A Probabilistic Definition of Burstiness Characterization: A Systematic Approach*

Sami Ayyorgun Wu Feng

Research And Development in Advanced Network Technology (RADIANT)
Los Alamos National Laboratory
P.O. Box 1663, MS D451
Los Alamos, New Mexico 87545

{sami,feng}@lanl.gov

May 2003

Abstract—The burstiness of network traffic has a profound impact on the performance of many network protocols in areas such as congestion control (e.g., TCP), multiple-access (e.g., CSMA), routing (e.g., BGP), and switching/multiplexing in general. However, there does not exist a widely accepted definition of burstiness, be it either deterministic or probabilistic, in the networking community. Current deterministic definitions of burstiness, and hence traffic characterizations, inherently lack statistical gains, leading to an underutilization of network resources. The probabilistic characterizations, on the other hand, are either too complicated for tractable performance analyses, or lack the capability to effectively deal with the network traffic with recently observed behavior (e.g., *long-range dependency*).

We propose a probabilistic definition of burstiness based on *service curves*. This definition can address network traffic with arbitrary *heavy-tailed* distributions, including the *power-tailed* distributions, in performance analyses. It facilitates a simple systematic analysis for provable, probabilistic, performance guarantees. We show that the queue size, output traffic, virtual-delay, aggregate traffic, etc. at various points in a network can easily be characterized within the framework of our definition. The performance bounds provided by this characterization are tighter than what is currently available by other similar approaches.

1 Introduction

Service and traffic models which facilitate a tractable performance analysis is crucial for a widespread deployment of network services requiring *provable* Quality of Service (QoS) guarantees, such as real-time multimedia applications, by network service providers. Although the *best-effort service* provided by the Internet has brought the network to the masses, there is now a prevailing agreement among researchers and practitioners that services with provable QoS guarantees are needed, as well as in demand, for broadband networks.

The tractability—or, intractability—of performance analyses is greatly determined by the service and traffic models being employed. Service models provide an abstraction to incorporate QoS guarantees into both the analysis and applications. Traffic models, on the other hand, are needed to model the behavior of network traffic in real life, as well as to provide an understanding to alter their behavior as desired for a better network management and performance.

On the service part, a major challenge is to come up with definitions of service models to quantify the concept of "service" in analytically tractable ways. In the mid 90's, such a service model, called *service curve model*, has been proposed [1, 2] exhibiting analytical tractability similar to that in linear system theory.

On the traffic part, a major challenge is to come up with traffic models capturing the bursty nature of network traffic, and at the same time facilitating a tractable performance analysis. With

*Los Alamos National Laboratory, Technical Report LA-UR-03-3668, May 2003.

the rapid spread of data networks within the last decade, it has become apparent that the network traffic exhibits bursty behavior. However, a widely accepted definition of burstiness, be it either deterministic or probabilistic, still does not exist in networking community. Yet, it is known that the burstiness of network traffic has a profound impact on the performance of many network protocols in areas such as congestion control (e.g., TCP), multiple-access (e.g., CSMA), routing (e.g., BGP), and switching and multiplexing in general. Current deterministic definitions of burstiness, and hence traffic characterizations, inherently lack statistical gains, leading to an underutilization of network resources; e.g., [2, 3, 4]. The probabilistic characterizations, on the other hand, are either too complicated for tractable performance analyses (e.g., refer to the related work in [5] about *self-similar* traffic models), or lack capability to effectively deal with the network traffic with recently observed behavior (e.g., *long-range dependency*). Almost all of the traditional traffic models and/or characterizations (e.g., Markovian models) fall into the later category, as well as some of the new models (e.g., [6, 7, 8]).

We believe that service and traffic models are to be considered together and be defined in accordance, to facilitate a tractable performance analysis. This study attempts to do that by proposing a probabilistic definition of burstiness based on service curves, which facilitate a systematic treatment of stochastic performance guarantees. Our work here is strongly motivated by the work in [6, 7, 8]. We show that the performance bounds provided here by the proposed characterization are tighter than those in [6, 7, 8]. The characterization that we propose here also does not have the restriction on a bounding function (i.e., f) as specified in [6]. The new characterization is also defined more generally by basing it on service curves.

In coming up with this definition, we have had in mind the properties that a "good" traffic characterization should have in order to facilitate a tractable analysis. Some of these properties are listed below: Let $\mathcal{C}$ denote a "good" traffic characterization, then

1. if two traffic flows are characterized according to $\mathcal{C}$, then the aggregate of the flows should also be easily characterized according to $\mathcal{C}$,

2. if a traffic flow characterized according to $\mathcal{C}$ is fed into a network element commonly used in practice, then

   - the output flow should also be easily characterized according to $\mathcal{C}$,
   - and, both the queue size and the (virtual) delay should be easily characterized in the same framework as $\mathcal{C}$, as well,

3. the characterization $\mathcal{C}$ should be a "stationary" in the sense that the characterization of a time-shifted traffic does not change with respect to that of the unshifted traffic,

4. the characterization $\mathcal{C}$ should capture the behavior of traffic in real networks as closely as possible, without sacrificing neither the tractability nor the reality,

5. the characterization of a flow according to $\mathcal{C}$ should be measurable, and be easily measured.

We show that the burstiness definition that we propose here satisfies all of the above properties. By the first three properties, it facilitates a systematic treatment of performance guarantees. Our definition is also capable of addressing the network traffic with arbitrary *heavy-tailed* distributions, including the *power-tailed* distributions, in performance analyses. Hence, it entertains the fourth property as well. The last property is also satisfied.

The rest of the paper is organized as follows: Section 2 provides a background material. Section 3 introduces the new burstiness definition. Sections 3.1 and 3.2 examine the implications of the

new definition, by itself and for a single network element, respectively. Section 3.2.1 gives the mean performance guarantees for a single network element. Section 3.3 examines the performance guarantees according to the new burstiness definition, over a tandem of network elements. Section 4 provides a conclusion.

2 Background and Convention

We adopt a discrete-time formulation for the simplicity of exposition. Time is slotted into fixed-length intervals, and marked by the integers. The unit of transmission for communication is referred to as a *packet*, in this study. A *flow* is a non-decreasing function defined from the integers to the non-negative integers. The value $R(n)$ of a flow R at time n denotes the total number of packets that has arrived by time n (inclusive) for a *connection*. A *network element* is an input-output device that accepts packets at its input, processes them, and delivers them at its output. A network element is said to be *passive* if it does not generate any packet internally. Network elements are assumed to be passive in this study, for the simplicity of exposition. Packets are assumed to be able to instantaneously arrive and depart at a network element, i.e., a whole packet could arrive instantaneously at time k, and later depart at time n where $n \geqslant k$. Note that a packet could depart in the same interval in which it has arrived; this is sometimes referred to as *cut-through* operation. The capacity $c(n)$ of a network element at time n is the total number of packets that it could deliver (serve) at time n. The function c is called the *instantaneous capacity rate*, or just the *rate*, of the network element. Finally, all functions are assumed to be defined from the integers to the integers, unless otherwise noted from here on.

For convenience, the end of proofs in the text are marked by '■', and the end of examples are marked by '⌟', where both marks are flushed to the right margin.

We utilize the following definitions in this study, which have been previously introduced in the literature.

Definition 1 *Let f and g be any two functions. The* min-+ convolution *of f and g, denoted*[1] *as $f \bar{\otimes} g$, is defined as*

$$(f \bar{\otimes} g)(n) = \min_{k \leqslant n} \{f(k) + g(n-k)\} \qquad \textit{for all } n.$$

The convolution $f \bar{\otimes} g$ is read as "f min-convolved with g", or as "the min-plus convolution of f with g".

Consider the problem of finding a function X whose min-+ convolution with g is f, i.e., find an X such that $X \bar{\otimes} g = f$. One can proceed to find X as follows;

$$\begin{aligned}
(X \bar{\otimes} g)(n) &= f(n) \\
\min_{k \leqslant n} \{X(k) + g(n-k)\} &= f(n) \\
X(k) + g(n-k) &\geqslant f(n) && \text{for all } k \leqslant n \\
X(k) &\geqslant f(n) - g(n-k) && \text{for all } k \leqslant n \\
X(k) &\geqslant \max_{n \geqslant k} \{f(n) - g(n-k)\} \\
&= \max_{n \geqslant 0} \{f(n+k) - g(n)\}\,.
\end{aligned}$$

Hence, it is convenient to define the min-+ *deconvolution* of two functions f and g as follows.

[1] One reason to choose this notation over some others, for example '$*$', is that there are a companion and related other operators to this operator, which are employed later in the work. We believe that this choice of notation provides a better choice of notations for these other operators in a fitting manner.

