Stochastically Bounded Burstiness for Communication Networks

David Starobinski, *Student Member, IEEE,* and Moshe Sidi, *Senior Member, IEEE*

***Abstract*—A network calculus is developed for processes whose burstiness is stochastically bounded by *general* decreasing functions. This calculus is useful for a large class of input processes, including important processes exhibiting "subexponentially bounded burstiness" such as fractional Brownian motion. Moreover, it allows judicious capture of the salient features of real-time traffic, such as the "cell" and "burst" characteristics of multiplexed traffic. This accurate characterization is achieved by setting the bounding function as a sum of exponentials.**

***Index Terms*—Multiple time-scale traffic, network calculus, statistical bounds**

I. Introduction

IN high-speed networks [11], the stringent quality of service (QoS) requirements of real-time multimedia applications are met by special services providing guarantees. A distinction is made between deterministic and statistical services. The deterministic service [3], [10], also called guaranteed service within the Internet framework [12], is aimed at providing strict guarantees on the QoS requirements while the statistical service is intended to provide only statistical (probabilistic) guarantees. The main advantage of the statistical service over the deterministic service is that it can achieve a higher network utilization, at the expense of some minor quality degradation.

For services offering statistical guarantees, one is usually interested in bounding the mean delay of the packets through the network or the probability that the delay exceeds a certain value. For this reason, the concept of "exponentially bounded burstiness" (EBB) was introduced in [15] (see also the closely related works [2] and [7]). A continuous traffic stream with instantaneous rate $R(t)$, is EBB with upper rate ρ and bounding function $Ae^{-\alpha\sigma}$ if for *all* $\sigma \geq 0$ and $t > s \geq 0$

$$\Pr\left\{\int_s^t R(u)\,du \geq \rho(t-s)+\sigma\right\} \leq Ae^{-\alpha\sigma} \qquad (1)$$

where $\rho, A, \alpha > 0$. The EBB calculus is based on an appropriate EBB characterization of the external processes feeding the network. It is shown in [15] that if all the input processes of a feedforward communication network (and also some special nonfeedforward networks) are EBB, then, subject to throughput conditions, so are *all* the traffic streams within that network and the network is said to be stable. This model enables computation of exponential bounds (EB), at each buffer in the network, on various performance measures, such as delay survivor probabilities and mean delay.

The EBB model can be expected to provide relatively tight bounds when the delay probability distribution curve is linear on the logarithmic scale (the reasonable quality of the bounds, in such a case, has been validated in [8] and [15]). However, for a variety of models of real-time traffic, such as Markov modulated processes and other multiple time-scale processes, it has been observed that the delay distribution in a multiplexer is better described by two linear regions (see [11, Sec. 3.8] and references therein). These regions are commonly termed "cell region" and "burst region" in the literature. As shown schematically in Fig. 1, the EB bound for such kinds of traffic may be rather loose, since the rate of decrease of the EB bound cannot exceed the rate of decrease of the delay curve in the burst region. A call admission scheme based on EB bounds may therefore lead to significant network underutilization. However, it is not the only drawback of the EBB model. It turns out that a large class of traffic processes do *not* satisfy at all the characterization of exponentially bounded burstiness. This class includes all the processes exhibiting *subexponentially* bounded burstiness, i.e., burstiness stochastically bounded by functions decaying more slowly than any exponential. Many of the new models suggested in the literature for characterizing network traffic processes have subexponentially bounded burstiness. One of these models is the self-similar fractional Brownian motion (FBM) [5], [9], which has been shown to have good statistical fit with local area network (LAN) data. An FBM process has a subexponentially bounded burstiness since the distribution of the delay W in a buffer fed by such a process is bounded by a function behaving like a Weibull distribution [4], i.e., $\Pr(W \geq \sigma) \leq Ae^{-a\sigma^b}$, where $A, a > 0$ and $1 > b > 0$.

In this work, we develop a new calculus in order to deal with processes which do not satisfy the EBB characterization. We consider processes whose burstiness is stochastically bounded by *general* functions $f(\sigma)$. We refer to these processes as "stochastically bounded burstiness" (SBB) processes with bounding function $f(\sigma)$. We demonstrate the existence of a network calculus for SBB processes. This calculus allows us to prove the stability of feedforward communication networks fed by SBB processes (subject to some conditions on the bounding functions) and obtain upper bounds on the interesting performance measures. It turns out that the SBB calculus is also a powerful tool for obtaining much better bounds for multiple time-scale processes. Since the function $f(\sigma)$ is general, one can choose to bound the traffic burstiness with a sum of exponentials instead of a single exponential, as is the case in the EBB model. We develop a network calculus for sums of

Manuscript received June 4, 1998; revised July 22, 1999. This work was supported by the Israel Science Foundation administrated by the Academy of Science and Humanities.

The authors are with the Electrical Engineering Department, Technion–Israel Institute of Technology, Haifa 32000, Israel (e-mail: staro@tx.technion.ac.il; moshe@ee.technion.ac.il).

Communicated by R. Cruz, Associate Editor for Communication Networks.
Publisher Item Identifier S 0018-9448(00)00027-4.

