A Calculus for End-to-end Statistical Service Guarantees

Jörg Liebeherr, *Department of Computer Science, University of Virginia*
Stephen D. Patek, *Department of Systems and Information Engineering, University of Virginia*

Abstract— **The deterministic network calculus offers an elegant framework for determining delays and backlog in a network with deterministic service guarantees to individual traffic flows. A drawback of the deterministic network calculus is that it only provides worst-case bounds. Here we present a network calculus for statistical service guarantees, which can exploit the statistical multiplexing gain of sources. We introduce the notion of an *effective service curve* as a probabilistic bound on the service received by an individual flow, and construct an effective service curve for a network where capacities are provisioned exclusively to aggregates of flows. Numerical examples demonstrate that the calculus is able to extract a significant amount of multiplexing gain in networks with a large number of flows.**

I. Introduction

The deterministic network calculus recently evolved as a new theory for deterministic queueing systems, and has provided powerful tools for reasoning about delay and backlog in a network with service guarantees to individual traffic flows. Using the notion of arrival envelopes and service curves [12], several recent works have shown that delay and backlog bounds can be concisely expressed in a min-plus algebra [1], [6], [10].

However, the deterministic view of traffic only provides worst-case bounds and does not to take advantage of statistical multiplexing gain. The problem of trying to exploit the resource savings of statistical multiplexing while preserving the elegant formalism of the network calculus has been the subject of several studies. Kurose [17] uses the concept of stochastic ordering and obtains bounds on the distribution of delay and buffer occupancy of a flow in a network with FIFO scheduling. Chang [9] presents probabilistic bounds on output burstiness, backlog and delays in a network where the moment generating functions of arrivals are exponentially bounded. Different bounds for exponentially bounded arrivals are derived by Yaron and Sidi [24] and Starobinski and Sidi [23]. Results on statistical end-to-end delay guarantees in a network have been obtained for specific scheduling algorithms, such as EDF [21], [22], and GPS [14], and a class of coordinated scheduling algorithms [2], [18]. Several researchers have considered probabilistic formulations of service curves. Cruz defines a probabilistic service curve which violates a given deterministic service curve according to a certain distribution [13]. Chang (see [11], Chp. 7) presents a statistical network calculus for 'dynamic F-servers'. Finally, Knightly and Chiu [20] derive 'statistical service envelopes' as time-invariant lower bounds on the service received by an aggregate of flows.

This paper presents a network calculus for statistically multiplexed traffic, expressed in the min-plus algebra. Generally, we assume that network capacities are allocated to aggregates of flows. This is different from the per-flow capacity allocation generally applied in the deterministic network calculus. Within this context, we define an *effective service curve*, which is, with high certainty, a bound on the service received by a single flow. So, we will consider probabilistic per-flow service guarantees for networks where resources are reserved for aggregates. We will show that the main results of the deterministic network calculus carry over to the statistical framework we present.

The results in this paper are set in a continuous time model with fluid left-continuous traffic arrival functions, as is common for network delay analysis in the deterministic network calculus. We refer to [11] for the issues involved in relaxing these assumptions for application of the analysis in packet

This work is supported in part by the National Science Foundation through grants ANI-9730103, ECS-9875688 (CAREER), ANI-9903001, and ANI-0085955.

networks. A node represents a router (or switch) in a packet network. The transmission rate at a node corresponds to the capacity of an output link of a router. Packetization delays and other effects of discrete sized packets, such as the non-preemption of packet transmission, are ignored. When analyzing delays in a network, all processing overhead and propagation delays are ignored.

In the numerical examples, presented in this paper, we assume a 'regulated adversarial traffic' model where (1) arrivals from each flow into the network are constrained by a deterministic regulator and (2) traffic arrivals from different flows are statistically independent. The regulated adversarial traffic model has been used by several researchers, e.g., [15], [16], for modeling aggregates of sources, which are policed or shaped, but for which arrival distributions are not readily available.

The remaining sections of this paper are structured as follows. In Section II, we review the notation and key results of the deterministic network calculus. In Section III, we present the effective envelope from [5], and we define our notion of effective service curves. Next we present the results for a statistical network calculus in terms of effective service curves and effective envelopes. In Section IV we show how to construct effective service curves for individual flows at a node where service is allocated to an aggregate of flows. We discuss how these results can be used for end-to-end, per-flow service provisioning. In Section V, we show how to build 'effective envelopes' [5], which are used in our construction of effective service curves. In Section VI we discuss numerical examples for single node and multi-node networks and evaluate the statistical multiplexing gain achievable with effective service curves.

II. Network Calculus Preliminaries

The deterministic network calculus provides concise expressions for upper bounds on the backlog and delay experienced by an individual flow at one or more network nodes. An attractive feature of the network calculus is that end-to-end bounds can often be easily obtained from manipulations of the per-node bounds.

In this section we review some notation and results from the deterministic network calculus, as needed later in the paper. However, this section is not a comprehensive summary of the network calculus. For a complete discussion we refer to [1], [7], [11].

A. Operators

Much of the formal framework of the network calculus can be elegantly expressed in a min-plus algebra [3], complete with convolution and deconvolution operators for functions. Generally, the functions in this paper are non-negative, monotonically increasing and left-continuous, defined over time intervals $[0, t]$. We assume for a given function f that $f(t) = 0$ if $t < 0$.

The *convolution* $f * g$ of two functions f and g, is defined as

$$f * g(t) = \inf_{0 \le \tau \le t} \{f(t-\tau) + g(\tau)\} \ .$$

The *deconvolution* $f \oslash g$ of two functions f and g, is defined as

$$f \oslash g(t) = \sup_{\tau \ge 0} \{f(t+\tau) - g(\tau)\} \ .$$

We refer to [3], [7], [11] for a detailed discussion of the properties of the min-plus algebra and the properties of the convolution and deconvolution operators.

B. Arrival functions and Service Curves

Let us consider the traffic arrivals to a single network node. The arrivals of a flow in the time interval $[0, t)$ are given in terms of a function $A(t)$. The departures of a flow from the node in the time interval $[0, t)$ are denoted by $D(t)$, with $D(t) \le A(t)$. The backlog of a flow at time t, denoted by $B(t)$, is given by $B(t) = D(t) - A(t)$. The delay at time t, denoted as $W(t)$, is the delay experienced by an arrival which departs at time t, given by $W(t) = \inf\{d \ge 0 \mid A(t-d) \le D(t)\}$. If arrival and departure functions are plotted as functions of time, then $B(t)$ and $W(t)$, respectively, are the vertical and horizontal differences between arrival and departure functions. We will use $A(x, y)$ and $D(x, y)$ to denote the arrivals and departures in the time interval $[x, y)$, with $A(x, y) = A(y) - A(x)$ and $D(x, y) = D(y) - D(x)$.

We have the following assumptions on the arrival functions.

