1-1-2006

Statistical service bounds of regulated flows in a network

Jingwen Chen

Huazhong University of Science and Technology, jwchen@hust.edu.cn

Jiangtao Xi

University of Wollongong, jiangtao@uow.edu.au

Zongkai Yang

Huazhong University of Science and Technology

Follow this and additional works at: <https://ro.uow.edu.au/infopapers>

Part of the [Physical Sciences and Mathematics Commons](#)

Recommended Citation

Chen, Jingwen; Xi, Jiangtao; and Yang, Zongkai: Statistical service bounds of regulated flows in a network 2006.

<https://ro.uow.edu.au/infopapers/1799>

Statistical service bounds of regulated flows in a network

Abstract

Statistical service has been proposed for service differentiation networks to improve resource utilization. However, it has remained in a challenge to compute end-to-end statistical service bounds for aggregates of regulated flows in a network. In this paper, we develop a generalized statistical traffic envelope, global statistical envelope, which covers not only aggregated traffic of regulated flows, but also a large variety of traffic sources. Based on this characterization, we derive statistical bounds on delay and backlog in a service curve network. The general results are further applied to computing statistical delay bound of aggregated flows regulated by peak rate constrained leaky buckets in a network of rate-latency servers. The effectiveness of our theoretical results as verified by numerical evaluation in terms of providing significantly tight delay bounds.

Keywords

Statistical, service, bounds, regulated, flows, network

Disciplines

Physical Sciences and Mathematics

Publication Details

Chen, J., Xi, J. & Yang, Z. (2006). Statistical service bounds of regulated flows in a network. Proceedings of First International Conference on Communications and Networking in China (ChinaCom2006) USA: IEEE.

Statistical Service Bounds of Regulated Flows in a Network

Jingwen Chen*, Jiangtao Xi†, Zongkai Yang*

*Department of Electronics and Information Engineering
Huazhong University of Science and Technology, Wuhan, Hubei, China 430074

†School of Electrical, Computer and Telecommunications Engineering
University of Wollongong, Wollongong, NSW 2522, Australia

Abstract—Statistical service has been proposed for service differentiation networks to improve resource utilization. However, it has remained in a challenge to compute end-to-end statistical service bounds for aggregates of regulated flows in a network. In this paper, we develop a generalized statistical traffic envelope, Global Statistical Envelope, which covers not only aggregated traffic of regulated flows, but also a large variety of traffic sources. Based on this characterization, we derive statistical bounds on delay and backlog in a service curve network. The general results are further applied to computing statistical delay bound of aggregated flows regulated by peak rate constrained leaky buckets in a network of rate-latency servers. The effectiveness of our theoretical results is verified by numerical evaluation in terms of providing significantly tight delay bounds.

I. INTRODUCTION

With the explosive growth of multimedia applications and virtual private network services, quality of service provisioning is still a crucial and challenging issue. For service assurance networks, it's a common case that multiple flows are regulated at the ingress node, and then served uniformly as a whole, e.g., Differentiated Services (DiffServ) networks [1]. Also the service measures, e.g., delay and backlog, can be evaluated based on either deterministic or statistical model. Statistical model allows a small fraction of traffic to violate the service bounds, which results in significant improvement in terms of resource utilization upon deterministic model [2] [3].

For the statistical analysis of service in a network, the development of statistical network calculus [4] has attracted great attentions in recent years. An alternative approach is to use queueing theory, which can offers some accurate results. But it's difficult either to abstract various schedulers or to solve end-to-end service bounds. Moreover, the theoretical results lack of generality since they are associated with specific traffic source and node service. In contrast, statistical network calculus has been shown effective to provide general and tight results of end-to-end statistical service bounds with various traffic sources and schedulers. This is implemented mainly by extending the basic notions of network calculus [5], *Arrival Curve* and also sometimes *Service Curve*, to a probabilistic setting. However, existing studies in the context of statistical network calculus are inefficient or inapplicable to computing statistical service bounds of an aggregate of regulated flows.

