A Deterministic Characterization of Network Traffic for Average Performance Guarantees

Sami Ayyorgun* Wu-chun Feng*,§

Abstract—*We propose a deterministic characterization of network traffic, based on service curves. The proposed characterization facilitates 1) performance analyses for both average and scalar worst-case performance guarantees, 2) a systematic approach to performance analyses (specifically, we show that queue-size, output traffic, virtual-delay, aggregate traffic, etc. at various points in a network can easily be characterized within the framework of the proposed definition), and 3) a systematic approach to measurement-based analyses of probabilistic performance guarantees that are inferred via sample-path computations.*

We also discuss the notion of burstiness. We indicate that it is the decay-rate of the tail of the queue size distribution that we observe in deciding the degree of burstiness of a flow with respect to another one, after some appropriate normalizations of the flows. The faster the decay-rate is, the less bursty the traffic is, and vice versa.

Keywords—Traffic Characterization, Network Calculus, QoS, Measurement-Based Performance Analysis, Queueing.

1 Introduction

A few deterministic characterizations of network traffic had been introduced in the early 90s, e.g. see [1, 2, 3]. Deterministic characterizations of network traffic are appealing to researchers for two reasons: First, they provide insight into subtle issues in various performance analyses problems (e.g. see [1, 2, 3]). Second, they are easier to work with (i.e. they facilitate tractable analyses) compared to some complex probabilistic traffic models. However, this second appealing reason happens at the expense of having loose performance bounds, hence underutilization of network resources (e.g. link rates and buffer spaces).

The characterization introduced in [2] and its companion service model (the *service-curve model*, introduced in [4, 5]) have received considerable attention from the networking community, which recently have led to two books ([6] and [7]) on the subject matter. A primary reason for this attention is due to the systematic approach that they have facilitated towards performance analyses in communication networks. The facilitated systematicness is analogous to that in Linear System Theory.

Deterministic characterizations of network traffic introduced in the literature previously have not considered average performance guarantees, to the best of our knowledge. Their main focus was on some scalar worst-case performance metrics; such as delay, backlog, or jitter being less than or equal to a certain scalar quantity, at every point in time at a network element. Average performance guarantees, on the other hand, are more relevant to and demanded by many multimedia applications. Motivated partly by these demands, a few studies on the implications of the characterization in [2] on average-case performance guarantees have appeared in the literature in recent years, e.g. see [8, 9].

In this study, we propose a deterministic characterization of network traffic, which can address both average and scalar worst-case performance guarantees. Our goal in this study is threefold:

a. To clarify the notion of burstiness and how we might want to perceive it.

b. To come up with a deterministic characterization of network traffic for average performance guarantees such that a systematic approach to performance analyses, as presented by the characterization in [2], is retained.

c. To come up with a deterministic characterization which is directly applicable to measurement-based analysis of probabilistic performance guarantees that are inferred via sample-path computations.

The notion of burstiness and how we might want to perceive it will be discussed in Section 3. As for our second goal, we would like a traffic characterization to have the following properties: Let $\mathcal{C}$ denote a traffic characterization,

1. if two flows characterized according to $\mathcal{C}$ are aggregated, then the aggregate of the flows should also be easily characterized according to $\mathcal{C}$,
2. if a flow characterized according to $\mathcal{C}$ is fed into a network element commonly used in practice (such as *work-conserving servers*, *multiplexers*, *switches*, etc.) then
   (a) the output flow should also be easily characterized according to $\mathcal{C}$, and
   (b) 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 "stationary" in the sense that the characterization of a time-shifted traffic does not change with respect to that of the unshifted traffic,

*Research & Development In Advanced Network Technology, Los Alamos National Laboratory, P.O. Box 1663, M.S. D451, Los Alamos, NM 87545. E-mail: sami@lanl.gov

§Dept. of Computer & Information Science, The Ohio State University, Columbus, OH 43210. E-mail: feng@lanl.gov

4. the characterization of a flow according to $\mathcal{C}$ should be measurable and be easily measured,
5. the characterization $\mathcal{C}$ should allow a regulation of an arbitrary traffic to comply with some given specifications according to $\mathcal{C}$,
6. 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.

In this study, we show that the proposed characterization satisfies properties 1 through 4 above; our research is ongoing for the rest of the properties. The first five properties facilitate a systematic approach to performance analyses in communication networks.

As for our third goal, we would like the traffic characterization that we come up with be such that we would be able to view both the characterization itself and its implications towards performance guarantees, from the standpoint of relative frequency interpretation of probability. The characterization that we propose has this aspect inherently.

The rest of the paper is organized as follows: Section 2 provides a background. Section 3 discusses the notion of burstiness and how we might want to perceive it. Section 4 introduces the proposed new traffic characterization. Sections 4.1 and 4.2 show the implications of the new characterization, by itself and over a single network element, respectively. Section 4.3 clarifies the performance guarantees according to the new characterization, over a tandem of network elements. Section 4.4 gives the average performance guarantees. Section 4.5 shows the stationarity of the new characterization. Section 4.6 discusses the measurability of the proposed characterization. Section 5 gives a few remarks about the new characterization. Finally, Section 6 concludes the study.

