sum_{i=0}^{\tau-1} p_i \log \frac{p_i}{r_i}$$

is the Kullback-Leibler distance (or known as the relative entropy) that measures the distance between two probability mass functions $\underline{p}$ and $\underline{r}$ in $M_1(\Sigma)$.

(ii) *(Lower bound) For every open set $G \in M_1(\Sigma)$,*

$$\liminf_{n\to\infty} \frac{1}{n} \log \mathsf{P} (\underline{L}_n \in G) = - \inf_{\underline{p}\in G} \Delta(\underline{p}||\underline{r}).$$

Observe that

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n} A_i(\tau) - A_i(s) &= \frac{1}{n}\sum_{i=1}^{n} \sum_{t=s+1}^{\tau} a_i(t) \\
&= \frac{1}{n}\sum_{i=1}^{n} \sum_{t=s+1}^{\tau} a(t + U_i) \\
&= \frac{1}{n} \sum_{i=1}^{n} \sum_{t=s+1}^{\tau} \sum_{k=0}^{\tau-1} a(t + k) I(U_i = k) \\
&= \sum_{k=0}^{\tau-1} A(s + k, \tau + k) L_{n,k} \\
&= \langle \underline{A}(s, \tau), \underline{L_n} \rangle,
\end{aligned}$$

where

$$\underline{A}(s, \tau) = (A(s, \tau), A(s+1, \tau+1), \dots, A(s+\tau-1, 2\tau-1)),$$

and $\langle x, y\rangle$ is the dot product of two vectors x and y. Applying Sanov's theorem yields

$$\begin{aligned}
&\mathsf{P}(\sum_{i=1}^{n} A_i(\tau) - A_i(s) \geq nc(\tau - s)x) \\
&= \mathsf{P}(\langle \underline{A}(s, \tau), \underline{L_n} \rangle \geq c(\tau - s) + x) \\
&\approx \exp\Big(- n \inf_{\langle \underline{A}(s,\tau), \underline{p}\rangle \geq c(\tau-s)+x} \Delta(\underline{p}||\underline{r}) \Big).
\end{aligned}$$

The solution of the above minimization problem is the following exponentially twisted distribution:

$$p_k = \frac{e^{\theta(\tau-s)A(s+k,\tau+k)}}{\mathsf{E}\left[e^{\theta(\tau-s)(A_1(\tau)-A_1(s))}\right]} r_k, \quad k = 0, 1, \dots, \tau - 1, \quad (25)$$

where $\theta(s)$ is the solution of (24). Thus, under the critical time scale s^*, the change of measure we shall use is

$$p_k^* = \frac{e^{\theta(s^*)A\left(\tau-s^*+k,\tau+k\right)}}{\sum_{k=0}^{\tau-1} e^{\theta(s^*)A(k,k+\tau-s^*)}}, \quad (26)$$

where we use (22) and $r_k = \frac{1}{\tau}$ for all k in our setting.

Now we summarize our algorithm.

Algorithm 1:

(i) For $s = 1, \dots, \tau$, solve $\theta(s)$ in (24).

(ii) Find the critical time scale s^* by finding the minimum among $\Lambda_s^*(cs + x)$, $s = 1, \dots, \tau$, in (21).

(iii) For the j^{th} simulation, generate the i^{th} input by assigning $a_i(t) = a(t + \tilde{U}_{i,j})$, $i = 1, \dots, n$, where $\tilde{U}_{i,j}$'s are i.i.d. with the probability mass function

$$\mathsf{P}(\tilde{U}_{i,j} = k) = p_k^*, \quad , k = 0, 1, \dots, \tau - 1,$$

where p_k^* is defined in (26).

(iv) Form the estimate

$$\frac{1}{N} \sum_{j=1}^{N} \Big(\prod_{i=1}^{n} L(\tilde{U}_{i,j}) \Big) I(q_j(\tau) \geq nx),$$

where

$$L(k) = \frac{r_k}{p_k^*} = \frac{1}{\tau p_k^*}$$

is the likelihood ratio.

3.3 Multiplexing Multiple Classes of Inputs

In the previous section, we discussed the method to simulate a multiplexer with a single class of periodic inputs. In this section, we extend our method to the setting with multiple classes of periodic inputs.