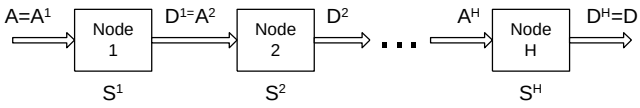

Fig. 1. Traffic of a flow through a set of H nodes. Let A^h and D^h denote the arrival and departures at the h-th node, with $A^i = A$, $D^H = D$, and $A^h = D^{h-1}$, for $h = 2, \ldots, H$.

(A1) Non-Negativity. The arrivals in any interval of time are non-negative. That is, for any $x < y$, we have $A(y) - A(x) \geq 0$.

(A2) Upper Bound. The arrivals A of a flow are bounded by a deterministic subadditive function A^*, called the *arrival envelope*,[1] such that $A(t+\tau) - A(t) \leq A^*(\tau)$ for all $t, \tau \geq 0$. [2]

The upper bound given by A^* assumes that the traffic of a flow is policed or shaped by a traffic conditioning function, such as a leaky bucket.

The service guaranteed to a flow is expressed in terms of 'service curves'. A *minimum service curve* for a flow is a function S which specifies a lower bound on the service given to a flow such that $D(t) \geq A * S(t)$ for all $t \geq 0$. A *maximum service curve* for a flow is a function $\overline{S}$ which specifies an upper bound on the service given to a flow by $D(t) \leq A * \overline{S}(t)$ for all $t \geq 0$.

The following theorem summarizes some key results of the deterministic network calculus. These results have been derived in [1], [6], [10]. We will follow the notation used in [1].

Theorem 1: **Deterministic Network Calculus.** Given a flow with arrival envelope A^* and with minimum and maximum service curves S and $\overline{S}$, the following hold:

1. *Output Envelope:* The function $D^* = A^* \oslash S$ is an envelope for the departures, in the sense that $D^*(t) \geq D(t+\tau) - D(\tau)$ for all $t, \tau \geq 0$.
2. *Backlog Bound:* An upper bound for the backlog, denoted by b_{max}, is given by $b_{max} = A^* \oslash S(0)$.
3. *Delay Bound:* An upper bound for the delay, denoted by d_{max}, is given by $d_{max} = \inf\{d \geq 0 \,;\, \forall t \geq 0 : A^*(t-d) \leq S(t)\}$.
4. *Network Service Curve:* If a flow passes through H nodes in series, as shown in Figure 1, and if the flow is offered minimum and maximum service curves S^h and $\overline{S}^h$, respectively, at each node $h = 1, \ldots, H$. Then, the sequence of nodes provides minimum and maximum service curves S^{net} and $\overline{S}^{net}$, referred to as network service curves, which are given by

$$\begin{aligned} S^{net} &= S^1 * S^2 * \ldots * S^H \,, \\ \overline{S}^{net} &= \overline{S}^1 * \overline{S}^2 * \ldots * \overline{S}^H \,. \end{aligned}$$

Thus, with Theorem 1, network service curves can be used to determine bounds on delay and backlog in a network. There are many additional properties and refinements that have been derived for the deterministic calculus. However, in this paper we will concern ourselves only with the results above.

III. Statistical Network Calculus

A drawback of the deterministic network calculus is that the deterministic view of traffic only yields pessimistic worst-case bounds, which do not take advantage of statistical multiplexing of flows when multiple flows are carried over the same link. We will now approach the network calculus in a probabilistic framework.

The statistical network calculus we present has two components. The first component, presented in Subsection III-A, relates to probabilistic service guarantees for aggregates of flows, where deterministic service is allocated to the aggregate. The second component of the calculus, presented in Subsection III-B, relates to probabilistic service guarantees for individual flows which are allocated probabilistic service in the form of 'effective service curves'. In Section IV, we show how to determine an effective service curve for an individual flow when deterministic service is allocated to the aggregate.

A. A Statistical Calculus for Groups of Flows

Let $\mathcal{C}$ denote a set of flows. The arrival and departure functions for each flow $j \in \mathcal{C}$ will be denoted by A_j and D_j, respectively. We use $A_\mathcal{C}$ and $D_\mathcal{C}$ to denote the aggregate arrivals and departures from class $\mathcal{C}$ at the network node, that is, $A_\mathcal{C}(t) = \sum_{j \in \mathcal{C}} A_j(t)$ and $D_\mathcal{C}(t) = \sum_{j \in \mathcal{C}} D_j(t)$. A deterministic arrival envelope for the aggregate is $A^*_\mathcal{C}(t) = \sum_{j \in \mathcal{C}} A^*_j(t)$. The backlog and delay for the set of flows are defined by $B_\mathcal{C}(t) = D_\mathcal{C}(t) - A_\mathcal{C}(t)$ and $W_\mathcal{C}(t) = \inf\{d \geq 0 \mid A_\mathcal{C}(t-d) \leq D_\mathcal{C}(t)\}$, respectively.

[1] A function E is called an *envelope* for a function f if $E(t) \geq f(t+\tau) - f(\tau)$ for all $t, \tau \geq 0$, or, equivalently, $f(t) \leq E * f(t)$, for all $t \geq 0$.

[2] A function f is subadditive if $f(x+y) \leq f(x) + f(y)$, for all $x, y \geq 0$.

If the above assumptions are satisfied, the following definition from [5] specifies a bound on the arrivals for an aggregate set of flows.

Definition 1: Given a set $\mathcal{C}$ of flows that satisfy assumptions (A1)–(A2), an *effective envelope* for $A_{\mathcal{C}}$ is a function $\mathcal{G}^{\varepsilon}_{\mathcal{C}}$ such that for all t: [3]

$$Pr\Big[A_{\mathcal{C}}(t) \leq \mathcal{G}^{\varepsilon}_{\mathcal{C}}(t)\Big] \geq 1-\varepsilon\,. \qquad (1)$$

Thus, an effective envelope provides a bound for arrivals in the time interval $[0,t)$, which is violated with probability ε.

With the effective envelope, we can state probabilistic bounds on properties of aggregates of flows, for example, bounds on the output from a node, the backlog and the delay at a node. Further, we can formulate deterministic network service curves for aggregates of flows.

Theorem 2: **Statistical Network Calculus for Groups of Flows.** Given a set $\mathcal{C}$ of flows, let $\mathcal{G}^{\varepsilon}_{\mathcal{C}}$ be an effective envelope for the arrivals from $\mathcal{C}$. Let $S_{\mathcal{C}}$ ($\overline{S}_{\mathcal{C}}$) be a minimum (maximum) service curve which gives a deterministic bound on the service allocated to the aggregate of the flows in $\mathcal{C}$. The following hold:

1. *Output Envelope:* The function $\mathcal{G}^{\varepsilon}_{\mathcal{C}} \oslash S_{\mathcal{C}}$ is an effective envelope for the departures, that is, for all $t,\tau \geq 0$, we have $Pr\left\{D_{\mathcal{C}}(t,t+\tau) \leq \mathcal{G}^{\varepsilon}_{\mathcal{C}} \oslash S_{\mathcal{C}}(\tau)\right\} \geq 1-\varepsilon$.
2. *Backlog Bound:* A probabilistic bound for the backlog is given by $b_{max} = \mathcal{G}^{\varepsilon}_{\mathcal{C}} \oslash S_{\mathcal{C}}(0)$, in the sense that $Pr\left\{B_{\mathcal{C}}(t) \leq b_{max}\right\} \geq 1-\varepsilon$ for all $t \geq 0$.
3. *Delay Bound:* A probabilistic bound for the delay is given by $d_{max} = \inf\left\{d \geq 0 \mid \forall t \geq 0 : \mathcal{G}^{\varepsilon}_{\mathcal{C}}(t-d) \leq S_{\mathcal{C}}(t)\right\}$, in the sense that $Pr\left\{W_{\mathcal{C}}(t) \leq d_{max}\right\} \geq 1-\varepsilon$ for all $t \geq 0$.
4. *Network Service Curve:* If the set of flows $\mathcal{C}$ passes through H network nodes in series and is offered minimum and maximum service curves $S^i_{\mathcal{C}}$ and $\overline{S}^i_{\mathcal{C}}$, respectively, at each node $i = 1,\ldots,H$, then minimum and maximum network service curves are given by

$$\begin{aligned} S^{net}_{\mathcal{C}} &= S^1_{\mathcal{C}} * S^2_{\mathcal{C}} * \ldots * S^H_{\mathcal{C}}\,, \\ \overline{S}^{net}_{\mathcal{C}} &= \overline{S}^1_{\mathcal{C}} * \overline{S}^2_{\mathcal{C}} * \ldots * \overline{S}^H_{\mathcal{C}}\,. \end{aligned}$$

[3] This definition corresponds to the local effective envelope in [5]. Since the global effective envelope defined in [5] will not be used in this paper, we drop the attribute.

The proof of the theorem, given in [19], follows the proof for the deterministic calculus (e.g., [1]), with an appropriate probabilistic argument inserted. Since service curves $S_{\mathcal{C}}$ and $\overline{S}_{\mathcal{C}}$ are deterministic service bounds for the aggregates of flows, but not for individual flows in $\mathcal{C}$, the service received by a given single flow can be worse than the service given to the set of flows as a whole. The calculus presented in the next subsection allows us to express probabilistic service guarantees to individual flows.

B. Statistical Network Calculus for Flows

We next introduce the notion of an effective service curve as a probabilistic bound on the service given to a single flow $j \in \mathcal{C}$. (To simplify notation, we will drop the subscript in A_j, A^*_j, and $\mathcal{S}^{\varepsilon}_j$.)

Definition 2: Given a flow with arrival function A, which satisfies assumptions (A1)–(A2), a *(minimum) effective service curve* is a function $\mathcal{S}^{\varepsilon}$ that satisfies for all $t \geq 0$,

$$\Pr\Big[D(t) \geq A * \mathcal{S}^{\varepsilon}(t)\Big] \geq 1-\varepsilon\,. \qquad (2)$$

A maximum effective service curve can be defined analogous to Definition 2.

The following theorem phrases the main results for the network calculus in terms of effective service curves.

Theorem 3: **Statistical Network Calculus for Flows.** Given the arrival function A of a flow with arrival envelope A^* and given an effective service curve $\mathcal{S}^{\varepsilon}$, the following hold:

1. *Output Envelope:* The function $A^* \oslash \mathcal{S}^{\varepsilon}$ is an effective envelope for the departures, that is, $Pr\left\{D(t,t+\tau) \leq A^* \oslash \mathcal{S}^{\varepsilon}(\tau)\right\} \geq 1-\varepsilon$ for all $t,\tau \geq 0$.
2. *Backlog Bound:* A probabilistic bound for the backlog is given by $b_{max} = A^* \oslash \mathcal{S}^{\varepsilon}(0)$, in the sense that $Pr\left\{B(t) \leq b_{max}\right\} \geq 1-\varepsilon$ for all $t \geq 0$.
3. *Delay Bound:* A probabilistic bound for the delay is given by $d_{max} = \inf\left\{d \geq 0\,;\ \forall t \geq 0 : A^*(t-d) \leq \mathcal{S}^{\varepsilon}(t)\right\}$, in the sense that $Pr\left\{W(t) \leq d_{max}\right\} \geq 1-\varepsilon$ for all $t \geq 0$.
4. *Network Service Curve:* If the flow passes through H network nodes in series and is offered minimum and maximum service curves $\mathcal{S}^{h,\varepsilon_h}$ ($\overline{\mathcal{S}}^{h,\varepsilon_h}$), respectively, at each node $h = 1,\ldots,H$.

Assuming that the events $D^1(t) < \mathcal{S}^{1,\varepsilon_1} * A^1(t)$, $D^2(t) < \mathcal{S}^{2,\varepsilon_2} * A^2(t), \ldots$, and $D^H(t) < \mathcal{S}^{H,\varepsilon_H} * A^H(t)$ are independent for all $t \geq 0$, then minimum and maximum effective network service curves are given by

$$\mathcal{S}^{net,\varepsilon_1+\varepsilon_2+\ldots+\varepsilon_H} = \mathcal{S}^{1,\varepsilon_1} * \mathcal{S}^{2,\varepsilon_2} * \ldots * \mathcal{S}^{H,\varepsilon_H},$$
$$\overline{\mathcal{S}}^{net,\varepsilon_1+\varepsilon_2+\ldots+\varepsilon_H} = \overline{\mathcal{S}}^{1,\varepsilon_1} * \overline{\mathcal{S}}^{2,\varepsilon_2} * \ldots * \overline{\mathcal{S}}^{H,\varepsilon_H}.$$

That is, for all $t \geq 0$,

$$\begin{aligned} Pr\{D(t) \geq & \\ & A^1 * (\mathcal{S}^{1,\varepsilon_1} * \mathcal{S}^{2,\varepsilon_2} * \ldots * \mathcal{S}^{H,\varepsilon_H})(t)\} \\ & \geq 1 - (\varepsilon_1 + \varepsilon_2 + \ldots + \varepsilon_H), \\ Pr\{D(t) \leq & \\ & A^1 * (\overline{\mathcal{S}}^{1,\varepsilon_1} * \overline{\mathcal{S}}^{2,\varepsilon_2} * \ldots * \overline{\mathcal{S}}^{H,\varepsilon_H})(t)\Big\} \\ & \geq 1 - (\varepsilon_1 + \varepsilon_2 + \ldots + \varepsilon_H). \end{aligned}$$

The proof of the theorem is given in [19]. We note that the additional assumption stated in the last part of the theorem may not hold, since service guarantee violations, as given by the effective service curves, at different nodes can be correlated. Assumptions, such as ours, which assume independence of events at different nodes are common in the analysis of network properties. The assumption is justified by the substantial effort that is otherwise required to keep track of the dependencies between events in a network. Such an effort is generally considered to be not practical.

In the next section we show how effective service curves can make probabilistic statements about service guarantees for individual flows in a network where deterministic service curves $\mathcal{S}_\mathcal{C}$ (and $\overline{S}_\mathcal{C}$) are provisioned to flow aggregates. Due to statistical multiplexing, we expect the minimum effective service curve of a single flow in a set with N flows to be significantly greater than $\mathcal{S}_\mathcal{C}/N$ for large N.