2 Background

We adopt a discrete-time formulation for simplicity. 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, which belong to a stream of packets, that pass through a cross-section of a communication link by time n (inclusive). The *rate* r of a flow R is defined as

$$r(n) \triangleq R(n) - R(n-1) \quad \text{for all } n.$$

A *network element* is an input-output device or a medium that accepts packets at its input and delivers them at its output. Packets are assumed to instantaneously arrive and depart at a network element, i.e. a whole packet can arrive instantaneously at some time k and later depart at time n where $n \geqslant k$. Note that a packet can depart in the same interval in which it has arrived; this is sometimes referred to as *cut-through* operation. 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 simplicity. The capacity $c(n)$ of a network element at time n is the total number of packets that it can serve at that time.

We denote the set of all the integers by $\mathbb{Z}$ and the set of all the positive integers by $\mathbb{Z}^+$. Given a statement A which can be true or false, the notation $[A]$ stands for 1 if A is true, and 0 otherwise.[1] Finally, all functions are assumed to be defined from the integers to the integers, unless otherwise noted from here on.

We utilize the following definitions in this study, which have been introduced in the literature previously (e.g. see [2, 5, 6, 7, 11] and the references therein).

Definition 1 *The* min-+ convolution $f \overline{\otimes} g$ *of any two functions* f *and* g *is defined as*[2]

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

The convolution[3] $f \overline{\otimes} g$ is read as "f min-convolved with g".

Definition 2 *The* min-+ deconvolution $f \overline{\oslash} g$ *of any two functions* f *and* g *is defined as*

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

The deconvolution $f \overline{\oslash} g$ is read as "f min-deconvolved with g".

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

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

The convolution $f \underline{\otimes} g$ is read as "f max-convolved with g".

Definition 4 *The* max-+ deconvolution $f \underline{\oslash} g$ *of and two functions* f *and* g *is defined as*

$$(f \underline{\oslash} g)(n) = \min_{k \geqslant 0} \{f(n+k) - g(k)\} \quad \textit{for all } n.$$

The deconvolution $f \underline{\oslash} g$ is read as "f max-deconvolved with g".

[1] We have adopted this notation from [10].

[2] There are two other variants of min-+ convolution, depending on how the subscript of the minimum is delimited in the definition; one for '$0 \leqslant k \leqslant n$' and one for '$k \in \mathbb{Z}$'.

[3] One reason to choose this notation for min-+ convolution over some others, for example '$*$', is that this choice of notation provides a better selection of notations for other related operators introduced in Definitions 2, 3, and 4 in a fitting manner. All of these operators are employed together in this study.

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*

$$G(n) \geqslant (R \triangledown S)(n) \quad \textit{for all } n,$$

for any R*. A service curve* S *is a non-decreasing function defined from the integers to the non-negative integers, that* $S(n) = 0$ *for all* $n \leqslant 0$*.*

An S-server with equality *is an* S*-server such that the inequality in Definition 5 becomes an equality.*

An S-server with service curve S can be viewed as a generalization of a *work-conserving*[4] server. This view becomes clear if we note that the output flow G of a work-conserving server with input flow R and with a constant integer-capacity ρ is given[5] by

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

which can also be equivalently represented as

$$G(n) = (R \triangledown S)(n) \quad \text{for all } n,$$

where $S(n) = \max\{0, \rho \cdot n\}$. In other words, the work-conserving server in this case is an S-server with equality with service curve $S(n) = \max\{0, \rho \cdot n\}$.

Various applications of service curves, such as multimedia smoothing, in current data networks can be followed in [7].

3 The Notion of Burstiness

The burstiness of an arrival process has, roughly, to do with the proximity of arrival instances to each other and with the variation of arrival amounts from one arrival instance to another. A traffic characterization with respect to the burstiness of a traffic tries to restrict these two aspects of variations in arrivals, in a combined fashion so that some provable bounds on a specified set of performance metrics of interest could be given with ease.

A key point in coming up with a traffic characterization with *utility* is to have a focus on some performance metrics of interest. This often implies that one would need to base his/her perception of burstiness of a traffic on the behavior that the traffic induces on a network element, such a network element is typically a variant of a work-conserving server.

A good example to this perception which is squarely placed at the heart of the definition is the (σ, ρ) model which has led to the concept of *arrival curves* in the general case [2]. A flow R is said to be (σ, ρ) constrained if it satisfies the following condition

$$R(n+k) - R(k) \leqslant \sigma + \rho \cdot n \quad \forall\, k \text{ and non-negative } n.$$

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

[5] This is often referred to as Reich's result [12].

If a (σ, ρ) constrained flow R is fed into a work-conserving server with constant rate ρ, it is not difficult to show that the backlog $Q(n)$ at any time n is upper bounded by σ. A similar statement can also be given for a flow conforming to an arrival curve.

