Bounds for Independent Regulated Inputs Multiplexed in a Service Curve Network Element

Milan Vojnović and Jean-Yves Le Boudec

To Appear in IEEE Trans. on Communications, 2003.

***Abstract*— We consider the problem of bounding the probability of buffer overflow in a network node fed with independent arrival processes that are each constrained by arrival curves, but that are served as an aggregate. Existing results (for example [1] and [2]) assume that the node is a constant rate server. However, in practice, one finds complex network nodes that do not provide a constant service rate, and thus to which the existing bounds do not apply. Now many nodes can be adequately abstracted by a service curve property. We extend the results in [1] and [2] to such cases. As a by-product, we also provide a slight improvement to the bound in [2]. Our bounds are valid for both discrete and continuous time models.**

***Index Terms*—Statistical multiplexing, scheduling, queuing analysis**

I. Introduction

BOUNDS on the probability of buffer overflow in a network node (element) fed with independent arrival processes (inputs, flows) that are each constrained by arrival curves are obtained in [1], [2], [3], [4], [5], [6], [7], [8], [9], [10], [11], [12] under various assumptions. We say that a flow is regulated, or constrained, by an arrival curve $\alpha(\cdot)$, if the number of bits observed on the flow during any time interval of duration t is at most $\alpha(t)$. Leaky bucket regulation corresponds to an affine function $\alpha(\cdot)$. Existing results focus on work-conserving queuing systems that offer a constant service rate. However, in practice, the network nodes are often not work-conserving and do not offer the constant service rate at each instant of time. It turns out that many network nodes satisfy a service curve property [13], [14], [15], [16], [17]. In a deterministic context, a service curve property, with service curve β, means that at any time t, the total output traffic observed in $[0, t]$ is at least equal to $A(s) + \beta(t - s)$ for some s in $[0, t]$, where $A(s)$ is the total input traffic in $[0, s]$. Thus, it is of a practical importance to derive performance bounds for a service curve network element. In this note, on one hand, we extend the results by Kesidis and Konstantopoulos [1], and on the other hand, the results of Chang, Song, and Chiu [2] to hold for a service curve node. As a by-product, we also slightly improve the bound in [2], even for the case of a constant rate server. We also give a definition of service curve which is more adapted to a stochastic framework.

From the methodological viewpoint, a novelty of our approach is in that we systematically apply the following two steps: (1) we majorize the buffer overflow event with union of the events that are deviation of a sum of random variables from its mean, (2) under the given assumptions, these random variables are independent, with bounded support, and we know an upper bound on the summation mean; these properties allow us to use Hoeffding's inequalities [18]. In the first step, we often make use of sample-path results of deterministic network calculus (e.g. see [5], [17] and the references therein). Combined with the second step, where we apply Hoeffding's inequalities, it turns out that we are able to extend and recover the results of [1] and [2], and obtain some new ones.

Kesidis and Konstantopoulos [1], [3] consider a work-conserving constant rate server, and also assume that arrival curves are the combination of two leaky buckets (as is commonplace with ATM and in the Internet). In Section III (Theorem 1), we extend their results to a node that offers an arbitrary service curve, and to any arrival curve constraints. For this, we use a different proof; it is simpler, even for the original case considered in [1].

Chang, Song, and Chiu [2] consider the same problem as Kesidis and Konstantopoulos, but allow for arbitrary arrival curves. In Section IV (Theorem 3), we extend their result to a node that offers a service curve under a mild condition on the arrival and service curve (assumption (A6) in Section IV). Extending [2] to service curve is simple. However, by the virtue of stochastic comparisons and Hoeffding's inequalities we are able to obtain new bounds for the heterogeneous case, as explained later. We also slightly improve the bound in [2] (even for the original case), using an *under-sampling* argument. Incidentally, this makes the bound valid in continuous time, whereas [2] considers the discrete time case.

Both [1] and [2] give explicit results for the homogeneous case (all arrival curves are identical) and leave the heterogeneous case as an optimization problem to solve. For both cases, we also give simple formulas that apply to the heterogeneous case (Theorems 2 and 4). Of course, the bounds for the heterogeneous case also apply to the homogeneous case, but they are not as tight; this feature is inherited from Hoeffding's inequalities.

We also derive a variant for the heterogeneous case (Theorem 5), by combining the proof of Theorem 4 with a majorization similar to that found in [6]. The bound in Theorem 4 (as with Theorem 2) requires knowing the arrival curves of all flows. In contrast, Theorem 5 requires an incomplete knowledge about the arrival curves; it suffices to know the aggregate burstiness and aggregate sustainable rate. The bound is less tight than Theorem 4, but may be more useful in a context of differentiated services, where only aggregate information is available.

Chang, Song, and Chiu showed numerically that their bound is tighter than Kesidis and Konstantopoulos' bound. We confirm this also for our extensions by numerical computations: Theorems 3 and 4 seem to provide tighter bounds than Theorems 1 and 2, and should thus be preferred in practice. Section V shows a sample of numerical results. Another aspect would be to compare the bounds with empirical estimates, which goes beyond the scope of this paper.

The proofs of two lemmas are given in Appendix.

II. Notation and Assumptions

Consider a set $\mathcal{I} = \{1, 2, \ldots, I\}$ of flows input to a network element. Let A_i, for $i \in \mathcal{I}$, be a Borel counting measure on some probability space $(\Omega, \mathcal{F}, \mathsf{P})$. We interpret $A_i(s, t]$ as the number of bits observed on input flow i in the interval $(s, t]$. By convention, if $s > t$, $A_i(s, t] := -A_i((t, s])$. Likewise, define $A_i^*(s, t]$ for the output of the ith flow. Let $A(s, t] := \sum_{i=1}^{I} A_i(s, t]$ and $A^*(s, t] := \sum_{i=1}^{I} A_i^*(s, t]$.

