Statistical Service Assurances for Traffic Scheduling Algorithms

Robert Boorstyn, Almut Burchard, Jörg Liebeherr, Chaiwat Oottamakorn

Abstract— **Network services for the most demanding advanced networked applications which require absolute, per-flow service assurances can be deterministic or statistical. By exploiting statistical properties of traffic, statistical assurances can extract more capacity from a network than deterministic assurances. In this work we consider statistical service assurances for traffic scheduling algorithms. We present functions, so-called effective envelopes, which are, with high certainty, upper bounds of multiplexed traffic. Effective envelopes can be used to obtain bounds on the amount of traffic on a link that can be provisioned with statistical service assurances. We show that our bounds can be applied to a variety of packet scheduling algorithms. In fact, one can reuse existing admission control functions for scheduling algorithms with deterministic assurances. We present numerical examples which compare the number of flows with statistical assurances that can be admitted with our effective envelope approach to those achieved with existing methods.**

Key Words: Internet, Traffic control (communication), Packet switching, Statistical Multiplexing, Statistical Service, Scheduling, Quality-of-Service.

I. Introduction

The most demanding networked applications require absolute, per-flow service assurances. Such assurances are either deterministic or statistical. A *deterministic service* [16] guarantees that all packets from a flow satisfy given worst-case end-to-end delay bounds and no packets are dropped in the network [6], [9]. A deterministic service provides the highest level of QoS assurance, however, it leaves a significant portion of network resources on the average unused [37]. A *statistical service* [16] makes probabilistic service assurances, for example, of the form:

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

where ε is generally small, e.g., $\varepsilon = 10^{-6}$. By allowing a fraction of traffic to violate its QoS, one can significantly increase the achievable link utilization. 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.

R. Boorstyn and C. Oottamakorn are with the Department of Electrical Engineering, Polytechnic University, Brooklyn. Email: boorstyn@catt.poly.edu,cootta01@utopia.poly.edu. A. Burchard is with the Department of Mathematics, University of Virginia, Charlottesville. Email: burchard@virginia.edu. J. Liebeherr is with the Department of Computer Science, University of Virginia, Charlottesville. Email: jorg@cs.virginia.edu. 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). A short version of this paper has appeared in the Proceedings of IEEE Infocom 2000.

Since it is often not feasible to obtain a reliable statistical characterization of traffic sources, recent research on statistical QoS has attempted to define statistical services without making assumptions on properties of traffic sources. Starting with the seminal work in [15], 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 [13], [14], [15], [18], [19], [21], [26], [27], [31], [32], [33].

In this paper we attempt to provide new insights into the problem of determining the multiplexing gain of statistically independent, regulated, but otherwise adversarial 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 [6], [9]. 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., [9], [25], [40]. This is encouraging since only few results are available on statistical multiplexing of adversarial traffic which can consider non-trivial scheduling algorithms, e.g., [14], [21].

Let us make a few remarks on the scope of this paper. The scheduling algorithms considered here are First-Come-First-Served (FCFS), Static Priority (SP), and Earliest-Deadline-First (EDF) (see Footnotes 2 and 3). The first two scheduling algorithms were selected because of their (relative) simplicity, and because previous work exists on these algorithms to which our results can be compared to. The EDF algorithm was selected since it is known to be optimal for a deterministic service, in the sense that it can provide, among all scheduling algorithms, the highest level of delay assurances [17], [25]. Even though optimality of EDF scheduling may not hold for statistical QoS assurances, the results for EDF in a statistical environment may still serve as a benchmark for other scheduling algorithms. The results in this paper can also be applied to rate-based scheduling algorithms [39]. If such algorithms are rate-proportional, the derivations of statistical bounds with effective envelopes are quite simple [24].

The discussions in this paper only consider a single node. Since, in a network with multiple nodes, traffic from different flows may become correlated, the assumption of statistical independence of flows may not hold. In [24], a multi-node analysis is presented which uses the effective envelope approach in networks, where additional mechanisms, e.g., delay jitter control, restore the statistical independence of flows at downstream nodes.

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

II. Related Work

The literature on statistical services and statistical multiplexing in Quality of Service networks is extensive and a full discussion is beyond the scope of this paper. Excellent reviews of the state of the art of statistical multiplexing can be found in [22], [34], [35]. Here, we discuss only two groups of prior work on statistical multiplexing which we regard as particularly relevant to this paper. The first group studies the statistical multiplexing gain of statistically independent, regulated, adversarial traffic at a buffered multiplexer with fluid flow service. The second group extends deterministic QoS results to a probabilistic framework.

A. Regulated, Adversarial Traffic at a Multiplexer

Several researchers have studied the multiplexing gain of statistically independent, regulated, adversarial traffic at a buffered multiplexer, where each flow is allocated a fixed amount of link bandwidth and buffer capacity. In these studies, it is assumed that traffic is served in a fluid flow fashion, without consideration of scheduling at the multiplexer. The allocated rate, sometimes called *effective bandwidth*, and the buffer capacity for a flow are selected such that the probability of losses due to buffer overflows is smaller than some small number ε.

Elwalid, Mitra, and Wentworth [15] consider a multiplexer which sees arrivals from regulated peak-rate constrained leaky buckets. The adversarial traffic pattern used in [15] is a periodic on-off source, which is known to maximize the overflow probability in a bufferless multiplexer. In [14], the solution approach of [15] is applied to the GPS [29] scheduling algorithm. The work of [15] has been extended by LoPresti, Zhang, Towsley, and Kurose [26] and by Rajagopal, Reisslein, and Ross [31].

The question of the adversarial traffic pattern at a buffered multiplexer with (P, σ, ρ) regulators has received much interest. Even though specific definitions of an adversarial traffic pattern may vary, in general, it is a feasible arrival scenario which maximizes QoS violations. As suggested in [13] and by others, and supported by numerical data presented in [27], on-off traffic sources are adversarial for bufferless multiplexers, but not for buffered multiplexers. Kesidis and Konstantopoulos [18], [19] address the problem of finding explicit expressions for the adversarial traffic patterns at a buffered multiplexer, and show that the adversarial traffic pattern is periodic, with multiple on-phases and different rates in each 'on' phase.

Reisslein, Ross, and Rajagopal [32], [33] analyze the multiplexing gain of regulated adversarial traffic for a particular switch architecture, called *bufferless multiplexer*, where arriving traffic on a flow is shaped at a dedicated buffer for this flow. The output rate of each buffer is set to a fixed rate, such that no traffic experiences a violation of its delay bound in the buffer. The output from the per-flow buffers is multiplexed at a bufferless multiplexer.

Our work has similar goals as the studies cited above, in that we investigate statistical multiplexing for statistically independent, regulated, and adversarial traffic. On the other hand, our approach deviates from the above papers in several ways. We do not use a particular adversarial traffic pattern in our analysis. In fact, for the scheduling algorithms and for the general class of traffic regulators with subadditive deterministic envelope functions considered in this paper, an explicit derivation of an adversarial pattern may be a formidable task. Furthermore, we do not assume a fluid flow service, and, instead, explicitly consider scheduling at the multiplexer.

B. Probabilistic Extensions of Deterministic Calculus

Inspired by Cruz's deterministic network calculus [9], [10], [11], several researchers have made probabilistic extensions to deterministic service models.

Chang [6] derives probabilistic bounds for the delay and loss in a multiplexer with a shared buffer, where each flow is served in a fluid-flow fashion at an allocated rate. A key difference to our work is an assumption that there exists an *a priori* bound on the moment generating function of the arrivals.

Yaron and Sidi [38] prove that if the incoming flows to a multiplexer satisfy exponentially bounded burstiness constraints, then the output of the multiplexer has exponentially bounded burstiness. The work was recently extended in [36] to more general bounding functions.

Cruz [12] computes probabilistic bounds on the delay and backlog at a scheduler, assuming that probabilistic bounds on traffic arrivals and service curves [11] are available. Qiu and Knightly [30] extend this approach and develop a framework for statistical service envelopes.

Andrews [2] provides probabilistic bounds for delay violations at an EDF scheduler, assuming that the arrival distribution of traffic is known. Specific results are derived for on-off traffic as assumed in [15].

Kurose [23] explores bounds on the distribution of the delay and buffer occupancy of a flow in a network environment by characterizing the traffic on a flow by a family of random variables, which describe the traffic over time intervals. For FIFO scheduling, Kurose also provides bounds on the output of a node, and, thus, can obtain bounds for networks with multiple nodes. Zhang and Knightly [41] extend this work for specific arrival distribution and

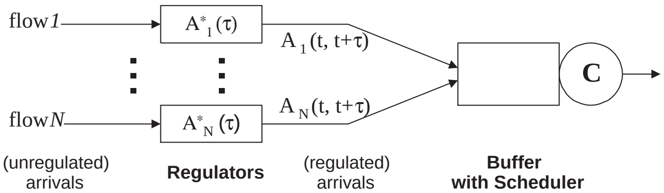

Fig. 1. Regulators and scheduler at a link.

scheduling algorithms. Both [23] and [41] calculate the arrival distribution of aggregate sources directly, without resorting to large deviations results.