We also base our perception of burstiness of a flow on the behavior that it induces on a network element, specifically on the queue-size behavior, as explained in the following section.

3.1 Burstiness via Level-crossing

Consider the rate r of a given a flow R. Let us trim r at a certain level σ as given below where the trimmed rate function is denoted by r_1,

$$r_1(n) \triangleq \min\{r(n), \sigma\} \quad \text{for all } n.$$

Let us also consider the part of the rate function which has been trimmed off, i.e.

$$r_2(n) \triangleq r(n) - r_1(n) \quad \text{for all } n.$$

An example for r, r_1, and r_2 can be viewed in Figure 1 where time-slots are drawn very closely to each other for convenience.

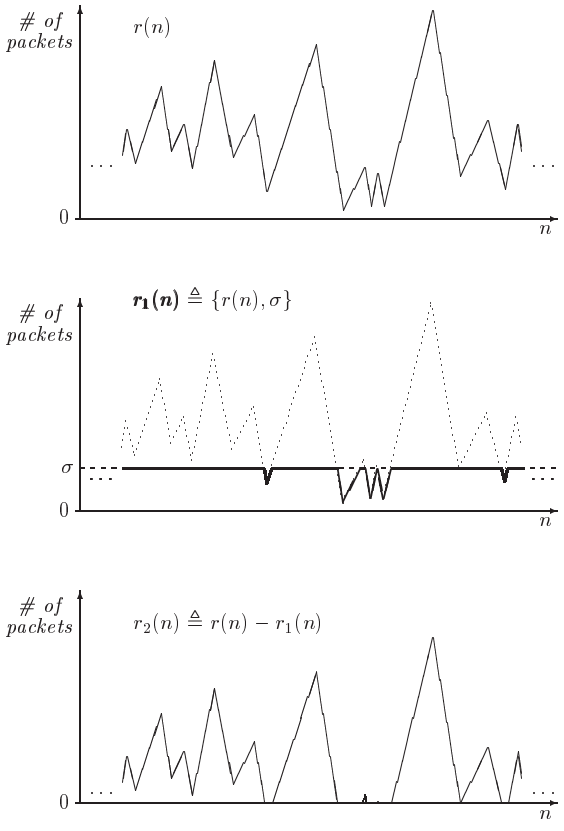

Figure 1: Given a rate function $r(n)$ and a trim-off level σ, the trimmed rate function $r_1(n)$ and the trimmed-off rate function $r_2(n)$.

Vaguely, we tend to perceive for some values of σ that r_1 is *at most* as bursty as r while r_2 is *at least* as bursty as r.

We can put the vague statement above more precisely as follows: Multiply both r_1 and r_2 by some constants a_1 and a_2, respectively, such that the rate functions $\lfloor a_1 \cdot r_1(n) \rfloor$ and $\lceil a_2 \cdot r_2(n) \rceil$ have the same long-term average rates as $r(n)$ has. Feed $r(n)$, $\lfloor a_1 \cdot r_1(n) \rfloor$, and $\lceil a_2 \cdot r_2(n) \rceil$ into identical network elements with labels 0, 1, and 2, respectively, where the average rate that each network element serves its packets is greater than that of its input. We would then observe for some trim-off values σ that the queue-size behavior in network element 1 is "less erratic" than that of element 0 while the queue-size behavior in network element 2 is "more erratic" than that of element 0, when we consider the entire duration of observations.

The statement in the last paragraph, i.e. the one given by the last sentence above, is still vague. What we precisely mean by this vague statement can be put as follows: Select a rate function r from a random sample space. For each selection of r; feed r, $\lfloor a_1 \cdot r_1(n) \rfloor$, and $\lceil a_2 \cdot r_2(n) \rceil$ into network elements 0, 1, and 2, respectively, where the rate functions and the network elements are as explained in the previous paragraph. Observe the queue-size distributions in each network element. For some trim-off values σ, the decay-rate of the tail of the queue-size distribution (i.e. the complementary cumulative distribution function of the queue-size) in network element 1 is faster than that of network element 0 beyond a certain queue-size value, while it is slower in network element 2 than that of network element 0 again beyond a certain queue-size value.

Thus, it is the decay-rate of the tail of the queue-size distribution that we observe in deciding the degree of burstiness of a flow with respect to another one, after some appropriate normalizations of the flows as indicated earlier. The faster the decay-rate is, the less bursty the traffic is, and vice versa.

In summary, we decide the degree of burstiness of a traffic source A with respect to another one B, from the perspective of a network element of interest. We do this as follows: Normalize the sources such that the average rate of traffic coming out of source A is equal to that of source B. Feed the traffic generated by each normalized source A and B into identical network elements labeled 1 and 2 respectively, where the average rate that each network element serves its packets is greater than that of its input. Observe the queue-size distribution in each network elements 1 and 2. If there exists a queue-size level σ_0 beyond which the decay-rate of the tail of the queue-size distribution in network element 1 is smaller (i.e. slower) than that of network element 2, then we say that source A is more bursty than source B, and vice versa.