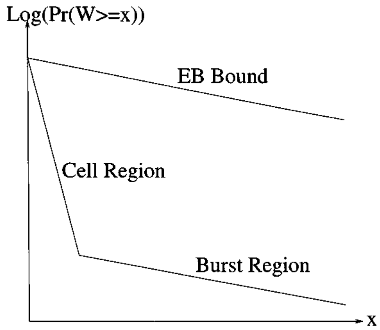

Fig. 1. Cell/burst regions, typical delay curve distribution, and EB bound.

exponentials and show that it leads to tighter bounds than those provided by the EBB calculus.

II. Model and Definitions

We consider the same network model as introduced in [15]. We assume that the system is started at time $t=0$ and that all the network queues are empty at this time. The network queues are assumed to be of infinite size. Throughout this work we derive results only for discrete-time processes $\{R(t)\}_{t\in\mathbb{N}}$ (although similar results hold for the continuous-time setting [13]). We introduce the notation $(\int_\sigma^\infty du)^n f(u)$ for denoting the n-fold integration of the function $f(u)$. Note that the result of the operation $(\int_\sigma^\infty du)^n f(u)$ is a function of the variable σ. As an example, taking $f(u)=e^{-\alpha u}$ leads to $\left(\int_\sigma^\infty du\right)^n f(u) = (1/\alpha)^n e^{-\alpha\sigma}$. We present now the following definitions.

Definition 1: Let $\mathcal{F}$ represent the set of all the functions $f(\sigma)$ such that for any order n, the n-fold integral $(\int_\sigma^\infty du)^n f(u)$ is bounded for any $\sigma \geq 0$.

Definition 2: A stochastic process $W(t)$ is stochastically bounded (SB) with bounding function $f(\sigma)$ if

i) $f(\sigma) \in \mathcal{F}$.
ii) $\Pr(W(t) \geq \sigma) \leq f(\sigma)$ for all $\sigma \geq 0$ and all $t \geq 0$.

Definition 3: A discrete-time process $R(t)$ has a stochastically bounded burstiness (SBB) with upper rate ρ and bounding function $f(\sigma)$ if

i) $f(\sigma) \in \mathcal{F}$.
ii) $\Pr\left\{\sum_{u=s+1}^{t} R(u) \geq \rho(t-s)+\sigma\right\} \leq f(\sigma)$ for all $\sigma \geq 0$ and all $0 \leq s < t$.

In this case, we say that $R(t)$ is SBB, and refer to it as an SBB process.

The requirement that the function $f(\sigma)$ belongs to $\mathcal{F}$ comes from a relation, proven in the sequel, between the bounding function of the output process from a network element and the bounding function of the input process of that network element. This relation involves the integral $\int_\sigma^\infty f(u)\,du$ which is required to be bounded to avoid triviality. Since we do not make, *a priori*, any assumption on the size of the network, it is required that for any order n the multiple integral $(\int_\sigma^\infty du)^n f(u)$ be bounded. It can be shown that all the exponentially decaying functions as well as important functions (distributions) exhibiting subexponential decay, such as Weibull, belong to $\mathcal{F}$ [13]. Finally, we can assume, without any loss of the generality, that $f(\sigma)$ is not increasing with σ since $\Pr(\cdot \geq \sigma)$ cannot increase with σ.

III. Basic Rules of the SBB Calculus

The first step is to provide an SBB characterization for the external processes feeding the network. The following theorem states that if the amount of unfinished work (workload) in a work-conserving system, transmitting at rate ρ, is SB with bounding function $f(\sigma)$ then the input process of the system is SBB with upper rate ρ and bounding function $f(\sigma)$.

Theorem 1 (Characterization): Consider a system that transmits at rate ρ, and assume it is work-conserving. Suppose that it is fed with a single stream of traffic rate $R(t)$, and let $W(t)$ be the amount of data stored in the system at time t. If $W(t)$ is SB with bounding function $f(\sigma)$ then the input process of the system is SBB with upper rate ρ and bounding function $f(\sigma)$.

Proof: By the assumptions of the theorem, there exists a bounding function $f(\sigma)$ such that $\Pr(W(t) \geq \sigma) \leq f(\sigma)$, for all $\sigma \geq 0$ and all $t \geq 0$. Let $0 \leq s < t$, then

$$\left\{\sum_{u=s+1}^{t} R(u) \geq \rho(t-s)+\sigma\right\} \subset \{W(t) \geq \sigma\}$$

since the system can transmit no more than $\rho(t-s)$ information units during the time from s to t. Hence,

$$\Pr\left\{\sum_{u=s+1}^{t} R(u) \geq \rho(t-s)+\sigma\right\} \leq f(\sigma)$$

which establishes the theorem. □

Next, we show that the sum of two SBB processes is itself an SBB process. The case of any finite number of summed processes is easily dealt by applying iteratively the theorem. Note that the following statement holds regardless of the statistical dependencies between the summed processes.

Theorem 2 (Sum): Let $R_1(t)$ be SBB with upper rate ρ_1 and bounding function $f_1(\sigma)$ and $R_2(t)$ be SBB with upper rate ρ_2 and bounding function $f_2(\sigma)$. Then $R_1(t)+R_2(t)$ is SBB with upper rate $\rho_1+\rho_2$ and bounding function $g(\sigma) = f_1(p\sigma) + f_2((1-p)\sigma)$. The value of p is any real number such that $0 < p < 1$.