IV. Construction of Effective Service Curves

In this section, we present the construction of an effective service curve for a single flow in a network where bandwidth is allocated to aggregates of flows. The effective service curve is determined from the unused bandwidth that is allocated to the aggregate of flows.

We consider a set $\mathcal{C} = \{1, 2, \ldots, N\}$ of N flows which satisfy assumptions (A1)–(A2). We will construct an effective service curve for flow j with arrival function A_j. We will use the following notation: $A_N = \sum_{k\in\mathcal{C}} A_k$, $A_{N-j} = \sum_{k\in\mathcal{C},k\neq j} A_k$.

We do not require that the flows have the same arrival envelopes. However, if flows are homogeneous, one can interpret A_j and A_{N-j} as "any single flow" and "any subset of $N-1$ flows", respectively, from the set $\mathcal{C}$.

We assume that the aggregate set of N flows is allocated a (deterministic) minimum service curve, denoted by S_N, and the set $\mathcal{C} - \{j\}$ is allocated a maximum service curve of $\overline{S}_{N-j}$. We let $\mathcal{G}^\varepsilon_{N-j}$ denote an effective envelope for the arrivals from $\mathcal{C} - \{j\}$, With this notation, an effective service curve for flow i is given by the next theorem.

Theorem 4: The function

$$\mathcal{S}^\varepsilon_j(t) = [S_N - \mathcal{G}^\varepsilon_{N-j} * \overline{S}_{N-j}]_+(t)$$

is a (minimum) effective service curve for flow $j \in \mathcal{C}$.[4]

A proof of the theorem is given in the Appendix. The above effective service curve does not assume knowledge of the scheduling algorithm used to determine the order of transmission of the aggregate of N flows. Thus, the effective service curve is expected to be pessimistic for most scheduling algorithms, including FIFO. The effective service curve is least conservative if flows in the set $\mathcal{C} - \{j\}$ are transmitted with higher priority than flow j.

The effective service curve in the theorem has a corresponding version in the deterministic network calculus, which is given by $S_1(t) = [S_N - A^*_{N-1} * \overline{S}_{N-1}]_+(t)$ with $A^*_{N-1} = \sum_{i_k\in\mathcal{C},i_k\neq i_1} A^*_k$. However, this deterministic service curve will be positive only for large values of t [7].

The following corollary states looser bounds on the effective service to flow j.

Corollary 1: Using the same notation as in Theorem 4, and assuming that $\overline{S}_N(t) \geq \overline{S}_{N-j}(t)$, the following are (minimum) effective service curves for flow $j \in \mathcal{C}$.

1. $\mathcal{S}^\varepsilon_j = [S_N - \mathcal{G}^\varepsilon_{N-j} * \overline{S}_N]_+$,
2. $\mathcal{S}^\varepsilon_j = [S_N - \mathcal{G}^\varepsilon_N * \overline{S}_N]_+$,
3. $\mathcal{S}^\varepsilon_j = [S_N - \mathcal{G}^\varepsilon_{N-j}]_+$,
4. $\mathcal{S}^\varepsilon_j = [S_N - \mathcal{G}^\varepsilon_N]_+$.

[4] We use "$[f]_+(t) = \max\{f(t), 0\}$".

Proof. The first two service curves follow from $\overline{S}_N(t) \geq \overline{S}_{N-j}(t)$ and since $\mathcal{G}_N^\varepsilon$ can always be used instead of $\mathcal{G}_{N-j}^\varepsilon$. The last two service curves in addition exploit that $f(t) \geq f * g(t)$, which follows from the definition of the convolution operator. □

Thus, effective service curves for single flows can be determined even if only information about the aggregate reservations to a set of N flows is available. In our numerical examples, we will only work with the last and most pessimistic effective service curve. We will show that even with these very loose bounds, we are able to extract a significant amount of the multiplexing gain if the number of flows is large.

V. Effective Envelopes for Heterogeneous Traffic

The presentation of the effective envelopes $\mathcal{G}_\mathcal{C}^\varepsilon$ in Subsection III-A and Section IV does not depend on a specific arrival model, but also does not offer any guidance for constructing $\mathcal{G}_\mathcal{C}^\varepsilon$.

For our numerical examples, we construct $\mathcal{G}_\mathcal{C}^\varepsilon$ as in [5], adopting an adversarial traffic model [15], where flows can individually exhibit a worst-case arrival pattern as allowed by (A2), but sources do not conspire to construct a joint worst-case.

We take a probabilistic view of traffic, where the arrivals of a flow in the time interval $[0, t)$ are given by a random process $A(t)$. In addition to assumptions (A1) and (A2) from Section II, we assume that the following hold for the arrival processes.

(A3) Stationarity. The arrival processes are *stationary*, i.e., $\forall \tau_1, \tau_2 \geq 0$ we have $Pr[A(\tau_1, \tau_1 + t) \leq x] = Pr[A(\tau_2, \tau_2 + t) \leq x]$.

(A4) Independence. The arrivals from two flows $i, j \in \mathcal{C}$, A_i and A_j, are stochastically independent.

The construction of effective envelopes $\mathcal{G}_\mathcal{C}^\varepsilon$ for a set $\mathcal{C}$ of flows in [5] uses the moment generating function of A_j, denoted as $M_j(s,t) = E[e^{A_j(\tau,\tau+t)s}]$. As shown in [5], if assumptions (A1)–(A4) hold, we obtain $M_j(s,t) \leq \overline{M}_j(s,t)$, where

$$\overline{M}_j(s,t) \;=\; 1 + \frac{\rho_j\, t}{A_j^*(t)}\big(e^{s\,A_j^*(t)} - 1\big)\;, \qquad (3)$$

and where $\rho_j := \lim_{t\to\infty} A_j^*(t)/t$ is assumed to exist.

Since the effective envelopes in [5] assume homogeneity of flows, that is, all flows in $\mathcal{C}$ have the same arrival envelope, we briefly discuss an adaptation of the results to heterogeneous flows, with different arrival envelopes [8].

With assumption (A4) and with the bound in Eqn. (3), we obtain from the Chernoff bound that

$$Pr[A_\mathcal{C}(t) \geq x] \leq e^{-xs} \prod_{j\in\mathcal{C}} \overline{M}_j(s,t)\;.$$

Setting the right hand side equal to ε and solving for x gives

$$x = \frac{1}{s}\big(\sum_{j\in\mathcal{C}} \log \overline{M}_j(s,t) + \log \varepsilon^{-1}\big)\;. \qquad (4)$$

Any choice of s yields a point of an effective envelope for the arrivals from $\mathcal{C}$. We select the value of the effective envelope at t to be

$$\mathcal{G}_\mathcal{C}^\varepsilon(t) \;=\; \min_s \frac{1}{s}\big(\sum_{j\in\mathcal{C}} \log \overline{M}_j(s,t) + \log \varepsilon^{-1}\big)\;.$$

With this choice, $\mathcal{G}_\mathcal{C}^\varepsilon(t) \leq A_\mathcal{C}^*(t)$ is always satisfied. Since the right hand side of Eqn. (4), as a function of s, has at most one minimum, which can be found by searching for the zero of the derivative [8].