Definition 2 *Let f and g be two functions. The* min-+ deconvolution *of f and g, denoted as $f \overline{\oslash} g$, is defined as*

$$(f \overline{\oslash} g)(n) = \max_{k \geqslant 0} \{f(n+k) - g(k)\} \qquad \text{for all } n.$$

The deconvolution $f \overline{\oslash} g$ is read as "f min-deconvolved with g", or as "the min-plus deconvolution of f with g"

A companion operator to the min-+ convolution is called the max-+ *convolution*, and defined as follows.

Definition 3 *Let f and g be two functions. The* max-+ convolution *of f and g, denoted as $f \underline{\otimes} g$, is defined as*

$$(f \underline{\otimes} g)(n) = \max_{k \leqslant n} \{f(k) + g(n-k)\} \qquad \text{for all } n.$$

The convolution $f \underline{\otimes} g$ is read as "f max-convolved with g", or as "the max-plus convolution of f with g".

Again, consider the problem of finding a function X whose max-+ convolution with g is f; that is, find an X such that $X \underline{\otimes} g = f$. One can proceed to find X as follows;

$$\begin{aligned}
(X \underline{\otimes} g)(n) &= f(n) \\
\max_{k \leqslant n} \{X(k) + g(n-k)\} &= f(n) \\
X(k) + g(n-k) &\leqslant f(n) && \text{for all } k \leqslant n \\
X(k) &\leqslant f(n) - g(n-k) && \text{for all } k \leqslant n \\
X(k) &\leqslant \min_{n \geqslant k} \{f(n) - g(n-k)\} \\
&= \min_{n \geqslant 0} \{f(n+k) - g(n)\}\,.
\end{aligned}$$

Hence, it is also convenient to define the max-+ *deconvolution* of two functions f and g as follows.

Definition 4 *Let f and g be two functions. The* max-+ deconvolution *of f and g, denoted as $f \overline{\oslash} g$, is defined as*

$$(f \overline{\oslash} g)(n) = \min_{k \geqslant 0} \{f(n+k) - g(k)\} \qquad \text{for all } n.$$

The deconvolution $f \overline{\oslash} g$ is read as "f max-deconvolved with g", or as "the max-plus deconvolution of f with g".

Definition 5 *An S-server* with service curve *S is a network element that when fed with an input flow R, the corresponding output flow G satisfies*

$$G(n) \geqslant (R \overline{\otimes} S)(n) \quad \text{for all } n$$

for any arbitrary R. A service curve S is a non-decreasing function defined from the integers to the non-negative integers, and takes on the value zero for non-positive values (i.e., $S(n) = 0$ for all $n \leqslant 0$).

Note that a *work-conserving*[2] server with the capacity of serving packets with a constant integer rate ρ is an S-server with service curve $S(n) = \max\{0, \rho \cdot n\}$, since the output G of a work-conserving server to an input flow R is given[3] by

$$G(n) = \min_{k \leqslant n} \{R(k) + \rho\,(n-k)\} \quad \text{for all } n$$

[2] A network element is said to be *work-conserving* if it serves packets whenever it has packets to serve, unconditionally of any other criteria.

[3] This is often referred to as Reich's result [9]. See [10], for example, for a simple derivation.

which could also be equivalently represented as

$$G(n) = (R \overline{\nabla} S)(n) \quad \text{for all } n.$$

3 A Probabilistic Definition of Burstiness Characterization, and Its Implications

We propose the following probabilistic definition for burstiness characterization of network traffic, and study some of its implications.

Definition 6 *A traffic flow R is said to be* bursty *with* service curve *S and bounding function f, and denoted as $R \sim (S, f)$, if*

$$\mathrm{P}\Big(R(n) - R(k) > S(n-k) + \sigma, \quad \text{for some } k < n\Big) \leqslant f(\sigma) \tag{1}$$

for all σ and for all n, where f is defined from the integers to the non-negative real numbers.

We assume the following properties for any bounding function f, without loss of generality:

1. f is non-increasing, since the probability corresponding to a σ in the above definition is non-increasing with σ.
2. $f(\sigma) \leqslant 1$ for all σ, since the probability of an event can not be larger than 1. We assume for mathematical convenience that $f(\sigma) = 1$ for all $\sigma < 0$, unless otherwise noted from here on.
3. $\lim_{\sigma \to \infty} f(\sigma) = 0$, since any cumulative distribution function F satisfies $\lim_{x \to \infty} F(x) = 1$.

In the remainder of this study, we examine some of the properties/implications of this traffic characterization. Specifically, we will show that it satisfies all of the properties of a "good" traffic characterization that we have stated earlier in the introduction.

3.1 Implications on Aggregate Flows and Average Rate

We first show that the traffic characterization provided by definition 6 satisfies the property 1 of a "good" traffic characterization stated earlier in the introduction. That is, we show that the aggregate of flows where each flow is characterized according to this characterization could also be easily characterized by the same characterization. This is stated more precisely in the following theorem, and proved thereafter.

Theorem 1 *Let R_1 and R_2 be two flows that $R_1 \sim (S_1, f_1)$ and $R_2 \sim (S_2, f_2)$. The aggregate flow $R_1 + R_2$ is bursty with service curve $S_1 + S_2$ and bounding function $f_1 \overline{\nabla} f_2$. In other words, $R_1 + R_2 \sim (S_1 + S_2, f_1 \overline{\nabla} f_2)$.*

Proof: The proof follows by considering the following events for any n, σ, and $u \leqslant \sigma$;

$$A = \Big\{(R_1 + R_2)(n) - (R_1 + R_2)(k) > (S_1 + S_2)(n-k) + \sigma, \quad \text{for some } k < n\Big\}$$
$$A_1 = \Big\{R_1(n) - R_1(k) > S_1(n-k) + u, \quad \text{for some } k < n\Big\}$$
$$A_2 = \Big\{R_2(n) - R_2(k) > S_2(n-k) + \sigma - u, \quad \text{for some } k < n\Big\}.$$

Notice that $A \subseteq A_1 \cup A_2$, since the intersection of the complements of both of the events A_1 and A_2 is clearly a subset of the complement of the event A. Thus, by the union bound and the fact that $R_1 \sim (S_1, f_1)$ and $R_2 \sim (S_2, f_2)$, we get

$$\begin{aligned} \mathrm{P}(A) &\leqslant \mathrm{P}(A_1 \cup A_2) \\ &\leqslant \mathrm{P}(A_1) + \mathrm{P}(A_2) \\ &\leqslant f_1(u) + f_2(\sigma - u) \end{aligned}$$

since the last inequality holds for any $u \leqslant \sigma$, we also get

$$\begin{aligned} \mathrm{P}(A) &\leqslant \min_{u \leqslant \sigma} \{f_1(u) + f_2(\sigma - u)\} \\ &= (f_1 \overline{\triangledown} f_2)(\sigma)\,. \end{aligned}$$

One might think that we would have had to use infimum 'inf' above instead of minimum 'min', since f_i's are real-valued. However, it turns out that this is not the case since the above minimum if effectively taken over a set of finite number of elements due to the fact that $f_i(s) = 1$ for all $s < 0$, for i equals to both 1 and 2. Moreover, since $\lim_{\sigma \to \infty} f(\sigma) = 0$ for any f, note that we have $(f_1 \overline{\triangledown} f_2)(\sigma) = 1$ for all $\sigma < 0$.[4] This completes the proof. ■

The burstiness characterization provided by definition 6 has also an implication on the long-term average rate of a flow. The long-term average rate of a flow $R \sim (S, f)$ is upper-bounded by $\limsup_{n \to \infty} \frac{S(n)}{n}$ if the area under the bounding function f (for non-negative values of σ) is finite, i.e., $\sum_0^\infty f(\sigma) < \infty$. Let us adopt the convention that whenever we refer to 'the area under the bounding function f, we mean 'the sum of the bounding function f over non-negative values of its argument'. This relation on the long-term average rate is stated in the following theorem, and proved thereafter.

Theorem 2 *Given a flow $R \sim (S, f)$, the long-term average rate μ of flow R satisfies*

$$\mu = \limsup_{(n-k) \to \infty} \frac{\mathrm{E}[R(n) - R(k)]}{n - k} \leqslant \limsup_{n \to \infty} \frac{S(n)}{n}$$

if the area under the bounding function f is finite, i.e., $\sum_0^\infty f(\sigma) < \infty$.

[4] This is why we have assumed for mathematical convenience that a bounding function in burstiness definition 6 satisfies $f(s) = 1$ for all $s < 0$.