Proof: Set $\sigma \geq 0$, $0 \leq s < t$ and let $0 < p < 1$ (in Section IV-B, we present one of the possible criteria for setting the value of p). Then, we have

$$\begin{aligned}
&\Pr\left\{\sum_{u=s+1}^{t} R_1(u) + \sum_{u=s+1}^{t} R_2(u) \geq (\rho_1+\rho_2)(t-s)+\sigma\right\}\\
&\leq \Pr\left(\left\{\sum_{u=s+1}^{t} R_1(u) \geq \rho_1(t-s)+p\sigma\right\}\right.\\
&\qquad \left.\bigcup\left\{\sum_{u=s+1}^{t} R_2(u) \geq \rho_2(t-s)+(1-p)\sigma\right\}\right)\\
&\leq \Pr\left\{\sum_{u=s+1}^{t} R_1(u) \geq \rho_1(t-s)+p\sigma\right\}\\
&\quad + \Pr\left\{\sum_{u=s+1}^{t} R_2(u) \geq \rho_2(t-s)+(1-p)\sigma\right\}\\
&\leq f_1(p\sigma) + f_2((1-p)\sigma) = g(\sigma).
\end{aligned}$$

It is easy to show that $g(\sigma)$ belongs to the set of functions $\mathcal{F}$. □

Now, we show that the output process from *any* work-conserving network element (buffer) fed by an SBB input process is SBB and find a very simple relation between the bounding functions of these processes. We also show that in such a case the workload in the buffer is SB. In the following theorem, we consider a single input process. Note that if there are many input processes, they can be aggregated to a single one by employing Theorem 1. Nevertheless, the theorem can be easily extended such that an SBB characterization is derived for each input–output stream (session) apart as we show in [13].

Theorem 3 (Input–Output Relation): Let $R_{in}(t)$ be the aggregate input traffic stream of a work-conserving element, transmitting at rate C. Let $W(t)$ be the amount of data stored in the system and let $R_o(t)$ be the output traffic stream of the system. If $R_{in}(t)$ is SBB with upper rate $\rho < C$ and bounding function $f(\sigma)$, then

i) $R_o(t)$ is SBB with upper rate ρ and bounding function

$$g(\sigma) = f(\sigma) + \frac{1}{C-\rho}\int_\sigma^\infty f(u)\,du.$$

ii) $W(t)$ is SB with bounding function

$$g(\sigma) = f(\sigma) + \frac{1}{C-\rho}\int_\sigma^\infty f(u)\,du.$$

Proof:

i) Let $d(s)$ be the random variable defined by

$$d(s) = \min\{0 \le u \le s : W(s-u) = 0\}.$$

The quantity $d(s)$ equals the time (number of slots) that has passed since the last time the queue was empty prior to s. Letting

$$R^{s,t} = \sum_{u=s+1}^{t} R(u)$$

we have

$$\Pr\left\{R_o^{s,t} \ge \rho(t-s)+\sigma\right\} = \sum_{i=0}^{s} \Pr\left(\left\{R_o^{s,t} \ge \rho(t-s)+\sigma\right\} \cap \{d(s)=i\}\right). \quad (2)$$

Notice that $\{d(s)=i\}$ implies that in each one of the i time slots $s-i+1, s-i+2, \cdots, s$, C units of information have been transmitted, and $\{R_o^{s,t} \ge \rho(t-s)+\sigma\}$ implies that at least additional $\rho(t-s)+\sigma$ information units have been transmitted from s to t. Moreover, $\{d(s)=i\}$ also implies that the queue of the work-conserving element was empty at time $s-i$, and then all the above information units have entered it from $s-i$ to t. Therefore,

$$\left(\left\{R_o^{s,t} \ge \rho(t-s)+\sigma\right\} \cap \{d(s)=i\}\right) \subset \left\{R_{in}^{s-i,t} \ge \rho(t-s)+\sigma+Ci\right\}. \quad (3)$$

Substituting (3) into (2), we get

$$\Pr\left\{R_o^{s,t} \ge \rho(t-s)+\sigma\right\} \le \sum_{i=0}^{s} \Pr\left\{R_{in}^{s-i,t} \ge \rho(t-s+i)+\sigma+(C-\rho)i\right\}. \quad (4)$$

By the assumption of the theorem

$$\Pr\left\{R_{in}^{s,t} \ge \rho(t-s)+\sigma\right\} \le f(\sigma) \quad (5)$$

holds for all integer $0 \le s < t$ and $\sigma \ge 0$. Thus substituting (5) into (4), we get

$$\Pr\left\{R_o^{s,t} \ge \rho(t-s)+\sigma\right\} \le \sum_{i=0}^{\infty} f(\sigma+(C-\rho)i) = f(\sigma) + \frac{1}{C-\rho}\sum_{i=1}^{\infty} f(\sigma+(C-\rho)i)(C-\rho). \quad (6)$$

Reminding that $f(\sigma)$ is not increasing with σ, the following inequality holds:

$$\sum_{i=1}^{\infty} f(\sigma+(C-\rho)i)(C-\rho) \le \int_\sigma^\infty f(u)\,du \quad (7)$$

based on the assumption that $C-\rho > 0$. Inserting (7) into (6), we obtain

$$\Pr\left\{R_o^{s,t} \ge \rho(t-s)+\sigma\right\} \le f(\sigma) + \frac{1}{C-\rho}\int_\sigma^\infty f(u)\,du = g(\sigma).$$

The function $g(\sigma)$ belongs to $\mathcal{F}$ because that $f(\sigma)$ belongs to $\mathcal{F}$.

ii) The proof of this part is very similar to the proof of part i). For details, see [13]. □

By applying inductively the above theorem, one can show that if all the processes feeding a feedforward network are SBB, then so are all the processes within the network (assuming that the stability condition is satisfied) and the workload in each network element is SB.