VI. Evaluation

We now present numerical examples which demonstrate different applications of effective service curves, and evaluate the statistical multiplexing gain feasible with effective service curves from Section IV.

We assume that individual flows are regulated at the entrance to the network, using a peak rate limited leaky bucket with arrival envelope $A_j^*(\tau) = \min\{P_j\tau, \sigma_j + \rho_j\tau\}$ for flow j, where $P_j \geq \rho_j$ is the peak rate, ρ_j is the average rate, and σ_j is a burst size parameter. We consider two types of flows with parameters as given in the following table.

Type	Peak Rate P_j (Mbps)	Mean Rate ρ_j (Mbps)	Burst Size σ_j (bits)
Type 1	1.5	0.15	95400
Type 2	6.0	0.15	10345

The parameters are selected to be equal to those in [4], [15] and other studies.

We assume that the arrivals satisfy assumptions (A1)–(A4), and we construct effective envelopes as shown in Section V.

We assume that capacities are allocated to aggregates of flows, in terms of deterministic service

curves S_N and $\overline{S}_N$ for a set of N flows. We will assume that service curves for the aggregate have a very simple constant-rate form, such as $S_N(t) = Nc\ t$ ($c > 0$), where c is referred to as 'per-flow capacity'. For the construction of effective service curves $\mathcal{S}_j^\varepsilon$, we use the most conservative bound from Corollary 1, i.e., $\mathcal{S}_j^\varepsilon = [S_N - \mathcal{G}_N^\varepsilon]_+$. This bound does not require a maximum service curve (as used in Theorem 4) and merely requires us to calculate the effective envelope $\mathcal{G}_N^\varepsilon$.

We compare the results obtained with effective service curves to the following non-statistical per-flow service provisioning schemes.

- A *peak rate* allocation, where each flow j has a service curve of $S_j(t) = P_j\ t$, provides an upper bound for the amount of resources reserved for a flow.
- An *average rate* allocation, where each flow j has a service curve of $S_j(t) = \rho_j\ t$, is a lower bound for the amount of resources reserved.
- A *deterministic* allocation delivers worst-case delay guarantees. The resources allocated to a flow are determined by the smallest (deterministic) constant-rate service curve $S_j(t) = \hat{c}_j\ t$ that satisfies the delay bound d, i.e.,

$\hat{c}_j = \inf\{c \geq 0 \mid \forall t \geq 0 : A^*(t-d) \leq c\ t\}$.

A. Example 1: Single Node

We investigate arrivals from a group of N Type-1 flows at a single node. The delay guarantee of the flows is given by $d = 50$ ms.

We first compare the shape of effective service curves for different values of N and for $\varepsilon = 10^{-9}$, with the deterministic service curves.

We assume that the capacity allocated for the aggregate of N flows, is $S_N(t) = N\hat{c}\,t$, where $\hat{c} \approx 0.8785$ is the constant-rate service curve required by a Type-1 flow to satisfy a delay bound of $d = 50$ ms according to the deterministic allocation given above. Then, according to Corollary 1, the effective service curve of any single flow from the set is given by $\mathcal{S}_1^\varepsilon(t) = [N\hat{c}\ t - \mathcal{G}_N^\varepsilon(t)]_+$. In Figure 2 we plot effective service curves $\mathcal{S}_1^\varepsilon$ for different values of N, and compare it to the deterministic service curve $S_1 = \hat{c}\ t$. The figure shows that for large N ($N \geq 100$), the effective service curve is significantly larger than the deterministic service curve. For small values of N, i.e., $N \leq 30$, there is not sufficient multiplexing gain. In those cases, the conservative bound of the effective service curve is inferior to a deterministic service curve, or improves upon the deterministic service curve only for large values of t. To explain the change of slope of the curve for $N = 30$ at $t \approx 70$ ms, we note that at $\sigma/(P-\rho) \approx 70$ ms, the arrival envelope of Type-1 flows changes from $P\ t$ to $\rho\ t$.

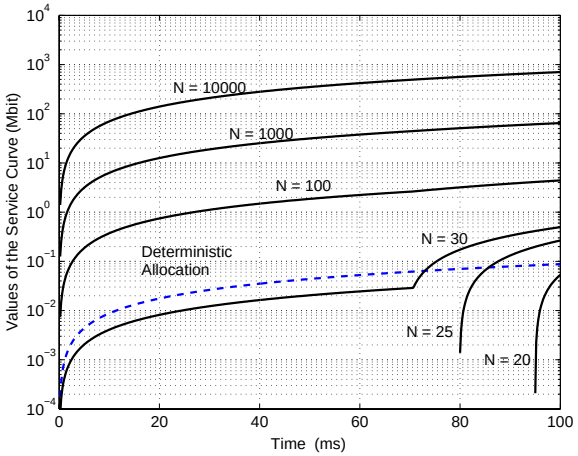

Fig. 2. Example 1: Effective vs. deterministic service curve as a function of time. Effective service curves are shown for different values of N, and are calculated for $\varepsilon = 10^{-9}$.

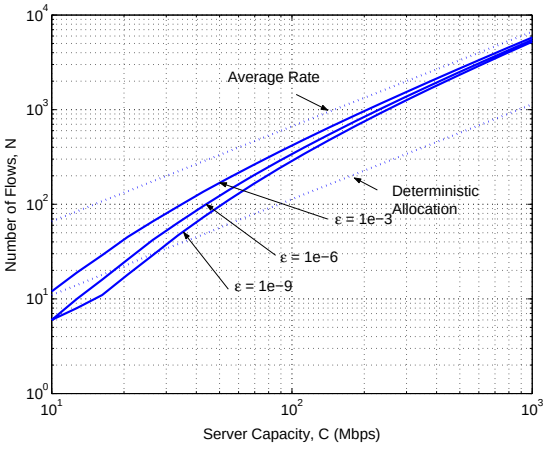

Fig. 3. Example 1: Number of flows admitted on a link with capacity C to satisfy a delay bound of $d = 50$ ms.

Next we compare the number of flows that can be provisioned on a link with capacity C, again using a delay bound of $d = 50$ ms. For deterministic allocation, we use the same service curve as before, i.e., $S_1(t) \approx 0.8785\ t$. For the effective service curve we find the largest N such that $\mathcal{S}_1^\varepsilon(t) = [C\ t - \mathcal{G}_N^\varepsilon(t)]_+$ assures, via Theorem 3, the delay bound d with prob-

ability $1-\varepsilon$. The results are shown in Figure 3, where we include plots for effective service curves with $\varepsilon = 10^{-3}, 10^{-6}, 10^{-9}$. We also include results for an average rate allocation (which does not satisfy the delay bound). The graphs show that, even for ε very small, the effective service curve shows significant statistical multiplexing gain, as C is increased, and for $C \geq 30$ Mbps we can admit more flows than through deterministic allocation. For large C, the plots for effective service curves and average rate allocation become close. For small C, on the other hand, the number of flows is too small to extract multiplexing gain, and, consequently, the effective service curve can be inferior to a deterministic service curve.

