# Effective Envelopes: Statistical Bounds on Multiplexed Traffic in Packet Networks

*Robert Boorstyn* * *Almut Burchard* ** *Jörg Liebeherr* † *Chaiwat Oottamakorn* *

* Department of Electrical Engineering, Polytechnic University, Brooklyn, NY 11201

** Department of Mathematics, University of Virginia, Charlottesville, VA 22903

† Department of Computer Science, University of Virginia, Charlottesville, VA 22903

*Abstract*—**A statistical network service which allows a certain fraction of traffic to not meet its QoS guarantees can extract additional capacity from a network by exploiting statistical properties of traffic. Here we consider a statistical service which assumes statistical independence of flows, but does not make any assumptions on the statistics of traffic sources, other than that they are regulated, e.g., by a leaky bucket. Under these conditions, we present functions, so-called** *local effective envelopes* **and** *global effective envelopes***, which are, with high certainty, upper bounds of multiplexed traffic. We show that these envelopes can be used to obtain bounds on the amount of traffic on a link that can be provisioned with statistical QoS. A key advantage of our bounds is that they can be applied with a variety of scheduling algorithms. In fact, we show that one can reuse existing admission control functions that are available for scheduling algorithms with a deterministic service. We present numerical examples which compare the number of flows with statistical QoS guarantees that can be admitted with our empirical envelope approach to those achieved with existing methods.**

*Key Words: Statistical Multiplexing, Statistical Service, Scheduling, Quality-of-Service.*

## I. Introduction

Performance guarantees in QoS networks are either deterministic or statistical. A *deterministic service* guarantees that all packets from a flow satisfy given worst-case end-to-end delay bounds and no packets are dropped in the network [2], [4], [8], [15]. A deterministic service provides the highest level of QoS guarantees, however, it leaves a significant portion of network resources on the average unused [22].

A *statistical service* makes probabilistic service guarantees, for example, of the form:

$$Pr[Delay > X] < \varepsilon \quad \text{or} \quad Pr[Loss] < \varepsilon \ .$$

By allowing a fraction of traffic to violate its QoS guarantees, one can improve the statistical multiplexing gain at network links and increase the achievable link utilization. The key assumption that leads to the definition of statistical services is that traffic arrivals are viewed as random processes. With this assumption a statistical service can improve upon a deterministic service by (1) taking advantage of knowledge about the statistics of traffic sources, and (2) by taking advantage of the statistical independence of flows.

Since it is often not feasible to obtain a reliable statistical characterization of traffic sources, recent research on statistical QoS has attempted to exploit statistical multiplexing without assuming a specific source model. Starting with the seminal work in [8], researchers have investigated the statistical multiplexing gain by only assuming that flows are statistically independent, and that traffic from each flow is constrained by a deterministic regulator, e.g., by a leaky bucket [5], [8], [7], [9], [10], [12], [16], [17], [19], [20], [21]. Henceforth, we will refer to traffic which satisfies these assumptions as *regulated adversarial traffic*.

In this paper we attempt to provide new insights into the problem of determining the multiplexing gain of statistically independent, regulated, but otherwise arbitrary traffic flows at a network link. We introduce the notion of *effective envelopes*, which are, with high certainty, upper bounds on the aggregate traffic of regulated flows. We use effective envelopes to devise admission control tests for a statistical service for a large class of scheduling algorithms. We show that with effective envelopes, admission control for a statistical service can be done in a similar fashion as with deterministic envelopes for a deterministic service [2], [4] . In fact, we show that one can reuse admission control conditions derived for various packet scheduling algorithms in the context of a deterministic service, e.g., [4], [15], [23]. Note that only few results are available on statistical multiplexing of adversarial traffic, which can consider scheduling algorithms other than a simple multiplexer [7], [12].

Related work, which, due to space constraints, cannot

This work is supported in part by the National Science Foundation through grants NCR-9624106 (CAREER), ANI-9730103, and DMS-9971493, and by the New York State Center for Advanced Technology in Telecommunications (CATT).

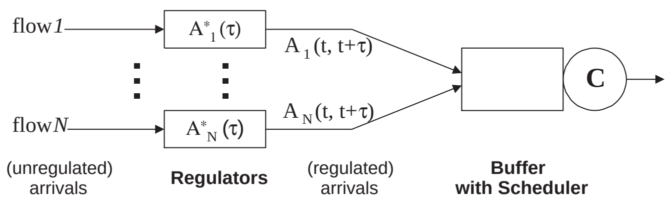



Fig. 1. Regulators and Scheduler at a Link.

be fully discussed, are all attempts to consolidate the deterministic network calculus [4] with statistical multiplexing (e.g., [2], [6], [10], [11], [12], [14]). In addition, of particular relevance to this paper are all previous results on statistical multiplexing gain with adversarial regulated traffic, as cited above.

The results derived in this paper only apply to a single node. Since traffic from multiple flows passing through the same sequence of congested nodes may become correlated, the assumption of statistical independence of flows may not hold in such a setting. Only few results are currently available on end-to-end QoS guarantees for adversarial regulated traffic [7], [20], [21].

The remaining sections of this paper are structured as follows. In Section II we specify our assumptions on the traffic and define the effective envelopes. In Section III we derive sufficient schedulability conditions for a general class of packet schedulers, which can be used for a deterministic and (two types of) statistical QoS guarantees. In Section IV, we use large deviations results to derive bounds for effective envelopes. In Section V we compare the statistical multiplexing gain attainable with the effective envelopes approach to those obtained with other methods ([8], [12], [19]). In Section VI we present conclusions of our work.

## II. Traffic Arrivals and Envelope Functions

We consider traffic arrivals to a single link with transmission rate $C$. As shown in Figure 1, the arrivals from each flow are policed by a regulator, and then inserted into a buffer. A scheduler determines the order in which traffic in the buffer is transmitted. In the following, we view traffic mainly as continuous-time fluid-flow traffic. Note, however, that our discussion applies, without restrictions, to discrete-time or discrete-size (packetized) views of traffic arrivals.

QoS guarantees for a flow $j$ are specified in terms of a delay bound $d_j$. A QoS violation occurs if traffic from flow $j$ experiences a delay exceeding $d_j$. (We assume that delays consist only of waiting time in the buffer and transmission time.)

### A. Traffic Arrivals

Traffic arrivals to the link come from a set of flows which is partitioned into $Q$ classes $\mathcal{C}_q$, each containing $N_q$ flows. (Each flow may itself be an aggregate of the traffic from multiple sessions.)

The traffic arrivals from flow $j$ in an interval $[t_1, t_2)$ are denoted as $A_j(t_1, t_2)$. We assume that a traffic flow is characterized by a family of random variables $A_j(t_1, t_2)$ which is characterized as follows:

*(A1)* **Additivity.** For any $t_1 < t_2 < t_3$, we have $A_j(t_1, t_2) + A_j(t_2, t_3) = A_j(t_1, t_3)$.

*(A2)* **Subadditive Bounds.** Traffic $A_j$ is regulated by a deterministic subadditive envelope $A_j^*$ as

$$A_j(t, t+\tau) \leq A_j^*(\tau) \qquad \forall t \geq 0, \forall \tau \geq 0 \,. \tag{1}$$

*(A3)* **Stationarity.** The $A_j$ are *stationary* random variables, i.e., $\forall t, t' > 0$

$$Pr[A_j(t, t+\tau) \leq x] = Pr[A_j(t', t'+\tau) \leq x] \,. \tag{2}$$

In other words, all time shifts of $A_j$ are equally probable.

*(A4)* **Independence.** The $A_i$ and $A_j$ are stochastically independent for all $i \neq j$.

*(A5)* **Homogeneity within a Class.** Flows in the same class have identical deterministic envelopes and identical delay bounds. So, $A_i^* = A_j^*$ and $d_i = d_j$ if $i$ and $j$ are in the same class. Henceforth, we denote by $d_q$ the delay bound associated with traffic from class $q$. By $A_{\mathcal{C}_q}$ we denote the arrivals from class $q$, that is, $A_{\mathcal{C}_q}(t, t+\tau) = \sum_{j \in \mathcal{C}_q} A_j(t, t+\tau)$.