Knightly [20], [21] characterizes flow arrivals using first and second moment information on the sources. The notion of a *rate-variance envelope* is introduced to describe the variance of the arrivals of a flow over a time interval. Appealing to the Central Limit Theorem (CLT), a bound for the probability of a delay bound violation is derived for an SP scheduler. In Subsection IV-B.1, we relate the rate-variance envelope from [21] to our framework.

Our work can be viewed as extending the approach pursued in [20], [21], [23]. The most significant generalization of our work is the presentation of a formal framework which allows us to consider different scheduling algorithms, traffic characterizations, and probabilistic bounds. Moreover, our approach enables us to derive schedulability conditions for deterministic and statistical QoS assurances in a uniform fashion.

III. 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. Throughout this paper, we view traffic as continuous-time fluid-flow traffic.

QoS assurances 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 the waiting time in the buffer and the transmission time.

A. Traffic Arrivals

Traffic arrivals to the link come from a set of flows which is partitioned into Q classes. We use $\mathcal{C}_q$ to denote the set of flows in class q and N_q to denote the number of flows in class q.

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 nonnegative random variables $A_j(t_1, t_2)$ 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.** Each A_j is regulated by a deterministic subadditive envelope A^*_j such that $A_j(t, t+\tau) \leq A^*_j(\tau)$ for all $\tau \geq 0, t \geq 0$.

(A3) **Stationarity.** The A_j are stationary random variables, i.e., $Pr[A_j(t, t+\tau) \leq x] = Pr[A_j(t', t'+\tau) \leq x]$ for all $t \geq 0, t' \geq 0$.

(A4) **Independence.** Arrivals 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 point out that the above assumptions on the shape of the traffic envelopes and on the randomness of flows are quite general. Note that we do not require ergodicity.
- The traffic regulators most commonly used in practice are *leaky buckets* with a peak rate enforcer [1], [4]. Here, traffic on flow j is characterized by three parameters (P_j, σ_j, ρ_j) with a deterministic envelope given by

$$A^*_j(\tau) = \min\{P_j\tau, \sigma_j + \rho_j\tau\} \qquad \forall \tau \geq 0 \ , \tag{1}$$

where $P_j \geq \rho_j$ is the peak traffic rate, ρ_j is the average traffic rate, and σ_j is a burst size parameter. We will use this type of regulators in our numerical examples in Section VI.
- 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 assume without loss of generality that this bound is saturated, that is, that for all t,

$$\lim_{\tau\to\infty} \frac{A_j(t, t+\tau)}{\tau} \; = \; \rho_j \ . \tag{2}$$

- Stationarity has the useful consequence that expected values can be computed as long-time averages. For example, for any function F,

$$E\left[F(A_j(t, t+\tau))\right] \; = \; \lim_{T\to\infty} E\left[\frac{1}{T}\int_0^T F(A_j(s, s+\tau))\, ds\right] \ . \tag{3}$$

Similar relations hold for the joint distributions of several random variables.

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 are 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 $\mathcal{G}_{\mathcal{C}}$ that satisfies for all $\tau \geq 0$ and all t

$$Pr\Big[A_{\mathcal{C}}(t, t+\tau) \leq \mathcal{G}_{\mathcal{C}}(\tau; \varepsilon)\Big] \geq 1 - \varepsilon \ . \tag{4}$$

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 τ.

Under the stationarity assumption (A3), Eqn. (4) 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 need not be subadditive in τ, but satisfy the weaker property $\mathcal{G}_\mathcal{C}(\tau_1+\tau_2, \varepsilon_1+\varepsilon_2) \leq \mathcal{G}_\mathcal{C}(\tau_1,\varepsilon_1) + \mathcal{G}_\mathcal{C}(\tau_2,\varepsilon_2)$.

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 τ. 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 [6], [37]. Consider a time interval I_β of length β. The **empirical envelope** $\mathcal{E}_\mathcal{C}(\ .\ ;\beta)$ of a collection $\mathcal{C}$ of flows is the maximum traffic in any time subinterval of I_β of length $\tau \leq \beta$ as follows:

$$\mathcal{E}_\mathcal{C}(\tau;\beta) = \sup_{[t,t+\tau)\subseteq I_\beta} A_\mathcal{C}(t,t+\tau)\ . \tag{5}$$

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

$$Pr\Big[\mathcal{E}_\mathcal{C}(\tau;\beta) \leq \mathcal{H}_\mathcal{C}(\tau;\beta,\varepsilon),\ \ \forall 0\leq\tau\leq\beta\Big] \geq 1-\varepsilon\ . \tag{6}$$

The attribute 'global' is justified since $\mathcal{H}_\mathcal{C}(\tau;\beta,\varepsilon)$ is a bound for traffic for all intervals of length $\tau \leq \beta$ in I_β. Note that we can always choose $\mathcal{E}_\mathcal{C}(\tau;\beta) \leq \sum_{j\in\mathcal{C}} \mathcal{E}_j(\tau)$ for all $\tau \leq \beta$, where $\mathcal{E}_j(\tau;\beta) = \sup_{[t,t+\tau)\subseteq I_\beta} A_j(t,t+\tau)$ is the empirical envelope of a single flow j.

Due to stationarity of the A_j, Eqn. (6) holds for all intervals of length β, if it holds for one specific interval I_β. When applied to scheduling, we will select β to exceed the longest busy period.[1]

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

Lemma 1: Given a set of flows that is partitioned into Q classes $\mathcal{C}_q$, with aggregate arrival functions $A_{\mathcal{C}_q}$. Let $\mathcal{G}_{\mathcal{C}_q}(.;\varepsilon)$ and $\mathcal{H}_{\mathcal{C}_q}(.;\beta,\varepsilon)$ be local and global effective envelopes for class q. Then the following inequalities hold.

(a) If $\sum_q \mathcal{G}_{\mathcal{C}_q}(\tau,\varepsilon) \leq x$, then, for all t,

$Pr\left[\sum_q A_{\mathcal{C}_q}(t,t+\tau) > x\right] < Q\varepsilon.$

(b) If $\sum_q \mathcal{H}_{\mathcal{C}_q}(\tau,\beta;\varepsilon) \leq x(\tau)$ for some function $x(\dot{})$ then, for all t,

$Pr\left[\exists q\, \exists\tau :\ \sum_q \mathcal{E}_{\mathcal{C}_q}(\tau,\beta) > x(\tau)\right] < Q\varepsilon.$

Proof: We only prove part (a) of the lemma. The proof for (b) is almost identical as for (a).

Fix t and τ. Assume $\sum_q \mathcal{G}_{\mathcal{C}_q}(\tau,\varepsilon) \leq x$. If the event that $\sum_q A_{\mathcal{C}_q}(t,t+\tau) > x$ occurs, then $\sum_q A_{\mathcal{C}_q}(t,t+\tau) > x \geq \sum_q \mathcal{G}_{\mathcal{C}_q}(\tau,\varepsilon)$. This implies that there exists a q such that $A_{\mathcal{C}_q}(t,t+\tau) > \mathcal{G}_{\mathcal{C}_q}(\tau,\varepsilon)$. Using the defining property of the local effective envelope, we can bound the probability that this happens by

$$\begin{aligned} Pr\left[\exists q : A_{\mathcal{C}_q}(t,t+\tau) > \mathcal{G}_{\mathcal{C}_q}(\tau,\varepsilon)\right] &\leq \\ \leq \sum_q Pr\left[A_{\mathcal{C}_q}(t,t+\tau) > \mathcal{G}_{\mathcal{C}_q}(\tau,\varepsilon)\right] &< Q\varepsilon\ . \qquad \square \end{aligned} \tag{7}$$

Our derivations in the next section will make it clear that for ε small enough, both $\mathcal{G}_{\mathcal{C}_q}(\tau;\varepsilon)$ and $\mathcal{H}_{\mathcal{C}_q}(\tau;\beta,\varepsilon)$ are not very sensitive with respect to ε, so that the bounds for ε and $Q\varepsilon$ are comparable.

IV. Construction of Effective Envelopes

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

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:

$$\mathcal{G}_\mathcal{C}(\tau;\varepsilon) \leq \mathcal{H}_\mathcal{C}(\tau;\beta,\varepsilon) \leq \mathcal{G}_\mathcal{C}(\tau';\varepsilon')\ . \tag{8}$$

where $\tau'/\tau > 1$ and $\varepsilon'/\varepsilon < 1$ depend on β. We claim that for ε sufficiently small and β not too large, $\tau'/\tau \approx 1$, the 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:

$$M_\mathcal{C}(s,\tau) \ := \ E[e^{A_\mathcal{C}(t,t+\tau)s}]\ , \tag{9}$$

$$M_j(s,\tau) \ := \ E[e^{A_j(t,t+\tau)s}]\ . \tag{10}$$

Stationarity (A3) guarantees that the moment generating functions do not depend on t. Due to the stochastic independence (A4) and homogeneity (A5) of the flows, we can write:

$$M_\mathcal{C}(s,\tau) \ = \ \prod_{j=1}^N M_j(s,\tau) \ = \ \Big(M_j(s,\tau)\Big)^N\ . \tag{11}$$

Thus, to obtain a bound on $M_\mathcal{C}(s,\tau)$, it is sufficient to bound the moment generating function of a single flow $A_j(t,t+\tau)$.