Some studies (see [6] and references therein) are based

on the notion of arrival curve rather than its probabilistic version. In [6], the authors computed probabilistic backlog bound at a Packet Scale Rate Guarantee (PSRG) [7] node. Statistical delay bounds at each node were solved by the delay-from-backlog property [7] of *PSRG* model, and then summed up as the end-to-end one. This approach normally leads to overestimation of end-to-end delay bound [8], and is available for *PSRG* node rather than more general service curve ones.

Unlike the work in [6], most of studies are based on various types of statistical traffic envelope, the extension of arrival curve model (i.e., deterministic traffic envelope). A statistical traffic envelope bounds in the probabilistic sense the amount of flow traffic within arbitrary time duration. Let $A(t, t + \tau)$ denote the accumulated traffic of a flow during time period $[t, t + \tau]$. A statistical traffic envelope $\mathcal{G}^e(\tau)$ of this flow indicates $Pr\{A(t, t + \tau) \leq \mathcal{G}^e(\tau)\} \geq 1 - \varepsilon$ for any $t, \tau \geq 0$, where ε is the violation probability with $\mathcal{G}^e(\tau)$. One class of statistical traffic envelopes [8] [9] [10] [11] [12] [13], referred to as *Explicit Burstiness based (EB)* envelopes below, can be expressed as $\mathcal{G}^e(\tau) = \alpha(\tau) + \sigma$. Here $\alpha(\tau)$ is the base function, the burstiness σ is arbitrary, and ε is determined only by the burstiness σ , i.e., $\varepsilon = f(\sigma)$. In contrast, another class of statistical traffic envelopes [3], referred to as *General* envelope below, don't have the above limits to formulation with *EB* ones.

EB envelopes have been shown effective for both the characterization of a variety of traffic sources and the computation of statistical service bounds. However, *EB* envelopes cannot be used for an aggregate of regulated flows. If the regulators are peak rate constrained leaky buckets, an *EB* envelope cannot express the aggregated traffic of flows since its definition includes no factor of peak rate. If the regulators are alternatively leaky buckets, there had been no answer for an *EB* envelope to express the aggregated traffic, until it was mentioned in [8] to use *Exponentially Bounded Burstiness (EBB)* [13]. To our best knowledge, such statement is supported only by Eqn. (60) in [14]. Because the envelope function given by that formula is concave, it can be certainly replaced with an *EB* form. But it's impossible either to prove that ε decays exponentially with σ (i.e., *EBB*) or obtain $\varepsilon = f(\sigma)$ and $\alpha(\tau)$, which are needed for computing statistical service bounds. Thus, in the case of regulators being of leaky bucket type, it is also infeasible to apply an *EB* envelope for statistical service analysis.

The only two *General* envelopes are presented in [3] originally to express aggregated traffic of regulated flows. When applying them to computing statistical service bounds, one needs to assume Gaussian traffic process [3], or to estimate the bound on busy period [3] [14] [15]. The former doesn't hold in general, and the latter is normally intractable in a realistic network [5]. Thus, these two *General* envelopes are actually restricted in the application.

As is introduced above, it remains in a challenge to compute end-to-end statistical service bounds in the common case of traffic source being an aggregate of regulated flows. In this paper, we propose a generalized statistical traffic envelope, i.e., *Global Statistical Envelope*. Its is available for the characterization of not only aggregated traffic of regulated flows, but also a large variety of traffic sources that existing studies have done. Based on the notion of global statistical envelope, we derive statistical bounds on delay and backlog in a service curve [5] system, which is commonly used to abstract either a single node or a network. These general results are further applied to attaining statistical delay bound of an aggregate of regulated flows served by a network of rate-latency server [16], which has been proposed to model an Internet router.

The contribution of this paper consists in the generality of statistical traffic envelope, and the effectiveness of theoretical analysis in terms of providing tight end-to-end statistical service bounds. It is expected that our analytical model can be applied to statistical performance analysis in more realistic networks for efficient statistical service provisioning.