Paper approved by G. S. Kuo, the Editor for Network Architecture of the IEEE Communications Society. Manuscript received May 9, 2001; revised September 30, 2002.

The authors are with EPFL-IC (ISC/LCA), CH-1015 Lausanne, Switzerland (e-mail: milan.vojnovic@epfl.ch, jean-yves.leboudec@epfl.ch).

We make the following assumptions:

(A1) $A_1, A_2, \ldots, A_I$ are independent.

(A2) For all $i \in \mathcal{I}$, A_i has α_i as an arrival curve, i.e. for all $s, t \in \mathbb{R}$,

$$A_i(s,t] \leq \alpha_i(t-s),\ \mathsf{P}-\text{a.s.},$$

where α_i is a non-negative, wide-sense increasing function[1] such that $\alpha(t) = 0$, for all $t < 0$. We assume, without loss of generality, that α_i is sub-additive, i.e. $\alpha_i(t+s) \leq \alpha_i(t)+\alpha_i(s)$ for all $t, s \in \mathbb{R}$ [14], [15], [16], [17].

(A3) For each $i \in \mathcal{I}$, and any $s, t \in \mathbb{R}$,

$$\mathsf{E}[A_i(s,t]] \leq \bar{\alpha}_i \times (t-s), \tag{1}$$

where $\bar{\alpha}_i = \lim_{t\to\infty} \frac{\alpha_i(t)}{t} = \inf_{t>0} \frac{\alpha_i(t)}{t}$. (The last equality is by sub-additivity of α_i [20].)

(A4) There exists a sequence of random points ("the construction points"):

$$\ldots < S_{-2} < S_{-1} < S_0 \leq 0 < S_1 < S_2 < \ldots$$

such that $\lim_{n\to-\infty} S_n = -\infty$ and $\lim_{n\to\infty} S_n = \infty$, and for all $n \in \mathbb{Z}$, $A(S_n, S_{n+1}] = A^*(S_n, S_{n+1}]$, $\mathsf{P}-$a.s..

(A5) Define $\mathcal{S}(t) = \{S_n,\ n \in \mathbb{Z} :\ S_n \leq t\}$. The network element offers the service curve β to the aggregate of all flows, if for all $t \in \mathbb{R}$, and any $u \in \mathcal{S}(t)$,

$$\exists s \in [u,t] :\ A^*(u,t] - A(u,s] \geq \beta(t-s),\ \mathsf{P}-\text{a.s.},$$

where β is a non-negative wide-sense increasing function.

Let $Q(t)$ be the number of bits in the network element at time t (it is the unfinished work; we call it backlog). We assume that the element has buffer capacity that is sufficient to ensure no losses. Then, indeed, $Q(t) = A(u,t] - A^*(u,t]$, for any $u \in \mathcal{S}(t)$. From (A5), it follows that, for any $t \in \mathbb{R}$,

$$Q(t) \leq \sup_{-\infty<s\leq t} \{A(s,t] - \beta(t-s)\}. \tag{2}$$

In the next two sections we give upper bounds on $\mathsf{P}(Q(0) > q)$, for an arbitrary time instant 0. Before that, in the remainder of this section, we first introduce some additional definitions and then discuss assumptions (A3)–(A5).

For two functions f and g, we define the *vertical* and *horizontal deviations* by $v(f,g) = \sup_{t\geq 0}\{f(t) - g(t)\}$, $h(f,g) = \sup_{t\geq 0}\{\inf\{u \geq 0 :\ f(t) \leq g(t+u)\}\}$ [17]. Note that $v(f,g)$ is the worst-case backlog for a network element that offers the service curve g to the aggregate arrival process that has f as an arrival curve. Similarly, $h(f,g)$ is the worst-case virtual delay (equal to the worst-case delay if the node would be FIFO). We also define $\lambda_a(t) = at$ for $t \geq 0$ and $\lambda_a(t) = 0$ for $t < 0$, $a \in \mathbb{R}$. Let $\bar{\alpha} = \sum_{i=1}^I \bar{\alpha}_i$ and $\alpha = \sum_{i=1}^I \alpha_i$.

We discuss (A3) first. Note that (A3) is true for $A_1, A_2, \ldots, A_I$ stationary and ergodic in their intensities. Indeed, by stationarity $\mathsf{E}[A_i(s,t]] = \mathsf{E}[A_i(0,1]](t-s)$ and by ergodicity $\mathsf{E}[A_i(0,1]] = \lim_{u\to\infty} A_i(0,u]/u \leq \lim_{u\to\infty} \alpha_i(u)/u = \bar{\alpha}_i$.

Regarding (A4), it follows from a known result (see, e.g., Lemma 1 in [21]) that for (A4) to hold it is sufficient that

(A4-a) $v(\alpha,\beta) < \infty$,

(A4-b) $\liminf_{t\to\infty} \{\alpha(t) - \beta(t)\} = -\infty$.

For instance, for the rate-latency service curve $\beta(t) = c\max\{t-e, 0\}$, $c, e \geq 0$, the second condition is the intuitive stability condition $\bar{\alpha} < c$. In the general case, roughly speaking, conditions (A3) and (A4) are weak stability conditions.