*Remarks:*

- We want to point out that the above assumptions are quite general. The class of subadditive deterministic traffic envelopes is the most general class of traffic regulators [4], [2]. The assumptions on the randomness of flows are also quite general. Note that, different from [9], [10], we do not require ergodicity.
- The traffic regulators most commonly used in practice are *leaky buckets* with a peak rate enforcer. Here, traffic on flow $j$ is characterized by three parameters $(P_j, \sigma_j, \rho_j)$ with a deterministic envelope given by

$$A_j^*(\tau) = \min\{P_j \tau, \sigma_j + \rho_j \tau\} \qquad \forall \tau \geq 0 \,, \tag{3}$$

where $P_j \geq \rho_j$ is the peak traffic rate, $\rho_j$ is the average traffic rate, and $\sigma_j$ is a burst size parameter. We will use this type of regulators in our numerical examples in Section V.

- A consequence of subadditivity of the $A_j^*$ is that the limit $\rho_j := \lim_{\tau \to \infty} A_j^*(\tau)/\tau$ exists, and that it provides an upper bound for the longterm arrival rate for $A_j$. We will assume without loss of generality, that for all $t$,

$$\lim_{\tau\to\infty} \frac{A_j(t,t+\tau)}{\tau} \;=\; \rho_j \; . \tag{4}$$

### B. Definition of Effective Envelopes

We next define *local effective envelopes* and *global effective envelopes* which are, with high certainty, upper bounds on aggregate traffic from a given class $q$. The envelopes will be defined for a set of flows $\mathcal{C}$ with arrival functions $A_j$ and aggregate traffic $A_{\mathcal{C}}(t,t+\tau) = \sum_{j\in\mathcal{C}} A_j(t,t+\tau)$.

*Definition 1:* A **local effective envelope** for $A_{\mathcal{C}}(t,t+\tau)$ is a function $\mathcal{G}_{\mathcal{C}}(\,\cdot\,;\varepsilon)$ that satisfies for all $\tau\geq 0$ and all $t$

$$Pr\Big[A_{\mathcal{C}}(t,t+\tau)\leq \mathcal{G}_{\mathcal{C}}(\tau;\varepsilon)\Big]\geq 1-\varepsilon \; . \tag{5}$$

In other words, a *local effective envelope* provides a bound for the aggregate arrivals $A_{\mathcal{C}}(t,t+\tau)$ for *any specific* ('local') time interval of length $\tau$. Under the stationarity assumption (A3), Eqn. (5) holds for all times $t$, provided that it only holds for one value $t=t_o$.

It is easy to see that there exists a smallest local effective envelope, since the minimum of two local effective envelopes is again such an envelope. Note, however, that local effective envelopes are in general not subadditive in $\tau$, but satisfy the weaker property

$$\mathcal{G}_{\mathcal{C}}(\tau_1+\tau_2,\varepsilon_1+\varepsilon_2) \;\leq\; \mathcal{G}_{\mathcal{C}}(\tau_1,\varepsilon_1)+\mathcal{G}_{\mathcal{C}}(\tau_2,\varepsilon_2) \; . \tag{6}$$

A local effective envelope $\mathcal{G}_{\mathcal{C}}(\tau;\varepsilon)$ is a bound for the traffic arrivals in an arbitrary, but fixed interval of length $\tau$. Global effective envelopes, to be defined next, are bounds for the arrivals in all subintervals $[t,t+\tau)$ of a larger interval.

For the definition of global effective envelopes, we take advantage of the notion of empirical envelopes, as used in [2], [22]. Consider a time interval $I_\beta$ of length $\beta$. The **empirical envelope** $\mathcal{E}_{\mathcal{C}}(\,\cdot\,;\beta)$ of a collection $\mathcal{C}$ of flows is the maximum traffic in subintervals of $I_\beta$ as follows:

$$\mathcal{E}_{\mathcal{C}}(\tau;\beta)=\sup_{[t,t+\tau)\subseteq I_\beta} A_{\mathcal{C}}(t,t+\tau) \; . \tag{7}$$

*Definition 2:* A **global effective envelope** for an interval $I_\beta$ of length $\beta$ is a subadditive function $\mathcal{H}_{\mathcal{C}}(\,\cdot\,;\beta)$ which satisfies

$$Pr\Big[\mathcal{E}_{\mathcal{C}}(\tau;\beta)\leq \mathcal{H}_{\mathcal{C}}(\tau;\beta,\varepsilon),\;\; \forall 0\leq\tau\leq\beta\Big]\geq 1-\varepsilon \; . \tag{8}$$

The attribute 'global' is justified since $\mathcal{H}_{\mathcal{C}}(\,\cdot\,;\beta,\varepsilon)$ is a bound for traffic for all intervals of length $\tau\leq\beta$ in $I_\beta$. Now, due to stationarity of the $A_j$, Eqn. (8) holds for *all* intervals of length $\beta$, if it holds for one specific interval $I_\beta$. When applied to scheduling, we will select $\beta$ such that it has at least the length of the longest busy period.[1]

Assuming that one has obtained local or global effective envelopes separately for each traffic class, the following lemma helps to obtain bounds for the traffic from all classes.

*Lemma 1: Given a set of flows that is partitioned into $Q$ classes $\mathcal{C}_q$, with arrival functions $A_{\mathcal{C}_q}$. Let $\mathcal{G}_{\mathcal{C}_q}$ and $\mathcal{H}_{\mathcal{C}_q}$ be local and global effective envelopes for class $q$. Then the following inequalities hold.* Correction: 7-20-99

*(a) If $\sum_q \mathcal{G}_{\mathcal{C}_q}(\tau,\varepsilon)\leq x$, then, for all $t$,*
$Pr\Big[\sum_q A_{\mathcal{C}_q}(t,t+\tau)>x\Big]<Q\cdot\varepsilon.$

*(b) If $\sum_p \mathcal{H}_{\mathcal{C}_q}(\tau,\beta;\varepsilon)\leq x(\tau)$ for all $\tau$, then*
$Pr\Big[\exists\tau:\;\sum_p \mathcal{E}_{\mathcal{C}_q}(\tau,\beta)>x(\tau)\Big]<Q\cdot\varepsilon.$

The rather simple proof of the lemma can be found in [1]. Our derivations in Section IV will make it clear that for $\varepsilon$ small enough, neither $\mathcal{G}_{\mathcal{C}_q}$ nor $\mathcal{H}_{\mathcal{C}_q}$ are very sensitive with respect to $\varepsilon$, so that the bounds for $\varepsilon$ and $Q\cdot\varepsilon$ are comparable.

## III. Deterministic and Statistical Schedulability Conditions

In this section, we present three schedulability conditions for a general class of work-conserving scheduling algorithms. The first condition, expressed in terms of deterministic envelopes, ensures deterministic guarantees. The second and third conditions, which use the local and global effective envelopes, respectively, yield statistical guarantees. All three schedulability conditions will be derived from the same expression for the delay of a traffic arrival in an arbitrary work-conserving scheduler (Eqn. (14) in Section III-A).

In our discussions, we will not take into consideration that packet transmissions on a link cannot be preempted. This assumption is reasonable when packet transmission times are short. For the specific scheduling algorithms considered in this paper, accounting for non-preemptiveness of packets does not introduce principal difficulties, however, it requires additional notation (see [15]). Also, to keep notation minimal, we assume that the transmission rate of the link is normalized, that is $C=1$.

### A. Schedulability

Suppose a (tagged) arrival from a flow $j$ in class $q$ ($j\in\mathcal{C}_q$) arrives to a work-conserving scheduler at time $t$.

[1] For arrival functions $A_j$ and regulators with deterministic envelopes $A_j^*$, the longest busy period in a work-conserving scheduler is given by: $\inf\{\tau>0\;;\;\sum_{j\in\mathcal{C}} A_j^*(\tau)\leq\tau\}$.

Without loss of generality we assume that the scheduler is empty at time 0. We will derive a condition that must hold so that the arrival does not violate its delay bound $d_q$.

