A Systematic Approach for Providing End-to-End Probabilistic QoS Guarantees[*]

Sami Ayyorgun[§]   Wu-chun Feng[§]

Abstract—*We propose a probabilistic characterization of network traffic. This characterization can handle traffic with heavy-tailed distributions in performance analysis. 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 our characterization. This characterization is measurable and allows for a simple probabilistic method for regulating network traffic. All of these properties of the proposed characterization enable a systematic approach for providing end-to-end probabilistic QoS guarantees.*

1 Introduction

For nearly two decades, Quality of Service (QoS) guarantees in communication networks have been actively researched. However, network services with provable QoS guarantees have not been widely deployed for reasons ranging from economic issues to technical problems; see [1] for a recent discussion on this subject.

One reason that services with explicit QoS guarantees have not been widely deployed is that an *effective* systematic approach to end-to-end probabilistic QoS guarantees has yet to be developed. Most probabilistic performance analyses in the literature, specifically queueing analyses, focus on single-hop networking problems, in part because the assumptions about the adopted traffic models, such as independence or conformance to a specific stochastic model, fail for the multi-hop cases (i.e. such assumptions are lost after the first network element). As a result, multi-hop analyses become intractable. This holds for both *Markovian* and *self-similar* traffic models [2–4]. The desired analytic tractability is provided by some deterministic traffic models, e.g. [5–7], but they lead to severe underutilization of network resources.

In classical queueing theory, a particular stochastic process is first adopted to model traffic, and later the implications of this model on queueing behavior over various network elements are examined. This is often difficult to do even for a single network element with even a remotely realistic stochastic model, and becomes intractable for multiple network elements in tandem.

We propose a probabilistic characterization of network traffic by basing the characterization *directly* on the queue-size behavior that a traffic induces on a network element, rather than picking a particular stochastic model first and then examining its implications on the queue-size behavior. In this way, we eliminate many mathematical complications forbidding a tractable performance analysis. Our characterization does not depend on or adopt any particular stochastic process model. Hence, it

can characterize traffic belonging to any stochastic process model including those with *long-range dependency*.

The goal of this study is to examine the properties of the proposed characterization, whereby to show that it enables a systematic analysis of end-to-end probabilistic QoS guarantees in communication networks. To see the effectiveness of the facilitated analysis method, we carried out various simulation studies some of which we present in this text. The characterization that we propose is capable of addressing network traffic with *heavy-tailed* distributions, including *power-tailed* distributions, in performance analyses.

The rest of the paper is organized as follows: Sec. 2 provides background. Sec. 3 introduces the proposed traffic characterization. Sections 3.1 and 3.2 provide the implications of the new characterization on aggregate flows, average rate, over a single network element, and over tandem networks. Sec. 3.3 gives the mean performance guarantees. Sections 3.4 and 3.5 show the stationarity and measurability of the new characterization, respectively. Sec. 3.6 provides some simulation results. Sec. 4 discusses the viability of the proposed analysis scheme. Sec. 5 provides a brief comparison with some related work. Finally, Sec. 6 concludes the study.

2 Background

We adopt a discrete-time formulation in the context of packet networks. A *flow* is a discrete random process whose sample-paths are non-decreasing functions 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, belonging to a stream of packets, that pass through a cross-section of a communication link by time n (inclusive). Packets are assumed to instantaneously arrive and depart at a network element, i.e. a whole packet can arrive instantaneously at time k and depart at time n where $n \geqslant k$. All functions are defined from the integers to the integers, unless otherwise noted.

We use the following definitions in this study, which were previously introduced in the literature [8–12].

Definition 1 *The* min-+ convolution $f \triangledown g$ *of any two functions f and g is defined as*

$$(f \triangledown g)(n) = \min_{k \leqslant n}\{f(k) + g(n - k)\} \qquad \text{for all } n,$$

'$f \triangledown g$' is read as "f min-convolved with g".

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

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

'$f \triangledown g$' is read as "f min-deconvolved with g".

Definition 3 *The* max-+ *convolution* $f \triangle g$ *of and two functions* f *and* g *is defined as*

$$(f \triangle g)(n) = \max_{k \leqslant n}\{f(k) + g(n-k)\} \qquad \text{for all } n,$$

'$f \triangle g$' *is read as* "f *max-convolved with* g".

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

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

'$f \overline{\triangle} g$' *is read as* "f *max-deconvolved with* g".

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

$$G(n) \geqslant (R \triangledown S)(n) \quad \text{for all } n. \tag{1}$$

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 *that inequality (1) becomes an equality.*

An S-server can be viewed as a generalization of a *work-conserving server*[1]. 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 of serving ρ packets per time-slot is given[2] by

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

which can be equivalently represented also as

$$G(n) = (R \triangledown S)(n) \quad \text{for all } n, \text{ where } S(n) = \max\{0, \rho \cdot n\},$$

i.e. such a work-conserving server is an S-server with equality with service curve $S(n) = \max\{0, \rho \cdot n\}$.