Proof: The proof follows by considering the mean $\mathrm{E}[R(n) - R(k)]$ of flow R in any interval $(k, n]$, and upper-bounding it as;

$$\begin{aligned}
\mathrm{E}[R(n) - R(k)] &= \sum_{\sigma=0}^{\infty} \mathrm{P}(R(n) - R(k) > \sigma) \\
&= \sum_{\sigma=0}^{S(n-k)} \mathrm{P}(R(n) - R(k) > \sigma) + \sum_{\sigma=S(n-k)}^{\infty} \mathrm{P}(R(n) - R(k) > \sigma) \\
&= \sum_{\sigma=0}^{S(n-k)} \mathrm{P}(R(n) - R(k) > \sigma) + \sum_{\sigma=0}^{\infty} \mathrm{P}(R(n) - R(k) > S(n-k) + \sigma) \\
&\leqslant \sum_{\sigma=0}^{S(n-k)} 1 + \sum_{\sigma=0}^{\infty} \mathrm{P}(R(n) - R(k) > S(n-k) + \sigma) \\
&= S(n-k) + 1 + \sum_{\sigma=0}^{\infty} \mathrm{P}(R(n) - R(k) > S(n-k) + \sigma) \\
&\leqslant S(n-k) + 1 + \sum_{\sigma=0}^{\infty} \mathrm{P}\Big(R(n) - R(u) > S(n-u) + \sigma, \quad \text{for some } u < n\Big) \\
&\leqslant S(n-k) + 1 + \sum_{\sigma=0}^{\infty} f(\sigma)\,.
\end{aligned}$$

Thus, the mean rate μ of R is upper-bounded as

$$\begin{aligned}
\mu &= \limsup_{(n-k)\to\infty} \frac{\mathrm{E}[R(n) - R(k)]}{n-k} \\
&\leqslant \limsup_{(n-k)\to\infty} \frac{S(n-k) + 1 + \sum_0^{\infty} f(\sigma)}{n-k} \\
&= \limsup_{n\to\infty} \frac{S(n)}{n}
\end{aligned}$$

since $\sum_0^{\infty} f(\sigma)$ is finite. ■

Example 1 For service curves of the form $S(n) = \max\{0, \rho(n - D)\}$, the mean rate μ of R is upper-bounded by ρ, again if the area under the bounding function f is finite. ⌟

Whenever we refer to a service curve of the form $S(n) = \max\{0, \rho(n - D)\}$, we assume ρ to be a positive integer, and D to be a non-negative finite integer, just for the sake of simplicity. We could have adopted that $S(n) = \max\{0, \lfloor \rho(n - D) \rfloor\}$, however that would have cluttered the arguments in examples.

In the following section, we study the implications of definition 6 for an S-server, whereby we show some of the other properties of this characterization pursuant to the properties of a "good" characterization mentioned earlier in the introduction.

3.2 Implications for an S-server

If a flow $R \sim (S, f)$ is fed into an S-server, the complementary cumulative distribution function of the queue size Q at the server is also upper-bounded by f. This is stated in the following theorem, and proved thereafter. The queue size $Q(n)$ is the total number of packets which resides in the server at time n; that is, if R and G denote the aggregates of the flows at the input and at the output of the server, respectively, then $Q(n) = R(n) - G(n)$.

Theorem 3 *If an input flow $R \sim (S, f)$ is fed into an S-server with service curve S, the distribution of the queue size Q at the server satisfies*

$$\mathrm{P}(Q(n) > \sigma) \leqslant f(\sigma)$$

for all $\sigma \geqslant 0$, and for all n.

Proof: Let the corresponding output flow be denoted by G. The proof follows by considering the following events for any n and $\sigma \geqslant 0$;

$$\begin{aligned}
\{Q(n) > \sigma\} &= \{R(n) - G(n) > \sigma\} \\
&\subseteq \Big\{ R(n) - \min_{k \leqslant n} \{R(k) + S(n-k)\} > \sigma \Big\} \qquad (*) \\
&= \Big\{ \max_{k \leqslant n} \{R(n) - R(k) - S(n-k)\} > \sigma \Big\} \\
&= \Big\{ R(n) - R(k) - S(n-k) > \sigma, \quad \text{for some } k < n \Big\} \\
&= \Big\{ R(n) - R(k) > S(n-k) + \sigma, \quad \text{for some } k < n \Big\}
\end{aligned}$$

thus, we obtain the desired result by taking the probability of both sides

$$\begin{aligned}
\mathrm{P}(Q(n) > \sigma) &\leqslant \mathrm{P}\Big(R(n) - R(k) > S(n-k) + \sigma, \quad \text{for some } k < n \Big) \qquad (*) \\
&\leqslant f(\sigma)\,. \qquad \blacksquare
\end{aligned}$$

Notice that this result could in fact be used to measure a bounding function f of an input traffic R for any given service curve S. To do this, we would need to construct an S-server with equality, i.e., we construct an S-server for which the output is given by equality in definition 5 (that is, $G(n) = (R \triangledown S)(n)$ for all n). Then, upon feeding a flow R into the S-server, if we observe the cumulative distribution function of the queue size Q at the server, the empirical values of the complementary cumulative distribution function would give the tightest (i.e., the smallest possible) bounding function f in characterizing the flow R as $R \sim (S, f)$. This is facilitated by the fact that when we have an S-server with equality, the relations in lines tagged $(*)$ above in the proof of the theorem 3 become an equality.

Given a flow R with unknown characterization according to definition 6, we can find a tight characterization of flow R by utilizing theorem 1 and theorem 3, if the long-term average rate μ of R is know—it is often not difficult to get a fairly good approximation of long-term average rate of a flow, by a variant of the Law of Large Numbers. By theorem 1, any service curve S to be used in characterizing R according to definition 6 should satisfy $\limsup_{n \to \infty} \frac{S(n)}{n} \geqslant \mu$. Hence, the service curve $S(n) = \max\{0, \lceil \mu \rceil \cdot n\}$ is a good candidate in order to find a characterization of R according to definition 6, as well as service curves of the from $S(n) = \max\{0, \lceil \mu \rceil\,(n - D)\}$. Then, by theorem 3, an approximation for the tightest bounding function f could be found to complete the characterization.

The output flow of an S-server fed by a flow $R \sim (S, f)$ could also be easily characterized by definition 6. This is stated in the following theorem, and proved thereafter.

Theorem 4 *The output flow G of an S-server with service curve S, fed by a flow $R \sim (S, f)$, is bursty with service curve $S \oslash S$ and the bounding function f. In other words, $G \sim (S \oslash S, f)$.*

Proof: The proof follows by considering the following events for any n, $k < n$, and σ;

$$\begin{aligned}
\Big\{G(n) - G(k) > (S \oslash S)(n-k) + \sigma\Big\} &\subseteq \Big\{G(n) - (R \otimes S)(k) > (S \oslash S)(n-k) + \sigma\Big\} \\
&\subseteq \Big\{R(n) - (R \otimes S)(k) > (S \oslash S)(n-k) + \sigma\Big\} \\
&= \Big\{R(n) - \min_{l \leqslant k}\{R(l) + S(k-l)\} > (S \oslash S)(n-k) + \sigma\Big\} \\
&= \Big\{\max_{l \leqslant k}\{R(n) - R(l) - S(k-l)\} > (S \oslash S)(n-k) + \sigma\Big\} \\
&= \Big\{R(n) - R(l) - S(k-l) > (S \oslash S)(n-k) + \sigma, \quad \text{for some } l \leqslant k\Big\} \\
&= \Big\{R(n) - R(l) > (S \oslash S)(n-k) + S(k-l) + \sigma, \quad \text{for some } l \leqslant k\Big\} \\
&\subseteq \Big\{R(n) - R(l) > S(n-l) - S(k-l) + S(k-l) + \sigma, \quad \text{for some } l \leqslant k\Big\} \\
&= \Big\{R(n) - R(l) > S(n-l) + \sigma, \quad \text{for some } l \leqslant k\Big\} \\
&= \Big\{R(n) - R(l) > S(n-l) + \sigma, \quad \text{for some } l \leqslant k < n\Big\} \\
&\subseteq \Big\{R(n) - R(l) > S(n-l) + \sigma, \quad \text{for some } l < n\Big\}.
\end{aligned}$$

Taking the union of both sides over all $k < n$ (note that the union of the right-hand-side is equal to itself), we obtain the desired result as shown below

$$\begin{aligned}
\mathrm{P}\Big(G(n) - G(k) > (S \oslash S)(n-k) + \sigma, \quad \text{for some } k < n\Big) \leqslant& \\
&\mathrm{P}\Big(R(n) - R(l) > S(n-l) + \sigma, \quad \text{for some } l < n\Big) \\
&\leqslant f(\sigma)\,.
\end{aligned}$$

■

We would actually need to slightly rectify the above result by replacing the service curve $S \oslash S$ in the characterization of the output flow by S_o given below

$$S_o(n) = \begin{cases} 0 & \text{if } n \leqslant 0 \\ (S \oslash S)(n) & \text{else.} \end{cases}$$

We would like to have this rectification for two reasons: (1) A service curve is defined to take on the value zero for non-negative values of its argument. (2) We would only need to have $S \oslash S$ for positive values of its argument (as this could be noted from the first line of the proof of the above result). We have not done this in the body of the theorem in order not to clutter the result.