Let us use $A^{q,t}(t_1, t_2)$ to denote the traffic arrivals in the time interval $[t_1, t_2)$ which will be served before a class $q$ arrival at time $t$. Let $A^{q,t}_{\mathcal{C}_p}(t_1, t_2)$ denote the traffic arrivals from flows in $\mathcal{C}_p$ which contribute to $A^{q,t}(t_1, t_2)$.

Suppose that $t - \hat{\tau}$ is the last time before $t$ when the scheduler does not contain traffic that will be transmitted before the tagged arrival from class $q$. That is,

$$\hat{\tau} = \inf\{x \geq 0 \mid A^{q,t}(t-x, t) \leq x\} . \tag{9}$$

So, in the time interval $[t - \hat{\tau}, t)$ the scheduler is continuously transmitting traffic which will be served before the tagged arrival. (Note that $\hat{\tau}$ is a function of $t$ and $q$. To keep notation simple, we do not make the dependence explicit.)

Given $\hat{\tau}$, the tagged class-$q$ arrival at time $t$ will leave the scheduler at time $t + \delta$ if $\delta > 0$ is such that

$$\delta = \inf\left\{\tau_{out} \mid A^{q,t}(t - \hat{\tau}, t + \tau_{out}) \leq \hat{\tau} + \tau_{out}\right\} . \tag{10}$$

Hence, the tagged class-$q$ arrival does not violate its delay bound $d_q$ if and only if

$$\forall \hat{\tau}\ \exists \tau_{out} \leq d_q\ : \left\{A^{q,t}(t - \hat{\tau}, t + \tau_{out}) \leq \hat{\tau} + \tau_{out}\right\} . \tag{11}$$

Then, the traffic arrival does not have a deadline violation if $d_q$ is selected such that

$$\sup_{\hat{\tau}} \left\{A^{q,t}(t - \hat{\tau}, t + d_q) - \hat{\tau}\right\} \leq d_q . \tag{12}$$

In general, Eqn. (12) is a sufficient condition for meeting a delay bound. For FIFO and EDF schedulers, the condition is also necessary [15]. [2]

For a specific work-conserving scheduling algorithm, let $\bar{\tau}_p$ (with $-\hat{\tau} \leq \bar{\tau}_p \leq d_q$ ) denote the smallest values for which

$$A_{\mathcal{C}_p}(t - \hat{\tau}, t + \bar{\tau}_p) \geq A^{q,t}_{\mathcal{C}_p}(t - \hat{\tau}, t + d_q) . \tag{13}$$

*Remark:* For most work-conserving schedulers one can easily find $\bar{\tau}_p$ such that equality holds in Eqn. (13). For example, for FIFO, SP,[3] and EDF schedulers, we have:

[2] A FIFO scheduler transmits traffic in the order of arrival times. An EDF (Earliest-Deadline-First) scheduler tags traffic with a deadline which is set to the arrival time plus the delay bound $d_q$, and transmits traffic in the order of deadlines.

[3] An SP (Static Priority) scheduler assigns each class a priority level (we assume that a lower class index indicates a higher priority), and has one FIFO queue for traffic arrivals from each class. SP always transmits traffic from the highest priority FIFO queue which has a backlog.

FIFO: $\bar{\tau}_p = 0$

SP: $\bar{\tau}_p = \begin{cases} -\hat{\tau} & , p > q \\ 0 & , p = q \\ d_q & , p < q \end{cases}$

EDF: $\bar{\tau}_p = \max\{-\hat{\tau}, d_q - d_p\}$

With Eqn. (13), the arrival from class $q$ at time $t$ does not have a violation if $d_q$ is selected such that

$$\sup_{\hat{\tau}} \left\{\sum_p A_{\mathcal{C}_p}(t - \hat{\tau}, t + \bar{\tau}_p) - \hat{\tau}\right\} \leq d_q . \tag{14}$$

Next, we show how Eqn. (14) can be used to derive schedulability conditions for deterministic and statistical services, using deterministic envelopes, local effective envelopes, and global effective envelopes. For a deterministic service, the delay bound $d_q$ must be chosen such that Eqn. (14) is never violated. For a statistical service, $d_q$ is chosen such that a violation of Eqn. (14) is a rare event.

### *B. Schedulability with Deterministic Envelopes*

Exploiting the property of deterministic envelopes in Eqn. (1), we can relax Eqn. (14) to

$$\sup_{\hat{\tau}} \left\{\sum_p \sum_{j \in \mathcal{C}_p} A^*_j(\bar{\tau}_p + \hat{\tau}) - \hat{\tau}\right\} \leq d_q . \tag{15}$$

Since, $\bar{\tau}_p + \hat{\tau}$ is not dependent on $t$, we have obtained a sufficient schedulability condition for an arbitrary traffic arrival. We refer the reader [15] to verify that for FIFO and EDF scheduling algorithms the condition in Eqn. (15) is also necessary, in the sense that if it is violated, then there exist arrival patterns conforming with $A^*_j$ leading to deadline violations for class $q$. For SP scheduling, the condition is necessary only if the deterministic envelopes are concave functions.

Next we present bounds on the likelihood of a violation of Eqn. (14), using local and global effective envelopes.

### *C. Schedulability with Local Effective Envelopes*

With Eqn. (14), the probability that the tagged arrival from time $t$ experiences a deadline violation is less than $\varepsilon$ if $d_q$ is selected such that

$$Pr\left[\sup_{\hat{\tau}} \left\{\sum_p A_{\mathcal{C}_p}(t - \hat{\tau}, t + \bar{\tau}_p) - \hat{\tau}\right\} \leq d_q\right] \geq 1 - \varepsilon . \tag{16}$$

Let us, for the moment, make the convenient assumption that

$$Pr\left[\sup_{\hat{\tau}}\left\{\sum_p A_{\mathcal{C}_p}(t-\hat{\tau},t+\bar{\tau}_p)-\hat{\tau}\right\}\leq d_q\right]\approx \sup_{\hat{\tau}} Pr\left[\sum_p A_{\mathcal{C}_p}(t-\hat{\tau},t+\bar{\tau}_p)-\hat{\tau}\leq d_q\right]. \quad (17)$$

Assuming that equality holds in Eqn. (17), we can re-write Eqn. (16) as

$$\sup_{\hat{\tau}} Pr\left[\sum_p A_{\mathcal{C}_p}(t-\hat{\tau},t+\bar{\tau}_p)-\hat{\tau}\leq d_q\right]\geq 1-\varepsilon\ . \quad (18)$$

*Remark:* The assumption in Eqn. (17) requires further justification, since, in general, the right hand side is larger than the left hand side. On the other hand, several works on statistical QoS have used Eqn. (17) with equality [3], [11], [12], [13], [14], and, in several cases, have supported the assumption with numerical examples.

Recall from the definition of the local effective envelope that $\mathcal{G}_{\mathcal{C}_p}(\tau,\varepsilon)\leq x$ implies $Pr\left[A_{\mathcal{C}_p}(t,t+\tau)>x\right]<\varepsilon$. Then, with Lemma 1(a) and assuming that Eqn. (17) holds with equality, we have that a class-$q$ arrival has a deadline violation with probability $<\varepsilon$ if $d_q$ is selected such that

$$\sup_{\hat{\tau}}\left\{\sum_p \mathcal{G}_{\mathcal{C}_p}(\bar{\tau}_p+\hat{\tau},\varepsilon/Q)-\hat{\tau}\right\}\leq d_q\ . \quad (19)$$

With Eqn. (19) we have found an expression for the probability that an arbitrary traffic arrival results in a violation of delay bounds. This condition can be viewed as a general formulation of the schedulability conditions for statistical QoS from [11], [12], [14].

The drawback of the condition in Eqn. (19) is its dependence on the assumption in Eqn. (17). Empirical evidence from numerical examples, including those presented in this paper, as well as numerical evidence from previous work which employed this assumption [3], [12], suggests that Eqn. (19) is not overly optimistic. However, it should be noted that the bound in Eqn. (19) is not a rigorous one.