The concept of service curves was introduced in [8,9]. Service curves together with an appropriate traffic characterization, such as the deterministic ones in [6,14] or the probabilistic one in this study, enable a systematic approach to performance analyses in communication networks, e.g. see [6,9–12]. This systematic approach is analogous to that in Linear Systems Theory. For various applications of the service-curve model, such as multimedia smoothing, see for example [12].

3 Traffic Characterization

We propose the following probabilistic characterization of network traffic.

Definition 6 *A flow* R *is said to be* bursty with a service curve S and a bounding function f, *and denoted as* $R \sim (S, f)$, *if the following inequality holds for all* n *and* σ;

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

where f *is defined from the integers to the non-negative real numbers.*

---

[1] 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.

[2] This is often referred to as Reich's result [13].

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

1. f is non-increasing, as the probability corresponding to a σ in Definition 6 is non-increasing with σ.

2. $f(\sigma) \leqslant 1$ for all σ, as the probability of an event can not be larger than 1. We assume for mathematical convenience that $f(\sigma) = 1$ for all negative σ.

3. $\lim_{\sigma \to \infty} f(\sigma) = 0$, as any cumulative distribution function F satisfies $\lim_{x \to \infty} F(x) = 1$.

In the rest of this section, we show some of the implications/properties of Definition 6. The proofs of the theorems and corollaries can be found in Appendix or [15].

3.1 Implications on Agg. Flow and Ave. Rate

The implications of Definition 6 on aggregate flows and average rate are given by Theorems 1 and 2, respectively.

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 bounding function* $f_1 \triangledown f_2$. *In other words,* $R_1 + R_2 \sim (S_1 + S_2, f_1 \triangledown f_2)$.

By a repeated application of Theorem 1, we can easily find the characterization of an aggregate flow that has an arbitrary number of flows being aggregated.

Theorem 2 *Given a flow* $R \sim (S, f)$, *where* $\sum_{\sigma=0}^{\infty} f(\sigma)$ *is finite, the long-term average rate* μ *of* R *satisfies*

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

3.2 Implications over an S-server

The next theorem is key to understand the characterization given by Definition 6. The following explanations are needed for the theorem: The queue-size $Q(n)$ denotes the total number of packets that reside in a network element at time n. That is, if R and G denote the aggregates of the flows at the input and at the output of the network element, respectively, then $Q(n) = R(n) - G(n)$.

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

$$\text{P}\big(Q(n) > \sigma\big) \leqslant f\big(\sigma + (S \overline{\triangle} S^*)(0)\big) \quad \text{for all } n \text{ and } \sigma \geqslant 0.$$

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

The output flow of an S-server can also be easily characterized in accordance with Definition 6. This is given by the following theorem.

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

2

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

Definition 7 *The* virtual-delay $D(n)$ *at time* n *at a network element is defined with respect to an input flow* R *and the corresponding output flow* G, *as*

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

The virtual-delay $D(n)$ is the delay experienced by a packet arriving at time n, if the packets were to be served in the order in which they arrive.

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

$$\mathrm{P}\big(D(n) > \sigma\big) \leqslant f\big((S \,\overline{\boxtimes}\, S^*)(\sigma)\big) \quad \text{for all } n \text{ and } \sigma \geqslant 0.$$

All of the theorems presented in this section enable a systematic analysis of end-to-end probabilistic QoS guarantees over a tandem of network elements. This is made explicit by the following corollary.

Corollary 1 *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*

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

   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*

$$\mathrm{P}\big((Q_1+Q_2)(n) > \sigma\big) \leqslant (g\,\triangledown\, h)(\sigma) \quad \text{for all } \sigma, \text{ where}$$

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

$$h(\sigma) = f\big(\sigma + (S_2 \,\overline{\boxtimes}\, (S^* \,\overline{\triangledown}\, S_1))(0)\big)$$
$$= f\big(\sigma + ((S_1 \,\triangledown\, S_2) \,\overline{\boxtimes}\, S^*)(0)\big).$$

3. *The total virtual-delay* $D_T(n) = D_1(n) + D_2(n+\delta)$ *experienced by a packet arriving at 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* $\delta \geqslant 0$

$$\mathrm{P}\big(D_T(n) > \sigma\big) \leqslant (g\,\triangledown\, h)(\sigma) \quad \text{for all } \sigma, \text{ where}$$

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

$$h(\sigma) = f\big((S_2 \,\overline{\boxtimes}\, (S^* \,\overline{\triangledown}\, S_1))(\sigma)\big)$$
$$= f\big(((S_1 \,\triangledown\, S_2) \,\overline{\boxtimes}\, S^*)(\sigma)\big).$$

We can actually slightly improve the above results in items 2 and 3 in Corollary 1, and further emphasize the main result that we want to convey by the corollary. It is easy to show that the bounds we would obtain by replacing $(g\,\triangledown\, h)(\sigma)$ in items 2 and 3 in Corollary 1 by $h(\sigma)$ also hold; see [15] for a proof. Hence, the bounds in the last two items in Corollary 1 can actually be replaced by

$$\min\{h(\sigma), (g\,\triangledown\, h)(\sigma)\}.$$

Thus, it becomes clear by Corollary 1 and the above remark that we can actually view the tandem network as a single S-server with service curve $S_1 \,\triangledown\, S_2$, and obtain valid performance results.