Returning to the trimming of rate function r, we might want to lower- and/or upper- bound the rate of the overshoot of r above each trim-off level σ, in an attempt to restrict its burstiness, in light of the above discussions.

However, recalling our perception of burstiness of a flow via the queue-size behavior that it induces on a network element, we should actually lower- and/or upper-bound the rate of overshoot of not r but rather the queue-size *itself* directly. With this view, we propose a new traffic characterization in the following section.

4 A New Traffic Characterization

Motivated by the discussions in Section 3, we propose the following deterministic characterization of traffic.

Definition 6 *A flow R is said to be bursty with service curve S and level-crossing function $U(\sigma, n)$, and denoted as $R \sim (S, U)$, if the following inequality holds for all k, positive n, and σ*

$$\frac{1}{n} \sum_{k < i \leqslant n+k} \Big[R(i) - R(j) > S(i-j) + \sigma, \quad \textit{for some } j < i \Big] \leqslant U(\sigma, n) \quad (1)$$

where $U(\sigma, n)$ is defined from $\mathbb{Z} \times \mathbb{Z}^+$ to the non-negative real numbers.

We assume without loss of generality that the following properties hold for any level-crossing function $U(\sigma, n)$:

1. $U(\sigma, n)$ is non-increasing in σ, as the quantity corresponding to a σ on the left-hand-side of inequality (1) is non-increasing with σ.
2. $U(\sigma, n) \leqslant 1$ for any σ and positive n. We assume for mathematical convenience that $U(\sigma, n) = 1$ for all negative σ and for any positive n.
3. $\lim_{\sigma \to \infty} U(\sigma, n) = 0$ for any positive n (this is certainly the case if flow R is bounded, which we can assume without loss of generality for almost all practical purposes).

In the rest of this section, we examine some of the properties/implications of Definition 6. Specifically, we show that it satisfies properties 1 through 4 of a traffic characterization that we have sought to have as stated in the introduction. *The proofs of the theorems and corollaries can be found in [13].*

4.1 Implications on Aggregate Flows and Average Rate

We first show that the traffic characterization provided by Definition 6 satisfies property 1. This is given by the following theorem.

Theorem 1 *Given any two flows $R_1 \sim (S_1, f_1)$ 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 level-crossing function $U_1 \triangledown U_2$ where the convolution is carried out over the first arguments (i.e. σ) of U_i's. In other words, $R_1 + R_2 \sim (S_1 + S_2, U_1 \triangledown U_2)$.*

Let us adopt a convention from here on that whenever we refer to a min-+ (max-+) convolution (deconvolution) of any two bivariate functions, we mean the min-+ (max-+) convolution (deconvolution) carried out over their first arguments.

The characterization provided by Definition 6 has also an implication on the long-term average rate of a flow. This is given by the following theorem.

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

$$\mu \triangleq \limsup_{(n-k)\to\infty} \frac{R(n) - R(k)}{n-k} \leqslant \limsup_{n\to\infty} \frac{S(n)}{n} + \limsup_{n\to\infty} \sum_{\sigma \geqslant 0} U(\sigma, n)\,.$$

In the following section, we state the implications of Definition 6 over an S-server, whereby we show some of the other properties satisfied by this characterization as stated in the introduction.

4.2 Implications over an S-server

If a flow $R \sim (S^*, f)$ is fed into an S-server, the queue-size Q at the server is also similarly upper-bounded by U. This is given more precisely by the following theorem. 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 $Q(n) \triangleq R(n) - G(n)$.

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

$$\frac{1}{n} \sum_{k<i\leqslant n+k} [Q(i) > \sigma] \leqslant U\big((S \oslash S^*)(0) + \sigma, n\big)$$

for all k, $n > 0$, and σ.

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 queue-size distribution becomes $U(\sigma, n)$.

Theorem 3 provides the real essence behind, hence is key to understand, the characterization given by Definition 6.

The output flow of an S-server fed by a flow $R \sim (S^*, U)$ can also be easily characterized according to Definition 6. This is given by the following theorem.

Theorem 4 *The output flow G of an S-server with service curve S, fed by a flow $R \sim (S^*, U)$, is bursty with service curve $S^* \overline{\otimes} S$ and level-crossing function U. In other words, $G \sim (S^* \overline{\otimes} S, U)$.*

We might actually need to slightly rectify the above result by replacing the service curve $S^* \overline{\otimes} S$ in characterizing the output flow by S_o which is given below

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

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 $S^* \overline{\otimes} S$ for positive values of its argument (this could be seen clearly via the proof of the above result, see [13]). We have not done this rectification in the body of the theorem in order not to clutter the result.

The characterization provided by Definition 6 has also an implication on the *virtual-delay* at an S-server. The definition of *virtual-delay* is given below, which is 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 of a flow, which arrive at time n, through the network element that serves them, if the packets are to be served in the order in which they arrive.