### D. Schedulability with Global Effective Envelopes

We next use global effective envelopes to express the probability of a deadline violation in a time interval. We will see that this bound, while more pessimistic, can be made rigorous.

Consider again the traffic arrival from class $q$ which occurs at time $t$. The arrival time $t$ lies in a busy period of the scheduler $I_\beta$ of length at most $\beta$, which starts at time $\leq t-\hat{\tau}$ and which ends at a time after the tagged arrival has departed.

Using the properties of the empirical envelope $\mathcal{E}_{\mathcal{C}_p}$, as defined in Section II, we have that, for all $t$ and $\bar{\tau}_p+\hat{\tau}\geq 0$, Correction: 7-20-99

$$\mathcal{E}_{\mathcal{C}_p}(\bar{\tau}_p+\hat{\tau};\beta)\geq A_{\mathcal{C}_p}(t-\hat{\tau},t+\bar{\tau}_p). \quad (20)$$

Thus, we can only have a deadline violation if

$$\exists\hat{\tau}:\left\{\sum_p \mathcal{E}_{\mathcal{C}_p}(\bar{\tau}_p+\hat{\tau};\beta)-\hat{\tau}\right\}\geq d_q\ . \quad (21)$$

With Lemma 1(b), the probability that an arrival from class $q$ experiences a deadline violation in the interval $I_\beta$ is $<\varepsilon$, if $d_q$ is selected such that

$$\sup_{\hat{\tau}}\left\{\sum_p \mathcal{H}_{\mathcal{C}_p}(\bar{\tau}_p+\hat{\tau};\beta,\varepsilon/Q)-\hat{\tau}\right\}\leq d_q\ . \quad (22)$$

Note that the nature of the statistical guarantees derived with local effective envelopes (in Subsection III-C) and with global effective envelopes (in Subsection III-D) are quite different. Local effective envelopes are (under the assumption in Eqn. (17)) concerned with the probability that a particular traffic arrival results in a deadline violation. Global effective envelopes address the probability that a deadline violation occurs for some arrival in a certain time interval. Clearly, a service which guarantees the latter is more stringent, and will lead to more conservative admission control.

Lastly, we want to point to the structural similarities of the conditions in Eqs. (15), (19), and (22). Thus, schedulability conditions which have been derived for a deterministic service can be reused, without modification, for a statistical service if effective envelopes are available.

## IV. Construction of Effective Envelopes

In this section we will construct the local and global effective envelopes $\mathcal{G}_\mathcal{C}$ and $\mathcal{H}_\mathcal{C}$ for the aggregate traffic from a set of flows as described in (A1)-(A5). Throughout this section, we will work only with flows from a single class. So, we will drop the index '$q$'; and $\mathcal{C}$ and $N$, respectively, will denote the set of flows and the number of flows. We denote by $A^*(\tau)$ the common deterministic envelope for the flows in $\mathcal{C}$, and by $A_\mathcal{C}(t,t+\tau)$ the aggregate traffic.

Our derivations proceed in the following steps:

**Step 1.** We compute bounds for the moments of the individual flows $A_j(t,t+\tau)$. Since the flows are independent, this directly leads to bounds for the moments of

$A_{\mathcal{C}}(t, t+\tau)$.[4]

**Step 2.** We use the Chernoff bound to determine a local effective envelope $\mathcal{G}_{\mathcal{C}}$ directly from our bounds on the moments.

**Step 3.** We use a geometric argument to construct $\mathcal{H}_{\mathcal{C}}$ from any local effective envelope $\mathcal{G}_{\mathcal{C}}$. Specifically, we will provide bounds of the following nature:

$$\mathcal{G}_{\mathcal{C}}(\tau;\varepsilon) \leq \mathcal{H}_{\mathcal{C}}(\tau;\beta,\varepsilon) \leq \mathcal{G}_{\mathcal{C}}(\tau';\varepsilon') . \qquad (23)$$

where $\tau'/\tau > 1$ and $\varepsilon'/\varepsilon < 1$ depend on $\beta$. We claim that for $\varepsilon$ sufficiently small and $\beta$ not too large, $\tau'/\tau \approx 1$, and resulting global effective envelope is reasonably close to the local effective envelope.

### *A. Moment bounds*

The moment generating functions of the distributions of $A_{\mathcal{C}}$ and the $A_j$ are defined as follows:

$$M_{\mathcal{C}}(s,\tau) \;:=\; E[e^{A_{\mathcal{C}}(t,t+\tau)s}] , \qquad (24)$$
$$M_j(s,\tau) \;:=\; E[e^{A_j(t,t+\tau)s}] . \qquad (25)$$

Due to the stochastic independence of the flows, we can write:

$$M_{\mathcal{C}}(s,\tau) \;=\; \prod_{j=1}^{N} M_j(s,\tau) . \qquad (26)$$

Thus, to obtain a bound on $M_{\mathcal{C}}(s,\tau)$, it is sufficient to bound the moment-generating function of a single flow $A_j(t,t+\tau)$. The following lemma provides such a bound. We refer to [1] for a proof.

*Lemma 2: Assume that $A(t,t+\tau)$ satisfies Conditions (A1), (A2), and (A3). Then,*

$$M(s,\tau) \;\leq\; 1 + \frac{\rho\tau}{A^*(\tau)}\left(e^{sA^*(\tau)} - 1\right) . \qquad (27)$$

Combining Eqn. (26) with (27) of Lemma 2 yields the bound

$$M_{\mathcal{C}}(s,\tau) \;\leq\; \left(1 + \frac{\rho\,\tau}{A^*(\tau)}\left(e^{sA^*(\tau)} - 1\right)\right)^N . \qquad (28)$$

### *B. Local Effective Envelopes*

#### *B.1 Using the Central Limit Theorem*

The bound in Eqn (28) can be strengthened to bounds for individual moments. A case of particular interest is the bound for the variance

$$\underbrace{Var\left[A_{\mathcal{C}}(t,t+\tau)\right]}_{=:\tilde{\mathbf{s}}^2} \;\leq\; N \;\underbrace{\rho\tau(A^*(\tau)-\rho\tau)}_{=:\hat{\mathbf{s}}^2} , \qquad (29)$$

where we have used the bound on the second moment together with the assumption that $E\left[A_{\mathcal{C}}(t,t+\tau)\right] = \rho\tau$.

An application of the Central Limit Theorem, will now yield a bound which is equivalent to Knightly's bound on the *rate variance* in [12].

Using first the Central Limit Theorem and then the bound on the variance in Eqn.(29), we see that for $x > \rho\tau$ Correction: 7-20-99

$$Pr\left[A_{\mathcal{C}}(t,t+\tau) \geq Nx\right]$$
$$\approx 1 - \Phi\left(\frac{Nx - N\rho\tau}{\tilde{\mathbf{s}}}\right) \qquad (30)$$
$$\leq 1 - \Phi\left(\sqrt{N}\,\frac{x-\rho\tau}{\hat{\mathbf{s}}}\right) , \qquad (31)$$

where $\Phi$ is the cumulative normal distribution. Here, $\tilde{\mathbf{s}}$ and $\hat{\mathbf{s}}$, respectively, are the square roots of the left hand and right hand sides of Eqn. (29).

To find $\mathcal{G}_{\mathcal{C}}$ so that

$$Pr[A_{\mathcal{C}}(0,\tau) \geq \mathcal{G}_{\mathcal{C}}(\tau;\varepsilon)] \;\leq\; \varepsilon , \qquad (32)$$

we set $Pr\left[A_{\mathcal{C}}(t,t+\tau) \geq Nx\right] \approx \varepsilon$ in Eqn. (31) and solve for $Nx$. This gives us an (approximate) local effective envelope as

$$\mathcal{G}_{\mathcal{C}}(\tau;\varepsilon) \;\approx\; N\rho\tau + z\sqrt{N}\rho\tau\sqrt{\frac{A^*(\tau)}{\rho\tau} - 1} , \qquad (33)$$

where $z \approx \sqrt{|\log(2\pi\varepsilon)|}$ is defined by $1 - \Phi(z) = \varepsilon$.