Next, note that the definition in (A5) is different than the classical service curve definition (e.g., see [17], Section 1.3.1), which is in the framework of deterministic network calculus; there it would be $A(-\infty, 0] = A^*(-\infty, 0] = 0$. It can be easily observed that the two definitions are compatible. However, in contrast to the classical definition, we do not assume that the system is empty at time 0.

[1] We say that function $\alpha(\cdot)$ is wide-sense increasing if $s \leq t$ always implies $\alpha(s) \leq \alpha(t)$. This is also called "non-decreasing".

III. Extending Kesidis and Konstantopoulos' bound

We extend [1] in the following two theorems, the proofs of which are given at the end of this section.

Theorem 1—Homogeneous Case: Assume (A1)–(A5), $v(\alpha,\beta) < \infty$, $h(\alpha,\beta) < \infty$, and $\alpha_i = \alpha_1$, for all $i \in \mathcal{I}$. Then, for $\bar{\alpha}h(\alpha,\beta) < q < v(\alpha,\beta)$,

$$\mathsf{P}(Q(0) > q) \leq \exp\left(-I\frac{q}{v}\ln\frac{q}{\bar{\alpha}h} + I\left(1-\frac{q}{v}\right)\ln\frac{v-\bar{\alpha}h}{v-q}\right),$$

where for brevity $v = v(\alpha,\beta)$ and $h = h(\alpha,\beta)$.

The theorem gives us a bound for $q \in (\bar{\alpha}h(\alpha,\beta), v(\alpha,\beta))$. Otherwise, for $q \leq \bar{\alpha}h(\alpha,\beta)$, use $\mathsf{P}(Q(0) > q) \leq 1$, and for $q \geq v(\alpha,\beta)$, $\mathsf{P}(Q(0) > q) = 0$.

We can apply Theorem 1 to the original case in [1] by letting $\alpha_1(t) = \min(\pi_1 t, \bar{\alpha}_1 t + \sigma_1)$ and $\beta(t) = ct$. It can be found later in the proof of Theorem 1 that the bound is obtained by computing $\sup_{\theta>0} F(\theta)$, where in this special case, $F(\cdot)$ reads as

$$F(\theta) = \theta q - I\ln\left(1 - \frac{\bar{\alpha}_1}{c} + \frac{\bar{\alpha}_1}{c}e^{\theta\frac{\pi_1-c}{\pi_1-\bar{\alpha}_1}\sigma_1}\right),$$

which is exactly the result in Theorem 1 of [1]; this shows that we do have an extension of that result. It has to be mentioned that, in fact, [1] proves a tighter bound than that of Theorem 1 [1], but which is not expressible in a closed-form (see discussion in Sec. III [1]).

Next, we provide a looser bound than in Theorem 1, but which holds for the heterogeneous case.

Theorem 2—Heterogeneous Case: Assume (A1)–(A3) and (A5). Let $\mathcal{G} = \{(\gamma_1, \gamma_2, \ldots, \gamma_I) \in \mathbb{R}_+^I :\ \forall i \in \mathcal{I},\ v_i, h_i < \infty,\ \sum_{i=1}^I \gamma_i \leq 1\}$, where for brevity $v_i := v(\alpha_i, \gamma_i\beta)$ and $h_i := h(\alpha_i, \gamma_i\beta)$. Assume, in addition, that for each $i \in \mathcal{I}$, (A4) holds for a virtual node that offers the service curve $\gamma_i\beta$ fed with the arrival process A_i. Then, for any $\underline{\gamma} \in \mathcal{G}$, and $\sum_{i=1}^I \bar{\alpha}_i h(\alpha_i, \gamma_i\beta) < q < v(\alpha,\beta)$,

$$\mathsf{P}(Q(0) > q) \leq \exp\left(-F(\underline{\gamma})\right), \tag{3}$$

where

$$F(\underline{\gamma}) = \frac{2(q - \sum_{i=1}^I \bar{\alpha}_i h(\alpha_i, \gamma_i\beta))^2}{\sum_{i=1}^I v(\alpha_i, \gamma_i\beta)^2}.$$

Proof: [Theorem 1] Define, for each $i \in \mathcal{I}$, and all $t \in \mathbb{R}$,

$$Q_i(t) = \sup_{-\infty<s\leq t} \{A_i(s,t] - \gamma_i\beta(t-s)\}.$$

Now from (2), for any $(\gamma_1, \gamma_2, \ldots, \gamma_I) \in \mathbb{R}_+^I$ such that $\sum_{i=1}^I \gamma_i \leq 1$, we have, $Q(t) \leq \sum_{i=1}^I Q_i(t)$, for any $t \in \mathbb{R}$. Hence,

$$\mathsf{P}(Q(0) > q) \leq \mathsf{P}(\sum_{i=1}^I Q_i(0) > q). \tag{4}$$

We note the following properties.

1) For any $t \in \mathbb{R}$,

$$Q_1(t), Q_2(t), \ldots, Q_I(t) \text{ are independent.} \tag{5}$$

2) For any $t \in \mathbb{R}$, and each $i \in \mathcal{I}$,

$$0 \leq Q_i(t) \leq v(\alpha_i, \gamma_i \beta). \tag{6}$$

3) For any $t \in \mathbb{R}$,

$$\mathsf{E}[Q(t)] \leq \bar{\alpha} h(\alpha, \beta). \tag{7}$$

The first property is obvious from (A1); the second from (A2) and the definition of the vertical deviation. We prove the third property next. To that end, define, for any $t \in \mathbb{R}$,

$$V(t) = \inf\{v \in [0, s] : \ s \in \mathcal{S}(t), \ A^*(s, t] \geq A(s, t - v]\}.$$

Note that $V(t)$ is the virtual delay (sojourn time) of a bit that departs at time t. If the system would be FIFO, then $V(t)$ is the delay of a bit that departs at t. It can easily be shown that for any $t \in \mathbb{R}$, $V(t) \leq h(\alpha, \beta)$.