[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: $B_{max} = \inf\{\tau > 0 \mid \sum_{j\in\mathcal{C}} A^*_j(\tau) \leq \tau\}$.

The k-th moments of $A_j(t, t+\tau)$ and $A_\mathcal{C}(t, t+\tau)$ are defined by

$$m_\mathcal{C}^{(k)}(\tau) \quad := E[(A_\mathcal{C}(t,t+\tau))^k] \ , \qquad (12)$$
$$m_j^{(k)}(\tau) \quad := E[(A_j(t,t+\tau))^k] \ . \qquad (13)$$

The moments are related with the moment generating functions by

$$M_\mathcal{C}(s,\tau) \ = \ \sum_{k=0}^{\infty} m_\mathcal{C}^{(k)}(\tau)\frac{s^k}{k!} \ , \qquad (14)$$
$$M_j(s,\tau) \ = \ \sum_{k=0}^{\infty} m_j^{(k)}(\tau)\frac{s^k}{k!} \ . \qquad (15)$$

The following lemma will be used to provide bounds on the moment generation function and the moments of the arrivals on a flow $A_j(t, t+\tau)$. A proof of the lemma is given in Appendix I.

Lemma 2: Assume that $A_j(t, t+\tau)$ satisfies assumptions (A1), (A2), and (A3). Then, for every convex increasing function F,

$$E\left[F(A_j(t,t+\tau)\right] \ \le \\ \le \left(\frac{\rho\tau}{A^*(\tau)}\right) F(A^*(\tau)) + \left(1-\frac{\rho\tau}{A^*(\tau)}\right) F(0) \ . \ (16)$$

With Lemma 2, we can easily obtain bounds for the moment generating function $M_j(s,\tau)$ and the k-th moments $m_j^{(k)}$. These bounds are formulated in Theorems 1 and 2.

Theorem 1: Given a set of flows $\mathcal{C}$ from a single class which satisfy assumptions (A1)–(A5). Let $A_j(t, t+\tau)$ denote the arrivals from a flow $j \in \mathcal{C}$, let $A_\mathcal{C}(t, t+\tau)$ denote the aggregate traffic, and let $A^*(\tau)$ denote the subadditive envelope for each flow in $\mathcal{C}$. Then,

$$M_j(s,\tau) \le 1 + \frac{\rho\tau}{A^*(\tau)}\left(e^{sA^*(\tau)}-1\right) \ , \qquad (17)$$
$$M_\mathcal{C}(s,\tau) \ \le \ \left(1+\frac{\rho\,\tau}{A^*(\tau)}\left(e^{sA^*(\tau)}-1\right)\right)^N \ . \ (18)$$

Proof: Eqn. (17) is obtained by setting $F(y) = e^{sy}$ in Lemma 2. Combining Eqn. (17) with Eqn. (11) yields Eqn. (18). □

The bound in Eqn. (18) can be strengthened to bounds for individual moments.

Theorem 2: Under the assumptions of Theorem 1, for $k \ge 1$,

$$m_j^{(k)}(\tau) \ \le \ \rho\tau\left(A^*(\tau)\right)^{k-1} \qquad (19)$$
$$m_\mathcal{C}^{(k)}(\tau) \ \le \ k!\left(\textit{coeff. of } s^k \textit{ in } 1+\frac{\rho\tau}{A^*(\tau)}\left(e^{sA^*(\tau)}-1\right)\right) (20)$$

Proof: Lemma 2 with $F(y) = y^k$ yields Eqn. (19). Using the formula for the moments in Eqn. (14), we compute

$$m_\mathcal{C}^{(k)}(\tau) =$$
$$= k!\cdot\left(\text{coeff. of } s^k \text{ in } M_\mathcal{C}(s,\tau)\right) \qquad (21)$$
$$= k!\cdot\left(\text{coeff. of } s^k \text{ in } \left(M_j(s,\tau)\right)^N\right) \qquad (22)$$
$$= \sum_{k_1+\cdots+k_n=k}\ \prod_{j=1}^{N} m_j^{(k_j)}(\tau) \qquad (23)$$
$$\le k!\cdot\left(\text{coeff. of } s^k \text{ in } 1+\frac{\rho\tau}{A^*(\tau)}\left(e^{sA^*(\tau)}-1\right)\right) (24)$$

Here, the first line is from Eqn. (14), the second follows from independence, the third combines the Cauchy product formula for power series with Eqn. (15). The inequality in the last line follows from the bounds in Eqn. (19) and the positivity of the moments $m_j^{(k)}(\tau)$. □

B. Local Effective Envelopes

B.1 Using the Central Limit Theorem

Eqs. (2) and (3) imply that $E\left[A_j(t,t+\tau)\right] = \rho\tau$. Combining this with the bound for the second moment from Theorem 2 yields the bound

$$Var\left[A_j(t,t+\tau)\right] \ \le \ \underbrace{\rho\tau(A^*(\tau)-\rho\tau)}_{=:\hat{\mathbf{s}}^2} \qquad (25)$$

for the variance of the individual flows. As indicated, we define

$$\hat{\mathbf{s}} \ = \ \rho\tau\sqrt{\frac{A^*(\tau)}{\rho\tau}-1} \ . \qquad (26)$$

By the independence and homogeneity of the flows, it follows that

$$Var\left[A_\mathcal{C}(t,t+\tau)\right] \ = \ N\,Var\left[A_j(t,t+\tau)\right] \qquad (27)$$
$$\le \ N\,\rho\tau(A^*(\tau)-\rho\tau) \ . \qquad (28)$$

Using first the Central Limit Theorem and then the bound on the variance in Eqn.(28), we see that for $x > \rho\tau$

$$Pr\left[A_\mathcal{C}(t,t+\tau) \ge Nx\right] \approx$$
$$\approx 1-\Phi\left(\frac{Nx-N\rho\tau}{\sqrt{Var\left[A_\mathcal{C}(t,t+\tau)\right]}}\right) \qquad (29)$$
$$\le 1-\Phi\left(\sqrt{N}\,\frac{x-\rho\tau}{\hat{\mathbf{s}}}\right) \ , \qquad (30)$$

where Φ is the cumulative normal distribution. To find $\mathcal{G}_\mathcal{C}$ so that

$$Pr[A_\mathcal{C}(0,\tau) \ge \mathcal{G}_\mathcal{C}(\tau;\varepsilon)] \ \le \ \varepsilon \ , \qquad (31)$$

we set $Pr\left[A_\mathcal{C}(t,t+\tau) \ge Nx\right] \approx \varepsilon$ in Eqn. (30) and solve for Nx. This produces an (approximate) local effective envelope of the form

$$\mathcal{G}_\mathcal{C}(\tau;\varepsilon) \ \approx \ N\rho\tau + z\sqrt{N}\rho\tau\sqrt{\frac{A^*(\tau)}{\rho\tau}-1} \ , \qquad (32)$$

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

We remark that our bound in Eqn. (25) is equivalent to Knightly's bound on the *rate variance* in [21]. The rate variance in [21] is defined by

$$RV\left[A_j(t,t+\tau)\right] \ := \ Var\left(\frac{A_j(t,t+\tau)}{\tau}\right) . \qquad (33)$$

Knightly's bound states that

$$RV\left[A_j(t,t+\tau)\right] \ \leq \ \frac{A^*(\tau)}{\tau}\rho - \rho^2 \ , \qquad (34)$$

which is Eqn. (25) with both sides multiplied by τ^{-2}.

B.2 Using the Chernoff Bound

While the estimate in Eqn. (32) 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 (see [28]):

$$Pr[Y \geq y] \leq e^{-sy}\, E[e^{sY}] \qquad \forall s \geq 0 \ . \qquad (35)$$

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

$$Pr[A_{\mathcal{C}}(0,\tau) \geq Nx] \ \leq \ e^{-Nxs} M_{\mathcal{C}}(s,\tau) \ \leq \qquad (36)$$

$$\leq \ \left[e^{-xs}\left(1+\frac{\rho\,\tau}{A^*(\tau)}(e^{sA^*(\tau)}-1)\right)\right]^N . \qquad (37)$$

Here, Eqn. (36) simply uses the Chernoff bound, and Eqn. (37) uses Eqn. (18). We want to find the value of s which makes the bound in Eqn. (37) as tight as possible. For $x < A^*(\tau)$, the right hand side of (37) is minimal when s is chosen so that

$$e^{sA^*(\tau)} \ = \ \frac{x}{\rho\tau}\,\frac{A^*(\tau)-\rho\tau}{A^*(\tau)-x} \ . \qquad (38)$$

Substituting this value of s into Eqn. (37) yields

$$Pr\big[A_{\mathcal{C}}(0,\tau) \geq Nx\big] \leq$$

$$\leq \left[\left(\frac{\rho\tau}{x}\right)^{\frac{x}{A^*(\tau)}}\left(\frac{A^*(\tau)-\rho\tau}{A^*(\tau)-x}\right)^{1-\frac{x}{A^*(\tau)}}\right]^N . \qquad (39)$$

Again, our goal is to find $\mathcal{G}_{\mathcal{C}}$ satisfying Eqn. (31). Using the bound in Eqn. (39) and enforcing that $\mathcal{G}_{\mathcal{C}}(\tau;\varepsilon)$ is never larger than $NA^*(\tau)$ we may set

$$\mathcal{G}_{\mathcal{C}}(\tau;\varepsilon) \ = \ N\min(x, A^*(\tau)) \ , \qquad (40)$$

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

$$\left(\frac{\rho\tau}{x}\right)^{\frac{x}{A^*(\tau)}}\left(\frac{A^*(\tau)-\rho\tau}{A^*(\tau)-x}\right)^{1-\frac{x}{A^*(\tau)}} \ \leq \ \varepsilon^{1/N} \ . \qquad (41)$$

It can be verified that for N sufficiently large, the bound in Eqn. (40) matches closely the CLT bound of Eqn. (32).
Remark: For deterministic envelopes with a peak-rate constraint $A^*(\tau) \leq P\tau$, the expressions for $\mathcal{G}_{\mathcal{C}}$ in Eqn. (40) and Eqn. (32) describe lines, with slopes which depend on ρ, P, N, and ε. 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 ρ, P, and ε).