The remainder of this paper is organized as follows. In Section II, we present the notion of global statistical envelope. Section III offers general results of statistical bounds on backlog, delay and traffic departure, which are further applied in Section IV to attaining statistical delay bound of an aggregate of regulated flows in a rate-latency server network. The numerical experiments are stated in Section V. Section VI finally concludes the paper.

II. STATISTICAL TRAFFIC ENVELOPE OF A FLOW

In this section, we will present the notion of global statistical envelope, and an instance for an aggregate of regulated flows.

A. Definition

We use fluid traffic processes with continuous time model. Let $A(t)$ denote the accumulated traffic of a flow at time t , and $A(t_1, t_2)$ is particularly used for the one during the time period $[t_1, t_2]$.

For a flow $A(t)$, its *Global Statistical Envelope* (GSE in abbr.) $\mathcal{H}^\varepsilon(\tau)$ with violation probability ε is defined as a non-decreasing function of time interval τ that meets

$$\Pr\{\forall \tau \geq 0 : A(t, t + \tau) \leq \mathcal{H}^\varepsilon(\tau)\} \geq 1 - \varepsilon \quad (1)$$

for any $t \geq 0$, and $\mathcal{H}^\varepsilon(\tau) = 0$ for any $\tau < 0$.

The definition of GSE allows any non-decreasing function probabilistically bounding the traffic, and the associated violation probability being a function of any parameters in the envelope function. Hence, GSE is a general type of statistical

traffic envelope. Moreover, since the event in Eqn. (1) is formulated as holding for all $\tau \geq 0$, a GSE actually indicates a probabilistic upper bound on all sample paths.

B. Aggregate of Regulated Flows

To obtain the GSE of an aggregate of regulated flows, we need to use its global effective envelope (referred to as *GEE* below) [3] and deterministic traffic envelope. For a flow $A(t)$, a GEE $\mathcal{H}^{l,\varepsilon}(\tau)$ is defined as a function of τ that meets $\Pr\{\forall \tau \in [0, l] : A(t, t + \tau) \leq \mathcal{H}^{l,\varepsilon}(\tau)\} \geq 1 - \varepsilon$ for any $t \geq 0$, where $l > 0$ is constant. Given a GEE and a deterministic envelope of a flow, the following lemma shows how to construct a GSE.

Lemma 1: If a flow with GEE $\mathcal{H}^{l,\varepsilon}(\tau)$ also has a deterministic envelope $A^*(\tau)$, and $\mathcal{H}^{l,\varepsilon}(\tau) \geq A^*(\tau)$ at $\tau = l$, then

$$\mathcal{H}^\varepsilon(\tau) = \begin{cases} \min\{\mathcal{H}^{l,\varepsilon}(\tau), A^*(\tau)\}, & \tau \leq l \\ A^*(\tau), & \tau > l \end{cases} \quad (2)$$

is a GSE of this flow with violation probability ε .

Proof: For any t , divide $[t, \infty)$ into two non-overlapped parts T_s and T_d , in which $\mathcal{H}^\varepsilon(\tau)$ is taken respectively as $\mathcal{H}^{l,\varepsilon}(\tau)$ and $A^*(\tau)$. It's obvious that, for any $t \geq 0$, $\Pr\{\exists \tau \in T_s : A(t, t + \tau) > \mathcal{H}^{l,\varepsilon}(\tau)\} \leq \varepsilon$, and $\Pr\{\exists \tau \in T_d : A(t, t + \tau) > A^*(\tau)\} = 0$. By Eqn. (2), we then have $\Pr\{\exists \tau \geq 0 : A(t, t + \tau) > \mathcal{H}^\varepsilon(\tau)\} = \Pr\{\exists \tau \in T_s : A(t, t + \tau) > \mathcal{H}^\varepsilon(\tau)\} + \Pr\{\exists \tau \in T_d : A(t, t + \tau) > A^*(\tau)\} \leq \varepsilon$, which finishes the proof. ■