IV. Network Calculus for Sums of Exponentials

A. Motivation: Tighter Bounds for Multiple Time-Scale Traffic

Markov modulated processes have been extensively used for representing real-time traffic such as packetized voice traffic and video traffic [11]. For these processes, it has been observed that, on the logarithmic scale, the delay distribution in a multiplexer can be roughly broken into two linear regions (cell and burst regions). This behavior is a typical characteristic of multiple time-scale traffic [6], [14]. Consider, for example, a stationary Markov modulated arrival process $R(t), t \in \mathbb{N}$, feeding a buffer with service rate $C = 1$. The modulating chain is assumed to have two states (say 1, 2) with transition probabilities $p_{12} = 10^{-6}$ and $p_{21} = 1/100$. When in state 1, the source is producing independent and identically distributed (i.i.d.) arrivals $\mathcal{A}_1$ distributed according to $\Pr(\mathcal{A}_1 = 0) = 7/8$ and $\Pr(\mathcal{A}_1 = 2) = 1/8$. When in state 2, the source is producing i.i.d. arrivals $\mathcal{A}_2$ distributed according to $\Pr(\mathcal{A}_2 = 0) = 11/20$ and $\Pr(\mathcal{A}_2 = 2) = 9/20$. The parameters of this Markov chain are representative of multiple time-scale traffic [6]. Using standard z-transform techniques we can find the steady-state buffer occupancy distribution

$$\Pr(W=\sigma) = 0.857 \cdot e^{-1.946\sigma} + 2.410 \cdot 10^{-5} \cdot e^{-0.273\sigma} \quad (8)$$

where σ represents the buffer size. For this example, the bound on the survivor probability $\Pr(W \geq \sigma)$ computed following the EBB approach of [15] is

$$\Pr(W \geq \sigma) \leq e^{-0.273\sigma}, \qquad \text{for all } \sigma \geq 0. \quad (9)$$

This bound is loose. Using only two exponentials, we can provide a much tighter bound (corresponding, even, to the true probabilities in this case)

$$\Pr(W \geq \sigma) \leq (1 - 10^{-4}) \cdot e^{-1.946\sigma} + 10^{-4} \cdot e^{-0.273\sigma}, \qquad \text{for all } \sigma \geq 0. \quad (10)$$

The comparison of the two bounds with the true probabilities is shown in Fig. 2(a). The question is whether there exists a network algebra for sums of exponentials. The answer is, obviously, in the affirmative if we resort to our general definition of stochastically bounded burstiness and choose the bounding function as a sum of exponentials. Back to the example, one can see that $W(t)$ is SB with bounding function

$$f(\sigma) = e^{-1.946\sigma} + 10^{-4} \cdot e^{-0.273\sigma}, \qquad \text{for all } \sigma \geq 0 \quad (11)$$

and, by Theorem 1, $R(t)$ is SBB with upper rate $\rho = 1$ and bounding function as defined in (11) (the factor -10^{-4} appearing before the first exponential in (10) has been omitted in (11), since it is meaningless). Note that the random variable $W(t)$ is stochastically smaller than the random variable W, corresponding to the steady-state workload, since the buffer is empty at time 0 and the arrival process is stationary (see, e.g., [1]).

B. Basic Rules of the Calculus for Sums of Exponentials

The basic rules of the network calculus for processes with burstiness stochastically bounded by sums of exponentials are easily derived from the theorems presented in Section III. The first rule is related to the addition of processes, i.e., if $R_1(t)$ is SBB with upper rate ρ_1 and bounding function