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

3.3 Average Performance Guarantees

In the framework of our characterization, average performance guarantees at various points in a network follow immediately from the results in Section 3.2. The next two corollaries show this for average queue size and virtual delay at an S-server, respectively.

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

$$\mathrm{E}[Q(n)] \leqslant \sum_{\sigma=0}^{\infty} f\big(\sigma + (S \,\overline{\boxtimes}\, S^*)(0)\big).$$

Corollary 3 *If a flow* $R \sim (S^*, f)$ *is fed into an* S-server *with service curve* S, *then the mean virtual-delay* $D(n)$ *at the server is upper-bounded at any time* n *as*

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

3.4 Stationarity

The proposed characterization has also a *stationarity* property in the sense that the characterization of a time-shifted traffic does not change with respect to that of the unshifted traffic. This is what we show next. Let $R'(n) \triangleq R(n-t)$ for all n and for some t, i.e. R' is a time-shifted version of R. We have for all n, σ, and t;

$$\Big\{ R'(n) - R'(k) > S(n-k) + \sigma, \text{ for some } k < n \Big\}$$

$$= \Big\{ R(n-t) - R(k-t) > S(n-k) + \sigma, \text{ for some } k < n \Big\}$$

$$= \Big\{ R(n-t) - R(k-t) > S(n-k) + \sigma,\ \exists\, k-t < n-t \Big\},$$

setting $u \equiv n - t$ and $v \equiv k - t$, we get

$$= \Big\{ R(u) - R(v) > S(u-v) + \sigma, \text{ for some } v < u \Big\}.$$

Taking the probabilities of the both sides of the last equality above, we get the desired result; i.e. if R is bursty with service curve S and bounding function f, then so is its time-shifted version R', and vice versa. Due to the propagation delay in communication networks, realistic traffic characterizations need to have this property.

3.5 Measurability

We can determine a characteristic of a flow in accordance with Definition 6 by using Theorems 2 and 3.

First, we need to pick an appropriate service curve S. For that, we need to know the long-term average rate of flow R that we want to characterize. A fairly good approximation of the long-term average rate of a flow can be found by using a variant of the Law of Large Numbers [16]. Let μ be the estimated long-term average rate of flow R. 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 following from

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

are good candidates to find a characteristic of R. The parameter D can be chosen according to i) the end-to-end delay requirement of R, and ii) how this delay requirement is to be distributed over each network element on the path of flow R before it reaches its destination.

Once a service curve is picked as described above, we can then find a bounding function f in characterizing R as $R \sim (S, f)$ by feeding it into an S-server with equality with service curve S. Let the S-server be initially empty; i.e. no packet is stored at the server at time 0. It can be noted via the proof of Thm. 3 that in this case we have

$$P\Big(R(n) - R(k) > S(n-k) + \sigma, \ \exists \ k < n\Big) = P(Q(n) > \sigma).$$

Thus, simply by measuring the queue-size probabilities $P(Q(n) > \sigma)$, we can find a tight bounding function f corresponding the S being chosen. By using the monotonicity result on the queue-size distribution in G/G/1 queues with certain basic properties (e.g. *stationary* or *ergodic* input traffic, initially empty queue size, *stable* server, etc.), it suffices to observe the stationary queue-size distributions to determine f; see Ch. 6 of [17].

3.6 Simulations

We carried out various simulation studies to check the validity of our findings. Due to the page limitation, we report only a small part of these studies.

The main simulation setup that we consider here is shown in Figure 1(a). Input flow R_{in} was obtained from a real traffic trace collected on the incoming/outgoing link of the Los Alamos National Laboratory. The link connected the Laboratory subnet to outside networks. Hence, R_{in} represents a Wide-Area Network (WAN) traffic. The underlying network technology was FDDI (Fiber Distributed Data Interface).
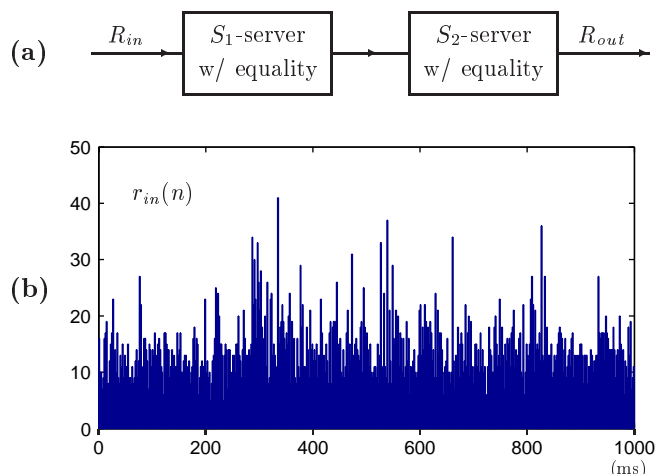
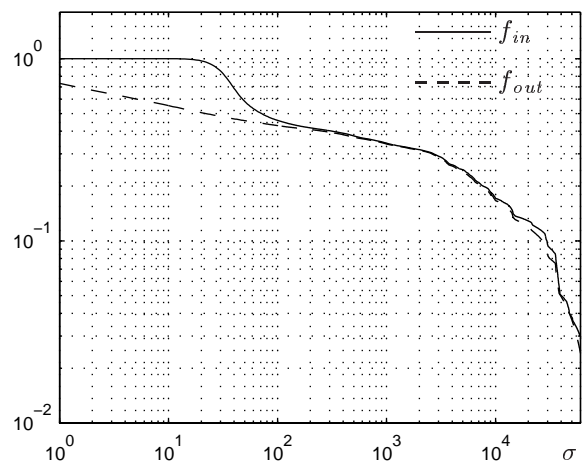