Example 2 For service curves of the form $S(n) = \max\{0, \rho(n - D)\}$, the min-+ deconvolution of S with itself is equal to $\max\{0, \rho \cdot n\}$; whereby also note that for service curves of the form $S(n) = \max\{0, \rho \cdot n\}$ (i.e., when $D = 0$), the deconvolution of S with itself is equal to itself. ⌟

We would like to note by example 2 that for service curves of the form $S(n) = \max\{0, \rho \cdot n\}$ and for S-servers with equality (i.e., $G(n) = (R \overline{\nabla} S)(n)$ for all n) with this type of service curves (i.e., an S-server which is in fact a work-conserving server with rate ρ), the result proven above for the burstiness characterization of the output flow of an S-server shows that the bound we have obtained here is tighter than the corresponding result provided in [6], and hence in [7] too. Consequently, all the bounds that we would obtain in analyzing the performance guarantees in tandem networks, in the framework of definition 6, would also be tighter too.

The burstiness characterization provided by definition 6 has also an implication on the *virtual-delay* at an S-server. The *virtual-delay* is defined below, which has been previously introduced in the literature.

Definition 7 *The* virtual-delay $D(n)$ *at any time n for an input flow R at a network element is defined as*

$$D(n) = \min\{\delta : \delta \geqslant 0,\ G(n+\delta) \geqslant R(n)\}$$

where G is the corresponding output flow.

The virtual-delay $D(n)$ is basically the delay experienced by the packets arriving at time n, through the network element, if the packets are to be served in the order in which they have arrived.

If a flow $R \sim (S, f)$ is fed into an S-server, the complementary cumulative distribution function of the virtual-delay at the server is upper-bounded by the composite function $f(S \overline{\wedge} S)$. This is stated in the following theorem, and proved thereafter.

Theorem 5 *If a flow $R \sim (S, f)$ is fed into an S-server with service curve S, the distribution of the virtual-delay $D(n)$ at the server satisfies*

$$\mathrm{P}(D(n) > \sigma) \leqslant f\big((S \overline{\wedge} S)(\sigma)\big)$$

for all $\sigma \geqslant 0$, and for all n.

Proof: The proof follows by considering the following events for any n and $\sigma \geqslant 0$;

$$\begin{aligned}
\{D(n) > \sigma\} &= \{G(n+\sigma) < R(n)\} \\
&\subseteq \{(R \overline{\nabla} S)(n+\sigma) < R(n)\} \\
&= \{R(n) - (R \overline{\nabla} S)(n+\sigma) > 0\} \\
&= \Big\{R(n) - \min_{k \leqslant n+\sigma} \{R(k) + S(n+\sigma-k)\} > 0\Big\} \\
&= \Big\{\max_{k \leqslant n+\sigma} \{R(n) - R(k) - S(n+\sigma-k)\} > 0\Big\} \\
&= \Big\{R(n) - R(k) - S(n+\sigma-k) > 0, \quad \text{for some } k < n\Big\}
\end{aligned}$$

(note that a k above can not be greater than or equal to n, since in that case the left-hand-side of the inequality could not become positive)

$$\begin{aligned}
&= \Big\{R(n) - R(k) > S(n+\sigma-k), \quad \text{for some } k < n\Big\} \\
&= \Big\{R(n) - R(k) > S(n-k) + S(n-k+\sigma) - S(n-k), \quad \text{for some } k < n\Big\} \qquad (*) \\
&\subseteq \Big\{R(n) - R(k) > S(n-k) + (S \overline{\wedge} S)(\sigma), \quad \text{for some } k < n\Big\}
\end{aligned}$$

hence by taking the probability of both sides, and applying the burstiness characterization of R, we get the desired result

$$\begin{aligned} \mathrm{P}(D(n) > \sigma) &\leqslant \mathrm{P}\Big(R(n) - R(k) > S(n-k) + (S \overline{\wedge} S)(\sigma), \quad \text{for some } k < n\Big) \\ &\leqslant f\big((S \overline{\wedge} S)(\sigma)\big)\,. \end{aligned}$$

■

We could actually slightly improve the above result. This could be done if we would replace the subscript '$k \geqslant 0$' in taking the max-+ deconvolution of S with itself by '$k > 0$', as it could be noted by the line tagged $(*)$ in the derivation of the above result, since in that line we have '$k < n$'.

This result could actually be further improved if we set a time origin for flows. That is; if we assume for almost all practical purposes that flow R satisfies $R(-\infty) = 0$, i.e., there is a certain point in time before which no packet has arrived in flow R, and call that point as the origin (i.e., $n = 0$), then we could replace '$(S \overline{\wedge} S)(\sigma)$' in the above result by

$$\min_{0<k\leqslant n} \{S(k+\sigma) - S(k)\}$$

which is greater than or equal to $(S \overline{\wedge} S)(\sigma)$. This again could be noted by the line tagged $(*)$ in the derivation. However, if we do that, we would obtain a time-dependent result (i.e., the probability on the right-hand-side of the inequality in theorem 5 depends on time n), whereas the result we have here is time-independent (i.e., holds for any n, and specifically holds for the steady-state).

Example 3 For service curves of the form $S(n) = \max\{0, \rho(n-D)\}$, the max-+ deconvolution of S with itself is given by

$$(S \overline{\wedge} S)(n) = \begin{cases} -\rho \cdot n & \text{if } n < 0 \\ 0 & \text{if } 0 \leqslant n < D \\ \rho \cdot (n - D) & \text{else} \end{cases}$$

which is also equal to $\min\{\rho \cdot n, S(n)\}$ if we were to represent it more compactly. Hence for service curves of this form, the bound on the distribution of the virtual-delay $D(n)$ is given by[5]

$$f\big((S \overline{\wedge} S)(\sigma)\big) = \begin{cases} 1 & \text{if } \sigma < 0 \\ f(0) & \text{if } 0 \leqslant \sigma < D \\ f(\rho\,(\sigma - D)) & \text{else.} \end{cases}$$

⌟

These results with a little bit of more work, facilitate a systematic treatment of probabilistic performance guarantees in tandem networks, which is the subject of a later section.

3.2.1 Mean Performance Guarantees at an S-server

It immediately follows by the results in section 3.2 that we could also give mean performance guarantees at an S-server with this burstiness characterization provided by definition 6. Specifically, by the theorems 3 and 5 we could see that the mean virtual-delay and backlog are also bounded. These are pointed out by the following corollaries.

[5] With the first improvement that we have mentioned before, we could actually replace 'D' by '$D-1$' below, and obtain a tighter bound on the distribution of the virtual-delay.

Corollary 1 *If an input flow $R \sim (S, f)$ is fed into an S-server with service curve S, the mean queue size $Q(n)$ at the server at any time n is upper-bounded by the area under the bounding function f, i.e.,*

$$\mathrm{E}[Q(n)] \leqslant \sum_{\sigma=0}^{\infty} f(\sigma)\,.$$

Proof: The proof follows immediately by theorem 3, as shown below;

$$\begin{aligned}\mathrm{E}[Q(n)] &= \sum_{\sigma=0}^{\infty} \mathrm{P}(Q(n) > \sigma) \\ &\leqslant \sum_{\sigma=0}^{\infty} f(\sigma)\,.\end{aligned}$$

■

Note that, given a flow $R \sim (S, f)$ where f is the tightest possible bounding function, if the area under the bounding function f is not finite, then when flow R is fed into an S-server with equality (i.e., the output G of the server satisfies $G(n) = (R \oslash S)(n)$ for all n) the expected value of the queue size at any time n would be infinite.