Next, note, for any $t \in \mathbb{R}$, $s \in \mathcal{S}(t)$,

$$\begin{aligned} Q(t) &= A(s, t] - A^*(s, t] \leq A(s, t] - A(s, t - V(t)] \\ &\leq A(s, t] - A(s, t - h(\alpha, \beta)] = A(t - h(\alpha, \beta), t]. \end{aligned}$$

Taking expectation in the above display and combining with (A3) we recover (7).

Let $\gamma_i = 1/I$. By (4)-(7) and using (4.5) in the proof of Hoeffding's inequality (Theorem 1, [18]), we obtain that for any $\theta > 0$,

$$\mathsf{P}(Q(0) > q) \leq e^{-\theta q} \left(1 - \frac{\mathsf{E}[Q_1(0)]}{v(\alpha_1, \beta/I)} + \frac{\mathsf{E}[Q_1(0)]}{v(\alpha_1, \beta/I)} e^{\theta v(\alpha_1, \beta/I)}\right)^I.$$

The right-hand side in the last inequality is increasing with $\mathsf{E}[Q_1(0)]$. Now, by (7) applied to Q_i, we obtain $\mathsf{E}[Q_i(0)] \leq \bar{\alpha}_i h(\alpha_i, \gamma_i \beta) = \bar{\alpha}_1 h(\alpha_1, \beta/I)$. It is simple to observe $h(\alpha_1, \beta/I) = h(\alpha, \beta)$, and $v(\alpha_1, \beta/I) = v(\alpha, \beta)/I$. We showed

$$\mathsf{P}(Q(0) > q) \leq \exp\left(-\sup_{\theta > 0} F(\theta)\right),$$

where

$$F(\theta) = q\theta - I \ln\left(1 - \bar{\alpha} \frac{h(\alpha, \beta)}{v(\alpha, \beta)} + \bar{\alpha} \frac{h(\alpha, \beta)}{v(\alpha, \beta)} e^{\theta \frac{v(\alpha, \beta)}{I}}\right). \tag{8}$$

Computing $\sup_{\theta > 0} F(\theta)$ yields the desired result. ∎

Note that we could immediately apply Hoeffding's inequality (Theorem 1, [18]) to (4)-(7). However, the last part of the proof is given for the sake of a comparison with [1] made earlier.

Proof: [Theorem 2] The proof builds upon the proof of Theorem 1. Given (4)-(7), the problem is equivalent to deriving an upper bound on the complementary distribution (4) of a sum of independent *non-uniformly* bounded random variables. From Hoeffding's inequality (Theorem 2, [18]) it follows that, for any $\underline{\gamma} \in \mathcal{G}$, and $q > \sum_{i=1}^I \mathsf{E}[Q_i(0)]$,

$$\mathsf{P}(Q(0) > q) \leq \exp\left(-\frac{2(q - \sum_{i=1}^I \mathsf{E}[Q_i(0)])^2}{\sum_{i=1}^I v(\alpha_i, \gamma_i \beta)^2}\right).$$

The right-hand side is increasing with $\sum_{i=1}^I \mathsf{E}[Q_i(0)]$, hence we can replace it with its upper bound $\sum_{i=1}^I \bar{\alpha}_i h(\alpha_i, \gamma_i \beta)$ and still have a bound. This recovers the inequality in (3), which completes the proof. ∎

IV. Extending Chang, Song, and Chiu's bound

We extend [2] in three theorems, the proofs of which are given at the end of this section.

Assume in addition to (A1)–(A5);

(A6) There exists $\tau < \infty$ such that for all $s \geq \tau$, $\beta(s) \geq \alpha(s)$.

(A6) is a stronger form of (A4-b), which holds in practice (for example, but not only, when α is concave and β is convex) when the natural stability conditions are met. Notice that τ replaces, in the context of service curve, the concept of an upper bound on the duration of a busy period, which is useful only for work-conserving servers.

For any $K \in \mathbb{N}$, and any $t \geq 0$, let $\mathcal{T}_K(t)$ be the set of partitions of $[0, t]$ in K intervals, in other words

$$\mathcal{T}_K(t) = \{(t_0, t_1, ..., t_K) : \ 0 = t_0 \leq t_1 \leq ... \leq t_K = t\}.$$

(if time would be discrete, we require that the partition $\mathcal{T}_K(t)$ is uniform, i.e. $t_k = kt/K$, $k = 0, \ldots, K$).

Theorem 3—Homogeneous Case: Assume (A1)–(A6) and $\alpha_i = \alpha_1$, for all $i \in \mathcal{I}$. Then, for any $K \in \mathbb{N}$ and any $\underline{t} \in \mathcal{T}_K(\tau)$,

$$\mathsf{P}(Q(0) > q) \leq \sum_{k=0}^{K-1} \exp\left(-I g(t_k, t_{k+1})\right), \tag{9}$$

where, for $q > \alpha(v) - \beta(u)$, $g(u, v) = +\infty$, else for $q < \bar{\alpha} v - \beta(u)$, $g(u, v) = 0$, else

$$g(u, v) = \frac{\beta(u) + q}{\alpha(v)} \ln \frac{\beta(u) + q}{\bar{\alpha} v} + \left(1 - \frac{\beta(u) + q}{\alpha(v)}\right) \ln \frac{\alpha(v) - \beta(u) - q}{\alpha(v) - \bar{\alpha} v}.$$

If time would be discrete, and we let $\beta(s) = c(s + 1)$, $K = t$, $t_k = k$, then the theorem gives the same bound as [2]. However, even for the original scenario in [2], we have a slight improvement: if τ is large (which may happen simply because our time unit is very small), we expect the bound in [2] to be large, because it relies on the union bound. We expect to have a better bound by allowing K to be smaller than τ (under-sampling). This is verified in Section V. Note that the theorem implies that for any $K \in \mathbb{N}$ and $\underline{t} \in \mathcal{T}_K(\tau)$, the right hand-side in (9) *is* a bound; hence, we can take infimum over all possible partitions of $[0, \tau]$.