$$f(\sigma) = \sum_{i=1}^{N} A_i e^{-\alpha_i \sigma}$$

and $R_2(t)$ is SBB with upper rate ρ_2 and bounding function

$$g(\sigma) = \sum_{i=1}^{M} B_i e^{-\beta_i \sigma}$$

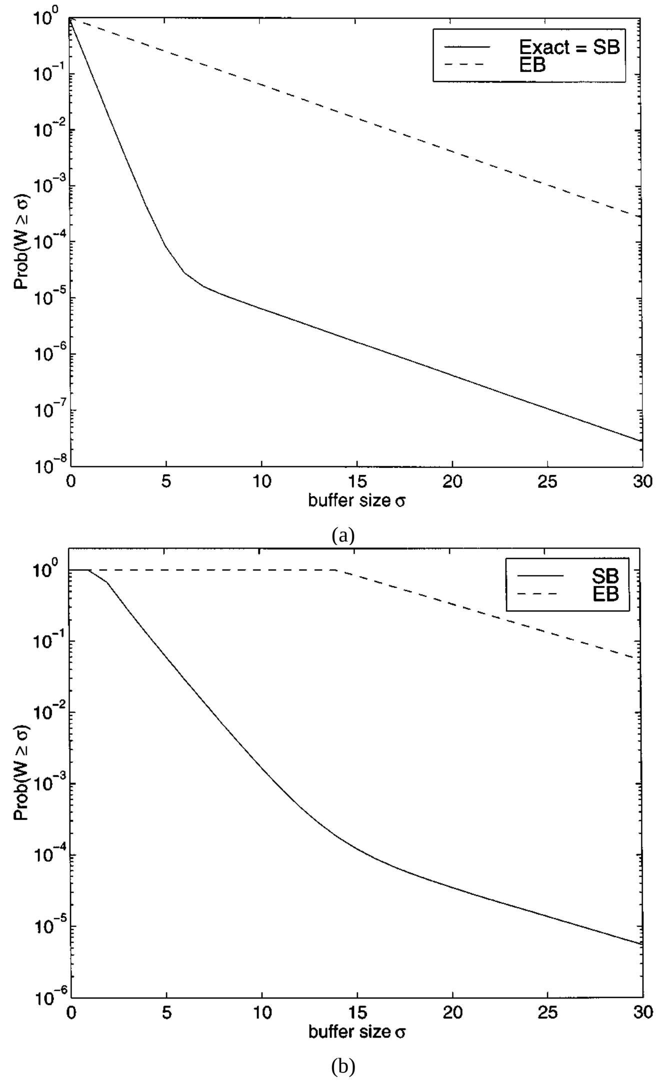

Fig. 2. (a) Bounding a distribution (exact) with a single exponential (EB) and with a sum of two exponentials (SB). (b) Bounds on the workload distribution of a multiplexer fed by two dependent processes: EB bound versus SB bound.

then, according to Theorem 2, $R_1(t)+R_2(t)$ is SBB with upper rate $\rho_1 + \rho_2$ and bounding function

$$h(\sigma) = \sum_{i=1}^{N+M} C_i e^{-\gamma_i \sigma}$$

where

$$C_i \triangleq \begin{cases} A_i, & N \geq i \geq 1 \\ B_{i-N}, & N+M \geq i \geq N+1 \end{cases}$$

$$\gamma_i \triangleq \begin{cases} p\alpha_i, & N \geq i \geq 1 \\ (1-p)\beta_{i-N}, & N+M \geq i \geq N+1 \end{cases}$$

and where the value of p is any real number such that $1 > p > 0$. The usual criterion for choosing p is to maximize the asymptotic decay rate of $h(\sigma)$. Assuming that $\alpha_N = \min_i \alpha_i$ and $\beta_M = \min_i \beta_i$, then one should choose $p = \beta_M/(\alpha_N + \beta_M)$

in order to maximize the smallest γ_i. The second rule deals with the calculus for an isolated network element. If a general work-conserving multiplexer with service rate C is fed by an SBB input process $R_{in}(t)$ with upper rate $\rho < C$ and bounding function $f(\sigma) = \sum_{i=1}^{N} A_i e^{-\alpha_i \sigma}$ then, according to Theorem 3, the output process $R_o(t)$ is SBB with upper rate ρ and bounding function

$$g(\sigma) = \sum_{i=1}^{N} A_i (1 + (C - \rho)^{-1} \alpha_i^{-1}) e^{-\alpha_i \sigma}$$

and the amount of data stored in the multiplexer $W(t)$ is SB with the same bounding function $g(\sigma)$.

C. An Example

As an application of these results, consider a general multiplexer with service capacity $C = 3$. The multiplexer is fed by the sum of two input streams with unknown degree of dependency. The first stream $R_{1, in}(t)$ corresponds to the Markov modulated process described in Section IV-A. The second stream, $R_{2, in}(t)$, is a similar Markov modulated process that is EBB with upper rate 1 and bounding function $e^{-0.548\sigma}$ and SBB with upper rate 1 and bounding function $e^{-2.197\sigma} + 10^{-4} \cdot e^{-0.543\sigma}$. Using the technique of [15], one finds that the summed process $R_{in}(t) = R_{1, in}(t) + R_{2, in}(t)$ is EBB with upper rate 2 and bounding function $2 \cdot e^{-0.182\sigma}$ and that the workload distribution $W(t)$ in the multiplexer is EB with bounding function $12.989 \cdot e^{-0.182\sigma}$. Using the network calculus for sums of exponentials, one deduces that $R_{in}(t)$ is SBB with upper rate 2 and bounding function

$$e^{-1.295\sigma} + e^{-0.735\sigma} + 2 \cdot 10^{-4} \cdot e^{-0.182\sigma}$$

and $W(t)$ is SB with bounding function abovedisplayskip12pt

$$1.772 \cdot e^{-1.295\sigma} + 2.361 \cdot e^{-0.735\sigma} + 12.989 \cdot 10^{-4} \cdot e^{-0.182\sigma}.$$

As one can see, from Fig. 2(b), the SB bound is up to four orders of magnitude tighter than the EB bound. Finally, we obtain from Theorem 3 that the aggregate output stream from the multiplexer is SBB with upper rate 2 and the same bounding function as for $W(t)$.