B. Example 2: Multiple Nodes with Cross Traffic

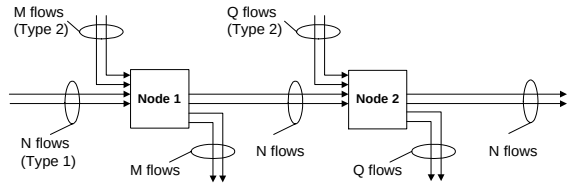

Fig. 4. Example 2: A network with 2 nodes and with cross traffic at each node. The cross-traffic consists of M Type-2 flows at the first node, and Q Type-2 flows at the second node.

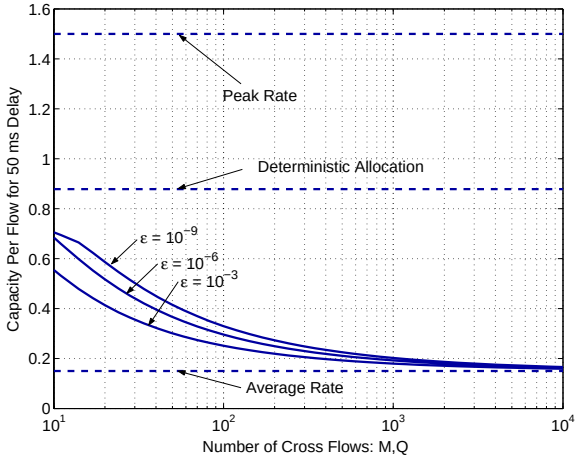

Fig. 5. Example 2: Required capacity for each Type-1 flow to reach an end-to-end delay bound of $d = 50$ ms, as a function of the number of Type-2 cross flows.

We assume a network with two nodes, as shown in Figure 4, and determine the multiplexing gain attainable with effective service curves for a set of N flows through these two nodes, with end-to-end delay bound of $d = 50$ ms. There is cross traffic from M flows at the first node, and from Q flows at the second node. The flows with traffic through both nodes are assumed to be of Type 1, and cross traffic at both nodes is of Type 2.

Consider one Type-1 flow which passes through Node 1 and Node 2. Using Corollary 1, the effective service curves of this flow, denoted by, $\mathcal{S}_1^{1,\varepsilon}$ and $\mathcal{S}_1^{2,\varepsilon}$, are given by $\mathcal{S}_1^{1,\varepsilon} = [S_{N+M}^1 - \mathcal{G}_{N+M}^{1,\varepsilon}]_+$ and $\mathcal{S}_1^{2,\varepsilon} = [S_{N+Q}^2 - \mathcal{G}_{N+Q}^{2,\varepsilon}]_+$, respectively. $\mathcal{G}_{N+M}^{1,\varepsilon}$ is the effective envelope of the $N+M$ flows at the first node, and $\mathcal{G}_{N+Q}^{2,\varepsilon}$ is the effective envelope of the $N+Q$ flows at the second node.

We assume that $S_{N+M}^1(t) = \tilde{c}(N+M)\ t$ and $S_{N+Q}^2(t) = \tilde{c}(N+Q)\ t$ are the deterministic service curves allocated to the aggregate of Type-1 and Type-2 flows, at Node 1 and Node 2, respectively. Here, $\tilde{c} > 0$ is selected as the smallest value such that the effective network service curve of a Type-1 flow, $\mathcal{S}_1^{1,\varepsilon} * \mathcal{S}_1^{2,\varepsilon}$, satisfies a probabilistic end-to-end delay bound of $d = 50$ ms, according to Theorem 3. Recall, from Theorem 3 that the probability of an end-to-end delay bound violation is 2ε.

Calculating $\mathcal{G}_{N+Q}^{2,\varepsilon}$ raises a technical problem. If we calculate the effective envelope using Eqn. (4), we need a (deterministic) arrival envelope for each Type-1 flow at the second node. From Theorem 3, we know that $A_1^* \oslash \mathcal{S}_1^{1,\varepsilon}$ is an effective envelope for the departures of a Type-1 flow from the first node.[5] To turn this envelope into a deterministic envelope, we require a policer for Type-1 flows in front of the second node, which discards all traffic exceeding $A_1^* \oslash \mathcal{S}_1^{1,\varepsilon}$.

Remark: An alternative approach, which does not require policing at the second node, sets $\mathcal{G}_{N+Q}^{2,\varepsilon} = \mathcal{G}_N^{2,\varepsilon} + \mathcal{G}_Q^{2,\varepsilon}$, where $\mathcal{G}_Q^{2,\varepsilon}$ is the effective envelope of the Q Type-2 flows and $\mathcal{G}_N^{2,\varepsilon} = N\ \left(A_1^* \oslash \mathcal{S}_1^{1,\varepsilon}\right)$. In a variation of this approach, we can calculate an effective service curve for all Type-1 flows at the first node, by $\mathcal{S}_N^{1,\varepsilon} = [S_{N+M}^1 - \mathcal{G}_M^{1,\varepsilon}]_+$, where $\mathcal{G}_M^{1,\varepsilon}$ is the effective envelope of the M Type-2 flows at the first node. In this variation we obtain $\mathcal{G}_N^{2,\varepsilon} = \left(N \cdot A_1^*\right) \oslash \mathcal{S}_N^{1,\varepsilon}$.

In Figure 5, we depict the required per-flow capac-

[5] In this example, we use A_1^* and A_2^* to denote the arrival envelope of a Type-1 and a Type-2 flow, respectively.

ity $\tilde{c}$ for Type-1 flows to satisfy a probabilistic delay bound as a function of the number of cross flows. We include results for $\varepsilon = 10^{-3}, 10^{-6}, 10^{-9}$. In the figure, we set $N = 1$; hence, the multiplexing gain is collected exclusively from cross traffic. The number of cross traffic flows is assumed to be identical at both nodes, that is, $M = Q$, and is varied from 10 to 10,000 flows. We also consider results for a peak rate allocation with $P = 1.5$ Mbps, average rate allocation with $\rho = 0.15$ Mbps, and a deterministic allocation with $\hat{c} \approx 0.8785$ Mbps.

Figure 5 illustrates the significant bandwidth savings attainable with effective service curves. Even when $\varepsilon = 10^{-9}$, the required bandwidth to satisfy an end-to-end delay bound of $d = 50$ ms is close to an average rate allocation when the number of cross flows is large.

C. Example 3: Multiple Nodes With No Cross Traffic

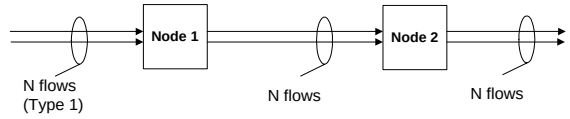

Fig. 6. Example 3: A network with 2 nodes and no cross traffic.