Figure 1: (a) Simulation setup. (b) A snapshot of r_{in}.

Figure 2: Plot for simulation setup in Fig. 1(a).

R_{in} was obtained by first slotting the time into 1 millisecond intervals, and then by counting the total number of packets arriving in each 1 ms interval over a duration of a traffic trace being collected. The value $r_{in}(n) \triangleq R_{in}(n) - R_{in}(n-1)$ is the total number of packets that arrive in interval n. The length of R_{in} that we ran in the simulations was 1,600,000 slots. A snapshot of r_{in} in the first second is shown in Figure 1(b).

The average rate of R_{in} over the 1,600,000 slots was 11.9974 packets per time slot. Hence, to find a characterization of R_{in}, we picked the following service curve

$$S^*(n) = \max\{0, 13(n-3)\}$$

By feeding R_{in} into an S-server with equality with service curve S^*, we found an estimate of a bounding function f_{in}, as explained in Section 3.5, which is shown in Fig. 2.

Next, we fed R_{in} into a tandem of two network elements shown in Figure 1, where

$$S_1(n) = \max\{0, 16(n-2)\} \quad S_2(n) = \max\{0, 14(n-1)\},$$

and obtained the output flow R_{out}. Corollary 1 states that $R_{out} \sim (S_3, f_{in})$ where $S_3(n) = (S^* \triangledown (S_1 \triangledown S_2))(n)$.

4

In order to check the validity of this result, we fed R_{out} into another S-server with equality with service curve S_3, and obtained an estimate f_{out} of its bounding function, as explained in Section 3.5. The estimate f_{out} is shown in Fig. 2. We expect by Corollary 1 that $f_{out}(n) \leqslant f_{in}(n)$. The proximity of f_{in} and f_{out} in Figure 2 is satisfying.

Finally, we would like to note that the queue-size distributions that we observed in this experimental study were all heavy-tailed, as this could be noted in Figure 2.

4 Viability

The analysis framework that we propose in this study to manage end-to-end probabilistic QoS guarantees requires that every input traffic at its entry point into a network has a known characteristic in accordance with Definition 6. To address this requirement, we came up with a simple probabilistic regulation method, called CpR, that regulates an arbitrary traffic to a given (S, f) specification, hence made the proposed framework viable [18]. We described CpR and showed its correctness in [18].

To show the effectiveness of CpR in regulating a flow R, we provide a simulation result here. Flow R being regulated was obtained from a real traffic trace in exactly the same way that R_{in} was obtained in Sec. 3.6. Both traffic traces were collected on the same network, by the same apparatus, as indicated in Sec. 3.6. The length of R that we ran in the simulations was 3,600,000 slots. The average rate of R over the 3,600,000 slots was 12.523 packets per time slot. Hence, we chose $S_{\mathrm{reg}}(n) = \max\{0, 13 \cdot n\}$ as the service curve that we would like the regulated flow to have as specified in its characteristic. The bounding function that we targeted for the regulated flow to have was a truncated Pareto distribution given below;

$$f_{\mathrm{pareto}}(\sigma) = \begin{cases} 1 & \text{if } \sigma < 0, \\ (\sigma + 1)^{-2} & \text{if } 0 \leqslant \sigma \leqslant 10^3, \\ 0 & \text{if } \sigma > 10^3. \end{cases}$$

We input S_{reg} and f_{pareto} into CpR and regulated flow R one thousand times with these parameters; i.e. obtained 1000 regulated flow sample-paths from R by CpR. We found an estimate of a bounding function for each one of these 1000 sample-paths by feeding a sample-path into an S-server with equality with service curve S_{reg}, as explained in Section 3.5. Later, we found the pointwise average of these 1000 bounding-function estimates. This average is denoted by f_{reg} and shown in Figure 3. As expected, Figure 3 shows that the regulated flow has the targeted bounding function f_{pareto}, i.e. the regulated flow conforms to $(S_{\mathrm{reg}}, f_{\mathrm{pareto}})$ specification; the proximity of f_{reg} and f_{pareto} in the figure is quite satisfying.