D. Limiting the Number of Exponentials to a Fixed Number

In this section, we present a refinement of the calculus for sums of exponentials. The main current drawback of this calculus is that the number of exponentials required to bound the burstiness of the processes within the network is growing each time that processes are summed. The example given in the previous section illustrates this fact. Each one of the two original SBB processes has a bounding function described by a sum of two exponentials. But, the aggregate process has a bounding function described by a sum of three exponentials. From both theoretical and practical points of view, one may rather be interested in limiting the number of exponentials to a fixed number (say two). Of course, there is a cost for such a convenience, that is, looser bounds. One should seek to minimize this cost. In the sequel of this section, we introduce an "ad hoc" procedure which provides a simple way to derive bounding functions described by sums of two exponentials. In [13], we prove that the SBB calculus combined with this procedure offers better bounds than the EBB calculus, at each node of a feedforward network, provided that the SBB bounding function of each external arrival process is upper-bounded by the EBB bounding function.

The procedure that we are interested in is the following: Given a (bounding) function consisting of a sum of an arbitrary number of exponentials

$$f(x) = \sum_{i=1}^{M} a_i e^{-\alpha_i x}$$

find a function consisting of a sum of two exponentials $g(x) = b_1 e^{-\beta_1 x} + b_2 e^{-\beta_2 x}$ such that $g(x) \geq f(x)$ for all $x \geq 0$. Clearly, there are an infinite number of functions that satisfy such a requirement. Among all these functions, we look for the optimal one $g_{\text{opt}}(x)$ that is the "closest" of $f(x)$. For instance, we may define $g_{\text{opt}}(x)$ as the function that minimizes the maximum of $g(x) - f(x)$, that is the absolute error, over the domain $[0, \infty)$. However, such a criterion does not ensure that $g_{\text{opt}}(x)$ will have the same asymptotic decay rate as $f(x)$. Therefore, we prefer to define the optimal function $g_{\text{opt}}(x)$ as the function that minimizes the maximum of $\log(g(x)) - \log(f(x))$ which is equivalent to minimizing the maximal *relative* error between $g(x)$ and $f(x)$. Unfortunately, it appears that the computation of $g_{\text{opt}}(x)$ is a difficult task. We suggest, instead, a simple heuristic which provides a suboptimal solution to this problem, $g_{\text{sub}}(x)$. The heuristic is based on the observation that, on a logarithmic scale, a function described by a sum of exponentials behaves almost as a piecewise-linear function, where each linear region corresponds to a dominating exponential. The first step of the ad hoc procedure is, thus, to bound the logarithm of the original bounding function by a piecewise-linear function $\ell(x)$. The function $\ell(x)$ is constructed by sampling some points of $\log(f(x))$ and connecting them by straight lines. The line starting from the last point toward the infinity is chosen to have a slope equal to the asymptotic decay of $f(x)$. We remark that $\ell(x) \geq \log(f(x))$ due to the convexity of $\log(f(x))$. The construction is illustrated in Fig. 3(a) for an example function $f_{\text{ex}}(x) = e^{-x} + 10^{-3} \cdot e^{-0.5x} + 10^{-6} \cdot e^{-0.25x}$. The accuracy of the fitting depends, obviously, on an appropriate selection of the sampling points. It turns out that only a few points are required for obtaining a good approximation, that is, two points for each linear region and two or three others for each region of transition from one dominating exponential to another. The next step is to bound $\ell(x)$ by another piecewise-linear function $q(x)$ consisting of only two linear regions, as shown in Fig. 3(b). The second segment of $q(x)$ has the same slope as the last segment of $\ell(x)$. The point of intersection between the two lines of $q(x)$ is chosen such that the maximum of $q(x) - \ell(x)$ is minimized. The computation of $q(x)$ is easily carried out due to the piecewise linearity of both $\ell(x)$ and $q(x)$. Now, let x_0 be the abscissa of the point of intersection between the two segments of $q(x)$. Then, for $x \leq x_0$ (respectively, $x \geq x_0$) we have that $\log(f(x))$