Theorem 5 *If an input flow $R \sim (S^*, U)$ is fed into an S-server with service curve S, then the virtual-delay $D(n)$ at the server satisfies*

$$\frac{1}{n} \sum_{k<i\leqslant n+k} [D(i) > \sigma] \leqslant U\big((S \oslash S^*)(\sigma), n\big)$$

for all k, $n > 0$, and σ.

We can actually slightly improve the results in theorems 3 and 5; see [13] for details.

All of the theorems presented in this section facilitate a systematic approach to performance analyses over a tandem of network elements. To further clarify this point, we provide the results in the following section.

4.3 Performance Guarantees Over A Tandem of Network Elements

Performance guarantees, within the framework of Definition 6, over a tandem of network elements follow directly from the results in section 4.2. This is given by the following corollary.

Corollary 1 *Let a flow $R_1 \sim (S^*, U)$ 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 level-crossing function U, where*

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

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

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

$$\frac{1}{n}\sum_{k<i\leqslant n+k}[(Q_1+Q_2)(i)>\sigma]\leqslant(g\,\overline{\nabla}h)(\sigma,n)$$

for all k, $n>0$, and σ, where

$$g(\sigma,n)=U\big((S_1\,\overline{\boxtimes}\,S^*)(0)+\sigma,n\big)$$

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

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

$$\frac{1}{n}\sum_{k<i\leqslant n+k}[D_1(i)+D_2(i+\delta(i))>\sigma]\leqslant(g\,\overline{\nabla}h)(\sigma,n)$$

for all k, $n>0$, and σ, where

$$g(\sigma,n)=U\big((S_1\,\overline{\boxtimes}S^*)(\sigma),n\big)$$

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

We can actually slightly improve the above results in items 2 and 3 in Corollary 1, and further emphasize the message that we are trying to convey by the corollary. It is not difficult to show that the bounds which we would obtain by replacing $(g\,\overline{\nabla}h)(\sigma)$ in items 2 and 3 in Corollary 1 by $h(\sigma)$, also hold; see [13] for a proof. Hence, the bounds in the last two items in Corollary 1 can actually be replaced by

$$\min\big\{h(\sigma,n),\,(g\,\overline{\nabla}h)(\sigma,n)\big\}$$

for all σ and positive n. Thus, it becomes clear by Corollary 1 and the above inequalities being pointed out that we could actually view the tandem network as a single S-server with service curve $S_1\,\overline{\nabla}S_2$, and obtain valid characterizations.

By a repeated application of Corollary 1 and the discussions in the previous paragraph, we could obtain similar results for any number of network elements in tandem.

4.4 Average Performance Guarantees

Average performance guarantees, within the framework of Definition 6, at various points inside a network follows immediately by the results in Sections 4.2 and 4.3. This is illustrated in this section specifically for average queue-size and virtual-delay at an S-server, by the following two corollaries. For simplicity, the corollaries are given for the case when the service curve of the S-server is equal to that of the input flow characterization.

Corollary 2 *If an input flow $R\sim(S,U)$ is fed into an S-server with service curve S, the average queue-size $Q(n)$ at the server is upper bounded as*

$$\frac{1}{n-k}\sum_{k<i\leqslant n}Q(i)\leqslant\sum_{\sigma=0}^{\infty}U(\sigma,n-k)$$

for all $k<n$. Hence, the long-term average queue-size is upper bounded as

$$\limsup_{(n-k)\to\infty}\frac{1}{n-k}\sum_{k<i\leqslant n}Q(i)\leqslant\limsup_{n\to\infty}\sum_{\sigma=0}^{\infty}U(\sigma,n)\,.$$

Proof: The proof follows immediately by Theorem 3, which is shown below. It holds for all $k<n$ that

$$\begin{aligned}\sum_{k<i\leqslant n}Q(i)&=\sum_{k<i\leqslant n}\sum_{\sigma\geqslant 0}[Q(i)>\sigma]\\&=\sum_{\sigma\geqslant 0}\sum_{k<i\leqslant n}[Q(i)>\sigma]\\&\leqslant\sum_{\sigma\geqslant 0}(n-k)U(\sigma,n-k)\quad\text{(by Thm. 3)}\end{aligned}$$

dividing both sides by $n-k$, we get

$$\frac{1}{n-k}\sum_{k<i\leqslant n}Q(i)\leqslant\sum_{\sigma\geqslant 0}U(\sigma,n-k)\,.$$

Consequently, the long-term average queue-size is upper bounded as

$$\limsup_{(n-k)\to\infty}\frac{1}{n-k}\sum_{k<i\leqslant n}Q(i)\leqslant\limsup_{n\to\infty}\sum_{\sigma=0}^{\infty}U(\sigma,n)\,.\qquad\blacksquare$$

Similarly, the time-averaged virtual-delay is also upper bounded, which is given by the following corollary.