Corollary 2 *If an input flow $R \sim (S, f)$ is fed into an S-server with service curve S, the mean virtual-delay $D(n)$ at the server at any time n is upper-bounded by the area under the composite function $f(S \overline{\oslash} S)$, i.e.,*

$$\mathrm{E}[D(n)] \leqslant \sum_{\sigma=0}^{\infty} f(S \overline{\oslash} S)(\sigma)\,.$$

Proof: The proof follows immediately by theorem 5, as shown below;

$$\begin{aligned}\mathrm{E}[D(n)] &= \sum_{\sigma=0}^{\infty} \mathrm{P}(D(n) > \sigma) \\ &\leqslant \sum_{\sigma=0}^{\infty} f(S \overline{\oslash} S)(\sigma)\,.\end{aligned}$$

■

Example 4 For service curves of the form $S(n) = \max\{0, \rho(n - D)\}$, we have pointed out earlier what the composite function $f(S \overline{\oslash} S)$ would be, in example 3. So, for service curves of this form, the mean virtual-delay $D(n)$ at an S-server at any time n is upper-bounded by the above corollary as

$$\begin{aligned}\mathrm{E}[D(n)] &\leqslant f(0) \cdot D + \sum_{\sigma=D}^{\infty} f(\rho(\sigma - D)) \\ &= f(0) \cdot D + \sum_{\sigma=0}^{\infty} f(\rho\sigma)\,.\end{aligned}$$

Since, $f(\sigma) \leqslant 1$, we would also have

$$\mathrm{E}[D(n)] \leqslant D + \sum_{\sigma=0}^{\infty} f(\rho\sigma)\,.$$

⌟

3.3 Performance Guarantees Over A Tandem of Network Elements

Performance guarantees, in the framework of definition 6, over a tandem of network elements follow almost directly from the results in section 3.2 where implications of the burstiness characterization provided by definition 6 is studied for an S-server. In order to uncover such performance guarantees, we would need to generalize the results in section 3.2 slightly. These generalizations are realized when we consider feeding a flow characterized according to definition 6 with a service curve S^*, into an S-server with service curve S where S is not necessarily the same as S^*. In the following subsection, we provide these generalizations. We also provide both time-dependent and time-independent results in this section, as pointed out earlier after the proof of theorem 5.

3.3.1 Generalized Implications for an S-server

The derivations to obtain the results in this section are similar to those of section 3.2, and can be followed in parallel. Some of the explanations that might be needed further for the results in this section could be found in section 3.2 where corresponding results are provided.

We first give the following definitions in order to compactly state some of the results in this section.

Definition 8 *Let f and g be two functions. The operator $\ominus$ for any given $n > 0$, operating on f and g, is defined as*

$$(f \ominus g)(\sigma; n) \triangleq \min_{0<k\leqslant n} \{f(\sigma + k) - g(k)\}$$

for all σ.

Notice that the operator $\ominus$ for any given $n > 0$ is in fact a variant of max-+ deconvolution. The only difference between the two is that the index of the terms over which the minimum is taken is limited to $0 < k \leqslant n$ in $\ominus$, whereas it is '$k \geqslant 0$' in max-+ deconvolution. Thus, notice also that we have

$$(f \ominus g)(\sigma; n) \geqslant (f \bar{\wedge} g)(\sigma) \quad \text{for all } \sigma, \text{ and } n > 0.$$

The function $(f \ominus g)(\sigma; n)$ is read as 'f min-diff with g by $(\sigma; n)$'.

We also define another variant of max-+ deconvolution, based on the above definition, which is given below.

Definition 9 *Let f and g be two functions. The operator $\ominus$, operating on f and g, is defined as*

$$(f \ominus g)(\sigma) \triangleq \min_{k>0} \{f(k + \sigma) - g(k)\}$$

for all σ.

Again notice that the only difference between $(f \ominus g)(\sigma)$ and $(f \bar{\wedge} g)(\sigma)$ is the subscripts of the minimums, which are '$k > 0$' and '$k \geqslant 0$', respectively. Thus, we also have

$$(f \ominus g)(\sigma) \geqslant (f \bar{\wedge} g)(\sigma) \quad \text{for all } \sigma. \tag{2}$$

The functional $(f \ominus g)(\sigma)$ is basically the value of the function $(f \ominus g)(\sigma; n)$ for infinitely large n. We read the function $f \ominus g$ as 'f min-diff with g', or as 'the minimum difference of f from g'.

In this section, we assume for almost all practical purposes that any flow R has its packets arriving after a finite point in time (i.e., $R(-\infty) = 0$), unless otherwise noted from here on. We call that finite point in time after which packets start arriving as the time origin, i.e., time 0. In

other words, we assume $R(n) = 0$ for all $n \leqslant 0$. We relax this assumption occasionally in some places in the text.

We use a function of the form $(f \ominus g)(0;n)$ in the following theorem, which stands for the minimum vertical distance between the functions f and g to the right of the origin and by time n, as this could be noted by definition 8. With that, we claim the following theorem.

Theorem 6 *If an input flow $R \sim (S^*, f)$ is fed into an S-server with service curve S, the distribution of the queue size Q at the server satisfies*

$$\mathrm{P}(Q(n) > \sigma) \leqslant f\big((S \ominus S^*)(0;n) + \sigma\big)$$

for all $\sigma \geqslant 0$, and for all $n > 0$.

Proof: Let the corresponding output flow be denoted by G. The proof follows by considering the following events for any $n > 0$ and $\sigma \geqslant 0$;

$$\begin{aligned}
\{Q(n) > \sigma\} &= \{R(n) - G(n) > \sigma\} \\
&\subseteq \Big\{R(n) - \min_{k \leqslant n}\{R(k) + S(n-k)\} > \sigma\Big\} \\
&= \Big\{\max_{k \leqslant n}\{R(n) - R(k) - S(n-k)\} > \sigma\Big\} \\
&= \Big\{R(n) - R(k) - S(n-k) > \sigma, \quad \text{for some } k < n\Big\} \\
&= \Big\{R(n) - R(k) > S(n-k) + \sigma, \quad \text{for some } k < n\Big\} \\
&= \Big\{R(n) - R(k) > S^*(n-k) + S(n-k) - S^*(n-k) + \sigma, \quad \text{for some } k < n\Big\}
\end{aligned}$$

notice that a k above is greater than or equal to 0, hence we have

$$\subseteq \Big\{R(n) - R(k) > S^*(n-k) + (S \ominus S^*)(0;n) + \sigma, \quad \text{for some } k < n\Big\}$$

thus, we obtain the desired result by taking the probability of both sides

$$\begin{aligned}
\mathrm{P}(Q(n) > \sigma) &\leqslant \mathrm{P}\Big(R(n) - R(k) > S^*(n-k) + (S \ominus S^*)(0;n) + \sigma \quad \text{for some } k < n\Big) \\
&\leqslant f\big((S \ominus S^*)(0;n) + \sigma\big).
\end{aligned}$$

■

Notice that this is a time-dependent result, i.e., the probability on the right-hand-side of the inequality in theorem 6 depends on time n.

Example 5 For service curves of the form

$$\begin{aligned}
S^*(n) &= \max\{0, \rho^*(n - D^*)\} \\
S(n) &= \max\{0, \rho(n - D)\}
\end{aligned}$$

note the following calculations:

1. If $\rho^* > \rho$, we have

$$\text{if } D^* \geqslant D, \qquad (S \ominus S^*)(0;n) = \min\big\{S(1), -(\rho^* - \rho)n + \rho^* D^* - \rho D\big\} \quad \text{for all } n$$

$$\text{if } D^* \leqslant D, \qquad (S \ominus S^*)(0;n) = \begin{cases} 0 & \text{if } 0 < n \leqslant D^* \\ -\rho^*(n - D^*) & \text{if } D^* < n \leqslant D \\ -(\rho^* - \rho)n + \rho^* D^* - \rho D & \text{if } n > D. \end{cases}$$

Since for any $\sigma \geqslant 0$, $(S \ominus S^*)(0;n)$ will eventually become smaller than $-\sigma$ for all n after a certain large enough n_o, the bound on the tail of the queue size distribution for $n > n_o$ would be of no use by this result. However, for times earlier than n_o, it would carry some useful information. Hence, for $\rho^* > \rho$, this result might be considered useful only for some flows of finite number of packets. Thus, one can assume for most practical purposes that $\rho^* \leqslant \rho$.[6]

2. For $\rho^* \leqslant \rho$, if $D^* \geqslant D$, then $(S \ominus S^*)(0;n) = S(1) - S^*(1)$ for any $n > 0$. In this case, the bound on the tail of the queue size distribution in theorem 6 becomes $f(\sigma + S(1) - S^*(1))$, and time-independent.

3. For $\rho^* \leqslant \rho$ and $D^* < D$, we have

$$(S \ominus S^*)(0;n) = \begin{cases} 0 & \text{if } 0 < n \leqslant D^* \\ -\rho^*(n - D^*) & \text{if } D^* < n \leqslant D \\ -\rho^*(D - D^*) & \text{if } n > D \end{cases}$$

hence, the bound on the tail of the queue size distribution in theorem 6 becomes

$$f\big((S \ominus S^*)(0;n) + \sigma\big) = \begin{cases} f(\sigma) & \text{if } 0 < n \leqslant D^* \\ f(\sigma - \rho^*(n - D^*)) & \text{if } D^* < n \leqslant D \\ f(\sigma - \rho^*(D - D^*)) & \text{if } n > D. \end{cases}$$

The time-independent version of theorem 6 is given below as a corollary. If we choose to relax the assumption on flow R that $R(n) = 0$ for all $n < 0$, as we have set before stating the theorem 6, then we would need the following corollary.