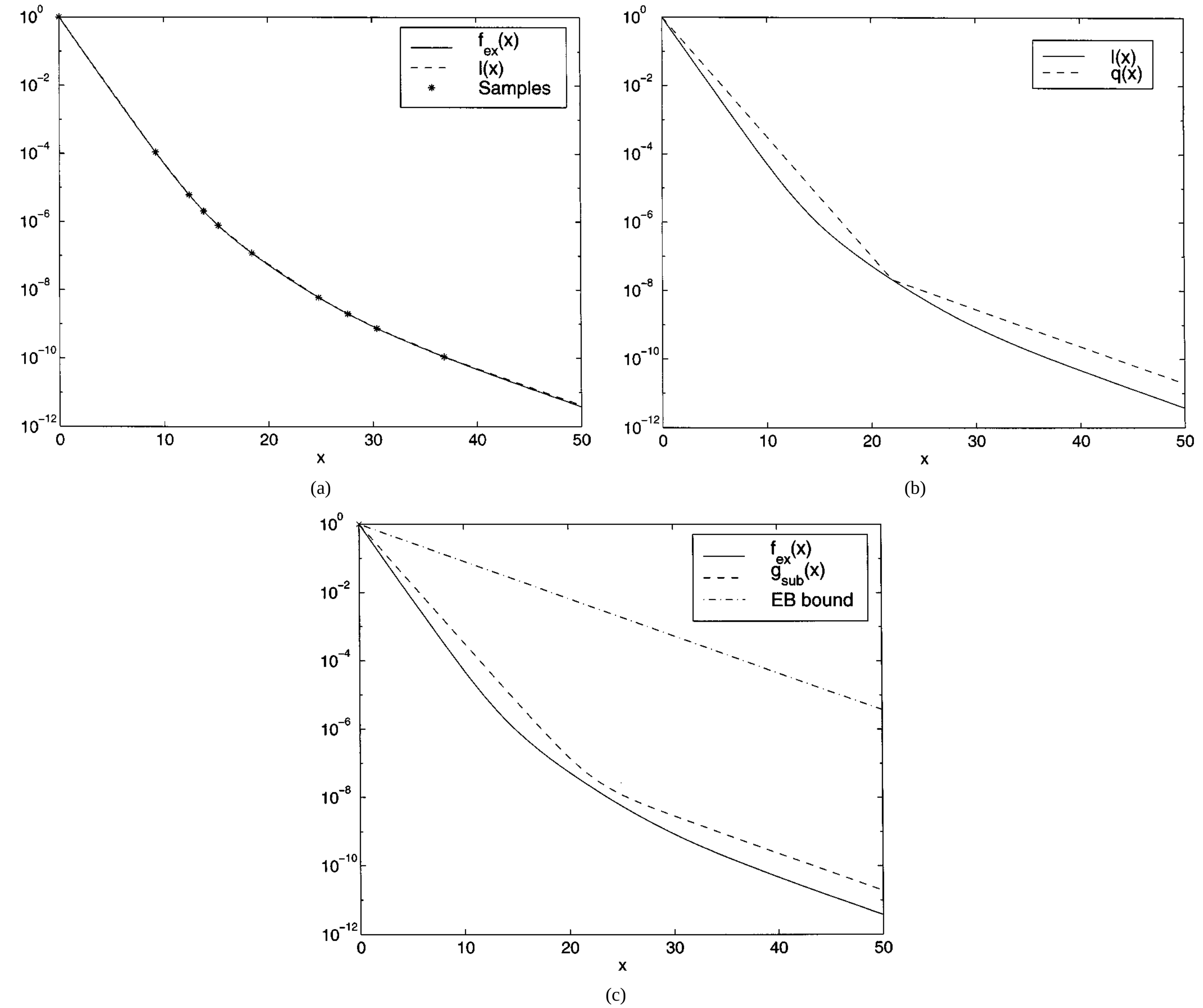

Fig. 3. (a) Plot, on a logarithmic scale, of a bounding function $f_{ex}(x)$ consisting of three exponentials and of its piecewise-linear fitting $\ell(x)$. The fitting is achieved by connecting sampled points of $f_{ex}(x)$. (b) the fitted function $\ell(x)$ and its bounding function $q(x)$ consisting of two linear regions. Here, the abscissa, x_0, of the point of intersection between the two linear regions of $q(x)$ is approximately equal to 22. (c) Replacing the original bounding function $f_{\rm ex}(x)$ with a bounding function $g_{\rm sub}(x)$ consisting of a sum of two exponentials and with a bounding function consisting of a single exponential (EB bound).

is bounded by the first (resp., second) segment of $q(x)$. Thus $f(x) \le Ae^{-\beta_1 x}$ for $x \le x_0$ and $f(x) \le Be^{-\beta_2 x}$ for $x \ge x_0$, where the parameters β_1 and B are directly obtained from the expression for $q(x)$. The value of the parameter A is equal to the value of $f(x)$ at $x = 0$, i.e., $A = \sum_{i=1}^{M} a_i$, whereas the parameter β_2 is equal to the asymptotic decay rate of $f(x)$, say α_M. We recall that our goal is to bound $f(x)$ by a function consisting of a sum of two exponentials. This goal is achieved by defining the suboptimal solution as follows:

$$g_{\rm sub}(x) = (A - B)e^{-\beta_1 x} + Be^{-\beta_2 x} \equiv b_1 e^{-\beta_1 x} + b_2 e^{-\beta_2 x}. \quad (12)$$

Since $A > B$ and $\beta_1 > \beta_2$, it is easy to see that $g_{\rm sub}(x) \ge f(x)$ for both $x \le x_0$ and $x \ge x_0$. We remark that this last step introduces an additional loss in terms of the tightness of the bounds. Fortunately, this loss is limited to the region around the intersection point. The result of the procedure for the function $f_{ex}(x) = e^{-x} + 10^{-3} \cdot e^{-0.5x} + 10^{-6} \cdot e^{-0.25x}$ is illustrated in Fig. 3(c). For this example, we obtain

$$g_{\rm sub}(x) = 1.001 \cdot e^{-0.8x} + 5.155 \cdot 10^{-6} \cdot e^{-0.25x}.$$

From Fig. 3(c), we see that the replacement of $f_{ex}(x)$ with a bounding function consisting of a single exponential alters significantly the shape of the original curve. A considerable improvement is achieved by replacing $f_{ex}(x)$ with a bounding function consisting of a sum of two exponentials.