Assume that there are m classes of inputs and within each class there are n periodic inputs. Also, assume all the inputs have the same period τ and the work conserving link has capacity $m \cdot n \cdot c$. Let $a_{i,j}(t)$ be the number of arrivals at time t from the i^{th} input of the j^{th} class, and $A_{i,j}(t) = \sum_{s=1}^{t} a_{i,j}(s)$ be its cumulative number of arrivals by time t. Generate the inputs by assigning

$$a_{i,j}(t) = b_j(t + U_{i,j}),$$

where $\{b_j(t), t \geq 1\}$, $j = 1, \dots, m$, are m periodic sequences with period τ and $U_{i,j}$'s are independent and uniformly distributed over $\{0, 1 \dots, \tau - 1\}$. Now the Lindley recursion can be rewritten as

$$q(t+1) = (q(t) + \sum_{i=1}^{n} \sum_{j=1}^{m} a_{i,j}(t+1) - m \cdot n \cdot c)^+.$$

Using a similar argument to that for the single class problem, we have from Crämer theorem that

$$\mathsf{P}(q(\tau) \geq mnx) \approx \exp\Big(- n \min_{1 \leq s \leq \tau} \Lambda_s^*(m(cs + x))\Big), \qquad (27)$$

where $\Lambda_s^*(\gamma)$ is the Legendre transform of the function $\Lambda_s(\theta) = \sum_{j=1}^{m} \Lambda_{s,j}(\theta)$ and

$$\Lambda_{s,j}(\theta) = \log \mathsf{E} e^{\theta(A_{1,j}(s))}.$$

The critical time scale s^* can then be found by finding the index of time that achieves the minimum in (27). Similarly, we can also apply Sanov's theorem to find the appropriate change of measure. The change of measure for the j^{th} class is

$$p^*_{j,k} = \frac{e^{\theta(s^*) B_j(\tau - s^* + k, \tau + k)}}{\sum_{k=0}^{\tau-1} e^{\theta(s^*) B_j(k, k+\tau-s^*)}}, \qquad (28)$$

where $B_j(s,t) = \sum_{\ell=s+1}^{t} b_j(\ell)$, and $\theta(s)$ is the solution of

$$m(cs + x) - \sum_{j=1}^{m} \frac{\sum_{k=0}^{\tau-1} B_j(k, k+s) e^{\theta(s) B_j(k,k+s)}}{\sum_{k=0}^{\tau-1} e^{\theta(s) B_j(k,k+s)}} = 0.$$

3.4 Numerical Results

In this section, we run simulations and compare the simulation results with theoretical results derived in the previous sections.

In our first experiment, we consider a work conserving link with a single class of periodic inputs as that in Section 3.2. In this experiment, we choose the (normalized) capacity $c = 1$, the period $\tau = 10$, and the periodic sequence

$$\{a(t), t = 1, \dots 10\} = \{9, 0, 0, 0, 0, 0, 0, 0, 0, 0\}.$$

We use the fast simulation algorithm in Section 3.2 to estimate the probabilities $\mathsf{P}(q(\tau) \geq nx)$ for $n = 1, 10$ and 100. The number of independent trials is 10^6. The results are shown in Figure 1, where we plot the curves $\frac{1}{n} \log \mathsf{P}(q(\tau) \geq nx)$ (as functions of buffer size x). As expected, all these curves move toward to the theoretical limit $\min_{1 \leq s \leq \tau} \Lambda_s^*(cs + x)$ in (19).

In our second experiment, we consider a work conserving link with two classed of periodic inputs ($m = 2$) as that in Section 3.3. In this experiment, we choose the (normalized) capacity $c = 1$, the period $\tau = 10$, and the periodic sequences

$$\{b_1(t), t = 1, \dots 10\} = \{9, 0, 0, 0, 0, 0, 0, 0, 0, 0\},$$

and

$$\{b_2(t), t = 1, \dots 10\} = \{5, 3, 1, 0, 0, 0, 0, 0, 0, 0\}.$$

We use the fast simulation algorithm outlined in Section 3.2 to estimate the probabilities $\mathsf{P}(q(\tau) \geq mnx)$ for $n = 1, 10$ and 100. The number of independent trials is also 10^6. The results are shown in Figure 2, where we also plot the curves $\frac{1}{n} \log \mathsf{P}(q(\tau) \geq mnx)$. As in the first experiment, all these curves move toward to the theoretical limit $\min_{1 \leq s \leq \tau} \Lambda_s^*(m(cs + x))$ in (27).