5 Related Work

Some of the related works that motivate this study can be found in [19–23]. Definition 6 is a generalization of the characterization in [23], which we have come up with
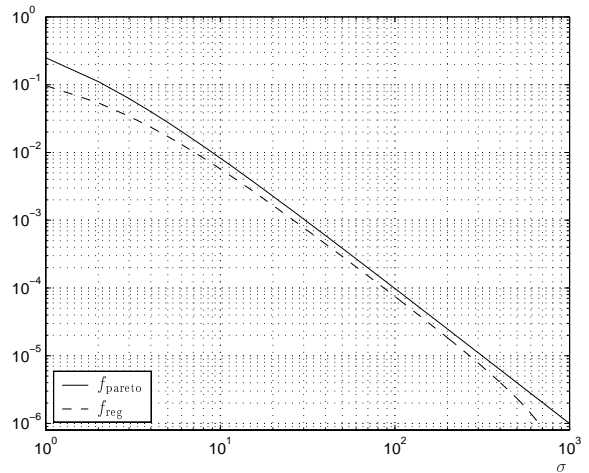

Figure 3: Plot for regulating flow R by CpR.

independently. The characterization in [22] is the most closely related one among the ones in [19–22] which motivate this study and [23]. The traffic characterization provided in [22] is stated below for clarity;

(Adapted from Def. 3 in [22]) *A flow R is said to have a* Stochastically Bounded Burstiness (SBB) *with upper-rate ρ and bounding function f if*

1. *for any $n \in \mathbb{Z}^+$, the n-fold integral $\left(\int_\sigma^\infty du\right)^n f(u)$ is bounded for any σ, and*

2. *the inequality $\mathrm{P}\big(R(t) - R(s) \geqslant \rho \cdot (t-s) + \sigma\big) \leqslant f(\sigma)$ holds for all σ and for all $0 \leqslant s < t$.*

The differences between the characterization in [22] and that of our study, and the differences between the enabled analysis methods are briefly as follows: ① Our characterization allows for a simple and tight regulation method indicated in Sec. 4—this is the single most important difference, as it determines the viability of a respective approach. ② Our characterization allows for a simple measurement scheme described in Sec. 3.5 to determine a tight characterization of a real traffic. In [22], a characterization of a flow is determined in traditional ways with often loose probabilistic bounding techniques. ③ The bounding function in our characterization does not have the restriction in [22], i.e. condition 1 in the definition stated above. As a result, our characterization can handle traffic that has a bounding function which belongs to a larger set (i.e. this larger set includes the set specified in condition 1 above), e.g. a piecewise-Pareto bounding function, while the characterization in [22] can not—such piecewise bounding functions were commonly observed in our simulations with real traffic. ④ Our characterization yields tighter performance bounds as it avoids some use of the *Union Bound* which is a loose probability bound. For example, if an SBB flow R with upper-rate ρ and bounding function f is fed into a work-conserving server with rate C, where $\rho < C$, the following bound on the queue-size distribution at the server is

given by Thm. 3 in [22];

$$\mathrm{P}\big(Q(n) > \sigma\big) \leqslant f(\sigma + 1) + \frac{1}{C - \rho} \int_{\sigma+1}^{\infty} f(u)\, du, \quad (2)$$

whereas the bound that we would give on $\mathrm{P}\big(Q(n) > \sigma\big)$ in this case would be just $f(\sigma)$ by our Thm. 3, where we would characterize R as $R \sim (S, f)$, $S(n) = \max\{0, \rho \cdot n\}$. The difference between the two bounds are due to the Union Bound used in deriving (2), which we did not use, hence our bound is tighter. See for example the comparison given by Figure 1 in [24]. ⑤ Our characterization is defined more generally by basing it on service curves.

The studies in [19,20] can be thought as a special case of the work in [22], thus the above five differences also apply for comparing our work with [19,20] too.

The above differences ①, ②, and ⑤ also apply for comparing our work with [21]. Moreover, the characterization in [21] is concerned with only steady-state distributions, hence would not address transient performance behaviors, whereas our characterization does. We are also concerned with transient behaviors since most of the communication sessions in todays networks are short-lived [2]. Finally, the computation of a characterization in [21] is a labor-intensive and complex task, hence is not suitable for real-time applications [21].

6   Conclusions

We proposed a probabilistic characterization of network traffic and examined some of its properties; Theorems 1 through 5, Corollaries 1 through 3, stationarity, measurability, and viability. Performance implications of Def. 6 over switches and multiplexers are examined in [24]. The properties examined in this study and the results in [24] show that the proposed characterization enables a systematic approach for providing end-to-end probabilistic QoS guarantees in communication networks analytically; e.g. see Corollary 1. This characterization can handle traffic with heavy-tailed distributions in performance analysis. Future work includes building a network testbed to demonstrate the applicability of the proposed analysis framework.