Corollary 3 *If an input flow $R\sim(S,U)$ is fed into an S-server with service curve S, the average virtual-delay $D(n)$ at the server is upper bounded as*

$$\frac{1}{n-k}\sum_{k<i\leqslant n}D(i)\leqslant\sum_{\sigma=0}^{\infty}U\big((S\,\overline{\boxtimes}S)(\sigma),n-k\big)$$

for all $k<n$. Hence, the long-term average virtual-delay is upper bounded as

$$\limsup_{(n-k)\to\infty}\frac{1}{n-k}\sum_{k<i\leqslant n}D(i)\leqslant\limsup_{n\to\infty}\sum_{\sigma=0}^{\infty}U\big((S\,\overline{\boxtimes}S)(\sigma),n\big)\,.$$

Proof: The proof follows immediately by Theorem 5, which is shown below. It holds for all $k<n$ that

$$\begin{aligned}\sum_{k<i\leqslant n}D(i)&=\sum_{k<i\leqslant n}\sum_{\sigma\geqslant 0}[D(i)>\sigma]\\&=\sum_{\sigma\geqslant 0}\sum_{k<i\leqslant n}[D(i)>\sigma]\\&\leqslant\sum_{\sigma\geqslant 0}(n-k)\,U\big((S\,\overline{\boxtimes}S)(\sigma),n-k\big)\end{aligned}$$

dividing both sides by $n-k$, we get

$$\frac{1}{n-k}\sum_{k<i\leqslant n} D(i) \leqslant \sum_{\sigma=0}^{\infty} U\big((S \barwedge S)(\sigma), n-k\big)\,.$$

The first inequality above holds by Theorem 5.

Consequently, the long-term average virtual-delay is upper bounded as

$$\limsup_{(n-k)\to\infty} \frac{1}{n-k}\sum_{k<i\leqslant n} D(i) \leqslant \limsup_{n\to\infty} \sum_{\sigma=0}^{\infty} U\big((S \barwedge S)(\sigma), n\big)\,.$$

This completes the proof. ■

4.5 Stationarity

The proposed characterization also has the "stationarity" property (i.e. property 3) stated in the introduction. This follows inherently from Definition 6 itself and can be shown very easily.

Let $R'(n) \triangleq R(n-t)$ for any given t. Note that R' is a time shifted version of R by t amount. The following relations hold for any k, $n>0$, σ, and t;

$$\sum_{k<i\leqslant n+k} \Big[R'(i) - R'(j) > S(i-j)+\sigma, \quad \text{for some } j<i\Big]$$
$$= \sum_{k<i\leqslant n+k} \Big[R(i-t) - R(j-t) > S(i-j)+\sigma, \quad \exists\ j<i\Big]$$
$$= \sum_{k<i\leqslant n+k} \Big[R(i-t) - R(j-t) > S(i-j)+\sigma,\ \exists j-t<i-t\Big]$$

replacing $i-t$ by u and $j-t$ by v, we get

$$= \sum_{k-t<u\leqslant n+k-t} \Big[R(u) - R(v) > S(u-v)+\sigma, \quad \exists\ v<u\Big]$$

since $R\sim(S,U)$ and (1) holds for any k, we finally have

$$\leqslant n\cdot U(\sigma, n).$$

Hence, if R is bursty with service curve S and level-crossing function U, then so is its time-shifted version R', and vice versa.

4.6 Measurability

The traffic characterization provided by Definition 6 is also measurable and can easily be measured. This is facilitated by Theorems 2 and 3.

Given a flow R and a service curve S, we can measure an estimate of a tight level-crossing function U in characterizing R as $R\sim(S,U)$, by utilizing Theorem 3. This can be done as follows: First, we need to have an S-server with equality with service curve S. Then, upon feeding flow R into such an S-server, we observe the queue-size Q at the server. The supremum of the empirical values of the quantity on the left-hand-side of the inequality in Theorem 3 would give the tightest (i.e. the smallest) possible level-crossing function U in characterizing the flow R as $R\sim(S,U)$. This could be followed better via the proof of Theorem 3; see [13].

Secondly, given a flow R with unknown characterization according to Definition 6, various service curves S can be suggested in light of Theorem 2 in order to characterize R as $R\sim(S,U)$.

To do this, we would first need to have a good estimate of the long-term average rate μ of R. Getting a fairly good approximation of the long-term average rate of a flow is often not difficult by applying a variant of the Law of Large Numbers [14].

Then by Theorem 2, we can pick a service curve S such that $\limsup_{n\to\infty}\frac{S(n)}{n} \geqslant \mu$. Thus, service curves of the from

$$S(n) = \max\{0, \rho\cdot(n-D)\} \quad \text{where integer } \rho>\mu,$$

are good candidates to find a characterization of R as $R\sim(S,U)$. A delay parameter D can be chosen considering (i) the end-to-end delay requirement of flow R, and (ii) how this delay requirement is to be distributed over each network element on the path of flow R before it reaches to its destination.