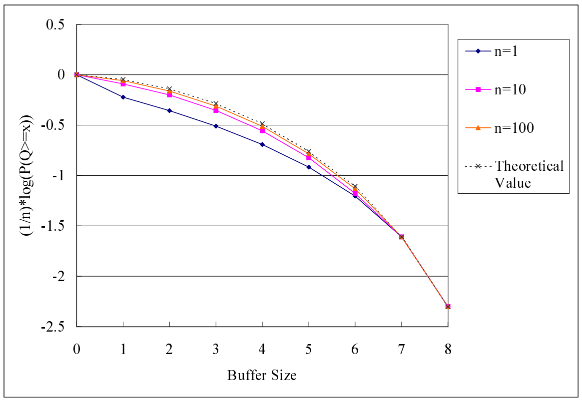

Figure 1: Multiplexing a single class of inputs

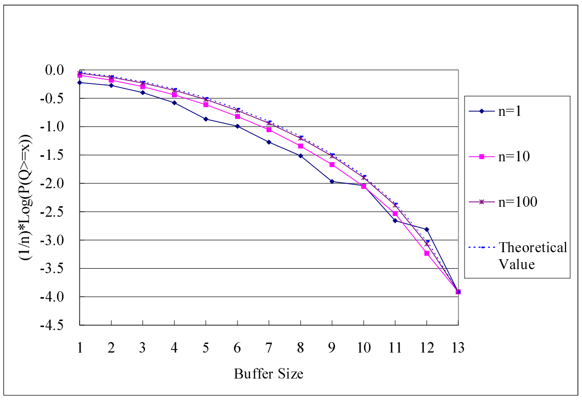

Figure 2: Multiplexing two classes of inputs

To gain some intuition on how buffer overflows, we consider a work conserving link with the (normalized) capacity $c = 1$. We choose the periodic sequence

$$\{a(t), t = 1, \dots, 10\} = \{9, 0, 0, 0, 0, 0, 0, 0, 0, 0\}.$$

We run the standard simulation for the event $\{q(\tau) \geq nx\}$ with $x = 4$. The number of independent trials is 10^6. We average the paths that lead to buffer overflow in these simulations. In Figure 3, we plot the average paths for $n = 10$, 20 and 30, respectively. All these three paths are quite close to the "most likely" path computed as follows (see e.g., [28] for continuous-time periodic on/off sources):

1. Let $A^*(t) = \sum_{k=0}^{\tau-1} (\sum_{s=1}^{t} a(s+k)) p_k^*$.
2. Let $a^*(t) = A^*(t) - A^*(t-1), t = 0, 1, \dots, \tau$.
3. Use Lindley's equation in (1) to compute the most likely queue length

   $$q^*(t+1) = (q^*(t) + a^*(t+1) - c)^+,$$

   starting from $q^*(0) = 0$.

	n=15					
	Crude Simulation			Importance Sampling		
x	Estimate of P(q(t)>=nx)	Standard Error	Relative Error	Estimate of P(q(t)>=nx)	Standard Error	Relative Error
1	3.21E-01	4.75E-04	0.04%	3.21E-01	5.13E-04	0.04%
2	6.48E-02	2.33E-04	0.09%	6.39E-02	2.92E-04	0.12%
3	5.89E-03	7.80E-05	0.34%	5.73E-03	6.74E-05	0.30%
4	2.16E-04	1.46E-05	1.74%	2.25E-04	4.39E-06	0.50%
5	2.00E-06	1.41E-06	18.20%	2.23E-06	6.02E-08	0.70%
6				2.13E-08	9.13E-09	11.02%

Table 1: Probability estimates for multiplexing 15 inputs

	n=25					
	Crude Simulation			Importance Sampling		
x	Estimate of P(q(t)>=nx)	Standard Error	Relative Error	Estimate of P(q(t)>=nx)	Standard Error	Relative Error
1	1.53E-01	3.45E-04	0.06%	1.53E-01	3.75E-04	0.06%
2	1.07E-02	1.05E-04	0.25%	1.06E-02	6.96E-05	0.17%
3	1.93E-04	1.35E-05	1.80%	2.02E-04	2.42E-06	0.31%
4	1.00E-06	1.00E-06	25.75%	6.54E-07	1.43E-08	0.56%
5				3.92E-10	2.21E-11	1.45%
6				2.48E-14	3.05E-15	3.16%

Table 2: Probability estimates for multiplexing 25 inputs

This suggests that there might be a conditional limit theorem for our setting. Note that in this experiment, $s^* = 3$ and the queue starts to build up at $\tau - s^* = 7$ in Figure 3.