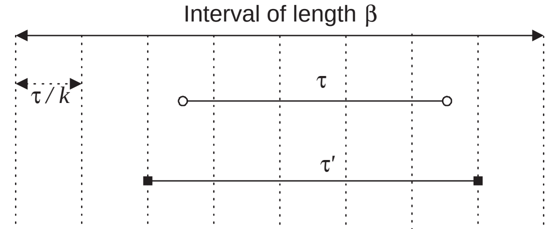

Fig. 2. Embedding intervals.

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 the empirical envelope $\mathcal{E}_{\mathcal{C}}(\tau;\beta)$ for a particular value of τ in terms of a given local effective envelope $\mathcal{G}_{\mathcal{C}}$. The second step fixes the value of the global effective envelope for a finite collection of values τ_i. Finally, we obtain the entire envelope via a subadditive extension.

Let us define two events:

$$B(x,t,\tau) \ = \ \{A_{\mathcal{C}}(t,t+\tau) \geq Nx\} \ , \qquad (42)$$
$$B_\beta(x,\tau) \ = \ \{\mathcal{E}_{\mathcal{C}}(\tau;\beta) \geq Nx\} \ . \qquad (43)$$

where I_β is an interval of length β. 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 I_β contains some interval of length τ where the arrivals exceed Nx.

With Eqn. (39), 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 $\mathcal{H}_{\mathcal{C}}(\tau;\beta,\varepsilon)$ from $\mathcal{G}_{\mathcal{C}}(\tau;\varepsilon)$.

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

$$Pr[B(x,t,\tau)] \ \leq \ Pr[B_\beta(x,\tau)] \ \leq \frac{\beta k}{\tau} Pr[B(x,t,\tau')] \ , \qquad (44)$$

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 $I_i = [t_i, t_{i+k+1}]$ 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 τ in I_β is contained in at least one of the I_i. The claim now follows with stationarity. □

Lemma 3 provides a bound on arrivals in all subintervals of length τ in I_β. One of its implications is that for every value of τ,

$$Pr\left[\mathcal{E}_{\mathcal{C}}(\tau;\beta) \geq \mathcal{G}_{\mathcal{C}}\big(\frac{k+1}{k}\tau;\varepsilon\big)\right] \ \leq \ \frac{\beta k}{\tau}\varepsilon \ , \qquad (45)$$

where $\mathcal{E}_{\mathcal{C}}(.;\beta)$ is the empirical envelope, and $\mathcal{G}_{\mathcal{C}}(\tau,\varepsilon)$ is any local effective envelope.

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

$$\mathcal{H}_{\mathcal{C}}(\tau_i;\beta,\varepsilon) \ = \ \mathcal{G}_{\mathcal{C}}(\tau_i';\varepsilon') \ , \qquad (46)$$

where

$$\tau_i' = \frac{k_i+1}{k_i}\tau_i \quad \text{and} \quad \varepsilon' = \varepsilon\left(\sum_{i=1}^{n}\frac{\beta k_i}{\tau_i}\right)^{-1} . \tag{47}$$

To justify this construction, note that by Eqn. (45) we have

$$Pr\Big[\exists i:\ \mathcal{E}_{\mathcal{C}}(\tau_i;\beta) \geq \mathcal{G}_{\mathcal{C}}(\tau_i',\varepsilon')\Big] \leq \sum_{i=1}^{n}\frac{\beta k_i}{\tau_i}\varepsilon' \leq \quad \varepsilon . \tag{48}$$

Subadditive Extension: We set

$$f(\tau) = \min\{NA^*(\tau),\ \mathcal{H}_{\mathcal{C}}(\tau_i;\beta,\varepsilon)\},\ \tau \in [\tau_{i-1},\tau_i), \tag{49}$$

where the values $\mathcal{H}_{\mathcal{C}}(\tau_i;\beta,\varepsilon)$ are given by Eqn. (46). Since the empirical envelope $\mathcal{E}_{\mathcal{C}}(\tau;\beta)$ increases with τ and cannot exceed $NA^*(\tau)$ by assumption (A2), we see that

$$Pr\left[\exists \tau \in I_\beta:\ \mathcal{E}_{\mathcal{C}}(\tau;\beta) \geq f(\tau)\right] \ \leq\ \varepsilon . \tag{50}$$

Since $\mathcal{E}_{\mathcal{C}}$ is subadditive, we may take $\mathcal{H}_{\mathcal{C}}(\tau;\beta,\varepsilon)$ to be the largest subadditive function which does not exceed $f(\tau)$, in formulas:

$$\mathcal{H}_{\mathcal{C}}(\tau;\beta,\varepsilon) \ = \inf_{\sum\theta_i=\tau}\sum_i f(\theta_i) . \tag{51}$$

Heuristic optimization: Since there exists no universal 'best' global effective envelope, it is clearly impossible to make an optimal choice for the values of τ_i and k_i in Eqs. (46) and (47). 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 ε is sufficiently small.

We will discuss only the case where the traffic regulators satisfy a peak rate constraint with peak rate P and average rate ρ. Our goal is to find an effective envelope satisfying a rate constraint

$$\mathcal{H}_{\mathcal{C}}(\tau;\beta,\varepsilon) \ \leq\ N\alpha\tau , \tag{52}$$

with $\alpha < P$ as close to ρ as possible. In this case we set

$$k_i = k \qquad \text{and} \qquad \tau_i = \gamma^i\tau_o \qquad i = 1,\ldots,n , \tag{53}$$

where τ_o is a small number, and

$$\gamma \ \approx\ 1 + \frac{1}{k+1} , \tag{54}$$

$$k \ \approx\ z\left(z + \sqrt{N}\,\frac{\rho\tau}{\hat{\mathbf{s}}}\right) = z\left(z + \frac{\sqrt{N}}{\sqrt{P/\rho - 1}}\right) . \tag{55}$$

Here, z is defined by $1 - \Phi(z) = \varepsilon$, and $\hat{\mathbf{s}} \ = \ \rho\tau\sqrt{P/\rho - 1}$ in accordance with Eqn. (26). This choice of the τ_i and k_i is used in all our numerical simulations. We motivate the choice in Appendix II by appealing to the Central Limit Theorem.

V. 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, which is expressed in terms of deterministic envelopes, provides deterministic assurances. The second and third conditions, which use the local and global effective envelopes, respectively, give statistical assurances. 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. (61) in Subsection V-A).

In our discussions, we neglect that packet transmissions on a link cannot be preempted. This 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 [25]). 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. 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,

$$\hat{\tau} = \inf\{x \geq 0 \mid A^{q,t}(t-x,t) \leq x\} . \tag{56}$$

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

$$\delta = \inf\left\{\tau_{out} \mid A^{q,t}(t-\hat{\tau},t+\tau_{out}) \leq \hat{\tau} + \tau_{out}\right\} . \tag{57}$$

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

$$\forall\hat{\tau}\ \exists\tau_{out} \leq d_q\ :\left\{A^{q,t}(t-\hat{\tau},t+\tau_{out}) \leq \hat{\tau} + \tau_{out}\right\} . \tag{58}$$

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

$$\sup_{\hat{\tau}}\left\{A^{q,t}(t-\hat{\tau},t+d_q) - \hat{\tau}\right\} \leq d_q . \tag{59}$$

In general, Eqn. (59) is a sufficient condition for meeting delay bound d_q. For FIFO and EDF schedulers, the condition is also necessary [25]. [2]

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

$$A_{\mathcal{C}_p}(t-\hat{\tau}, t+\bar{\tau}_p) \ge A^{q,t}_{\mathcal{C}_p}(t-\hat{\tau}, t+d_q) \ . \tag{60}$$

Remark: For most work-conserving schedulers one can easily find $\bar{\tau}_p$ such that equality holds in Eqn. (60). For example, for FIFO, SP,[3] and EDF schedulers, we have:

$$\begin{array}{ll} \text{FIFO:} & \bar{\tau}_p = 0 \\ \text{SP:} & \bar{\tau}_p = \begin{cases} -\hat{\tau}, & p > q \\ 0, & p = q \\ d_q, & p < q \end{cases} \\ \text{EDF:} & \bar{\tau}_p = \max\{-\hat{\tau}, d_q - d_p\} \ . \end{array}$$

By Eqn. (60), the arrival from class q at time t does not have a violation if d_q is selected such that

$$\sup_{\hat{\tau}} \left\{ \sum_p A_{\mathcal{C}_p}(t-\hat{\tau}, t+\bar{\tau}_p) - \hat{\tau} \right\} \le d_q \ . \tag{61}$$

Next, we show how Eqn. (61) 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 is chosen such that Eqn. (61) is never violated. For a statistical service, d_q is chosen such that a violation of Eqn. (61) is a rare event.