Finally, by the discussions given in the second paragraph in this section, we can find a tight level-crossing function U to complete the characterization of R as $R\sim(S,U)$.

Moreover, we would like to note that the above procedure to find an appropriate S can be applied recursively, depending on the decay-rate of the estimate of the level-crossing function U, in light of Theorem 2, to find a more fine-grained characterization of flow R.

5 Remarks About The New Characterization

We would like to give a few remarks about the traffic characterization provided by Definition 6.

1) The new characterization facilitates analyses for both average and scalar worst-case performance guarantees. Average performance guarantees can be obtained in general as exemplified by the results in Section 4.4. Scalar worst-case performance guarantees, on the other hand, are provided by the respective level-crossing functions (e.g. in Theorem 3) whose second arguments (i.e. n) are all set equal to 1. A relevant scalar worst-case performance bound is given by the minimum of the first argument of the corresponding level-crossing function $U(*,1)$ such that $U(*,1)=0$.

For example, a scalar worst-case performance bound on queue-size can be obtained as follows: If an input flow $R\sim(S^*,U)$ is fed into an S-server with service curve S, then we have by Theorem 3 that the queue-size Q at the server at any time n is upper bounded as

$$Q(n) \leqslant \min\{\sigma : U\big((S \barwedge S^*)(0)+\sigma, 1\big)=0\} \quad \text{for all } n.$$

2) The new characterization facilitates a systematic framework for measurement-based analyses of probabilistic performance guarantees that would be inferred via sample-path computations. This has become possible, since the implications of Definition 6 (such as Theorems 3 and 5) can also be viewed from the standpoint of relative frequency interpretation of probability. However, we would like to remark that the real starting point in getting Definition 6 is the discussions given in Section 3 which tries to clarify the notion of burstiness via the concept of level-crossing.

We have also proposed another traffic characterization that can be viewed as casting of Definition 6 in a probabilistic setting; see [15]. To see the applicability of the probabilistic version of Definition 6 in real networks, we have performed some preliminary experiments and simulations; see [16]. Our preliminary results show agreement with the theory. More results to this end will follow in the near future.

One of the main utilities of the deterministic characterization that we propose in this study is that in computing the average performance measures (such as average delay and average queue-size at a network element) it eliminates the need for knowing a probability measure for the observed data. Specifically, the quantities expressed in Theorems 3 and 5 for average queue-size and average delay, respectively, are what we typically compute in sample-path probabilistic analyses. The characterization that we propose leads formalizing this sample-path approach into a framework that we have examined in this study.

3) Lastly, we would like to point out another contrast between the characterization that we propose in this study and the characterization in [2] (i.e. other than the fact that studies in [1, 2, 3] are concerned with scalar worst-case performance metrics, whereas the characterization that we propose facilitates analyses for both average and scalar worst-case performance guarantees; see remark 1 in this section).

Consider feeding a flow $R \sim (S, U)$ into an S-server with service curve S. By Theorem 3, we know that the queue-size Q at the server satisfies

$$\frac{1}{n} \sum_{k<i\leqslant n+k} [Q(i) > \sigma] \leqslant U(\sigma, n)$$

for all k, $n > 0$, and σ. Multiplying both sides by n, we get

$$\sum_{k<i\leqslant n+k} [Q(i) > \sigma] \leqslant n \cdot U(\sigma, n)\,.$$

An example for $n \cdot U(\sigma, n)$, for a given specific n, is shown in Figure 2.

Now, rotate $n \cdot U(\sigma, n)$ by $+90$ degrees in the plane on which it is drawn, with respect to the origin at $(0, 0)$. Flip the rotated figure with respect to the vertical axis obtained after the rotation (i.e. this will be the axis tagged σ). Referring to Figure 2, now switch bullets '•' with circles '∘' and mark the horizontal axis appropriately as shown in Figure 2 (the bottom figure). At the end, what we obtain is an "extremal arrival pattern" for the queue-size as exemplified in Figure 2 where time-slots are drawn very close to each other for convenience.

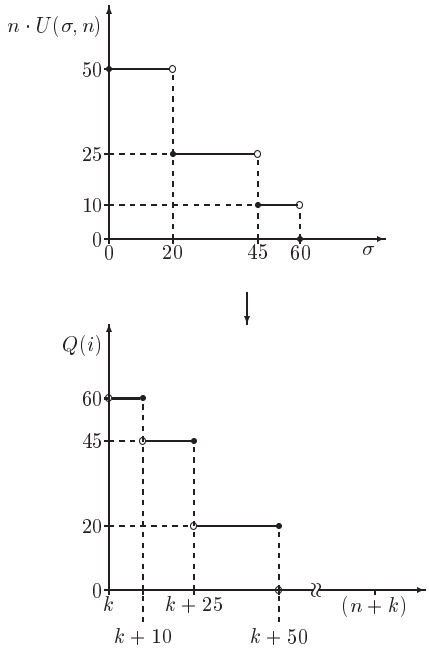

Figure 2: An "extremal arrival pattern" for the queue-size, for some k and n that is larger than 50.