#### *B.2 Using the Chernoff Bound*

While the estimate in Eqn. (33) is asymptotically correct, for finite values of $N$ it is only an approximation. To obtain a rigorous upper bound on $Pr\left[A_{\mathcal{C}}(0,\tau) \geq Nx\right]$, recall the Chernoff bound for a random variable $Y$[18]:

$$Pr[Y \geq y] \leq e^{-sy}\, E[e^{sY}] \qquad \forall s \geq 0 . \qquad (34)$$

In particular, for $A_{\mathcal{C}}$, this gives

$$Pr[A_{\mathcal{C}}(0,\tau) \geq Nx] \leq e^{-Nxs} M_{\mathcal{C}}(s,\tau) \qquad (35)$$
$$\leq \left[e^{-xs}\left(1 + \frac{\rho\,\tau}{A^*(\tau)}\left(e^{sA^*(\tau)} - 1\right)\right)\right]^N \qquad (36)$$

Here, Eqn. (35) simply used the Chernoff bound, and Eqn. (36) used Eqn. (28). Since we have a choice for selecting $s$ in Eqn. (36), we want to make the bound as small as possible. For $x < A^*(\tau)$, the right hand side is minimal when $s$ is chosen so that

$$e^{sA^*(\tau)} \;=\; \frac{x}{\rho\tau}\,\frac{A^*(\tau) - \rho\tau}{A^*(\tau) - x} . \qquad (37)$$

[4] Note that the moment generating function for arrival functions $A_j$ is also computed in [2]. However, different from [2], our arrivals $A_j$ are regulated by deterministic functions $A_j^*$.

Substituting this value of $s$ into Eqn. (36) yields

$$\begin{aligned} & Pr\big[A_{\mathcal{C}}(0,\tau) \geq Nx\big] \\ & \leq \left[ \left(\frac{\rho\tau}{x}\right)^{\frac{x}{A^*(\tau)}} \left(\frac{A^*(\tau)-\rho\tau}{A^*(\tau)-x}\right)^{1-\frac{x}{A^*(\tau)}} \right]^N \end{aligned} \tag{38}$$

Again, our goal is to find $\mathcal{G}_{\mathcal{C}}$ satisfying Eqn. (32). Using the bound in Eqn. (38) and enforcing that $\mathcal{G}_{\mathcal{C}}(\tau;\varepsilon)$ is never larger than $NA^*(\tau)$ we may set

$$\mathcal{G}_{\mathcal{C}}(\tau;\varepsilon) \;=\; N\min(x, A^*(\tau))\,, \tag{39}$$

where $x$ is set to be the smallest number satisfying the inequality

$$\left(\frac{\rho\tau}{x}\right)^{\frac{x}{A^*(\tau)}} \left(\frac{A^*(\tau)-\rho\tau}{A^*(\tau)-x}\right)^{1-\frac{x}{A^*(\tau)}} \;\leq\; \varepsilon^{1/N}\,. \tag{40}$$

It can be verified that for $N$ sufficiently large, this bound matches closely the CLT bound of Eqn. (33).

*Remark:* For deterministic envelopes with a peak-rate constraint $A^*(\tau) \leq P\tau$, both expressions for $\mathcal{G}_{\mathcal{C}}$ in Eqn. (39) and Eqn. (33) describe lines, with slopes which depend on $\rho$, $P$, $N$, and $\varepsilon$. In other words, the arrivals $A_{\mathcal{C}}(t, t+\tau)$ satisfy, with probability at least $1-\varepsilon$, again a rate constraint. The new rate differs from the mean rate $N\rho$ by an error of order $\sqrt{N}$ (for fixed values of $\rho$, $P$, and $\varepsilon$).

### *C. From Local to Global Effective Envelopes*

We use the results from the previous subsection to construct a global effective envelope $\mathcal{H}_{\mathcal{C}}$ for $A_{\mathcal{C}}$. The first step is a geometric estimate for $\mathcal{E}_{\mathcal{C}}$ for a particular value of $\tau$ in terms of the local effective envelope. The second step fixes the value of the global effective envelope for a finite collection of values $\tau_i$. Finally, we obtain the entire envelope by extrapolation.

Let us define two events:

$$B(x,t,\tau) \;=\; \{A_{\mathcal{C}}(t,t+\tau) \geq Nx\}\,. \tag{41}$$

$$B_\beta(x,\tau) \;=\; \{\mathcal{E}_{\mathcal{C}}(\tau;\beta) \geq Nx\}\,. \tag{42}$$

for an arbitrary interval $I_\beta$ of length $\beta$. The event $B(x,t,\tau)$ occurs if the arrivals in the specific time interval $[t, t+\tau]$ exceed $Nx$, while $B_\beta(x,\tau)$ occurs if there is some interval of length $\tau$ in the interval $I_\beta$ where the arrivals exceed $Nx$.

With Eqn. (38), we have a bound for the probability of events $B(x,t,\tau)$. The following bound for $B_\beta(x,\tau)$ in terms of $B(x,t,\tau)$ will be used to construct $\mathcal{H}_{\mathcal{C}}(\cdot;\beta,\varepsilon)$ from $\mathcal{G}_{\mathcal{C}}(\cdot;\varepsilon)$.

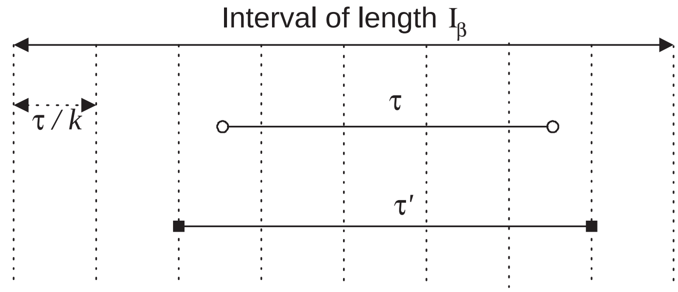



Fig. 2. Embedding Intervals.

*Lemma 3: Let $k \geq 2$ be a positive integer, $I_\beta$ an interval of length $\beta$, $t \in I_\beta$, and $0 \leq \tau \leq \beta$. Then*

$$\begin{aligned} Pr[B(x,t,\tau)] \;&\leq\; Pr[B_\beta(x,\tau)] \;\leq \\ &\leq \frac{\beta k}{\tau} Pr[B(x,t,\tau')]\,, \end{aligned} \tag{43}$$

*with $\tau'/\tau = (k+1)/k$.*

**Proof:** By stationarity, we may assume that $I_\beta = [0,\beta]$ and $t = 0$. The left inequality holds by definition, since $B(x,0,\tau) \subseteq B_\beta(x,\tau)$. To see the inequality on the right, let $t_i = i\tau/k$ ($i = 0,\ldots,\lceil \beta k/\tau \rceil$), and consider the intervals $[t_i, t_{i+k+1}]$ of length $\tau' = \frac{k+1}{k}\tau$ for $i = 1,\ldots,\lceil(\beta-\tau)k/\tau\rceil$ (all but possibly the last are subintervals of $[0,\beta]$.) See Figure 2 for an illustration of this construction. Clearly, every subinterval of length $\tau$ in $I_\beta$ is contained in at least one of the intervals of length $\tau'$. The claim now follows with stationarity. □

Lemma 3 provides a bound on arrivals in all subintervals of length $\tau$ in $I_\beta$. One of its implications is that for every value of $\tau$,

$$Pr\left[\mathcal{E}_{\mathcal{C}}(\tau;\beta) \geq \mathcal{G}_{\mathcal{C}}\big(\frac{k+1}{k}\tau;\varepsilon\big)\right] \;\leq\; \frac{\beta k}{\tau}\varepsilon\,, \tag{44}$$

where $\mathcal{E}_{\mathcal{C}}$ is the empirical envelope, and $\mathcal{G}_{\mathcal{C}}$ is any local effective envelope.