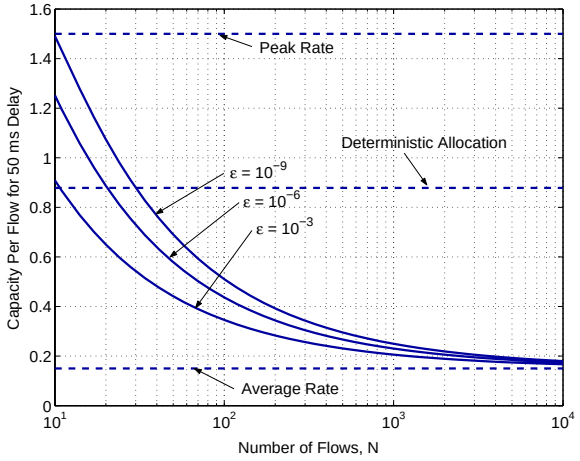

Fig. 7. Example 3: Capacity per flow needed to support a delay bound of $d = 50$ ms as a function of the total number of flows.

We consider the two-node network shown in Figure 6 with no cross traffic, with N Type-1 flows passing through both nodes. We will again evaluate the per-flow capacity needed at each node to satisfy a probabilistic or deterministic end-to-end delay bound of $d = 50$ ms. Similar to Example 2, we set $S_N^1(t) = S_N^2(t) = \tilde{c}N\ t$ to be the deterministic service curves allocated to the flows at the two nodes, where $\tilde{c} > 0$ is set to be the smallest rate such that the end-to-end delay bounds are satisfied.

Here, we could proceed as in the previous examples, that is, we could use the effective service curves given by $\mathcal{S}_1^{1,\varepsilon} = [S_N^1 - \mathcal{G}_N^{1,\varepsilon}]_+$ and $\mathcal{S}_1^{2,\varepsilon} = [S_N^2 - \mathcal{G}_N^{2,\varepsilon}]_+$, and determine $\tilde{c} > 0$ as the smallest number for which the delay bound of $d = 50$ ms is satisfied, according to Theorem 3. The functions $\mathcal{G}_N^{1,\varepsilon}$ and $\mathcal{G}_N^{2,\varepsilon}$ are effective envelopes for the arrivals of the N flows at the first and second node, respectively. Since there is no cross traffic, $\mathcal{G}_N^{2,\varepsilon} = \mathcal{G}_N^{1,\varepsilon} \oslash S_N^1$, is, with Theorem 3, an effective envelope for the arrivals at the second node. Hence, the difficulty in Example 2 with the construction of effective envelopes at downstream nodes, which required to assume policing of flows between nodes, does not arise.

However, the analysis of this example can be much simplified, by referring to the deterministic network calculus. From Theorem 1 and from $f = f * f$ for all functions f [3], we obtain that $S_N^{net} = S_N^1 * S_N^2$. With this deterministic network service curve, we can now proceed as in Example 1, and construct a network effective service curve from $\mathcal{S}_1^{net,\varepsilon}(t) = [\tilde{c}N\ t - \mathcal{G}_N^{1,\varepsilon}(t)]_+$. [6]

In Figure 7, we show the per-flow capacity $\tilde{c}$ required to satisfy a probabilistic delay bound, as a function of the number of flows. As in Example 2, the results illustrate that the bandwidth requirements of a flow approach the average rate, as the number of flows is increased.

VII. Conclusions

We have presented a network calculus for statistically multiplexed traffic, which introduces the notion of *effective service curves* as a probabilistic bound on the service received by individual flows in a network. We have shown that many of the results from the deterministic network calculus can be carried over to the statistical framework, by inserting an appropriate probabilistic arguments. Through the use of effective service curves, we are able to describe the service delivered to individual flows when capacities are allocated to aggregates of flows.

[6] We point out that the simplified analysis yields a better multiplexing gain, and we refer to [19] for a comparison.

As directions for future work, the calculus in this paper may be sufficient to provision end-to-end delays in feedforward networks, however, it is not directly applicable to general networks. Further, it is desirable to find explicit bounds for the effective envelope of heterogeneous arrivals (Section V) that can replace the numerical method used in this paper. The requirement for policing mechanisms between nodes (as in Example 2) is undesirable. It can be eliminated by determining an effective arrival envelope for an aggregate of flows from probabilistic envelopes for individual flows, i.e., without the need for a deterministic envelope for each flow at downstream nodes. Finally, the service curves considered in this paper are quite simple, and the benefits of more sophisticated service curves to the multiplexing gain have not been explored.

ACKNOWLEDGMENTS

The authors gratefully acknowledge the help from Almut Burchard with Section V, and numerous discussions with Jianping Wang and Erhan Yilmaz.

APPENDIX

I. PROOF OF THEOREM 4

We will show that $\mathcal{S}_j^\varepsilon$ in Theorem 4 satisfies Definition 2. Without loss of generality we will conduct the proof for the first flow in set $\mathcal{C}$, that is, we set $j = 1$.

$$\begin{aligned} D_1(t) &= D_N - D_{N-1}(t) && (5)\\ &\geq A_N * S_N(t) - A_{N-1} * \overline{S}_{N-1}(t) && (6)\\ &= \inf_{x\leq t}\left[A_N(t-x) + S_N(x)\right]\\ &\quad - \inf_{y\leq t}\left[A_{N-1}(t-y) + \overline{S}_{N-1}(y)\right], && (7)\end{aligned}$$

where Eqn. (6) follows from the definition of minimum and maximum service curves, and Eqn. (7) merely expands the operators.

Suppose the minimum of the left term in Eqn. (7) is attained at $x = \hat{x}$ and that the minimum of the right term is attained at $y = \hat{y}$. We now distinguish two cases: (1) $\hat{y} < \hat{x}$, and (2) $\hat{y} \geq \hat{x}$.

Case 1: $\hat{y} < \hat{x}$**.** With this assumption, we can write Eqn. (7) as

$$\begin{aligned} D_1(t) &\geq \inf_{x\leq t}\Big[A_N(t-x) + S_N(x)\\ &\quad - \inf_{y\leq t}\left[A_{N-1}(t-y) + \overline{S}_{N-1}(y)\right]\Big] && (8)\\ &= \inf_{x\leq t}\Big[A_N(t-x) + S_N(x)\\ &\quad - \inf_{y\leq x}\left[A_{N-1}(t-y) + \overline{S}_{N-1}(y)\right]\Big] && (9)\\ &= \inf_{x\leq t}\Big[(A_1(t-x) + A_{N-1}(t-x)) + S_N(x)\\ &\quad - \inf_{y\leq x}\left[(A_{N-1}(t-x)\right.\\ &\quad \left.+A_{N-1}(t-x,t-y)) + \overline{S}_{N-1}(y)\right]\Big] && (10)\\ &= \inf_{x\leq t}\Big[A_1(t-x) + S_N(x)\\ &\quad - \inf_{y\leq x}\left[A_{N-1}(t-x,t-y) + \overline{S}_{N-1}(y)\right]\Big], && (11)\end{aligned}$$