Next, we provide a looser bound than in Theorem 3, but which holds for the heterogeneous case.

Theorem 4—Heterogeneous Case: Assume (A1)–(A6). Then, for any $K \in \mathbb{N}$ and any $\underline{t} \in \mathcal{T}_K(\tau)$, for $q < v(\alpha, \beta)$,

$$\mathsf{P}(Q(0) > q) \leq \sum_{k=0}^{K-1} \exp\left(-g(t_k, t_{k+1})\right), \tag{10}$$

where

$$g(u, v) = \frac{2((q + \beta(u) - \bar{\alpha} v)^+)^2}{\sum_{i=1}^I \alpha_i(v)^2},$$

and $(\cdot)^+ = \max\{\cdot, 0\}$.

We can derive an additional bound for the heterogeneous case that requires only aggregate information about the arrival curves. We obtain this by using a majorization similar to [6] for leaky-bucket constrained processes. Note that this result (and this result only) is stated under a stronger assumption than (A3), namely,

(A3bis) $A_1, A_2, \ldots, A_I$ are stationary and ergodic.

Theorem 5—Heterogeneous Case: Assume (A1), (A2), (A3bis), (A4)–(A6). Then, the same bound as in (10) holds, with

$$g(u, v) = \frac{((q + \beta(u) - \bar{\alpha} v)^+)^2}{2 \sum_{i=1}^I v(\alpha_i, \lambda_{\bar{\alpha}_i})^2}.$$

The proofs of the above theorems require two lemmas, proved in appendix.

Lemma 1: Under (A2), (A5), and (A6), for any $q \geq 0$, it holds

$$\mathsf{P}(Q(0) > q) \leq \mathsf{P}(\sup_{0 \leq s \leq \tau} \{A(-s, 0] - \beta(s)\} > q).$$

Lemma 2: We have, for any $K \in \mathbb{N}$, $\underline{t} \in \mathcal{T}_K(\tau)$, and $q \geq 0$,

$$\mathsf{P}(\sup_{0\leq s\leq \tau}\{A(-s,0]-\beta(s)\} > q) \leq \sum_{k=0}^{K-1} \mathsf{P}(A(-t_{k+1},0] > q+\beta(t_k)). \tag{11}$$

Proof: [Theorem 3] By the hypothesis of the theorem, for any $s,t \in \mathbb{R}$, and any $i \in \mathcal{I}$, $A_i(s,t]$ is uniformly bounded with $\alpha_1(t-s)$ ((A2)). Thus, the kth summation term in (11) is the complementary distribution of a sum of independent uniformly bounded random variables. By Hoeffding's inequality (Theorem 1, [18]), and $\mathsf{E}[A_i(-t_{k+1},0]] \leq \bar{\alpha}_i t_{k+1}$ ((A3)), the kth summation term in (11) is upper-bounded by $\exp(-Ig(t_k,t_{k+1}))$, for $q > \bar{\alpha}t_{k+1} - \beta(t_k)$. This proves the result. ∎

Proof: [Theorem 4] By Hoeffding's inequality (Theorem 2, [18]), the kth summation term in (11) is upper-bounded with

$$\exp\left(-\frac{2(q+\beta(t_k)-\mathsf{E}[A(-t_{k+1},0]])^2}{\sum_{i=1}^I \alpha_i^2(t_{k+1})}\right).$$

For $q > \mathsf{E}[A(-t_{k+1},0]] - \beta(t_k)$, the last display is wide-sense increasing with $\mathsf{E}[A(-t_{k+1},0]]$. From (1), $\mathsf{E}[A(-t_{k+1},0]] \leq \bar{\alpha}t_{k+1}$, thus for $q > \bar{\alpha}t_{k+1} - \beta(t_k)$ we can replace $\mathsf{E}[A(-t_{k+1},0]]$ with $\bar{\alpha}t_{k+1}$ and still have an upper bound. ∎

Proof: [Theorem 5] Define, for any $t \in \mathbb{R}$, $\varepsilon > 0$,

$$\tilde{Q}_i^\varepsilon(t) := \sup_{-\infty<s\leq t}\{A_i(s,t] - (1+\varepsilon)\bar{\alpha}_i(t-s)\},$$

let, also, $Z_i^\varepsilon(t) := \tilde{Q}_i^\varepsilon(0) - \tilde{Q}_i^\varepsilon(-t)$.

Note, by (A3bis) and (A2), $\mathsf{E}[A_i(0,1]] < (1+\varepsilon)\bar{\alpha}_i$, and thus $\tilde{Q}_i^\varepsilon$, is stable. The last implies that for any $t \in \mathbb{R}$, $\mathsf{E}[Z_i^\varepsilon(t)] = 0$. Note, also from (A2), $-v(\alpha_i,\lambda_{\bar{\alpha}_i}) \leq Z_i(t) \leq v(\alpha_i,\lambda_{\bar{\alpha}_i})$, $t \in \mathbb{R}$.