We next assign a finite number of values for $\mathcal{H}_{\mathcal{C}}(\,\cdot\,;\beta,\varepsilon)$: Pick a collection of values $\tau_i$ and $k_i$ ($i = 1,\ldots,n$) and define

$$\mathcal{H}_{\mathcal{C}}(\tau_i;\beta,\varepsilon) \;=\; \mathcal{G}_{\mathcal{C}}(\tau_i';\varepsilon')\,, \tag{45}$$

where

$$\tau_i' = \frac{k+1}{k}\tau_i \quad \text{and} \quad \varepsilon' = \varepsilon\left(\sum_{i=1}^n \frac{\beta k_i}{\tau_i}\right)^{-1}. \tag{46}$$

To justify this construction, note that by Eqn. (44) we have

$$Pr\Big[\exists i:\; \mathcal{E}_{\mathcal{C}}(\tau_i;\beta) \geq \mathcal{G}_{\mathcal{C}}(\tau_i',\varepsilon')\Big] \;\leq\; \sum_{i=1}^n \frac{\beta k_i}{\tau_i}\varepsilon' \tag{47}$$

$$\leq\; \varepsilon\,. \tag{48}$$

To get values for the global effective envelope on intervals $(\tau_{i-1}, \tau_i)$ and $[0, \tau_1)$, we first extrapolate, using the bound $A^*$ and monotonicity, and then enforce subadditivity. More precisely, we set

$$\mathcal{H}_{\mathcal{C}}(\tau;\beta,\varepsilon) \;=\; \inf_{\sum \theta_i = \tau} \sum_i f(\theta_i)\;. \tag{49}$$

where $f$ is an auxiliary function defined by

$$f(\tau) \;=\; \begin{cases} \min\{\mathcal{H}_{\mathcal{C}}(\tau_{i-1};\beta,\varepsilon) + A^*(\tau - \tau_{i-1}), \\ \quad \mathcal{H}_{\mathcal{C}}(\tau_i;\beta,\varepsilon)\} \quad \tau \in [\tau_{i-1}, \tau_i), i = 2,\ldots,n \\ \min\{A^*(\tau),\ \mathcal{H}_{\mathcal{C}}(\tau_1;\beta,\varepsilon)\} \quad \tau \in [0, \tau_1) \end{cases} \tag{50}$$

In other words, $\mathcal{H}_{\mathcal{C}}$ is the largest subadditive function which does not exceed $f$.

Since there exists no universal "best" global effective envelope, it is clearly impossible to make an optimal choice for the values of $\tau_i$ and $k_i$. It is, however, possible to make good choices, which lead to global effective envelopes that approximate the given local effective envelope well, at least when $\varepsilon$ is sufficiently small.

In our numerical results, we use

$$k_i = k\;, \quad \tau_i = \gamma^i \tau_o \quad (i = 1,\ldots,n)\;, \tag{51}$$

where $\tau_o$ is a small number, and we choose

$$\gamma = 1 + \frac{1}{k+1}\;, \quad k \approx z\left(z + \sqrt{N}\,\frac{\rho\tau}{\hat{\mathrm{s}}}\right)\;, \tag{52}$$

where $z$ is defined by $1 - \Phi(z) = \varepsilon$ and $\hat{\mathrm{s}}$ by Eqn. (29). The choice of the $\tau_i$ in Eqn. (51) guarantees that

$$\mathcal{H}_{\mathcal{C}}(\tau;\beta,\varepsilon) \;\le\; \mathcal{G}_{\mathcal{C}}\big(\frac{k+1}{k}\,\gamma\,\tau, \varepsilon'\big)\;, \tag{53}$$

for all $\tau \in [\tau_o, \beta]$, where, by Eqn. (46),

$$\varepsilon' \;=\; \frac{\tau_o(\gamma-1)}{\beta k \gamma} \cdot \varepsilon\;. \tag{54}$$

In [1] we provide a justification for the choice of $k$ and $\gamma$ in Eqn. (52) for peak-rate constrained traffic with large burst sizes. This is done with a heuristic optimization which applies the CLT approximation from Eqn. (33).

## V. Evaluation

In this section, we evaluate the effective envelope approach, using the schedulability conditions from Section III and the bounds derived in Section IV. The key criteria for evaluation is the amount of traffic on a link which can be provisioned with QoS guarantees.

As benchmarks for statistical QoS provisioning we consider the following non-statistical methods:

- **Peak Rate:** Peak rate allocation, provides deterministic QoS guarantees, but, is an inefficient method for achieving QoS.
- **Deterministic:** We use admission control tests for deterministic QoS from Eqn. (15). The admissible traffic varies with the scheduling algorithm.
- **Average Rate:** Average rate allocation only guarantees finiteness of delays and average throughput.

We will evaluate the two methods for provisioning statistical QoS which are presented in this paper.

- **Local Effective Envelope:** Here we use Eqn. (18) to determine admissibility. We will evaluate the quality of the following two bounds, derived in Section IV:
  - **Local Effective Envelope (CB)**: Uses the bound from Eqn. (40), obtained with the Chernoff bound.
  - **Local Effective Envelope (CLT)**: Uses the bound from Eqn. (33), obtained with the Central Limit Theorem. Recall from our discussion in Section IV that the *local effective envelope (CLT)* results are equivalent to the rate-variance envelope method described in [12].
- **Global Effective Envelope:** We use Eqn. (22) to determine admissibility. The global effective envelope is constructed by first finding $\beta$ (see Footnote 1), and choosing a number $\tau_o$ which is small compared to the delay bounds. We determine the parameters $\gamma$, $k$, and $\tau_i$ according to (51) and (52). We then apply Eqn. (45) for each of the $\tau_i$, and complete the process by the extrapolation in Eqs. (49) and (50). (In Eqn. (45), we use the local effective envelope (CB) rather than the corresponding CLT bound, since the latter would yield only approximate bounds.)

We compare our results with the effective bandwidth approach for regulated adversarial traffic from the literature:

- **Effective Bandwidth [8], [16], [19]**:[5] The effective bandwidth approach assigns to each flow a fixed capacity, the *effective bandwidth*, and assumes that each flow is serviced at a rate which corresponds to the effective bandwidth.

The delay bounds will be indirectly derived from the buffer size. We set the delay bound $d$ to $d = B/C$, where $B$ is the buffer size at the scheduler and $C$ is the transmission rate of the link.

In our examples, we include the following results on effective bandwidth:

- **EB-EMW**: This is the result from the classical paper by Elwalid/Mitra/Wentworth (Eqn. (39) in [8]).
- **EB-RRR**: We use Eqn. (9) from [19] by Rajagopal/Reisslein/Ross which presents an improvement to the EB-EMW result.

[5] The cited works calculate effective bandwidth for regulated adversarial sources. The complete literature on effective bandwidth is much more extensive.

In all our experiments, we consider traffic regulators which are obtained from peak rate controlled leaky buckets with deterministic envelopes as given in Eqn. (3).[6] In all experiments, we consider a link with $C$ = 45 Mbps, and we consider two traffic classes. The traffic parameters of a flow in one of the classes are as follows:

| Class | Peak Rate $P$ (Mbps) | Mean Rate $\rho$ (Mbps) | Burst Size $\sigma$ (bits) |
|---|---|---|---|
| 1 | 1.5 | 0.15 | 95400 |
| 2 | 6.0 | 0.15 | 10345 |

To parameters are selected so as to match, at least approximately, the examples presented in [8], [19].

We will present three sets of examples. In the first example, we compare the deterministic envelopes with our bounds for the local and global effective envelopes for sets of homogeneous sources. In the second example, we compare the maximum number of admissible flows in a FIFO scheduler for a given delay bound $d$ and delay-violation probability $\varepsilon$. In the third example, we investigate the case of heterogeneous traffic with different QoS requirements, and we compare the admissible regions for different scheduling algorithms (SP, EDF).

### *A. Example 1: Comparison of Envelope Functions*

In the first example, we study the shape of local and global effective envelopes for homogeneous sets of flows, as functions of the lengths of time intervals. The envelopes are compared to the deterministic envelope $A_j^*(\tau) = \min\{P_j\tau, \sigma_j + \rho_j, \tau\}$, to the peak rate function $P_j\,\tau$, and to the average rate function $\rho_j\,\tau$. In our graphs, we plot the amount of traffic per flow for the various envelopes (e.g., we present $\sum_{j\in\mathcal{C}} \mathcal{G}_j(\tau;\varepsilon)/N$).