Acknowledgement

References

[1] A. K. Parekh. *Breaking the Great Internet Deadlock OR Why there isn't any Network QoS and what to do about it?* (invited talk), 11th International Workshop on Quality of Service (IWQoS), 2003. URL: iwqos03.cs.berkeley.edu/invited-talks/parekh.pdf

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

[3] I. Norros. *A storage model with self-similar input*, Queueing Systems, vol. 16, pp. 387–396, 1994.

[4] B. Tsybakov, N. D. Georganas. *Self-similar traffic and upper bounds to buffer overflow in an ATM queue*, Performance Evaluation, vol. 32, no. 1, pp. 57–80, Feb. 1998.

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

[6] 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.

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

[8] 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.

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

[10] H. Sariowan. *A Service-curve Approach to Performance Guarantees in Integrated-service Networks*. Ph.D. Thesis, Dept. of ECE, University of California, San Diego, 1996.

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

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

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

[14] S. Ayyorgun, W. Feng. *A Deterministic Characterization of Network Traffic for Average Performance Guarantees*, Proc. of the 38th Ann. Conference on Information Sciences and Systems (CISS), 2004. (Also available as Tech. Report LA-UR-03-4477, Los Alamos Natl. Lab., 2003.)

[15] S. Ayyorgun, W. Feng. *A Probabilistic Definition of Burstiness Characterization: A Systematic Approach*, Tech. Report LA-UR-03-3668, Los Alamos Natl. Lab., 2003.

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

[17] A. Müller, D. Stoyan. *Comparison Methods for Stochastic Models and Risks*, Wiley, 2002.

[18] S. Ayyorgun, S. Vanichpun, W. Feng. *Probabilistic Synthesis and Regulation of Traffic for Probabilistic QoS Guarantees*, Technical Report LA-UR-04-1270, Los Alamos National Laboratory, February 2004.

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

[20] 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.

[21] S. Chong, S. Q Li. *Probabilistic Burstiness-Curve-Based Connection Control for Real-Time Multimedia Services in ATM Networks*, IEEE Journal on Selected Areas in Communications, vol. 15, no. 6, pp. 1072–1086, 1997.

[22] D. Starobinski, M. Sidi. *Stochastically Bounded Burstiness for Communication Networks*, IEEE Trans. on Information Theory, vol. 46, no. 1, Jan. 2000.

[23] Q. Yin, Y. Jiang, S. Jiang, P. Y. Kong. *Analysis on Generalized Stochastically Bounded Bursty Traffic for Communication Networks*, Proc. of the 27th IEEE Conf. on Local Comp. Networks (LCN), 2002.

[24] S. Ayyorgun, W. Feng. *A New Traffic Model and Its Implications over Multiplexers and Switches*, Tech. Report LA-UR-03-7268, Los Alamos Natl. Lab., 2003.

Appendix

Proof of Theorem 1: The proof follows by considering the following events for any n, σ, and $u \leqslant \sigma$;

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

Note 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, i.e. $A_1^c \cap A_2^c \subseteq A^c$. Thus, by utilizing the Union Bound and the fact that $R_1 \sim (S_1, f_1)$ and $R_2 \sim (S_2, f_2)$, we get

$$P(A) \leqslant P(A_1 \cup A_2) \leqslant P(A_1) + P(A_2) \leqslant f_1(u) + f_2(\sigma - u),$$

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

$$P(A) \leqslant \min_{u \leqslant \sigma} \{ f_1(u) + f_2(\sigma - u) \} = (f_1 \, \triangledown f_2)(\sigma).$$

We allowed a slight abuse of notation in the last two relations above. To be strict, we should have used infimum 'inf' above instead of minimum 'min', since 1) f_i's are real-valued, and 2) the minimum could potentially be taken over an infinite number of terms. However, note that the minimum (or infimum) is taken effectively over a set of finite number of elements due to the fact that $f_i(s) = 1$ for all $s < 0$ and $\lim_{s \to \infty} f_i(s) = 0$ for i equals to both 1 and 2. More specifically, the min-+ convolution $(f_1 \, \triangledown f_2)(\sigma)$ is equal to the minimum of the terms corresponding to each integer u in $[0, \sigma]$ and the term 1. Finally note that we also have $(f_1 \, \triangledown f_2)(\sigma) = 1$ for all $\sigma < 0$, again due to the properties of a bounding function mentioned above. This completes the proof. ∎

Proof of Theorem 2: We use a simple relation about the mean of a random variable, that is given by the last equality in (3). This relation can derived as follows: Let X is a non-negative random variable, we have

$$E[X] = \sum_{a=0}^{\infty} a \cdot f_X(a) = \sum_{a=0}^{\infty} \left(\sum_{u=0}^{\infty} [u < a] \right) \cdot f_X(a)$$

where the notation '$[\texttt{Statement}]$' stands[3] for 1 if the $\texttt{Statement}$ is $\texttt{true}$, and 0 otherwise,

$$= \sum_{a=0}^{\infty} \sum_{u=0}^{\infty} [u < a] \cdot f_X(a) = \sum_{u=0}^{\infty} \left(\sum_{a=0}^{\infty} [u < a] \cdot f_X(a) \right)$$
$$= \sum_{u=0}^{\infty} \left(\sum_{a > u} f_X(a) \right) = \sum_{u=0}^{\infty} P(X > u). \qquad (3)$$