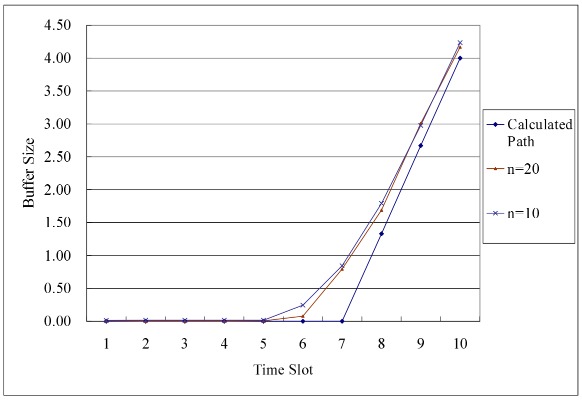

Figure 3: The most likely path that leads to buffer overflow

To show the significant variance reduction using our fast simulation algorithm for estimating small buffer overflow probabilities, we compare our fast simulation results with those from the standard approach (crude simulation). In Table 1 and Table 2, we report the estimates for the queue length distributions when the number of inputs $n = 15$ and $n = 25$, respectively (see e.g., [32] for a detailed discussion on the precision of simulation experiments). The confidence interval is calculated for 99% of confidence, and the relative error is calculated accordingly. The number of independent trials for each estimate is $N = 10^6$. Clearly, the standard error obtained from our fast simulation algorithm is much smaller than that from the crude simulation when the buffer overflow probabilities are small. Thus, we have a more accurate estimate for a small buffer overflow probability from the fast simulation method when the number of trials are the same. In fact, as shown in Table 2, the crude simulation even cannot render a quality estimate when the buffer size is more than 4.

To clearly illustrate the use of importance sampling in the fast simulation, we plot the biasing pmf of the queue size for a single input for different values of x. Notice that the probability distribution of the queue size under our assumption (the starting phase of the input is a uniformly distributed random variable ranging from 0 to $\tau - 1$) is uniformly distributed from 1 to $q_{max} = 8$. However, the distribution is different when using the biasing pmf (probability mass function) for the case of $x = 6$. The event of importance (queue size ≥ 6) has a much higher probability to occur. In other words, in the fast simulation, the rare event of interest will occur more often and thus speed up the simulation process.

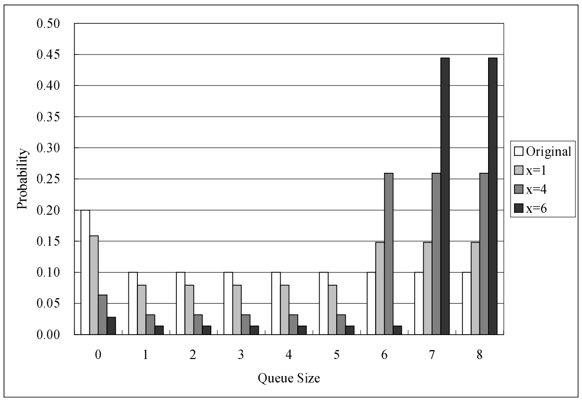

Figure 4: Probability mass functions (pmf) of the queue size for a single input using the biasing pmf for different values of x

To see whether the upper bound in Theorem 1 is tight, we consider the upper constraint function $f(t) = min(\pi t, \rho t + \sigma)$, where $\pi = 9, \rho = 1$, and $\sigma = 16$. For this upper constraint function, we consider the periodic sequence

$$\{a(t), t = 1, \ldots, 18\}$$
$$= \{9, 9, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0\}.$$

The periodic sequence has a period $\tau = 18$ and it is clearly f-upper constrained. As $\rho = 1$, we choose the (normalized) capacity $c = 2$. The number of periodic inputs n is chosen to be 50.

In Figure 5, we compare the fast simulation result for multiplexing the 50 periodic inputs with the theoretical upper bound in Theorem 1 and the upper bound in Kesidis and Konstantopoulos [23]. As expected, our bound is tighter than that in [23], and it is not too far away from the simulation result.