B. Schedulability with Deterministic Envelopes

Exploiting the property of deterministic envelopes in Assumption ($A2$), we can relax Eqn. (61) to

$$\sup_{\hat{\tau}} \left\{ \sum_p \sum_{j \in \mathcal{C}_p} A^*_j(\bar{\tau}_p + \hat{\tau}) - \hat{\tau} \right\} \le d_q \ . \tag{62}$$

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 to [25] to verify that for FIFO and EDF scheduling algorithms the condition in Eqn. (62) 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.

[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.

C. Schedulability with Local Effective Envelopes

By Eqn. (61), the probability that the tagged arrival from time t experiences a deadline violation is less than ε if d_q is selected such that

$$Pr\left[\sup_{\hat{\tau}} \left\{ \sum_p A_{\mathcal{C}_p}(t-\hat{\tau}, t+\bar{\tau}_p) - \hat{\tau} \right\} \le d_q \right] \ge 1-\varepsilon \ . \tag{63}$$

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

$$Pr\left[\sup_{\hat{\tau}} \left\{ \sum_p A_{\mathcal{C}_p}(t-\hat{\tau}, t+\bar{\tau}_p) - \hat{\tau} \right\} \le d_q \right] \approx$$
$$\approx \inf_{\hat{\tau}} Pr\left[\sum_p A_{\mathcal{C}_p}(t-\hat{\tau}, t+\bar{\tau}_p) - \hat{\tau} \le d_q \right] \ . \tag{64}$$

Assuming that equality holds in Eqn. (64), we can re-write Eqn. (63) as

$$\inf_{\hat{\tau}} Pr\left[\sum_p A_{\mathcal{C}_p}(t-\hat{\tau}, t+\bar{\tau}_p) - \hat{\tau} \le d_q \right] \ge 1-\varepsilon \ . \tag{65}$$

Remark: The assumption in Eqn. (64) does not hold in general, since the right hand side is an upper bound for the left hand side. Note that standard extreme-value theory [5] is not immediately applicable to the left hand side of Eqn. (64), since the supremum is taken over a family of random variables indexed by the continuous parameter $\hat{\tau}$. Thus, one must consider the correlations between the $A_{\mathcal{C}_p}(t-\hat{\tau}, t+\bar{\tau}_p)$ for different values of $\hat{\tau}$ in order to obtain a useful estimate for the distribution of the supremum. One way to provide a theoretical justification for the assumption in Eqn. (64) is to assume that arrivals follow a Gaussian process [22]. Other works on statistical QoS have supported the assumption in Eqn. (64) with numerical examples [7], [20], [21], [23].

Recall from the definition of the local effective envelope that $\mathcal{G}_{\mathcal{C}_p}(\tau, \varepsilon) \le x$ implies $Pr\left[A_{\mathcal{C}_p}(t, t+\tau) > x\right] < \varepsilon$. Then, with Lemma 1(a) and assuming that Eqn. (64) holds with equality, we have that a class-q arrival has a deadline violation with probability $< \varepsilon$ if d_q is selected such that

$$\sup_{\hat{\tau}} \left\{ \sum_p \mathcal{G}_{\mathcal{C}_p}(\bar{\tau}_p + \hat{\tau}, \varepsilon/Q) - \hat{\tau} \right\} \le d_q \ . \tag{66}$$

With Eqn. (66) 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 generalization of schedulability conditions for statistical QoS from [20], [21], [23].

The drawback of the condition in Eqn. (66) is its dependence on the assumption in Eqn. (64). Empirical evidence from numerical examples, including those presented in this paper, as well as numerical evidence from previous work which employed this assumption [7], [21], suggests that Eqn. (66) is not overly optimistic. However, it should be noted that the bound in Eqn. (66) 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, which starts at time $\leq t - \hat{\tau}$ and ends at a time after the tagged arrival has departed. The busy period is contained in an interval I_β of length β. Using the properties of the empirical envelope $\mathcal{E}_{\mathcal{C}_p}(.\,;\beta)$ as defined in Section III we have that, for all t and $\bar{\tau}_p + \hat{\tau} \geq 0$,

$$\mathcal{E}_{\mathcal{C}_p}(\bar{\tau}_p + \hat{\tau}; \beta) \geq A_{\mathcal{C}_p}(t - \hat{\tau}, t + \bar{\tau}_p). \tag{67}$$

Thus, we do not have any deadline violation of any class-p arrival in the time interval I_β, if

$$\sup_{\hat{\tau}} \left\{ \sum_q \mathcal{E}_{\mathcal{C}_p}(\bar{\tau}_p + \hat{\tau}; \beta) - \hat{\tau} \right\} \leq d_p \ . \tag{68}$$

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

$$\sup_{\hat{\tau}} \left\{ \sum_p \mathcal{H}_{\mathcal{C}_p}(\bar{\tau}_p + \hat{\tau}; \beta, \varepsilon/Q) - \hat{\tau} \right\} \leq d_q \ . \tag{69}$$

Note that the nature of statistical assurances derived with local effective envelopes (in Subsection V-C) and with global effective envelopes (in Subsection V-D) are quite different. Local effective envelopes are (under the assumption in Eqn. (64)) concerned with the probability that a deadline violation occurs at a certain time. Global effective envelopes address the probability that a deadline violation occurs 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. (62), (66), and (69). 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.

VI. Evaluation

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

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

- **Peak Rate Allocation:** The number of connections that can be supported with a peak rate allocation serves as a lower bound for any method for provisioning QoS.
- **Deterministic Allocation:** We use admission control tests for deterministic QoS from Eqn. (62). The admissible traffic depends on the scheduling algorithm.
- **Average Rate Allocation:** Since it only guarantees finiteness of delays and average throughput, an average rate allocation provides an upper bound for the number of connections that can be admitted on a link.

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

- **Local Effective Envelope:** We use Eqn. (65) to determine admissibility. We evaluate the quality of the following two bounds, derived in Section IV:
  - **Local Effective Envelope (CB)**: We use the bound from Eqn. (41), obtained with the Chernoff bound.
  - **Local Effective Envelope (CLT)**: We use the bound from Eqn. (32), 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 [21].
- **Global Effective Envelope:** We determine admissibility using the procedure developed in Eqs. (46)–(47) and (53)–(55) of Subsection V-D. We select β such that it is larger than the longest busy period (see Footnote 1). In Eqn. (46), 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 [15], [26], [31]**:[4] 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 [15]).
  - **EB-RRR**: We use Eqn. (9) from [31] by Rajagopal/Reisslein/Ross which presents an extension to the *EB-EMW* result. The delay bound is 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.
- **Bufferless Multiplexer (Bufferless MUX) [33]:** This method has been described in Section II. We use Eqn. (15) and the parameter selections provided in [33].

In all our experiments, we consider traffic regulators which are obtained from peak rate controlled leaky buckets with deterministic envelopes as given in Eqn. (1).[5] The link capacity is set to $C = 45$ Mbps and the number of traffic classes is $Q = 2$. The traffic parameters for single flows in the classes are as follows:

Class	Peak Rate P (Mbps)	Mean Rate ρ (Mbps)	Burst Size σ (bits)
1	1.5	0.15	95400
2	6.0	0.15	10345

The parameters are selected so as to match (approximately) the examples presented in [15], [31]. In [3] we

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

[5]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.

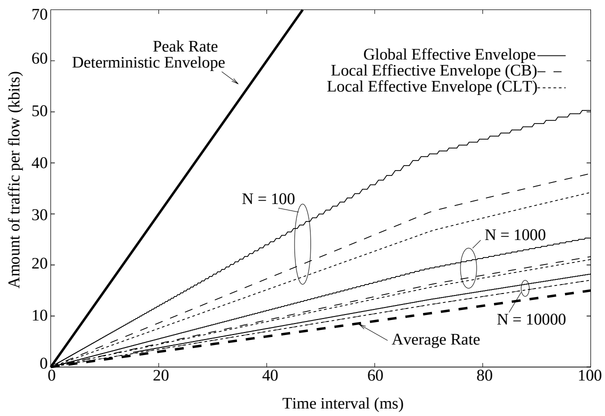

Fig. 3. Example 1: Comparison of envelopes for $\tau \leq 100\ ms$, $\varepsilon = 10^{-6}$, and for number of flows $N = 100, 1000, 10000$ (Class 1).