The proof of the theorem follows by considering the mean of $R(n) - R(k)$ in any interval $(k, n]$ and upper-bounding it as;

$$E[R(n) - R(k)] = \sum_{\sigma=0}^{\infty} P(R(n) - R(k) > \sigma) \qquad \text{by (3)}$$

$$= \sum_{\sigma=0}^{S(n-k)-1} P(R(n) - R(k) > \sigma) + \sum_{\sigma=S(n-k)}^{\infty} P(R(n) - R(k) > \sigma)$$

$$= \sum_{\sigma=0}^{S(n-k)-1} P(R(n) - R(k) > \sigma) + \sum_{\sigma=0}^{\infty} P\Big(R(n) - R(k) > S(n-k) + \sigma \Big)$$

$$\leqslant \sum_{\sigma=0}^{S(n-k)-1} 1 + \sum_{\sigma=0}^{\infty} P\Big(R(n) - R(k) > S(n-k) + \sigma \Big)$$

$$= S(n-k) + \sum_{\sigma=0}^{\infty} P\Big(R(n) - R(k) > S(n-k) + \sigma \Big)$$

$$\leqslant S(n-k) + $$
$$\sum_{\sigma=0}^{\infty} P\Big(R(n) - R(u) > S(n-u) + \sigma, \text{ for some } u < n \Big)$$

$$\leqslant S(n-k) + \sum_{\sigma=0}^{\infty} f(\sigma).$$

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

$$\mu \triangleq \limsup_{(n-k) \to \infty} \frac{E[R(n) - R(k)]}{n-k}$$

$$\leqslant \limsup_{(n-k) \to \infty} \frac{S(n-k) + \sum_{\sigma=0}^{\infty} f(\sigma)}{n-k}$$

$$= \limsup_{n \to \infty} \frac{S(n)}{n} \quad \text{(holds since } \sum_{\sigma=0}^{\infty} f(\sigma) < \infty \text{).} \quad \blacksquare$$

Note that if $\sum_{\sigma=0}^{\infty} f(\sigma)$ is infinite, then there is no apparent relation between μ and $\limsup_{n \to \infty} \frac{S(n)}{n}$.

Proof of Theorem 3: Let G denote the output flow corresponding to R in the theorem. The proof follows by considering the following events for any n and $\sigma \geqslant 0$;

---

[3] We adopted this notation from '*Concrete Mathematics: A Foundation for Computer Science*', Addison-Wesley, 2nd ed., 1994, by R. L. Graham, D. E. Knuth, O. Patashnik.

$$\{Q(n) > \sigma\} = \{R(n) - G(n) > \sigma\}$$
$$\subseteq \left\{ R(n) - \min_{k \leqslant n}\{R(k) + S(n-k)\} > \sigma \right\}$$
$$= \left\{ \max_{k \leqslant n}\{R(n) - R(k) - S(n-k)\} > \sigma \right\}$$
$$= \left\{ R(n) - R(k) - S(n-k) > \sigma, \ \text{ for some } k < n \right\}$$
$$= \left\{ R(n) - R(k) > S(n-k) + \sigma, \ \text{ for some } k < n \right\}$$
$$= \Big\{ R(n) - R(k) >$$
$$\quad S^*(n-k) + \sigma + S(n-k) - S^*(n-k), \ \exists\, k < n \Big\}$$
$$\subseteq \left\{ R(n) - R(k) > S^*(n-k) + \sigma + (S \,\overline{\triangle}\, S^*)(0), \ \exists\, k < n \right\}.$$

As R is given to be $R \sim (S^*, f)$, we obtain the desired result by taking the probabilities of the both sides in the last line above; we have for all n and non-negative σ that

$$\mathrm{P}(Q(n) > \sigma) \leqslant$$
$$\quad \mathrm{P}\Big(R(n) - R(k) > S^*(n-k) + \sigma + (S \,\overline{\triangle}\, S^*)(0), \ \exists\, k < n \Big)$$
$$\qquad \leqslant f\big(\sigma + (S \,\overline{\triangle}\, S^*)(0)\big). \qquad \blacksquare$$