V. Discussion and Open Problems

In this paper, we introduced a new network calculus, termed SBB calculus, which provides statistical upper bounds on various performance metrics, at each node of a feedforward network. We showed that the SBB methodology is superior to previous approaches from several points of views. First, it handles a larger class of input processes, including important processes like fractional Brownian motion. It can also provide much more accurate and realistic characterization of real-time traffic, when the bounding function is chosen as a sum of exponentials. This enhanced characterization can substantially improve the utilization of networks implementing services based on statistical guarantees. Finally, we showed that the calculus for sums of exponentials can be implemented at low complexity. For this purpose, we developed a new procedure which provides a simple and efficient way to derive bounding functions consisting of sums of two exponentials. This procedure ensures that the number of exponentials required to stochastically bound the burstiness of the processes within the network remains fixed and does not grow each time that SBB processes are summed.

The basic theorems presented in Section III and the techniques used for proving them should only be regarded as the building blocks of the SBB calculus. We show in [13] that these theorems can be further generalized and extended. For instance, the concept of "exponentially bounded fluctuation," developed in [8] for characterizing the fluctuation of variable rate links, can be extended to more general bounding functions. Also, if it is known that the processes feeding the network are independent, then tighter bounds can be derived [15]. Note that the analysis we presented here assumes nothing about the service discipline of the work-conserving element. For a specific element, a finer analysis may result in tighter bounds. This is, for example, the case for network elements implementing the WFQ (PGPS) service discipline [10], [16], [17].

This work leaves several issues open for further research. For instance, the problem of handling processes with a bounding function $f(\sigma) = 1/\sigma^{\alpha}$ for some $\alpha > 0$. In such a case, the calculus developed here cannot be used for feedforward networks of arbitrary size since $f \notin \mathcal{F}$ (see Definition 1). In [2], an efficient methodology for deriving tight exponential bounds for in-tree networks is developed. The extension of that approach to sums of exponentials and general bounding functions would be especially useful.

ACKNOWLEDGMENT

The authors wish to thank Prof. O. Zeitouni for fruitful discussions.

REFERENCES

[1] A. Borovkov, *Stochastic Processes in Queueing Theory*. New York, NY: Springer-Verlag, 1976.
[2] C. S. Chang and J. Cheng, "Computable exponential bounds for intree networks with routing," in *Proc. INFOCOM '95*, 1995, pp. 197–204.
[3] R. Cruz, "A calculus for network delay—Part I: Network elements in isolation," *IEEE Trans. Inform. Theory*, vol. 37, pp. 114–131, Jan. 1991.
[4] N. G. Duffield and N. O'Connell, "Large deviations and overflow probabilities for the general single server queue, with applications," *Mathematical Proc. Cambridge Philosophical Soc.* , no. 118, pp. 363–374, 1995.
[5] A. Erramilli, O. Narayan, and W. Willinger, "Experimental queueing analysis with long-range dependent packet traffic," *IEEE/ACM ToN*, vol. 4, no. 2, pp. 209–223, Apr. 1996.
[6] P. R. Jelenković, "The effect of multiple time scales and subexponentiality on the behavior of a broadband network multiplexer," Ph.D. dissertation, Columbia Univ., New York, 1996.
[7] J. Kurose, "On computing per-session performance bounds in high-speed multi-hop computer networks," in *Perform. 92*, Newport, CA, June 1992.
[8] K. Lee, "Performance bounds in communication networks with variable-rate links," in *Sigcomm '95*, 1995, pp. 126–136.
[9] I. Norros, "On the use of fractional brownian motion in the theory of connectionless networks," *IEEE J. Select. Areas Commun.*, vol. 13, pp. 953–962, Aug. 1995.
[10] A. Parekh and R. Gallager, "A generalized processor sharing approach to flow control in integrated services networks: The multiple node case," *IEEE/ACM ToN*, vol. 2, no. 2, pp. 137–150, Apr. 1994.
[11] M. Schwartz, *Broadband Integrated Networks*. Englewood Cliffs, NJ: Prentice-Hall, 1996.
[12] S. Shenker, C. Partridge, and R. Guérin, "Specification of guaranteed quality of service," IETF, RFC 2212,, Sept. 1997.
[13] D. Starobinski, "Quality of service in high speed networks with multiple time-scale traffic," Ph.D. dissertation, Technion–Israel Inst. Technol., Haifa, Israel, May 1999.
[14] D. N. Tse, R. G. Gallager, and J. H. Tsitsiklis, "Statistical multiplexing of multiple time-scale Markov streams," *IEEE J. Select. Areas Commun.*, vol. 13, pp. 1028–1038, Aug. 1995.
[15] O. Yaron and M. Sidi, "Performance and stability of communication networks via robust exponential bounds," *IEEE/ACM ToN*, vol. 1, no. 3, pp. 372–385, June 1993.
[16] ——, "Generalized processor sharing networks with exponentially bounded burstiness arrivals," *J. High Speed Networks*, vol. 3, pp. 375–387, 1994.
[17] Z. Zhang, D. Towsley, and J. Kurose, "Statistical analysis of the generalized processor sharing scheduling discipline," *IEEE J. Select. Areas Commun.*, vol. 13, pp. 1071–1080, Aug. 1995.