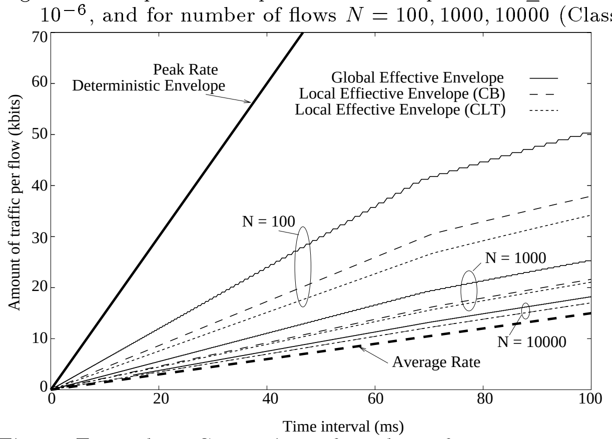

Fig. 4. Example 1: Comparison of envelopes for $\tau \leq 100\ ms$, number of flows $N = 1000$ and for different values of ε (Class 1).

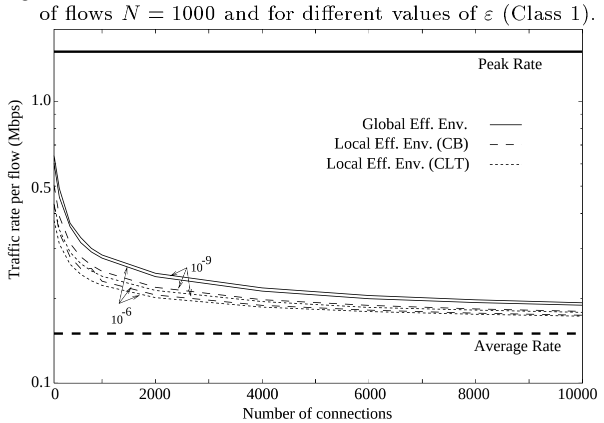

Fig. 5. Example 1: Traffic rates $\mathcal{G}_{\mathcal{C}}(\tau_0;\varepsilon)/(N\tau_0)$ and $\mathcal{H}_{\mathcal{C}}(\tau_0;\varepsilon)/(N\tau_0)$ for $\tau_0 = 50$ ms and $\varepsilon = 10^{-6}$ or 10^{-9} (Class 1).

present additional experiments, including experiments with MPEG video traces.

In this section we present three sets of examples. The first two examples only include results for Class-1 traffic.

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.

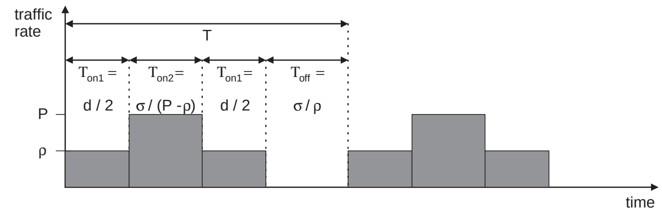

Fig. 6. Traffic pattern for (P, ρ, σ) flows used in simulations.

Figure 3 shows the results for multiplexed flows from Class 1. We set $\varepsilon = 10^{-6}$ for all types of effective envelopes. By depicting the amount of traffic per flow for different values of N, we can observe how the statistical multiplexing gain increases with the number of flows. Observe that local and global effective envelopes are much smaller than the deterministic envelope and the peak rate. 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). Figure 3 also shows that the difference between local and global effective envelopes narrows as the number of flows N is increased.

In Figure 4 we depict the sensitivity of the effective envelopes to the selection of the parameter ε. We use the same traffic parameters as before. We fix the value for the number of flows to $N = 1000$, and we show the effective envelopes for $\varepsilon = 10^{-3}, 10^{-6}$, and 10^{-9}. Figure 4 shows that the effective envelopes are not very sensitive to variations of the parameter ε.

In Figure 5 we show how the effective envelopes vary if the number of flows N is increased. We use the same parameters as before, but only consider the values of the envelopes at $\tau_0 = 50$ ms. For this value of τ_0, Figure 5 shows the values of the rates $\mathcal{G}_{\mathcal{C}}(\tau_0;\varepsilon)/(N\tau_0)$ and $\mathcal{H}_{\mathcal{C}}(\tau_0;\varepsilon)/(N\tau_0)$, as the number of flows N is varied. For comparison, we include the peak and average rates into the graph. There are three noteworthy observations to be made. First, as the number of flows N is increased, the effective envelopes, both local and global, approach the average traffic rate. Second, the difference between the two local effective envelopes diminishes when N is large. Third, for large values of N the difference between the local and global effective envelopes is quite small.

B. Example 2: Admissible Region for Homogeneous Flows

In this example, we investigate the number of flows admitted by various admission control methods at a link with a FIFO scheduler. Here, flows are homogeneous, that is, all flows belong to a single class. We compare the admissible regions, that is, the range of parameters which results in a positive admission control decision, of the local and global effective envelopes, to those of the effective bandwidth techniques (both *EB-EMW* and *EB-RRR*), the bufferless Multiplexer (*bufferless MUX*) and to deterministic QoS assurances.

We compare the results with those obtained from a discrete event simulation. For the simulation, we take a pat-

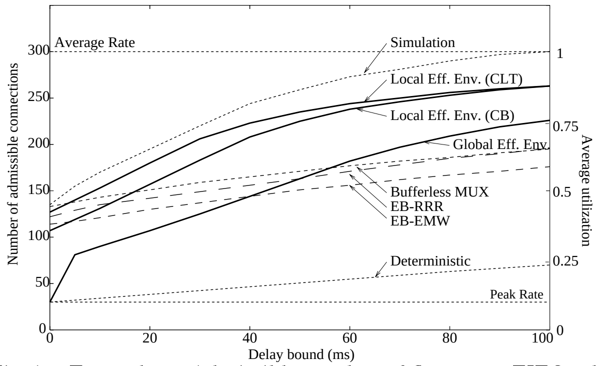

Fig. 7. Example 2: Admissible number of flows at a FIFO scheduler for homogeneous flows as a function of delay bounds ($\varepsilon = 10^{-6}$, Class 1).

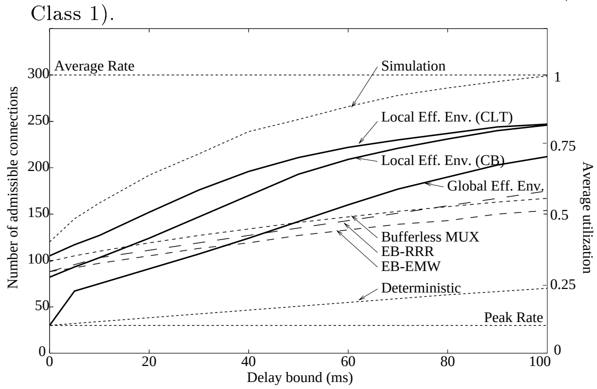

Fig. 8. Example 2: Admissible number of flows at a FIFO scheduler for homogeneous flows as a function of delay bounds ($\varepsilon = 10^{-9}$, Class 1).

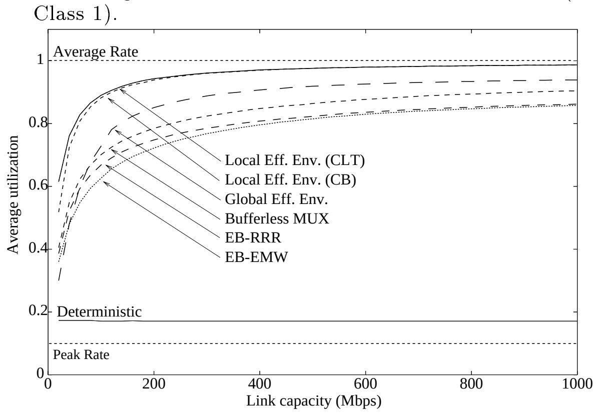

Fig. 9. Example 2: Achievable average utilization vs. link capacity at a FIFO scheduler ($\varepsilon = 10^{-6}$, $d = 50$ ms, Class 1).

tern which we expect, based on the simulations in [27], to be close to an adversarial traffic pattern for peak-rate controlled leaky buckets. However, we do not claim that the results from the simulation scenario are the worst possible.

In the simulations, the traffic for a flow with parameters given by (P, ρ, σ), is periodic with a pattern as shown in Figure 6. A flow transmits at the average rate ρ for a duration $T_{on1} = d/2$, where d is the delay bound. Then, the flow transmits at the peak rate P for a duration $T_{on2} = \sigma/(P - \rho)$, followed by another phase of length T_{on1} during which the flow transmits at rate ρ. Then, the source shuts off, waits for a duration $T_{off} = \sigma/\rho$ 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.

Figure 7 depicts the number of admitted flows as a function of the delay bound. Here, the probability of a violation of QoS assurances is set to $\varepsilon = 10^{-6}$. The figure shows that all methods for statistical QoS admit many more connections than a deterministic admission control test. The local effective envelopes (both CLT and CB) are closest to the simulation results. Note, however, that results obtained with local effective envelopes are approximate and are not guaranteed to be upper bounds on the admissible regions.

In Figure 8 we show the results for the same experiment with $\varepsilon = 10^{-9}$. A comparison of Figures 7 and 8 indicates that the admissible regions are not very sensitive to variations of ε.