Scalar worst-case performance bounds, such as those presented in [2], are achieved for extremal *input* flows whose arrival patterns in principal look like the bottom figure in Figure 2. Whereas, in our analyses we consider such "extremal arrival patterns" for *queue-sizes*.

6 Conclusions

We have proposed a deterministic characterization of network traffic, based on service curves. The new characterization is given by Definition 6 in Section 4. The proposed characterization facilitates performance analyses for both average and scalar worst-case performance guarantees. Average performance guarantees can be obtained as exemplified by the results in Section 4.4. Scalar worst-case performance guarantees can be found as explained by first remark in Section 5.

We have shown that this characterization satisfies properties 1 through 4 of a traffic characterization that we sought to have as stated in the introduction. Specifically, we have shown that the new characterization facilitates a systematic approach to performance analysis for both average and scalar worst-case Quality of Ser-

vice (QoS) guarantees, where the presented systematicness is similar to the one presented by the characterization in [2] and its companion service model (the *service-curve model*).

One of the main utilities of the proposed characterization is that it facilitates a systematic framework for measurement-based analysis of probabilistic performance guarantees that would be inferred via sample-path computations. In this sense, Definition 6 is directly applicable to such measurement studies. This is made possible, since the implications of Definition 6 (such as Theorems 3 and 5) can be viewed from the standpoint of relative frequency interpretation of probability.

We have also tried to clarify the notion of burstiness. We have pointed out that one might want to perceive the burstiness of a flow from the perspective of a network element; specifically, based on the queue-size behavior that the flow induces on a network element of interest.

To this end, we have explained how one might want to decide the degree of burstiness of a traffic source A with respect to another one B (see the explanations given at the end of Section 3.1). We have pointed out that it is the decay-rate of the tail of the queue-size distribution that we observe in deciding the degree of burstiness of traffic sources after appropriately normalizing them as indicated earlier. The faster the decay-rate is, the less bursty a source is, and vice versa.

We do not pretend that this work (i.e. the suggested analysis approach) at its current level of investigation is directly applicable to communication networks. This study merely constitutes a preliminary investigation.

A key issue that needs to be addressed to make this deterministic characterization viable for performance analysis in data networks is the discovery of a method to regulate arbitrary traffic so that the regulated traffic conforms to some given (S, U) specification. Recently, we have discovered a method to generate synthetic traffic which conforms to a given characterization according to the probabilistic version of Definition 6 (i.e. the characterization in [15]). This method facilitates a regulation of arbitrary traffic so that the regulated traffic conforms to a given characterization the probabilistic version of Definition 6. The details of this work will follow in the near future. Future work includes adapting these recent findings for the deterministic characterization that we propose in this study.

It would be also interesting to examine a similar traffic characterization where the new characterization would also have a lower-bound, as well as an upper-bound, on the quantity given in Definition 6.

References

[1] 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.

[2] R. L. Cruz. *A Calculus For Network Delay, Part I: Network Elements In Isolation*, IEEE Transactions on Information Theory, vol. 37, no. 1, pp. 114–131, Jan. 1991.

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

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

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

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

[7] J.-Y. Le Boudec, P. Thiran. *Network Calculus: A Theory of Deterministic Queuing Systems for the Internet*, Lecture Notes in Computer Science-2050, Springer Verlag, January 2002.

[8] G. Kesidis, T. Konstantopoulos. *Extremal shape-controlled traffic patterns in high-speed networks*, IEEE Transactions on Communications, vol. 48, no. 5, pp. 813–819, May 2000.

[9] F. Guillemin, N. Likhanov, R. Mazumdar, C. Rosenberg. *Extremal traffic and bounds on the mean delay of multiplexed regulated traffic streams*, Proceedings of the INFOCOM 2002, N.Y., June 2002.

[10] R. L. Graham, D. E. Knuth, O. Patashnik. *Concrete Mathematics: A Foundation for Computer Science*, Addison-Wesley, 2nd ed., 1994.

[11] H. Sariowan. *A Service-curve Approach to Performance Guarantees in Integrated-service Networks*. Ph.D. Dissertation, Department of Electrical and Computer Engineering, University of California, San Diego, 1996.

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

[13] S. Ayyorgun, W.-C. Feng. *A Deterministic Definition of Burstiness For Network Traffic Characterization*, Technical Report LA-UR-03-4477, Los Alamos National Laboratory.

[14] W. Feller. *An Introduction to Probability Theory and Its Applications, Volume 1*, John Wiley & Sons, 1968.

[15] S. Ayyorgun, W.-C. Feng. *A Probabilistic Definition of Burstiness Characterization: A Systematic Approach*, Technical Report LA-UR-03-3668, Los Alamos National Laboratory.

[16] S. Ayyorgun, W.-C. Feng. *A Systematic Approach to Probablistic Quality of Service Guarantees in Communication Networks*, Technical Report LA-UR-03-7267, Los Alamos National Laboratory, September 2003.