4. CONCLUSIONS

In this paper, we derived simple stochastic bounds for multiplexing independent regulated traffic. The bound for queue

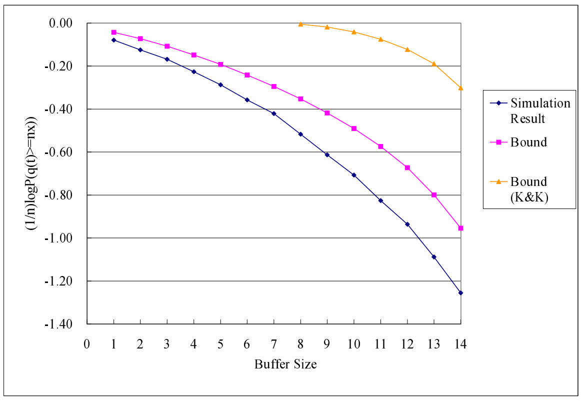

Figure 5: Comparison of stochastic bounds

length is shown to be a stochastic extension of the deterministic worst case bound in [8] and it is asymptotically tighter than the bound in Kesidis and Konstantopoulos [23]. Based on Sanov's theorem and importance sampling, we also proposed a fast simulation algorithm for simulating a work conserving link with a large number of periodic inputs. Our fast simulation algorithm shows significant variance reduction over the standard approach. Also, our simulation shows that our stochastic bound, though simple, is not too loose.

There are several problems that require further study.

(i) Kesidis and Konstantopoulos [23, 24] identified the worst-case traffic for a work conserving link with one input. The input is regulated by a leaky bucket and a peak rate regulator. It seems difficult to identify the worst-case traffic when there are more than one input regulated by a general constraint in our paper.

(ii) Though our fast simulation algorithm is much better than the standard approach, we did not prove that it is *asymptotically optimal* in the sense defined in [12]. The difficulty lies in the fact that the set for buffer overflow is a union of convex sets and may not be a convex set.

(iii) From our simulation results, there seems a conditional limit theorem for the most likely path that leads to buffer overflow. A formal proof might be derived by carefully examining the decay rate of each convex set that leads to buffer overflow.

Acknowledgments:

The first author would like to thank George Kesidis and Jörg Liebeherr for many valuable comments on an earlier draft of this paper.

5. REFERENCES

[1] R. Boorstyn, A. Burchard, J. Liebeherr, and C. Oottamakorn, "Effective Envelopes: Statistical Bounds on Multiplexed Traffic in Packet Networks," *Proc. IEEE INFOCOM 2000*, Vol. 3, pp. 1223-1232, Tel Aviv, Israel, 2000.

[2] D.D. Botvich and N.G. Duffield, "Large deviations, the shape of the loss curve, and economies of scale in large multiplexers," *Queueing Systems*, Vol. 20, pp. 293-320, 1995.

[3] J. Bucklew. *Large Deviation Techniques in Decision, Simulation and Estimation.* New York, NY: J. Wiley & Sons, Inc., 1990.

[4] C.S. Chang, "Stability, queue length and delay of deterministic and stochastic queueing networks," *IEEE Trans. Automat. Contr.*, vol. 39, pp. 913-931, 1994.

[5] C.S. Chang, *Performance Guarantees in Communication Networks*, Springer-Verlag, 2000.

[6] H. Cramér, "Sur un nouveau théoréme-limite de la théorie des probabilités." In *Actualités Scientifiques et Industrielles*, number 736 in Colloque consacré à la héorie des probabilités, pp. 5-23. Hermann, Paris, 1938.

[7] C. Courcoubetis, F. Kelly and R. Weber, "Measurement-based usage charges in communication networks," *preprint*, 1998. Available from http://www.statslab.cam.ac.uk/~frank.

[8] R.L. Cruz, "A calculus for network delay, Part I: Network elements in isolation," *IEEE Trans. Inform. Theory*, vol. 37, pp. 114-131, 1991.

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

[10] A. Elwalid, D. Mitra and R. Wentworth, "A new approach for allocating buffers and bandwidth to heterogeneous, regulated traffic in an ATM node," *IEEE Journal on Selected Areas in Communications*, Vol. 13, pp. 1115-1127, 1995.

[11] G. Hebuterne and J.Y. Le Boudec, "Comments on "a deterministic approach to the end-to-end analysis of packet flows in connection oriented networks"", *IEEE/ACM Transactions on Networking*, Vol. 8, No. 1, Feb., 2000.

[12] P. Heidelberger, "Fast simulation of rare events in queueing and reliability models," IBM Research Report RC 19028, Yorktown Heights, New York, 1993. Preliminary version appears in *Performance Evaluation of Computer and Communications Systems*, Springer Verlag, Lecture Notes in Computer Science, Volume 729, 165-202.