In Figure 9 we show how the achievable utilization at a FIFO scheduler increases with the link capacity. We fix the delay bound of traffic to $d = 50$ ms and we use $\varepsilon = 10^{-6}$. The average achievable link utilization is the sum of the average rates of flows which can be accepted according to a chosen schedulability conditions. Figure 9 illustrates that all statistical methods have a high statistical multiplexing gain at high link capacities. In contrast, the achievable utilization for deterministic QoS is (almost) constant when the link capacity is increased. Note that the difference between the admissible regions of the local and the global effective envelopes is small at high link capacities.

C. Example 3: Admissible Region for Heterogeneous Flows

In this example we consider different scheduling algorithms and heterogeneous traffic arrivals. As scheduling algorithms, we consider Static Priority (SP) and Earliest-Deadline-First (EDF).

In this example, we multiplex a number of flows from Class 1 and from Class 2 on 45 Mbps. The delay bound for Class-1 flows is selected relatively long, $d_1 = 100\ ms$, and the delay bound for Class-2 flows is selected 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 results are shown in Figure 10. The plot depicts the admissible region for SP and EDF schedulers using the results from the local effective envelope, global effective envelope, and deterministic envelope approaches, respectively. We also include the admissible regions using the effective bandwidth approaches (*EB-EMW* and *EB-RRR*). Note, however, that the effective bandwidth does not account for different scheduling algorithms. The results in Figure 10 show that the difference between SP and EDF schedulers is small in all cases. The global effective envelope is, again, more conservative than the local effective envelope method. Figure 10 indicates that methods which consider scheduling algorithms may have advantages over effective bandwidth methods (*EB-EMW*, *EB-RRR*) when dealing with heterogeneous traffic.

VII. Discussion

We have presented new results on evaluating the statistical multiplexing gain for traffic scheduling algorithms. We have introduced the notions of local and global effective envelopes, which are, with high probability, bounds on aggregated traffic flows, and we have derived admission con-

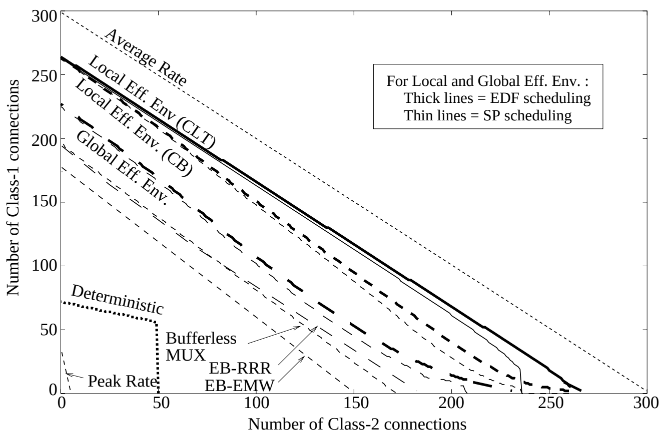

Fig. 10. Example 3: Admissible region of multiplexing Class 1 and Class 2 flows with $\varepsilon = 10^{-6}$, $d_1 = 100$ ms and $d_2 = 10$ ms.

trol tests for these bounds. We conclude with the following remarks:

• We have presented two schedulability conditions for a statistical service at a single node. The first condition, which uses local effective envelopes (Subsection V-C), provides a guarantee for the probability of a QoS violation for an arbitrary arrival. The second condition, which uses global effective envelopes (Subsection V-D), provides a guarantee for the probability of a QoS violation in an arbitrary time interval. Our motivation to introduce a second, more conservative, and possibly less intuitive, condition is the assumption in Eqn. (64). Without a verification of this assumption, an admission control conditions may be too optimistic.

• We believe that our approach which separates consideration of service models (deterministic, statistical), scheduling algorithms (FIFO, SP, EDF), and the choice of the large deviations tools (Central Limit Theorem, Chernoff bound) may prove to be useful, as it simplifies the task of testing new scheduling algorithms or large deviations results.

• Our work does not attempt to derive an adversarial traffic pattern. Even though results on adversarial patterns have been obtained recently for buffered multiplexers [18], [19], [31], it may not be feasible to derive adversarial traffic patterns for more complex scheduling algorithms. On the other hand, our results show that good bounds on the admissible regions are attainable even without knowledge of adversarial traffic patterns.

• A few years ago, a study addressed the question of the fundamental limits of a deterministic service [37], and found that (a) deterministic QoS drastically increases the admissible region over a peak-rate allocation, and (b) the choice of the scheduling algorithm has a noticeable impact on the size of the admissible region. With the results from this paper, we can now provide some insights in the fundamental limits of a statistical service, at least in the context of regulated, adversarial traffic arrivals.

1. The examples in this paper show that the difference between the admissible regions of statistical and deterministic QoS is significant, even if ε is selected very small, e.g., $\varepsilon = 10^{-9}$.

2. The results from Example 3 suggest that the selection of the scheduling algorithm (SP vs. EDF in our case) has a noticeable, but, in relative terms, small impact on the size of the admissible region. Additional numerical data are required to make more conclusive statements on the significance of scheduling algorithms for a statistical service.

3. The examples in this paper show that, for high data rates, the admissible region for a statistical service is sometimes close to that of an average rate allocation. In such a regime, the additional gain achievable by improving currently available methods appears small.

References

[1] ATM Forum, ATM Forum Traffic Management Specification Version 4.0, April 1996.

[2] M. Andrews. Probabilistic end-to-end delay bounds for earliest deadline first scheduling. In *Proc. IEEE Infocom 2000*, pages 603–612, Tel Aviv, March 2000.

[3] 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, August 1999.

[4] R. Braden, D. Clark, and S. Shenker. Integrated services in the internet architecture: an overview. IETF RFC 1633, July 1994.

[5] E. Castillo. *Extreme Value Theory in Engineering.* Academic Press, 1988.

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

[7] 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.

[8] C. Courcoubetis, F. P. Kelly, and R. R. Weber. Measurement-based usage charges in communication networks. Technical Report 1997-19, University of Cambridge, Statistical Laboratory, 1997.

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

[10] R. Cruz. A calculus for network delay, part II: Network analysis. *IEEE Transactions on Information Theory*, 37(1):132–141, January 1991.

[11] R. Cruz. Quality of service guarantees in virtual circuit switched networks. *IEEE Journal on Selected Areas in Communications*, 13(6):1048–1056, August 1995.

[12] R. Cruz. Quality of service management in integrated services networks. In *Proceedings of 1st Semi-Annual Research Review, CWC, UCSD*, June 1996.

[13] B. T. Doshi. Deterministic rule based traffic descriptors for broadband ISDN: Worst case behavior and connection acceptance control. In *International Teletraffic Congress (ITC 14)*, pages 591–600, Antibes Juan-les-Pins, France, June 1994.

[14] 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.

[15] 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.

[16] D. Ferrari and D. Verma. A scheme for real-time channel establishment in wide-area networks. *IEEE Journal on Selected Areas in Communications*, 8(3):368–379, April 1990.

[17] L. Georgiadis, R. Guérin, V. Peris, and K. N. Sivarajan. Efficient network QoS provisioning based on per node traffic shaping. *IEEE/ACM Transactions on Networking*, 4(4):482–501, August 1996.

[18] G. Kesidis and T. Konstantopoulos. Extremal shape-controlled traffic patterns in high-speed networks. IEEE Transactions on Communications. To appear.

[19] 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.

[20] 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.

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

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

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

[24] J. Liebeherr, S. Patek, and E. Yilmaz. Tradeoffs in designing networks with end-to-end statistical QoS guarantees. In *Proc. IEEE/IFIP Eigth International Workshop on Quality of Service (IWQoS '00)*, Pittsburgh, June 2000.

[25] 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.

[26] 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.

[27] P. Oechslin. On-off sources and worst case arrival patterns of the leaky bucket. Technical report, University College London, September 1997.

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

[29] A. K. Parekh and R. G. Gallager. A generalized processor sharing approach to flow control in integrated services networks: the single-node case. *IEEE/ACM Transactions on Networking*, 1(3):344–357, June 1993.

[30] J. Qiu and E. Knightly. Inter-class resource sharing using statistical service envelopes. In *Proc. IEEE Infocom '99*, pages 36–42, March 1999.

[31] 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.

[32] 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.

[33] 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 1061–1072, New York, March 1999.

[34] J. Roberts, U. Mocci, and J. Virtamo (Eds.). *Broadband Network Traffic: Performance Evaluation and Desgin of Broadband Multiservice Networks. Final Report of Action. COST 242. (Lecture Notes in Computer Science. Vol. 1152)*. Springer Verlag, 1996.

[35] M. Schwartz. *Broadband Integrated Networks*. Prentice Hall, 1996.

[36] D. Starobinski and M. Sidi. Stochastically bounded burstiness for communication networks. In *Proc. IEEE Infocom '99*, pages 36–42, New York, March 1999.

[37] 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.

[38] O. Yaron and M. Sidi. Performance and stability of communication networks via robust exponential bounds. *IEEE/ACM Transactions on Networking*, 1(3):372–385, June 1993.

[39] H. Zhang. Service Disciplines for Guaranteed Performance Service in Packet-Switching Networks. *Proceedings of the IEEE*, 83(10):1374–1399, October 1995.

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

[41] H. Zhang and E. Knightly. Providing end-to-end statistical performance guarantees with bounding interval dependent stochastic models. In *Proceedings of ACM SIGMETRICS'94*, pages 211–220, Nashville, May 1994.