Now, observe, for any $s,t \in \mathbb{R}$, $s \leq t$,

$$\tilde{Q}_i^\varepsilon(t) - \tilde{Q}_i^\varepsilon(s) \geq A_i(s,t] - (1+\varepsilon)\bar{\alpha}_i(t-s).$$

Hence, for any $t \in \mathbb{R}$, $Z_i^\varepsilon(t) \geq A_i(-t,0] - (1+\varepsilon)\bar{\alpha}_i t$, and thus

$$\begin{aligned}\mathsf{P}(A(-t,0] - (1+\varepsilon)\bar{\alpha}t > z) &\leq \mathsf{P}(\textstyle\sum_{i=1}^I Z_i^\varepsilon(t) > z)\\ &\leq \exp\left(-\frac{z^2}{2\sum_{i=1}^I v(\alpha_i,\lambda_{\bar{\alpha}_i})^2}\right),\end{aligned} \tag{12}$$

where the last inequality is by applying Hoeffding's inequality (Theorem 2, [18]) for a sum of independent zero-mean non-uniformly bounded random variables. Finally, from (12) and a simple variable substitution, we have, for any $u,v \geq 0$, $q \geq (1+\varepsilon)\bar{\alpha}(v) - \beta(u)$,

$$\mathsf{P}(A(-v,0] > q + \beta(u)) \leq \exp\left(-\frac{(q+\beta(u)-(1+\varepsilon)\bar{\alpha}v)^2}{2\sum_{i=1}^I v(\alpha_i,\lambda_{\bar{\alpha}_i})^2}\right).$$

By continuity of the right-hand side, we can let $\varepsilon \to 0$, and then combining with Lemmas 1 and 2, we complete the proof. ∎

V. Numerical Comparison of Bounds

We give numerical results for leaky-bucket constrained input flows, $\alpha_i(t) = \bar{\alpha}_i t + \sigma_i$, where $\bar{\alpha}_i$ is the sustainable rate, and σ_i is the burstiness of the ith flow. We assume packets of fixed-length equal to $L = 1500$ bytes. We consider the rate-latency service curve $\beta(t) = c\max\{t-e,0\}$, with rate $c = 150$ Mbps, and latency $e = L/c$.

We consider both homogeneous (Theorems 1 and 3) and heterogeneous case (Theorems 2 and 4), and Theorem 5 later. For the bounds of Theorems 3, 4, and 5, we uniformly partition the interval $[0,\tau]$, such that $t_k = k\tau/K$, for $k = 0,1,\ldots,K$, and then find $K \in \mathbb{N}$ that attains the minimum. In the homogeneous case, we set $\bar{\alpha}_1 = \rho c/I$ and $\sigma_1 = 8L$; here $\rho \in (0,1)$ is the load. We show numerical results in Figure 1 for $I = 100$, $\rho = 0.2$ and 0.8. In the heterogeneous case, we suppose two classes of the input flows each consisting of I_1 and I_2 flows, respectively. We set $\bar{\alpha}_1 = 2\bar{\alpha}_2$, $\sigma_1 = 8L$, and $\sigma_2 = 5L$. (Here the subscript 1 (2) refers to the first (second) class flow.) The results are shown in Figure 2, for $I_1 = I_2 = 50$, $\rho = 0.2$ and 0.8.

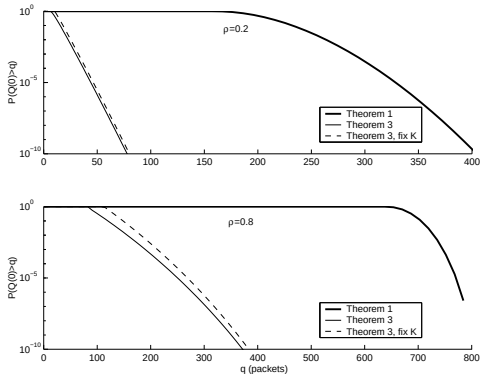

Fig. 1. Bounds of Theorems 1 and 3 for the homogeneous case of $I = 100$ input flows. The graphs are given for the loads (upper graph) $\alpha = 0.2$, and (lower graph) $\alpha = 0.8$. Bound of Theorem 3 is computed for uniform partition of $[0,\tau]$; for the optimum K and K fixed to $\lceil \tau/e \rceil$.

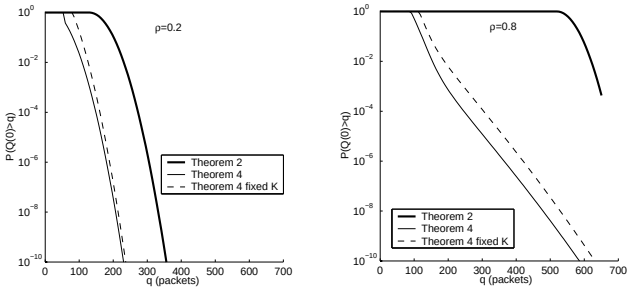

Fig. 2. Bounds of Theorems 2 and 4 for the heterogeneous case of two classes of the input flows each consisting of 50 flows. The graphs are for the loads (upper graph) $\alpha = 0.2$ and (lower graph) $\alpha = 0.8$. Bound of Theorem 4 is for uniform partition of $[0,\tau]$; for the optimum K and K fixed to $\lceil \tau/e \rceil$.