Corollary 3 *If an input flow $R \sim (S^*, f)$ is fed into an S-server with service curve S, the distribution of the queue size Q at the server satisfies*

$$\mathrm{P}(Q(n) > \sigma) \leqslant f\big((S \ominus S^*)(0) + \sigma\big)$$

for all $\sigma \geqslant 0$, and for all n.

Proof: The proof follows by theorem 6, and by letting n go to infinity. ■

Notice that we also have

$$\mathrm{P}(Q(n) > \sigma) \leqslant f\big((S \barwedge S^*)(0) + \sigma\big)$$

for all $\sigma \geqslant 0$ and n, since f is non-increasing and by relation (2).

Note that if service curve S of the S-server is greater than or equal to S^* of the flow at every point, then the bound on the tail of the queue size distribution becomes $f(\sigma)$.

Example 6 For service curves of the form as mentioned in example 5, and taking on from the last observation mentioned therein, we have

$$\begin{aligned} f\big((S \ominus S^*)(0) + \sigma\big) &= f(\sigma - \rho^*(D - D^*)) \\ &= \begin{cases} 1 & \text{if } \sigma < \rho^*(D - D^*) \\ f(\sigma - \rho^*(D - D^*)) & \text{else.} \end{cases} \end{aligned}$$

[6] Although the result that we have here does not directly imply, we suspect that the queue in the case $\rho^* > \rho$ would become unstable (i.e., with probability 1 it will be infinitely large, and remain that way, after a finite point in time) for any unbounded flow $R \sim (S^*, f)$ (i.e., $R(\infty) = \infty$).

The generalized result for the output flow of an S-server fed by a flow $R \sim (S^*, f)$ is given below. We would again leave the rectification of the service curve in the characterization of the output flow after the theorem, as we have done before for the corresponding result in theorem 4. We also choose to relax the assumption on flow R that $R(n) = 0$ for all $n < 0$, for setting a time origin as indicated before stating the theorem 6, for the following result as well.

Theorem 7 *The output flow G of an S-server with service curve S, fed by a flow $R \sim (S^*, f)$, is bursty with service curve $S^* \overline{\oslash} S$ and the bounding function f. In other words, $G \sim (S^* \overline{\oslash} S, f)$.*

Proof: The proof follows by considering the following events for any n, $k < n$, and σ;

$$\begin{aligned}
\Big\{G(n) - G(k) > (S^* \overline{\oslash} S)(n-k) + \sigma\Big\} &\subseteq \Big\{G(n) - (R \overline{\oslash} S)(k) > (S^* \overline{\oslash} S)(n-k) + \sigma\Big\} \\
&\subseteq \Big\{R(n) - (R \overline{\oslash} S)(k) > (S^* \overline{\oslash} S)(n-k) + \sigma\Big\} \\
&= \Big\{R(n) - \min_{l \leqslant k}\{R(l) + S(k-l)\} > (S^* \overline{\oslash} S)(n-k) + \sigma\Big\} \\
&= \Big\{\max_{l \leqslant k}\{R(n) - R(l) - S(k-l)\} > (S^* \overline{\oslash} S)(n-k) + \sigma\Big\} \\
&= \Big\{R(n) - R(l) - S(k-l) > (S^* \overline{\oslash} S)(n-k) + \sigma, \quad \text{for some } l \leqslant k\Big\} \\
&= \Big\{R(n) - R(l) > (S^* \overline{\oslash} S)(n-k) + S(k-l) + \sigma, \quad \text{for some } l \leqslant k\Big\} \\
&\subseteq \Big\{R(n) - R(l) > S^*(n-l) - S(k-l) + S(k-l) + \sigma, \quad \text{for some } l \leqslant k\Big\} \\
&= \Big\{R(n) - R(l) > S^*(n-l) + \sigma, \quad \text{for some } l \leqslant k\Big\} \\
&= \Big\{R(n) - R(l) > S^*(n-l) + \sigma, \quad \text{for some } l \leqslant k < n\Big\} \\
&\subseteq \Big\{R(n) - R(l) > S^*(n-l) + \sigma, \quad \text{for some } l < n\Big\}.
\end{aligned}$$

Taking the union of both sides over all $k < n$ (note that the union of the right-hand-side is equal to itself), we obtain the desired result as shown below

$$\begin{aligned}
\mathrm{P}\Big(G(n) - G(k) > (S^* \overline{\oslash} S)(n-k) + \sigma, \quad \text{for some } k < n\Big) \leqslant& \\
&\mathrm{P}\Big(R(n) - R(l) > S^*(n-l) + \sigma, \quad \text{for some } l < n\Big) \\
&\leqslant f(\sigma)\,. \qquad \blacksquare
\end{aligned}$$

We would again need to slightly rectify the above result by replacing the service curve $S^* \overline{\oslash} S$ in the characterization of the output flow by S_o given below

$$S_o(n) = \begin{cases} 0 & \text{if } n \leqslant 0 \\ (S^* \overline{\oslash} S)(n) & \text{else.} \end{cases}$$

We would like to have this rectification by the same reasons given for the corresponding rectification done after theorem 4.

Example 7 For service curves of the form

$$\begin{aligned}
S^*(n) &= \max\{0, \rho^*(n - D^*)\} \\
S(n) &= \max\{0, \rho(n - D)\}
\end{aligned}$$

note the following calculations:

1. If $\rho^* > \rho$, regardless of the values of D^* and D, $(S^* \overline{\oslash} S)(n)$ becomes infinite for any positive value of n. In this case, the result we have here does not provide any useful information for burstiness characterization of the output flow, since in that case its bounding function could be set to $f(\sigma) = 0$ for all $\sigma \geqslant 0$ without loss of generality. Also for the reasons given at the end of observation 1 in example 5, one can assume for most practical purposes that $\rho^* \leqslant \rho$.

2. With this assumption, regardless of the values of D^* and D, we have

$$(S^* \overline{\oslash} S)(n) = \min\left\{0, \rho^*(n - (D^* - D))\right\} \quad \text{for all } n.$$

Note, however, that the shape of S_o will be different for $D^* \geqslant D$ and $D^* < D$. ⌟

By setting $D^* = D = 0$ in example 7, and for S-servers with equality, note that the bound we have obtained here for characterizing the output flow is tighter than the corresponding result provided in [6], and hence in [7] too. Consequently, all the bounds that we obtain in analyzing the performance guarantees in tandem networks, in the framework of definition 6, are tighter as well.

The generalized result for a virtual-delay experienced at an S-server fed by a flow $R \sim (S^*, f)$ is given below. The definition of virtual delay is given by definition 7. The assumption on flow R that $R(n) = 0$ for all $n < 0$, for setting a time origin as we have indicated before stating the theorem 6, is intact here.

Theorem 8 *If a flow $R \sim (S^*, f)$ is fed into an S-server with service curve S, the distribution of the virtual-delay $D(n)$ at the server satisfies*

$$\mathrm{P}(D(n) > \sigma) \leqslant f\big((S \ominus S^*)(\sigma; n)\big)$$

for all $\sigma \geqslant 0$, and for all $n > 0$.

Proof: The proof follows by considering the following events for any $n > 0$ and $\sigma \geqslant 0$;

$$\begin{aligned}
\{D(n) > \sigma\} &= \{G(n+\sigma) < R(n)\} \\
&\subseteq \{(R \overline{\oslash} S)(n+\sigma) < R(n)\} \\
&= \{R(n) - (R \overline{\oslash} S)(n+\sigma) > 0\} \\
&= \Big\{R(n) - \min_{k \leqslant n+\sigma} \{R(k) + S(n+\sigma-k)\} > 0\Big\} \\
&= \Big\{\max_{k \leqslant n+\sigma} \{R(n) - R(k) - S(n+\sigma-k)\} > 0\Big\} \\
&= \Big\{R(n) - R(k) - S(n+\sigma-k) > 0, \quad \text{for some } k < n\Big\}
\end{aligned}$$

(note that a k above can not be greater than or equal to n, since in that case the left-hand-side of the inequality could not become positive)

$$\begin{aligned}
&= \Big\{R(n) - R(k) > S(n+\sigma-k), \quad \text{for some } k < n\Big\} \\
&= \Big\{R(n) - R(k) > S^*(n-k) + S(n-k+\sigma) - S^*(n-k), \quad \text{for some } k < n\Big\} \\
&\subseteq \Big\{R(n) - R(k) > S^*(n-k) + (S \ominus S^*)(\sigma; n), \quad \text{for some } k < n\Big\}
\end{aligned}$$

hence by taking the probability of both sides, and applying the burstiness characterization of R, we get the desired result

$$\begin{aligned} \mathrm{P}(D(n) > \sigma) &\leqslant \mathrm{P}\Big(R(n) - R(k) > S^*(n-k) + (S \oslash S^*)(\sigma; n), \quad \text{for some } k < n\Big) \\ &\leqslant f\big((S \oslash S^*)(\sigma; n)\big)\,. \end{aligned}$$

■

Notice that this is a time-dependent result (i.e., the probability on the right-hand-side of the inequality in theorem 8 depends on time n).