Appendix

I. Proof of Lemma 2

We begin the proof with Eqn. (3) which holds by the stationarity assumption (A3). Since the limit in Eqn. (3) exists, we may compute it by restricting T to be an integer multiple of τ. We compute the average over $[0,T]$ by partitioning $[0,T]$ into subintervals of length τ, and then averaging over the position of the subintervals.

$$E[F(A_j(t,t+\tau)] =$$

$$= \lim_{T\to\infty} E\left[\frac{1}{T}\sum_{i=1}^{T/\tau}\int_0^\tau F\big(A_j((i-1)\tau+\omega, i\tau+\omega)\big)d\omega\right] \quad (70)$$

$$\leq \lim_{T\to\infty} E\left[\max_{0\leq\omega\leq\tau}\frac{1}{T/\tau}\sum_{i=1}^{T/\tau} F\big(A_j((i-1)\tau+\omega, i\tau+\omega)\big)\right] \quad (71)$$

For a fixed value of T and a fixed arrival pattern $\{A_j(t,t+\tau)\}_{t\geq 0}$, let ω_0 be the shift for which the maximum is assumed in Eqn. (71). Set

$$y_i = A_j((i-1)\tau+\omega_0, i\tau+\omega_0)\ . \quad (72)$$

To obtain an upper bound for the limit in Eqn. (71), we consider the following optimization problem:

$$\text{maximize} \quad \frac{1}{n}\sum_{i=1}^{n} F(y_i) \qquad (n = T/\tau) \quad (73)$$

$$\text{subject to} \quad 0 \leq y_i \leq A^*(\tau) \qquad i = 1,\ldots,N \quad (74)$$

$$\sum_{i=1}^{N} y_i \leq A^*(T)\ . \quad (75)$$

By convexity, the maximal value is attained at some point on the boundary of the region defined by the side conditions. Moreover, since F is increasing, side condition (75) holds at this point with equality. Exploiting the symmetry of the problem under permutations of the y_i, we see immediately that the following is a maximizing solution:

$$y_i = \begin{cases} A^*(\tau) & \text{if } i \leq \lfloor \frac{A^*(T)}{A^*(\tau)} \rfloor \\ A^*(T) - i\,A^*(\tau) & \text{if } i = \lfloor \frac{A^*(T)}{A^*(\tau)} \rfloor + 1 \\ 0 & \text{otherwise}\ . \end{cases} \quad (76)$$

This assigns the maximum value $A^*(\tau)$ to as many y_i as possible, subject to the first side condition. So, the maximum of Eqn. (73), up to a rounding error of $O(1/n)$, is

$$\lambda(T)\,F\left(A^*(\tau)\right) \;+\; \left((1-\lambda(T)\right)F(0)\ , \quad (77)$$

where

$$\lambda(T) \;=\; \frac{A^*(T)}{nA^*(\tau)} \;<\; 1\ . \quad (78)$$

Inserting the maximum back into Eqn. (71) and recalling that $n = T/\tau$, we obtain the bound

$$E[F(A_j(t,t+\tau)] \leq$$

$$\leq \lim_{T\to\infty}\left\{\lambda(T)\,F\left(A^*(\tau)\right) + \left(1-\lambda(T)\right)F(0)\right\} \quad (79)$$

$$= \left(\frac{\rho\,\tau}{A^*(\tau)}\right)F\left(A^*(\tau)\right) + \left(1 - \frac{\rho\,\tau}{A^*(\tau)}\right)F(0)\ . \quad (80)$$

In the evaluation of the limit, we have used that

$$\lim_{T\to\infty} \lambda(T) \;=\; \lim_{T\to\infty} \frac{\tau}{T}\frac{A^*(T)}{A^*(\tau)} \;=\; \frac{\rho\tau}{A^*(\tau)} \qquad (81)$$

by the definition of ρ. This completes the proof. $\square$

II. Heuristic for Global Effective Envelope

Here we motivate the choice of the τ_i and the k_i for the heuristic optimization presented in Subsection IV-C. Let us for the moment accept Eqn. (53), fix τ_o and β, and optimize over the parameters k and γ. Eqn. (53) guarantees that

$$\mathcal{H}_{\mathcal{C}}(\tau;\beta,\varepsilon) \;\leq\; \mathcal{G}_{\mathcal{C}}\left(\frac{k+1}{k}\,\gamma\tau;\varepsilon'\right) \qquad (82)$$

for all $\tau \in [\tau_o,\beta]$, where summing the geometric series in Eqn. (47) gives

$$\varepsilon' \;=\; \frac{\tau_o(\gamma-1)}{\beta k}\cdot\varepsilon\,. \qquad (83)$$

We estimate

$$\begin{aligned} Pr\left[\exists\tau\in[\tau_o,\beta]:\; \mathcal{E}^{\beta}_{\mathcal{C}}(\tau)\;\geq\;N\alpha\tau\right] &\leq \\ \leq Pr\left[\exists i:\; \mathcal{E}^{\beta}_{\mathcal{C}}(\tau_i)\;\geq\;N\alpha\tau_i/\gamma\right] && (84)\\ \leq \sum_{i=1}^{n}\frac{\beta k}{\tau_o\gamma^i}Pr\left[A_{\mathcal{C}}\left(0,\frac{k+1}{k}\tau_i\right)\leq N\alpha\tau_i/\gamma\right] && (85)\\ \approx \frac{\beta k}{\tau_o(\gamma-1)}\left(1-\Phi\left(\sqrt{N}\,\frac{\alpha k/(\gamma(k+1))-\rho}{\rho\sqrt{P/\rho-1}}\right)\right) && (86)\end{aligned}$$

where the first step follows from monotonicity, the second step uses Lemma 3, and the third step invokes the Central Limit Theorem and a simple estimate for the geometric series. We next solve for α in the equation

$$\varepsilon \;=\; \frac{\beta k}{\tau_o(\gamma-1)}\left(1-\Phi\left(\sqrt{N}\,\frac{\alpha k/(\gamma(k+1))-\rho}{\rho\sqrt{P/\rho-1}}\right)\right)\,. \qquad (87)$$

For every integer k and $\gamma>1$, an (approximate) envelope is given by

$$\begin{aligned}\mathcal{H}_{\mathcal{C}}(\tau;\beta,\varepsilon) \;&\approx\; N\alpha\tau = && (88)\\ &= \frac{(k+1)\gamma}{k}\left(N\rho\tau + z'\sqrt{N}\rho\tau\sqrt{P/\rho-1}\right), && (89)\end{aligned}$$

where ε' is given by Eqn. (83), and $1-\Phi(z')=\varepsilon'$. This approximation is valid for τ in the interval $[\tau_o,\beta]$.

Our goal is to choose k and γ so that the right hand side of Eqn. (89) is as small as possible. The difficulty is that z' depends on the choice of k and γ. We can achieve our goal by minimizing instead the right hand of Eqn. (86). It is easy to see that the minimum value is achieved for some finite positive value of k and γ. Using the approximation $1-\Phi(z)\approx z^{-1}\phi(z)$, where ϕ is the density of standard normal distribution, differentiating with respect to k and γ, and solving approximately for the critical values, we see that the minimum is attained at a point satisfying Eqn. (54) and

$$k \approx z'\left(z' + \frac{\sqrt{N}}{\sqrt{P/\rho-1}}\right)\,, \qquad (90)$$

where z' is defined by $1-\Phi(z')=\varepsilon'$. Approximating z' by z we arrive at the conditions in Eqs. (54) and (55).

We turn to the basic choice made in Eqn. (53). The fact that the right hand side of Eqn. (86), and hence k and γ determined by Eqn. (54) and either Eqn. (55) or Eqn. (90), does not depend on τ_o and β indicates that optimal choices of k_i and $\gamma_i := \tau_i/\tau_{i-1}$ do not depend on the size of τ_i, provided that τ_i is small enough to lie in the region where $A^*(\tau_i)=P\tau_i$.

We propose the following analogous heuristic optimization for general regulators with subadditive deterministic envelope A^*. Assuming as above that β is given, we set τ_o to be a small number, and set z such that $1-\Phi(z)=\varepsilon$. We replace Eqs. (53) – (55) by the following method to recursively determine the k_i, γ_i, and τ_i for $1\leq i\leq n$, where n is the first time such that $\tau_n\geq\beta$.

$$\begin{aligned} k_i \;&=\; z\left(z + \frac{\rho\tau_{i-1}}{\hat{\mathbf{s}}_{i-1}}\sqrt{N}\right) && (91)\\ &=\; z\left(z+\sqrt{N}\left(\frac{A^*(\tau_{i-1})}{\rho\tau_{i-1}}-1\right)^{-1/2}\right)\,, && (92)\end{aligned}$$

where $\hat{\mathbf{s}}_i$ is as given in Eqn. (26) with the subscript i in $\hat{\mathbf{s}}_i$ corresponding to τ_i, and

$$\gamma_i = 1+\frac{1}{k_i+1} \quad\text{and}\quad \tau_i=\gamma_i\,\tau_{i-1}\,. \qquad (93)$$

When the algorithm terminates after n steps, we obtain ε' from Eqn. (47).