We make a few main observations. First, we find that the extensions of Chang, Song, and Chiu's bound (excluding Theorem 5, which is handled separately later) is substantially tighter than the extensions of Kesidis and Konstantopoulos' bound (see Figure 1 and 2). This confirms a similar observation in [2]. Second, the bound in Theorem 3 becomes tighter as we optimize with respect to K; this slightly improves upon [2]. We note that the bound of Theorem 2 reads as

$$F(\underline{\gamma}^*) = \frac{2((q-1/c(\sum_{i=1}^I \sqrt{\bar{\alpha}_i\sigma_i})^2)^+)^2}{\sum_{i=1}^I(\sigma_i+\bar{\alpha}_i e)^2}.$$

We next compare our exact bounds with the bounds obtained by neglecting the latency parameter (this would correspond if we would approximate the system with a constant rate server). In Figure 3, we show the bound of Theorem 3 for the latency parameter e equal to 0, 4, and $8\ L/c$. We observe that the bounds obtained for $e = 0$ (constant rate server) are over-optimistic. This is not negligible and is emphasized for lighter load; for $\rho = 0.2$, the discrepancy is about one order of magnitude for some backlog values.

Our next objective is to demonstrate how the bound of Theorem 4 (which holds for the heterogeneous case) compare with the bound of Theorem 3 (which holds only for the homogeneous case) in the homogeneous setting. We also compare with the bound of Theorem 5.

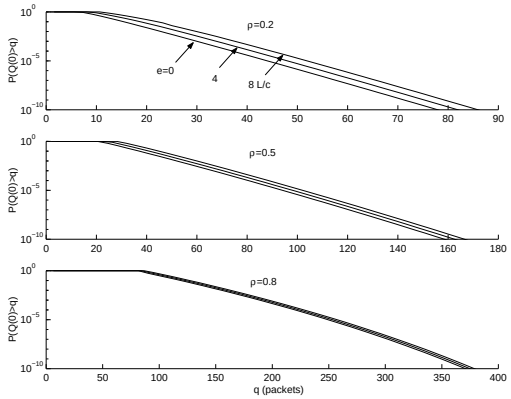

Fig. 3. Bound of Theorem 3 for the homogeneous case of $I = 100$ input flows and the latencies $e = \{0, 4, 8\}\ L/c$. The graphs are given for the loads (upper graph) $\rho = 0.3$, (middle graph) $\rho = 0.5$, and (lower graph) $\rho = 0.8$.

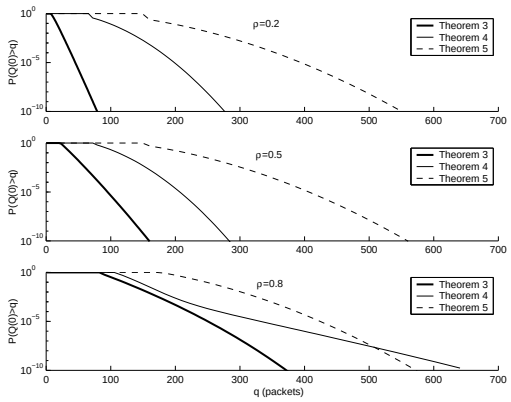

Fig. 4. Bounds of Theorems 3, 4, and 5, for the homogeneous case. The graphs are given for the loads $\rho = 0.2$, 0.5, and 0.8, top to bottom graphs, respectively.

We observe that for a light to moderate load the bound of Theorem 4 is substantially conservative with respect to the bound of Theorem 3. For high load, the bound of Theorem 4 is fairly close to the bound of Theorem 3, except for the buffer beyond certain value when it deviates in a conservative direction. One can combine the bounds obtained in the derivation of Theorem 4 and 5 to obtain a better bound [22].

VI. Concluding Remarks

Note that all the bounds in this paper, and thus the original bounds in [1] and [2], are applications of Hoeffding's inequalities [18]. The method used in this paper consists of stochastic comparisons and Hoeffding's inequalities; the method also extends to bounding probabilities of other events of interest, e.g. for delay and loss [22], [23].

References

[1] G. Kesidis and T. Konstantopoulos, "Worst-case performance of a buffer with independent shaped arrival processes," *IEEE Communications Letters*, vol. 4, no. 1, January 2000.

[2] C.-S. Chang, W. Song, and Y. Ming Chiu, "On the performance of multiplexing independent regulated inputs," in *Proc. of Sigmetrics 2001*, Massachusetts, USA, May 2001.

[3] G. Kesidis and T. Konstantopoulos, "Extremal traffic and worst-case performance for queues with shaped arrivals," *Fields Institute Communications/AMS, ISBN 0-8218-1991-7*, 2000.

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

[5] C.-S. Chang, *Performance Guarantees in communication networks*, Springer-Verlag, 2000.

[6] L. Massoulié and A. Busson, "Stochastic majorization of aggregates of leaky bucket-constrained traffic streams," preprint, http://www.research.microsoft.com/users/lmassoul/, 2000.

[7] A. Elwalid, D. Mitra, and R. H. Wentworth, "A new approach for allocating buffers and bandwidth to heterogeneous, regulated traffic in an ATM node," *IEEE Journal on Selected Areas in Communications*, vol. 13, no. 6, pp. 1115–1127, 1995.

[8] F. Lo Presti, Z. Zhang, D. Towsley, and J. Kurose, "Source time scale optimal buffer/bandwidth trade-off for regulated traffic in an ATM node," *IEEE Trans. on Networking*, vol. 7, no. 4, pp. 490–501, August 1999.

[9] D. Botvich and N. Duffield, "Large deviations, the shape of the loss curve, and economies of scale in large multiplexers," *Queueing Systems*, vol. 20, pp. 293–320, 1995.