Example 8 For service curves of the form

$$\begin{aligned} S^*(n) &= \max\{0, \rho^*(n - D^*)\} \\ S(n) &= \max\{0, \rho(n - D)\} \end{aligned}$$

note the following calculations:

1. If $\rho^* > \rho$, we have for any $\sigma \geqslant 0$

   $$\text{if } \sigma > D - D^*, \quad (S \oslash S^*)(\sigma; n) = \min\big\{S(\sigma+1), -(\rho^* - \rho)n + \rho^* D^* - \rho D + \rho\sigma\big\} \quad \text{for all } n$$

   $$\text{if } \sigma \leqslant D - D^*, \quad (S \oslash S^*)(\sigma; n) = \begin{cases} 0 & \text{if } 0 < n \leqslant D^* \\ -\rho^*(n - D^*) & \text{if } D^* < n \leqslant D - \sigma \\ -(\rho^* - \rho)n + \rho^* D^* - \rho D + \rho\sigma & \text{if } n > D - \sigma. \end{cases}$$

   For any given $\sigma \geqslant 0$, $(S \oslash S^*)(\sigma; n)$ will eventually become negative for all n larger than a certain large enough n_o. Thus, the bound on the tail of the virtual-delay distribution for $n > n_o$ would be of no use by this result. However, for $n \leqslant n_o$, the result would carry some useful information. For $\rho^* > \rho$, this result might be considered useful only for some flows of finite number of packets, as mentioned earlier at the end of calculation 1 in example 5. So, one can assume for most practical purposes that $\rho^* \leqslant \rho$.[7]

2. For $\rho^* \leqslant \rho$, if $D^* \geqslant D$, then $(S \oslash S^*)(\sigma; n) = S(\sigma+1) - S^*(1)$ for any $n > 0$. In this case, the bound on the tail of the distribution of virtual-delay in theorem 8 becomes $f(S(\sigma+1) - S^*(1))$, and time-independent.

3. For $\rho^* \leqslant \rho$ and $D^* < D$, we have for $\sigma \geqslant 0$

   $$\text{for } \sigma \leqslant D - D^* \quad (S \oslash S^*)(\sigma; n) = \begin{cases} 0 & \text{if } 0 < n \leqslant D^* \\ -\rho^*(n - D^*) & \text{if } D^* < n \leqslant D - \sigma \\ -\rho^*(D - D^* - \sigma) & \text{if } n > D - \sigma \end{cases}$$

   $$\text{for } \sigma > D - D^* \quad (S \oslash S^*)(\sigma; n) = S(\sigma+1) - S^*(1)\,.$$

[7] Although the result that we have here does not directly imply, we think that in the case $\rho^* > \rho$, almost all the packets would experience unbounded delays with probability 1, since we suspect that the queue in this case the would become unstable (i.e., with probability 1 it will be infinitely large, and remain that way, after a finite point in time) for any unbounded flow $R \sim (S^*, f)$ (i.e., $R(\infty) = \infty$).

Hence, the bound on the tail of the virtual-delay distribution in theorem 8 becomes

$$\text{for } \sigma < D - D^* \quad f\big((S \ominus S^*)(\sigma; n)\big) = \begin{cases} f(0) & \text{if } 0 < n \leqslant D^* \\ 1 & \text{if } n > D^* \end{cases}$$

$$\text{for } \sigma = D - D^* \quad f\big((S \ominus S^*)(\sigma; n)\big) = f(0) \quad \text{for all } n$$

$$\begin{aligned} \text{for } \sigma > D - D^* \quad f\big((S \ominus S^*)(\sigma; n)\big) &= f\big(S(\sigma + 1) - S^*(1)\big) \\ &= \begin{cases} f(-S^*(1)) & \text{if } \sigma < D \\ f\big(\rho(\sigma + 1 - D) - S^*(1)\big) & \text{else.} \end{cases} \end{aligned}$$

The time-independent version of the above result is given below as a corollary. If we choose to relax the assumption on flow R that $R(n) = 0$ for all $n < 0$, as we have set before stating the theorem 6, then we would need the following corollary.

Corollary 4 *If a flow $R \sim (S^*, f)$ is fed into an S-server with service curve S, the distribution of the virtual-delay $D(n)$ at the server satisfies*

$$\mathrm{P}(D(n) > \sigma) \leqslant f\big((S \ominus S^*)(\sigma)\big)$$

for all $\sigma \geqslant 0$, and n.

Proof: The proof follows by theorem 8, and by letting n go to infinity. ■

Notice that we also have

$$\mathrm{P}(D(n) > \sigma) \leqslant f\big((S \barwedge S^*)(\sigma)\big)$$

for all $\sigma \geqslant 0$ and n, since f is non-increasing and by relation (2).

All the theorems and corollaries in this section essentially facilitate a systematic treatment of probabilistic performance guarantees over a tandem of network elements. To further clarify this, we give the results in the following subsection.

3.3.2 Considering a Tandem of Network Elements

In order to clarify the claimed systematic treatment of performance guarantees over a tandem of network elements, we would like to give the following two theorems and two corollaries.

Theorem 9 *Let f, g, and h be any three functions. The following property holds*

$$((f \overline{\otimes} g) \overline{\otimes} h)(n) \leqslant (f \overline{\otimes} (g \overline{\otimes} h))(n) \quad \textit{for all } n.$$

Proof: The proof follows directly from the definitions of min-+ deconvolution and min-+ convolution, and is given below. It holds for all n that

$$\begin{aligned} ((f \overline{\otimes} g) \overline{\otimes} h)(n) &= \max_{k \geqslant 0} \{(f \overline{\otimes} g)(n + k) - h(k)\} \\ &= \max_{k \geqslant 0} \Big\{ \max_{l \geqslant 0} \big\{ f(n + k + l) - g(l) \big\} - h(k) \Big\} \\ &= \max_{k \geqslant 0} \Big\{ \max_{l \geqslant 0} \big\{ f(n + k + l) - g(l) - h(k) \big\} \Big\} \\ &= \max_{\substack{k \geqslant 0 \\ l \geqslant 0}} \big\{ f(n + k + l) - (g(l) + h(k)) \big\} \end{aligned}$$

notice that for all k and l such that $k+l$ would remain the same, $f(n+k+l)$ does not change its value for any given n, hence the maximum above for that fixed value of $k+l$ will occur for the minimum value of $g(l)+h(k)$ over all such k's and l's; that is, we have

$$\begin{aligned}
&= \max_{k+l\geqslant 0}\Big\{f(n+k+l) - \min_{0\leqslant u\leqslant k+l}\big\{g(u)+h(k+l-u)\big\}\Big\}\\
&\leqslant \max_{k+l\geqslant 0}\Big\{f(n+k+l) - \min_{u\leqslant k+l}\big\{g(u)+h(k+l-u)\big\}\Big\}\\
&= \max_{k+l\geqslant 0}\big\{f(n+k+l) - (g\,\overline{\nabla}\,h)(k+l)\big\}\\
&= \max_{u\geqslant 0}\big\{f(n+u) - (g\,\overline{\nabla}\,h)(u)\big\}\\
&= (f\,\overline{\nabla}\,(g\,\overline{\nabla}\,h))(n)\,.
\end{aligned}$$

■

Theorem 10 *Let f, g, and h be any three functions. The following property holds*

$$(f\ominus(g\,\underline{\nabla}\,h))(n)\geqslant((f\,\overline{\nabla}\,h)\ominus g)(n)\quad \textit{for all } n.$$

Proof: The proof follows directly from the definitions of the operator $\ominus$ and min-$+$ deconvolution, and is given below. It holds for all n that

$$\begin{aligned}
(f\ominus(g\,\underline{\nabla}\,h))(n) &= \min_{k>0}\{f(n+k) - (g\,\underline{\nabla}\,h)(k)\}\\
&= \min_{k>0}\Big\{f(n+k) - \max_{l\geqslant 0}\big\{g(k+l)-h(l)\big\}\Big\}\\
&= \min_{k>0}\Big\{f(n+k) + \min_{l\geqslant 0}\big\{-g(k+l)+h(l)\big\}\Big\}\\
&= \min_{\substack{k>0\\ l\geqslant 0}}\big\{f(n+k) - g(k+l) + h(l)\big\}\\
&= \min_{\substack{k>0\\ l\geqslant 0}}\big\{f(n+k) + h(l) - g(k+l)\big\}
\end{aligned}$$

notice that for all k and l such that $k+l$ would remain the same, $g(k+l)$ does not change its value, hence the minimum above for that fixed value of $k+l$ will occur for the minimum value of $f(n+k)+h(l)$ over all such k's and l's; that is, we have

$$\begin{aligned}
&= \min_{k+l>0}\Big\{\min_{0\leqslant u\leqslant n+k+l}\big\{f(n+k+l-u)+h(u)\big\} - g(k+l)\Big\}\\
&\geqslant \min_{k+l>0}\Big\{\min_{u\leqslant n+k+l}\big\{f(n+k+l-u)+h(u)\big\} - g(k+l)\Big\}\\
&= \min_{k+l>0}\big\{(f\,\overline{\nabla}\,h)(n+k+l) - g(k+l)\big\}\\
&= \min_{u>0}\big\{(f\,\overline{\nabla}\,h)(n+u) - g(u)\big\}\\
&= ((f\,\overline{\nabla}\,h)\ominus g)(n)\,.
\end{aligned}$$

■

The next result is a corollary of theorem 1.