Figures 3(a) and 3(b) show the results for multiplexed flows from Class 1 and Class 2, respectively. We set $\varepsilon$ = $10^{-6}$ for all envelopes. By depicting the amount of traffic per flow for different numbers of flows ($N$ denotes the number of flows), we can observe how the statistical multiplexing gain increases with the number of flows.

The first observation to be made is that the local and global effective envelopes are much smaller than the deterministic envelope or the peak rate. Another observation is that, for a fixed number of flows $N$, the global effective envelope is larger than the local effective envelopes, and the local effective envelope bound is smaller when using CLT (central limit theorem), as compared to CB (Chernoff bound). Note, however, that bounds for the envelopes with CLT are only approximate, and may be too optimistic, especially for small number of flows. Figure 3 also shows that local and global effective envelopes converge as the number of flows $N$ is increased.

[6] Most of the methods listed here can work with more complex regulators. However, since peak-rate enforced leaky buckets are widely used in practice, they serve as good benchmarks.

### *B. Example 2: Admissible Region for Homogeneous Flows*

In this example, we investigate the number of flows admitted by various admission control methods for guaranteeing QoS at a link with a FIFO scheduler. We assume that flows are homogeneous, that is, all flows belong to a single class. Again, the probability of a violation of QoS guarantees is set to $\varepsilon = 10^{-6}$.

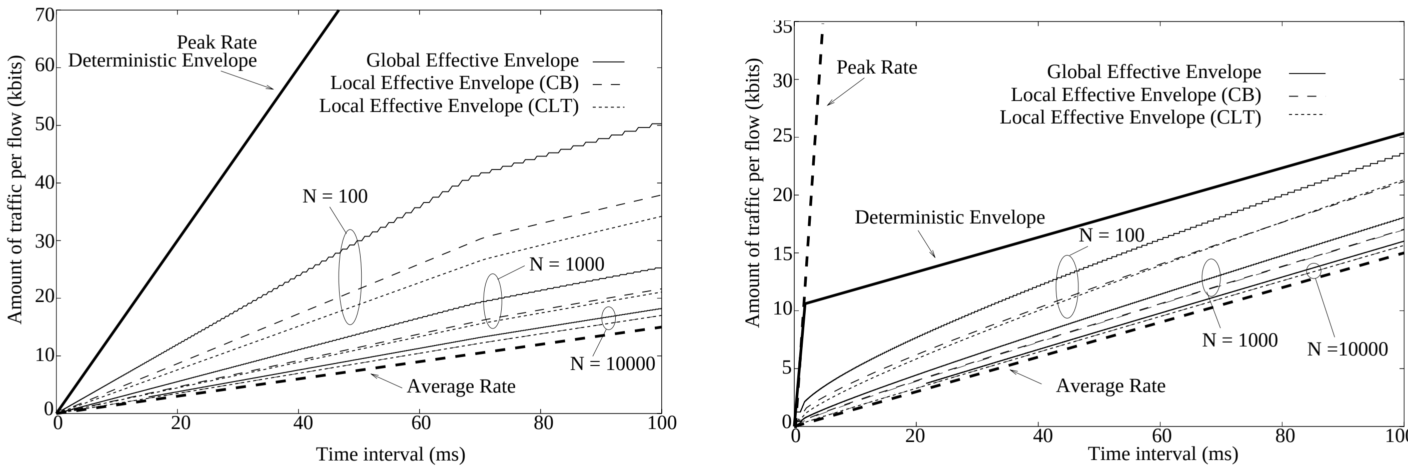
We compare the admissible regions of the local and global effective envelopes, to those of the effective bandwidth techniques (both EB-EMW and EB-RRR), and to a deterministic QoS guarantees.

We compare these results with those obtained from a discrete event simulation. For the simulation, we assume that the arrivals from a source have a pattern which is said to have adversarial patterns for peak-rate controlled leaky buckets [19]. If the parameters of a flow are given by $(P, \rho, \sigma)$, the adversarial pattern transmits at the peak rate $P$ for a duration $\sigma/(P-\rho)$, and then continues sending traffic at rate $\rho$ for a duration $dC/((C-\rho)\cdot 2)$. Then, the source shuts off, waits for a duration $\sigma/\rho$ and then repeats the pattern. The starting time of a pattern of the flows are uniformly and independently chosen over the length of its period. We refer to [1] for a detailed discussion of the simulations.

Figures 4(a) and (b) depict the number of admitted flows as a function of the delay bound. The figures show that all methods for statistical QoS admit many more connections than a deterministic admission control test. In both Figures, the effective envelopes (both CLT and CB) are closest to the simulation results. (Once again, we point out that the results using the local effective (CLT) bounds are identical to the rate-variance results presented in [12].) Note, however, that results obtained with local effective envelopes are approximate and are not guaranteed to be upper bounds on the admissible regions.

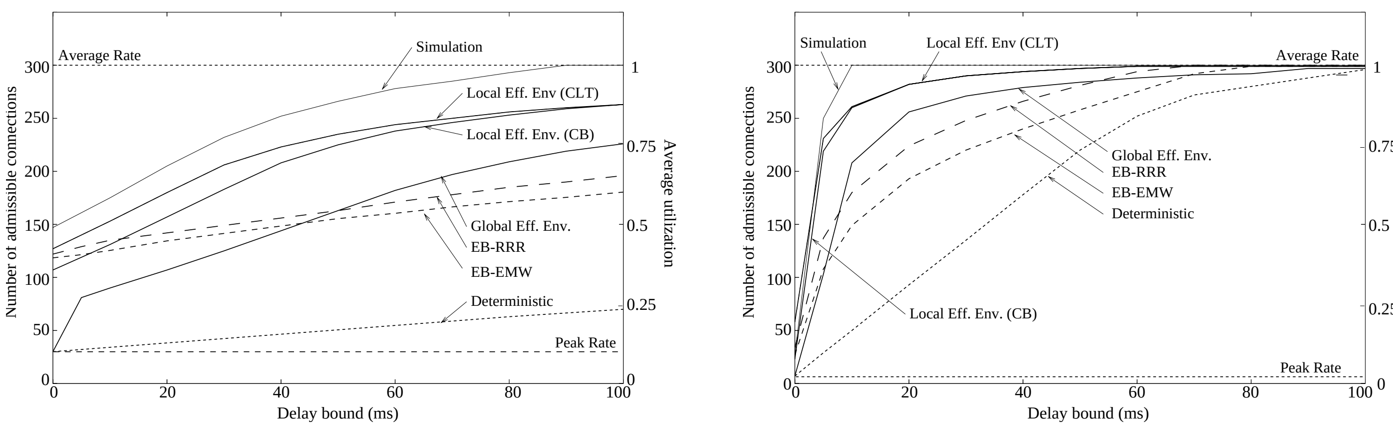
Comparing the results from effective envelopes to the effective bandwidth results, we observe that the effective envelope methods admits more connections than the effective bandwidth methods if delay bounds are large.

The difference of the admissible regions in Figure 4(a) to those in Figure 4(b) illustrate the high degree to which the size of the admissible region is dependent on the traffic parameters. The lower burst sizes of flows in Class 2 lead to larger admissible regions for all methods. Specifically, notice that deterministic QoS in Figure 4(b) yields similar




(a) Class 1. (b) Class 2.

Fig. 3. Example 1: Comparison of Envelope Functions for $\tau \leq 100\ ms$, $\varepsilon = 10^{-6}$, and for Number of Flows $N = 100, 1000, 10000$.




(a) Class 1. (b) Class 2.

Fig. 4. Example 2: Admissible Number of Connections at a FIFO Scheduler for Homogeneous Flows as a Function of Delay Bounds ($\varepsilon = 10^{-6}$, $0 < d \leq 100$ ms).

results to the statistical methods, if the delay bounds are large.

### *C. Example 3: Admissible Region for Heterogeneous Traffic*

Here we investigate an example with different scheduling algorithms and with heterogeneous traffic arrivals.