[13] (IETF Internet draft) Y. Bernet, J. Binder, S. Blake, M. Carlson, S. Keshav, B. Ohlman, D. Verma, Z. Wang, and W. Weiss, "A framework for differentied services," http://www.ietf.org/html.charters/diffserv-charter.html,
1998.

[14] (IETF Internet draft) S. Blake, D. Black, M. Carlson, E. Davies, Z. Wang and W. Weiss, "An architecture of differentiated services,"
http://www.ietf.org/html.charters/diffserv-charter.html,
1998.

[15] (IETF Internet draft) K. Nichols, S. Blake, F. Baker, and D. Black, "Definition of the differentiated services field (DS field) in the IPv4 and IPv6 headers," http://www.ietf.org/html.charters/diffserv-charter.html, 1998.

[16] (IETF Internet draft) V. Jacobson, K. Nichols, and K. Poduri, "An expedited forwarding PHB," http://www.ietf.org/html.charters/diffserv-charter.html, 1998.

[17] (IETF RFC 2205) L. Zhang, S. Berson, S. Herzog, and S. Jamin, "Resource ReSerVation Protocol (RSVP)," ftp://ds.internic.net/rfc/rfc2205.txt, 1997.

[18] (IETF RFC 2210) J. Wroclawski, "The use of RSVP with IETF integrated service," ftp://ds.internic.net/rfc/rfc2210.txt, 1997.

[19] (IETF RFC 2211) J. Wroclawski, "Specification of the controlled-load network element service," ftp://ds.internic.net/rfc/rfc2211.txt, 1997.

[20] (IETF RFC 2212) S. Shenker, C. Patridge and R. Guerin, "Specification of guaranteed quality of service," ftp://ds.internic.net/rfc/rfc2212.txt, 1997.

[21] (IETF RFC 2215) S. Shenker and J. Wroclawski, "General characterization parameters for integrated service network elements," ftp://ds.internic.net/rfc/rfc2215.txt, 1997.

[22] (IETF RFC 2216) S. Shenker and J. Wroclawski, "Network element service specification template," ftp://ds.internic.net/rfc/rfc2216.txt, 1997.

[23] G. Kesidis and T. Konstantopoulos, "Worst-case performance of a buffer with independent shaped arrival processes," to appear in *IEEE Communications Letters*. Available from http://cheetah.vlsi.uwaterloo.ca.

[24] G. Kesidis and T. Konstantopoulos, "Extremal traffic and worst-case performance for queues with shaped arrivals," to appear in *Fields Institute Communications*. Available from http://cheetah.vlsi.uwaterloo.ca.

[25] E. Knightly, "Enforceable quality of service guarantees for bursty traffic streams," *Proc. of IEEE INFOCOM'98*, pp. 635-642, San Francisco, March, 1998.

[26] J.F. Kurose, "On computing per-session performance bounds in high-speed multi-hop computer networks," *SIGMETRICS and PERFORMANCE'92*.

[27] F. LoPresti, Z. Zhang, J. Kurose and D. Towsley, "Source time scale and optimal buffer/bandwidth tradeoff for regulated traffic in an ATM node," *Proc. IEEE INFOCOM'97*, pp. 676-683, Kobe, Japan, April, 1997.

[28] M. Mandjes and A. Ridder, "Optimal trajectory to overflow in a queue fed by a large number of sources," *Queueing Systems*, Vol. 31, pp. 137-170, 1999.

[29] M. Reisslein, K.W. Ross, and S. Rajagopal, "A framework for guaranteeing statistical QoS," Feb., 1999. Available from http://www.eurecom.fr/~ross.

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

[31] I.N. Sanov, "On the probability of large deviations of random variables," in Russian, 1957. (English translation from *Mat. Sb. (42)) in Selected Translations in Mathematical Statistics and Probability I (1961), pp. 213-244).*

[32] W. Song and B. Schmeiser, "Reporting the Precision of Simulation Experiments," *New Directions in Simulation for Manufacturing and Communications*, pp. 402-407, 1994.

[33] A. Shwartz and A. Weiss. *Large Deviations for Performance Analysis: Queues, Communication and Computing*. 1994.

[34] A. Simonian and J. Guibert, "Large deviations approximations for fluid queues fed by a large number of on/off sources," *IEEE Journal on Selected Areas in Communications*, Vol. 13, pp. 1017-1027, 1995.

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