where Eqn. (10) uses the equalities $A_N(t-x) = A_1(t-x) + A_{N-1}(t-x)$ and $A_{N-1}(t-y) = A_{N-1}(t-x) + A_{N-1}(t-x,t-y)$. Using Definition 1, we obtain from Eqn. (11) that

$$\begin{aligned} 1-\varepsilon &\leq Pr\Big\{D_1(t) \geq \inf_{x\leq t}\Big[A_1(t-x) + S_N(x)\\ &\quad - \inf_{y\leq x}\left[\mathcal{G}^\varepsilon_{N-1}(y-x) + \overline{S}_{N-1}(y)\right]\Big]\Big\} && (12)\\ &= Pr\Big\{D_1(t) \geq \inf_{x\leq t}\left[A_1(t-x)\right.\\ &\quad \left.+\left[S_N - \mathcal{G}^\varepsilon_{N-1} * \overline{S}_{N-1}\right](x)\right]\Big\} && (13)\\ &= Pr\Big\{D_1(t) \geq A_1 * \left[S_N - \mathcal{G}^\varepsilon_{N-1} * \overline{S}_{N-1}\right](t)\Big\} && (14)\end{aligned}$$

Eqn. (12) merely applies Definition 1, and Eqs. (13) and (14) use the definition of the convolution operator.

Case 2: $\hat{y} \geq \hat{x}$**.** We rewrite Eqn. (7) as

$$\begin{aligned} D_1(t) &\geq \inf_{x\leq t}\Big[A_N(x) + S_N(t-x)\\ &\quad - \inf_{y\leq t}\left[A_{N-1}(y) + \overline{S}_{N-1}(t-y)\right]\Big] && (15)\\ &= \inf_{x\leq t}\Big[A_N(x) + S_N(t-x)\\ &\quad - \inf_{x\leq y\leq t}\left[A_{N-1}(y) + \overline{S}_{N-1}(t-y)\right]\Big] && (16)\\ &= \inf_{x\leq t}\Big[(A_1(x) + A_{N-1}(x)) + S_N(t-x) -\\ &\quad - \inf_{x\leq y\leq t}\left[(A_{N-1}(x) + A_{N-1}(x,y))\right.\\ &\quad \left.+\overline{S}_{N-1}(t-y)\right]\Big]. && (17)\end{aligned}$$

Eqn. (15) is a simple manipulation. Eqn. (16) follows with $\hat{y} \geq \hat{x}$. Eqn. (17) uses $A_N(x) = A_1(x) +$

$A_{N-1}(x)$ and $A_{N-1}(y) = A_{N-1}(x) + A_{N-1}(x,y)$. With the properties of $\mathcal{G}^{\varepsilon}_{N-1}$, we obtain from Eqn. (17) that

$$\begin{aligned} 1-\varepsilon \;\leq\; & Pr\Big\{ D_1(t) \geq \inf_{x\leq t} \big[A_1(x) + S_N(t-x) \\ & - \inf_{x\leq y\leq t} \left[\mathcal{G}^{\varepsilon}_{N-1}(y-x) + \overline{S}_{N-1}(t-y)\right]\Big]\Big\} \quad (18) \\ = & Pr\Big\{ D_1(t) \geq \inf_{x\leq t} \big[A_1(x) \\ & + \left[S_N - \mathcal{G}^{\varepsilon}_{N-1} * \overline{S}_{N-1}\right](t-x)\big]\Big\} \quad (19) \\ = & Pr\Big\{ D_1(t) \geq A_1 * \left[S_N - \mathcal{G}^{\varepsilon}_{N-1} * \overline{S}_{N-1}\right](t)\Big\}. \\ & \quad (20) \end{aligned}$$

Note that Eqn. (18) uses Definition 1 and Eqn. (19) follows from

$$\begin{aligned} & \inf_{x\leq y\leq t}\Big[\mathcal{G}^{\varepsilon}_{N-1}(y-x) + \overline{S}_{N-1}(t-y)\Big] \\ & \quad = \inf_{0\leq y\leq t-x}\Big[\mathcal{G}^{\varepsilon}_{N-1}(y) + \overline{S}_{N-1}(t-x-y)\Big], \end{aligned}$$

and the convolution operator. Eqn. (20) is another use of the convolution operator.

The above derivations also hold if the minima $\hat{y}$ and $\hat{x}$ are not attained by using minimizing sequences [19].

Finally, since $D_1(t) \geq 0$ with probability one, the theorem follows. □

REFERENCES

[1] R. Agrawal, R. L. Cruz, C. Okino, and R. Rajan. Performance bounds for flow control protocols. *IEEE/ACM Transactions on Networking*, 7(3):310–323, June 1999.

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

[3] F. L. Baccelli, G. Cohen, G. J. Olsder, and J.-P. Quadrat. *Synchronization and Linearity: An Algebra for Discrete Event Systems*. John Wiley and Sons, 1992.

[4] R. Boorstyn, A. Burchard, J. Liebeherr, and C. Oottamakorn. Effective envelopes: Statistical bounds on multiplexed traffic in packet networks. In *Proceedings of IEEE Infocom 2000*, pages 1223–1232, Tel Aviv, March 2000.

[5] R. Boorstyn, A. Burchard, J. Liebeherr, and C. Oottamakorn. Statistical service assurances for traffic scheduling algorithms. *IEEE Journal on Selected Areas in Communications. Special Issue on Internet QoS*, 18(12):2651–2664, December 2000.

[6] J. Y. Le Boudec. Application of network calculus to guaranteed service networks. *IEEE/ACM Transactions on Information Theory*, 44(3):1087–1097, May 1998.

[7] J. Y. Le Boudec and P. Thiran. *Network Calculus*. Springer Verlag, Lecture Notes in Computer Science, LNCS 2050, 2001.

[8] A. Burchard. Personal Communications, July 2001.

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

[10] C. S. Chang. On deterministic traffic regulation and service guarantees: a systematic approach by filtering. *IEEE/ACM Transactions on Information Theory*, 44(3):1097–1110, May 1998.

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

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

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

[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] 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.

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

[18] C. Li and E. Knightly. Coordinated network scheduling: A framework for end-to-end services. In *Proceedings of IEEE ICNP 2000*, Osaka, November 2000.

[19] J. Liebeherr and S. D. Patek. Provisioning end-to-end statistical service guarantees. Technical Report CS-2001-19, University of Virginia, Computer Science Department, July 2001. Available from http://www.cs.virginia.edu/~jorg/CS-2001-19.pdf.

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

[21] V. Sivaraman and F. M. Chiussi. Statistical analysis of delay bound violations at an earliest deadline first scheduler. *Performance Evaluation*, 36(1):457–470, 1999.

[22] V. Sivaraman and F. M. Chiussi. Providing end-to-end statistical delay guarantees with earliest deadline first scheduling and per-hop traffic shaping. In *Proceedings of IEEE Infocom 2000*, pages 603–612, Tel Aviv, March 2000.

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

[24] 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.