Proof of Theorem 4: The proof follows by considering the following events for any n, $k < n$, and σ;

$$\left\{ G(n) - G(k) > (S^* \,\overline{\triangledown}\, S)(n-k) + \sigma \right\}$$
$$\subseteq \left\{ G(n) - (R \,\overline{\triangledown}\, S)(k) > (S^* \,\overline{\triangledown}\, S)(n-k) + \sigma \right\}$$
$$\subseteq \left\{ R(n) - (R \,\overline{\triangledown}\, S)(k) > (S^* \,\overline{\triangledown}\, S)(n-k) + \sigma \right\}$$
$$= \left\{ R(n) - \min_{l \leqslant k}\{R(l) + S(k-l)\} > (S^* \,\overline{\triangledown}\, S)(n-k) + \sigma \right\}$$
$$= \left\{ \max_{l \leqslant k}\{R(n) - R(l) - S(k-l)\} > (S^* \,\overline{\triangledown}\, S)(n-k) + \sigma \right\}$$
$$= \left\{ R(n) - R(l) - S(k-l) > (S^* \,\overline{\triangledown}\, S)(n-k) + \sigma, \ \exists\, l \leqslant k \right\}$$
$$= \left\{ R(n) - R(l) > (S^* \,\overline{\triangledown}\, S)(n-k) + S(k-l) + \sigma, \ \exists\, l \leqslant k \right\}$$
$$\subseteq \left\{ R(n) - R(l) > S^*(n-l) - S(k-l) + S(k-l) + \sigma, \exists\, l \leqslant k \right\}$$
$$= \left\{ R(n) - R(l) > S^*(n-l) + \sigma, \ \text{ for some } l \leqslant k \right\}$$
$$= \left\{ R(n) - R(l) > S^*(n-l) + \sigma, \ \text{ for some } l \leqslant k < n \right\}$$
$$\subseteq \left\{ R(n) - R(l) > S^*(n-l) + \sigma, \ \text{ for some } l < n \right\}.$$

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

$$\mathrm{P}\Big(G(n) - G(k) > (S^* \,\overline{\triangledown}\, S)(n-k) + \sigma, \ \exists\, k < n \Big) \leqslant$$
$$\quad \mathrm{P}\Big(R(n) - R(l) > S^*(n-l) + \sigma, \ \exists\, l < n \Big)$$
$$\qquad \leqslant f(\sigma). \ \blacksquare$$

We might actually need to rectify the result in Theorem 4 slightly by replacing the service curve $S^* \,\overline{\triangledown}\, S$ in characterizing the output flow by S_o given below

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

We would like to have this rectification for two reasons: 1) A service curve is defined to take on the value zero for non-positive values of its argument. 2) In the proof, we utilize $S^* \,\overline{\triangledown}\, S$ only for positive values of its argument. We have not done this rectification in the body of the theorem in order not to clutter the result.

Proof of Theorem 5: The proof follows by considering the following events for any n and $\sigma \geqslant 0$;

$$\{D(n) > \sigma\} = \{G(n+\sigma) < R(n)\}$$
$$\subseteq \{(R \,\overline{\triangledown}\, S)(n+\sigma) < R(n)\}$$
$$= \{R(n) - (R \,\overline{\triangledown}\, S)(n+\sigma) > 0\}$$
$$= \left\{ R(n) - \min_{k \leqslant n+\sigma}\{R(k) + S(n+\sigma-k)\} > 0 \right\}$$
$$= \left\{ \max_{k \leqslant n+\sigma}\{R(n) - R(k) - S(n+\sigma-k)\} > 0 \right\}$$
$$= \left\{ R(n) - R(k) - S(n+\sigma-k) > 0, \ \text{ for some } k < n \right\}$$

(note that a k in the last line 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)

$$= \left\{ R(n) - R(k) > S(n+\sigma-k), \ \text{ for some } k < n \right\}$$
$$= \Big\{ R(n) - R(k) >$$
$$\quad S^*(n-k) + S(n-k+\sigma) - S^*(n-k), \ \exists\, k < n \Big\}$$
$$\subseteq \left\{ R(n) - R(k) > S^*(n-k) + (S \,\overline{\triangle}\, S^*)(\sigma), \ \exists\, k < n \right\}.$$

Hence, by taking the probabilities of both sides, and using the characterization of R, we get the desired result

$$\mathrm{P}(D(n) > \sigma) \leqslant$$
$$\quad \mathrm{P}\Big(R(n) - R(k) > S^*(n-k) + (S \,\overline{\triangle}\, S^*)(\sigma), \ \exists\, k < n \Big)$$
$$\qquad \leqslant f\big((S \,\overline{\triangle}\, S^*)(\sigma)\big). \qquad \blacksquare$$

Proof of Cor. 1 follows from Theorems 3 through 5 and a few algebraic properties of the min-+ algebra; see [15].

Proof of Corollary 2: The proof follows immediately by Theorem 3;

$$\mathrm{E}[Q(n)] = \sum_{\sigma=0}^{\infty} \mathrm{P}(Q(n) > \sigma) \leqslant \sum_{\sigma=0}^{\infty} f\big(\sigma + (S \,\overline{\triangle}\, S^*)(0)\big). \ \blacksquare$$

Proof of Corollary 3: The proof follows immediately by Theorem 5;

$$\mathrm{E}[D(n)] = \sum_{\sigma=0}^{\infty} \mathrm{P}(D(n) > \sigma) \leqslant \sum_{\sigma=0}^{\infty} f\big((S \,\overline{\triangle}\, S^*)(\sigma)\big). \ \blacksquare$$