Corollary 5 *Let Z_1 and Z_2 be two random processes that*

$$\mathrm{P}(Z_i(n)>\sigma)\leqslant f_i(\sigma)\quad \textit{for all } \sigma, \textit{ and for } i \textit{ equals to both } 1 \textit{ and } 2$$

where functions f_i's are as specified after definition 6. The following relation holds

$$\mathrm{P}((Z_1+Z_2)(n)>\sigma)\leqslant (f_1\,\overline{\nabla}\,f_2)(\sigma)\quad \textit{for all } \sigma.$$

Proof: The proof follows immediately from theorem 1. This could be seen from its proof if we were to shovel the service curves appearing in the definitions of the events A and A_i's to the other side of the inequalities, and view the left-hand-side of the inequalities as $Z_1 + Z_2$ and Z_i's respectively. ■

We now give the main result of this section as a corollary.

Corollary 6 *Let a flow $R_1 \sim (S^*, f)$ be fed into an S-server with service curve S_1. And let the output R_2 of this first server be fed into another S-server with service curve S_2. The following statements hold:*

1. *The output flow R_3 of the S-server with service curve S_2 is bursty with service curve S_3 and bounding function f, where*

$$S_3(n) = \begin{cases} 0 & \text{if } n \leqslant 0 \\ (S^* \overline{\nabla} (S_1 \nabla S_2)) & \text{else.} \end{cases}$$

   In other words, $R_3 \sim (S_3, f)$.

2. *The total number of packets, $Q_1(n) + Q_2(n)$, stored in the tandem network at any time n satisfies*

$$\mathrm{P}((Q_1 + Q_2)(n) > \sigma) \leqslant (g \nabla h)(\sigma)$$

   where

$$g(\sigma) = f\big((S_1 \ominus S^*)(0) + \sigma\big)$$

$$\begin{aligned} h(\sigma) &= f\big((S_2 \ominus (S^* \overline{\nabla} S_1))(0) + \sigma\big) \\ &= f\big((S_1 \nabla S_2) \ominus S^*)(0) + \sigma\big)\,. \end{aligned}$$

3. *The total virtual-delay, $D_1(n) + D_2(n+\delta)$, experienced by a packet arriving at any time n at the first network element, and at time $n + \delta$ for some $\delta \geqslant 0$ at the second network element, satisfies for any $\delta \geqslant 0$*

$$\mathrm{P}(D_1(n) + D_2(n+\delta)(n) > \sigma) \leqslant (g \nabla h)(\sigma)$$

   where

$$g(\sigma) = f\big((S_1 \ominus S^*)(\sigma)\big)$$

$$\begin{aligned} h(\sigma) &= f\big((S_2 \ominus (S^* \overline{\nabla} S_1))(\sigma)\big) \\ &= f\big((S_1 \nabla S_2) \ominus S^*)(\sigma)\big)\,. \end{aligned}$$

Proof: The proof follows for

- statement 1 by theorems 7 and 9, where notice for theorem 9 that the inequality in the theorem becomes and equality since S_1 and S_2 here are service curves,
- statement 2 by corollaries 5 and 3, and theorem 10, where notice for theorem 10 that the inequality in the theorem becomes and equality since S_1 and S_2 here are service curves,
- statement 3 by corollaries 5 and 4, and theorem 10, where again notice for theorem 10 that the inequality in the theorem becomes and equality since S_1 and S_2 here are service curves. ■

We could obtain a similar result for any number of network elements in tandem, by a repeated application of corollary 6. Thus, it has now become clear that the burstiness characterization 6 does indeed facilitate a systematic treatment of probabilistic performance guarantees over a tandem of network elements.

Finally, one might want to demonstrate the results stated in corollary 6 by an example set below, which is left as an exercise.

Example 9 A flow $R_1 \sim (S^*, f)$ is fed into an S-server with service curve S_1. The output R_2 of this first server is fed into another S-server with service curve S_2. The service curves S^*, S_1, and S_2 are as given below

$$S^*(n) = \max\{0, \rho^*(n - D^*)\}$$
$$S_1(n) = \max\{0, \rho_1(n - D_1)\}$$
$$S_2(n) = \max\{0, \rho_2(n - D_2)\}$$

where we assume $\rho^* \leqslant \min\{\rho_1, \rho_2\}$ for the practical reasons given in examples 5, 7, and 8. We are interested in finding the characterizations of the output flow R_3, the total queue size of the tandem network, and the total virtual-delay experienced through the tandem network in the framework of our definition, as given by corollary 6. One can work out the example for any choices of D^*, D_1, and D_2, again facilitated by the examples 6, 7, and 8 (where n is let go to infinity). □

Mean performance guarantees for a tandem of network elements could also be easily calculated via the results in this section, as we have done in section 3.2.1 similarly.

4 Conclusion

In this study, we have proposed a probabilistic definition of burstiness characterization for network traffic, based on service curves. We have shown that this characterization satisfies all the properties of a "good" characterization that we have stated in the introduction. Specifically, the new characterization facilitates a systematic treatment of stochastic performance guarantees, analytically. The characterization of a flow according to this definition is measurable by constructing an S-server with equality, and observing the queue size distribution upon feeding a flow into that server. It is also a realistic characterization, in the sense that it can address network traffic with arbitrary *heavy-tailed* distributions, including the *power-tailed* distributions (e.g., Pareto), which have been recently observed, in performance analyses.

The performance bounds provided by this characterization are tighter than what is currently available by other similar approaches, e.g., [6, 7, 8]). The proposed characterization here does not also have the restriction on a bounding function (i.e., f) as specified in [6]. The new characterization is also defined more generally by basing it on service curves.

Future work includes conducting experiments to characterize traffic in various networking topologies. We will first build various S-servers to conduct these experiments. Implications of this characterization for multiplexers and switches will also be examined.

It would be interesting to examine a similar burstiness characterization and its implications, where the new definition would also have a lower-bound on the probability given in definition 6 as well as an upper-bound. This could be viewed as being motivated by a desire to give more useful statements regarding the case $\rho^* > \rho$ which is mentioned in examples 5, 7, and 8.

References

[1] A. K. Parekh, R. G. Gallagher. *A generalized processor sharing approach to flow control in integrated services networks: the single-node case*, IEEE/ACM Transaction on Networking, vol. 1, pp. 344–357, 1993.

[2] R. L. Cruz. *Quality of Service Guarantees in Virtual Circuit Switched Networks*, IEEE Journal of Selected Areas in Communication, 13(6): 1048–1056, 1995.

[3] D. Ferrari, D. Verma. *A Scheme for Real-Time Channel Establishment in Wide-Area Networks*, IEEE Journal on Selected Areas in Communications, vol. 8, pp. 368–379, April 1990.

[4] Z. Wang, J. Crowcroft. *Analysis of Burstiness and Jitter in Real-Time Communications*, The Proceedings of SIGCOMM, pp. 13–19, 1993.

[5] K. Park, W. Willinger (editors). *Self-Similar Network Traffic and Performance Evaluation*, John Wiley & Sons, 2000.

[6] D. Starobinski, M. Sidi. *Stochastically Bounded Burstiness for Communication Networks*, IEEE Trans. on Information Theory, vol. 46, no. 1, Jan. 2000. (Also appeared in the Proceedings of Infocom 1999, pp. 36–42.)

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

[8] C.S.-Chang. *Stability, queue length and delay of deterministic and stochastic queuing networks*, IEEE Trans. on Automatic Control, vol. 39, pp. 913–931, 1994.

[9] E. Reich. *On the Integrodifferential Equation of Takasc, I*, Ann. Math. Stat., vol. 29, pp. 563–570, 1958.

[10] S. Ayyorgun, W.-C. Feng. *Notes on Burstiness and Bursty Traffic Characterization*, Technical Report, Los Alamos National Laboratory.