[10] S. Rajagopal, M. Reisslein, and K. W. Ross, "Packet multiplexers with adversarial regulated traffic," in *Proc. of IEEE INFOCOM 1998*, 1998, pp. 347–355.

[11] R. Boorstyn, A. Burchard, J. Liebeherr, and C. Ootamakorn, "Statistical service assurances for traffic scheduling algorithms," *IEEE Journal of Selected Areas in Communications*, vol. 18, no. 12, pp. 2651–2664, December 2000.

[12] K. Kumaran and M. Mandjes, "Multiplexing regulated traffic streams: Design and performance," in *Proc. of IEEE INFOCOM 2001*, March 2001.

[13] A. K. Parekh and R. G. Gallager, "A generalized processor sharing approach to flow control in integrated services networks: The single node case," *IEEE/ACM Trans. on Networking*, vol. 1-3, pp. 344–357, June 1993.

[14] J.-Y. Le Boudec, "Application of network calculus to guaranteed service network," *IEEE Trans. on Information Theory*, vol. 44, pp. 1087–1096, May 1998.

[15] C. S. Chang, "On deterministic traffic regulation and service guarantee: A systematic approach by filtering," *IEEE/ACM Trans. on Networking*, vol. 44, pp. 1096–1107, August 1998.

[16] R. Agrawal, R. L. Cruz, C. Okino, and R. Rajan, "Performance bounds for flow control protocols," *IEEE/ACM Trans. on Networking*, vol. 7, no. 3, pp. 310–323, June 1999.

[17] J.-Y. Le Boudec and P. Thiran, *Network Calculus*, Springer-Verlag, 2001.

[18] W. Hoeffding, "Probability inequalities for sums of bounded random variables," *American Statistical Association Journal*, pp. 13–30, March 1963.

[19] F. Baccelli and P. Brémaud, *Elements of Queueing Theory*, vol. 26, Applications of Mathematics, Springer-Verlag, 1991.

[20] J. F. C. Kingman, "Subadditive processes," in *Ecole d'été de probabilité de Saint-Flour*, vol. Lecture Notes in Mathematics (539), pp. 165–223. Springer Verlag, 1976.

[21] T. Konstantopoulos and G. Last, "On the dynamics and performance of stochastic fluid systems," *Journal of Applied Probability*, vol. 37, pp. 652–667, 2000.

[22] M. Vojnović and J.-Y. Le Boudec, "Stochastic analysis of some expedited forwarding networks," in *Proc. of IEEE INFOCOM 2002*, vol. 2, New-York, NY, June 2002, pp. 1004–1013.

[23] M. Vojnović and J.-Y. Le Boudec, "Elements of probabilistic network calculus for packet scale rate guarantee nodes," in *Proc. of MTNS 2002*, South Bend, IN, August 2002.

Appendix

Proof: From (2), for any $t \in \mathbb{R}$, $Q(t) \leq \max\{X(t), Y(t)\}$, where $X(t) := \sup_{-\infty<s\leq t-\tau}\{A(s,t] - \beta(t-s)\}$, and $Y(t) := \sup_{t-\tau<s\leq t}\{A(s,t] - \beta(t-s)\}$. By (A2), for any $t \in \mathbb{R}$, $X(t) \leq \sup_{-\infty<s\leq t-\tau}\{\alpha(t-s) - \beta(t-s)\}$. Now assume τ satisfies (A6), then we conclude $X(t) \leq 0$, for any $t \in \mathbb{R}$. It follows, for any $t \in \mathbb{R}$, $Q(t) \leq \max\{0, Y(t)\}$. Hence, for any $q \geq 0$, and $t \in \mathbb{R}$,

$$\mathsf{P}(Q(t) > q) \leq \mathsf{P}(\max\{0, Y(t)\} > q) = \mathsf{P}(Y(t) > q),$$

which by definition of Y recovers the stated claim. ■

Proof: Fix any $K \in \mathbb{N}$ and any $0 = t_0 \leq t_1 \leq \ldots \leq t_K = \tau$. Note, for any s such that $t_k \leq s < t_{k+1}$,

$$A(-\tau, -s] \geq A(-\tau, -t_{k+1}] \text{ and } \beta(s) \geq \beta(t_k).$$

Hence, $\sup_{0\leq s\leq\tau}\{A(-s,0] - \beta(s)\} =$

$$= \max_{k\in\{0,\ldots,K-1\}}\{\sup_{t_k\leq s\leq t_{k+1}}\{A(-s,0] - \beta(s)\}\}$$
$$\leq \max_{k\in\{0,\ldots,K-1\}}\{A(-t_{k+1},0] - \beta(t_k)\}.$$

For brevity, let $E := \{\sup_{0\le s\le \tau}\{A(-s,0]-\beta(s)\} > q\}$, $q \ge 0$. By the inequality above, we obtain, for any $q \ge 0$,

$$\begin{aligned} \mathsf{P}(E) &\le \mathsf{P}(\max_{k\in\{0,\dots,K-1\}}\{A(-t_{k+1},0]-\beta(t_k)\} > q) \\ &= \mathsf{P}(\bigcup_{k\in\{0,\dots,K-1\}}\{A(-t_{k+1},0] > q+\beta(t_k)\}) \\ &\le \sum_{k=0}^{K-1}\mathsf{P}(A(-t_{k+1},0] > q+\beta(t_k)). \end{aligned}$$

Since the latter inequality holds for any partition $0 \le t_1 \le \dots \le t_K = \tau$, we obtain (11). This completes the proof. ■