As scheduling algorithms, we consider Static Priority (SP) and Earliest-Deadline-First (EDF). For a deterministic service, EDF is optimal, in the sense that the admissible regions with EDF scheduling is maximal [15]. To our knowledge, results for a statistical service (with adversarial traffic), have not been reported for EDF.

In this example, we multiplex a number of flows from Class 1 and from Class 2 on 45 Mbps. We fix the delay bounds, such that the delay bound for Class-1 flows is relatively long, $d_1 = 100\ ms$, and the delay bound for Class-2 flows is relatively short, $d_2 = 10\ ms$. For any particular method, we determine the maximum number of Class-1 and Class-2 flows that can be supported simultaneously on the 45 Mbps link.

The result are shown in Figure 5. The plot depicts the admissible region for SP and EDF scheduling, using the results for the (two types of) local effective envelopes, effective envelopes, and deterministic envelopes. We also include the admissible regions for the effective bandwidth approaches (EB-EMW and EB-RRR). Note, however, that the shown effective bandwidth methods assume a simple multiplexer (with virtual buffer partitioning) and do not account for different scheduling algorithms.

The results in Figure 5 show that the difference between SP and EDF schedulers is small in all cases. The effective envelope is, again, more conservative than the local effective envelope method. Finally, Figure 5 illustrates that with heterogeneous flows and the effective bandwidth methods (EB-EMW, EB-RRR) may not perform as well as methods which consider scheduling algorithms.

We also performed a simulation for the EDF scheduling

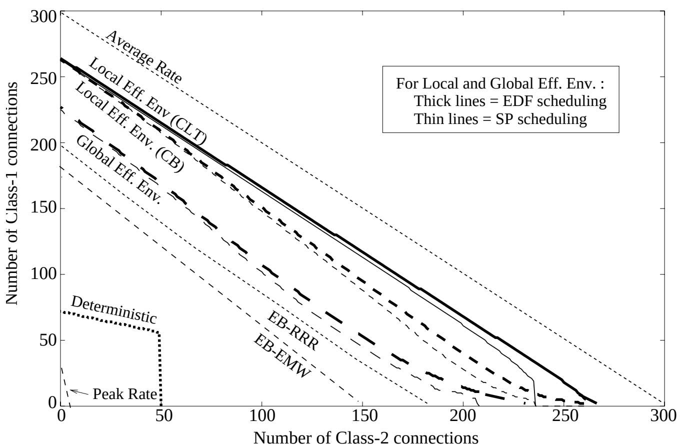



Fig. 5. Example 3: Admissible Region of Multiplexing Class 1 and Class 2 Flows with $\varepsilon = 10^{-6}$ and $d_1 = 100$ ms and $d_2 = 10$ ms.

algorithm. For the simulations, we used a source model which was shown to be adversarial for a simple multiplexer with buffer and bandwidth partitioning [19]. We do not know or claim that this source model is also adversarial for EDF scheduling. However, with this choice, the simulations give the same results as an average rate allocation.

## VI. CONCLUSIONS

We have presented new results on evaluating the statistical multiplexing gain for packet scheduling algorithms. A useful property of our approach is that it separates the consideration of the service definition (deterministic, statistical), the scheduling algorithm (FIFO, SP, EDF), and the mathematical methodology (Central Limit Theorem, Chernoff Bound). Thus, our work may be useful to researchers who want to determine the statistical multiplexing gain for other traffic regulators, scheduling algorithms, or large deviation results. As direction for future work, the admission control methodology presented in this paper needs to be extended to a network environment.

## REFERENCES


[1] R. Boorstyn, A. Burchard, J. Liebeherr, and C. Oottamakorn. Statistical multiplexing gain of link scheduling algorithms in QoS networks. Technical Report CS-99-21, University of Virginia, Computer Science Department, June 1999.

[2] C. Chang. Stability, queue length, and delay of deterministic and stochastic queueing networks. *IEEE Transactions on Automatic Control*, 39(5):913–931, May 1994.

[3] J. Choe and Ness B. Shroff. A central-limit-theorem-based approach for analyizing queue behvior in high-speed network. *IEEE/ACM Transactions on Networking*, 6(5):659–671, October 1998.

[4] R. Cruz. A calculus for network delay, Part I : Network elements in isolation. *IEEE Transaction of Information Theory*, 37(1):114–121, 1991.

[5] B. T. Doshi. Deterministic rule based traffic descriptors for broadband ISDN: Worst case behavior and connection acceptance control. In *International Teletraffic Congress (ITC)*, pages 591–600, 1994.

[6] N. G. Duffield and S. H. Low. The cost of quality in networks of aggregate traffic. In *Proceedings of IEEE INFOCOM'98*, San Francisco, March 1998.

[7] A. Elwalid and D. Mitra. Design of generalized processor sharing schedulers which statistically multiplex heterogeneous QoS classes. In *Proceedings of IEEE INFOCOM'99*, pages 1220–1230, New York, March 1999.

[8] 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*, 13(6):1115–1127, August 1995.

[9] G. Kesidis and T. Konstantopoulos. Extremal shape-controlled traffic patterns in high-speed networks. Technical Report 97-14, ECE Technical Report, University of Waterloo, December 1997.

[10] G. Kesidis and T. Konstantopoulos. Extremal traffic and worst-case performance for queues with shaped arrivals. In *Proceedings of Workshop on Analysis and Simulation of Communication Networks*, Toronto, November 1998.

[11] E. Knightly. H-BIND: A new approach to providing statistical performance guarantees to VBR traffic. In *Proceedings of IEEE INFOCOM'96*, pages 1091–1099, San Francisco, CA, March 1996.

[12] E. Knightly. Enforceable quality of service guarantees for bursty traffic streams. In *Proceedings of IEEE INFOCOM'98*, pages 635–642, San Francisco, March 1998.

[13] E. W. Knightly and Ness B. Shroff. Admission control for statistical QoS: Theory and practice. *IEEE Network*, 13(2):20–29, March/April 1999.

[14] J. Kurose. On computing per-session performance bounds in high-speed multi-hop computer networks. In *ACM Sigmetrics'92*, pages 128–139, 1992.

[15] J. Liebeherr, D. Wrege, and D. Ferrari. Exact admission control for networks with bounded delay services. *IEEE/ACM Transactions on Networking*, 4(6):885–901, December 1996.

[16] F. LoPresti, Z. Zhang, D. Towsley, and J. Kurose. Source time scale and optimal buffer/bandwidth tradeoff for regulated traffic in an ATM node. In *Proceedings of IEEE INFOCOM'97*, pages 676–683, Kobe, Japan, April 1997.

[17] P. Oechslin. Statistical multiplexing gain of link scheduling algorithms in QoS networks. Technical report, University College London, September 1999.

[18] A. Papoulis. *Probability, Random Variables, and Stochastic Processes. 3rd edition.* McGraw Hill, 1991.

[19] S. Rajagopal, M. Reisslein, and K. W. Ross. Packet multiplexers with adversarial regulated traffic. In *Proceedings of IEEE INFOCOM'98*, pages 347–355, San Francisco, March 1998.

[20] M. Reisslein, K. W. Ross, and S. Rajagopal. Guaranteeing statistical QoS to regulated traffic: The multiple node case. In *Proceedings of 37th IEEE Conference on Decision and Control (CDC)*, pages 531–531, Tampa, December 1998.

[21] M. Reisslein, K. W. Ross, and S. Rajagopal. Guaranteeing statistical QoS to regulated traffic: The single node case. In *Proceedings of IEEE INFOCOM'99*, pages 1061–1062, New York, March 1999.

[22] D. Wrege, E. Knightly, H. Zhang, and J. Liebeherr. Deterministic delay bounds for VBR video in packet-switching networks: Fundamental limits and practical tradeoffs. *IEEE/ACM Transactions on Networking*, 4(3):352–362, June 1996.

[23] H. Zhang and D. Ferrari. Rate-controlled service disciplines. *Journal of High Speed Networks*, 3(4):389–412, 1994.