Let C and N denote the set and the number of constituent flows within an aggregate. Assume that $A_j(t)$ are independent mutually and stationary, i.e., $\Pr[A_j(t, t + \tau) \leq x] = \Pr[A_j(t', t' + \tau) \leq x]$ for all $t, t' \geq 0$. For any $j \in C$, flow j is regulated to have a deterministic envelope $A_j^*(\tau)$, i.e., $A_j(t, t + \tau) \leq A_j^*(\tau)$ for any $t \geq 0$ and $\tau \geq 0$. For the type of regulator, we consider peak rate constrained leaky bucket. Regulated by a peak rate constrained leaky bucket with peak rate P_j , average rate ρ_j and burst size σ_j , flow j has a deterministic envelope as

$$A_j^*(\tau) = \min\{P_j\tau, \rho_j\tau + \sigma_j\}, \quad (3)$$

and the one of the flow aggregate can be obtained with

$$A_C^*(\tau) = \sum_{j \in C} \min\{P_j\tau, \rho_j\tau + \sigma_j\}. \quad (4)$$

We begin with computing the local effective envelope (referred to as *LEE* below) [3] of a flow aggregate based on Hoeffding bound [17]. For a flow $A(t)$, its LEE $\mathcal{G}^\varepsilon(\tau)$ is defined to satisfy $\Pr\{A(t, t + \tau) \leq \mathcal{G}^\varepsilon(\tau)\} \geq 1 - \varepsilon$ for any $t \geq 0$ and $\tau \geq 0$. Considering $A_j^*(\tau)$ and 0 respectively as the upper and lower bounds of independent traffic arrivals $A_j(t, t + \tau)$, by Hoeffding bound, we can readily obtain a LEE as $\mathcal{G}_C^{\varepsilon_g}(\tau) = \sqrt{\frac{-\log \varepsilon_g}{2} \sum_{j \in C} [A_j^*(\tau)]^2} + \sum_{j \in C} E[A_j(t, t + \tau)]$, where ε_g is the violation probability specified with $\mathcal{G}_C^{\varepsilon_g}(\tau)$. The stationarity of $A_j(t, t + \tau)$ and Eqn. (3) together indicate $\sum_{j \in C} E[A_j(t, t + \tau)] \leq \sum_{j \in C} \rho_j\tau$. Then, the above expression can be detailed with

$$\mathcal{G}_C^{\varepsilon_g}(\tau) = \sqrt{\frac{-\log \varepsilon_g}{2} \sum_{j \in C} [\min(P_j\tau, \rho_j\tau + \sigma_j)]^2} + \sum_{j \in C} \rho_j\tau. \quad (5)$$

For simplicity, we consider only homogeneously regulated flows, i.e., $A_j^* = \min(P_0\tau, \rho_0\tau + \sigma_0)$. In this case, Eqn. (5) can be simplified with

$$\mathcal{G}_C^{\varepsilon_g}(\tau) = \min\{P_e\tau, \rho_e\tau + \sigma_e\}, \quad (6)$$

where $P_e = N'P_0 + N\rho_0$, $\rho_e = (N + N')\rho_0$, $\sigma_e = N'\sigma_0$, and $N' = \sqrt{-(N \log \varepsilon_g)/2}$ in detail.

Then we will apply Lemma 1 in [15] to constructing a GEE from a LEE. That is, given a LEE $\mathcal{G}^{\varepsilon_g}$ of a flow, its GSE can be given by $\mathcal{H}^{1,\varepsilon}(\tau) = \mathcal{G}^{\varepsilon_g}(\gamma\tau + a)$, and

$$\varepsilon = \varepsilon_g \cdot \frac{l(\sqrt{\gamma} + 1)}{a(\sqrt{\gamma} - 1)}, \quad (7)$$

where $\gamma > 1$ and $a \in (0, l)$ are both arbitrary.

Referring to Eqn. (6), a GEE can be given by

$$\mathcal{H}_C^{1,\varepsilon}(\tau) = \min\{P_h\tau + M_h, \rho_h\tau + \sigma_h\},$$

where $P_h = \gamma P_e$, $M_h = P_e a$, $\rho_h = \gamma \rho_e$, and $\sigma_h = \rho_e a + \sigma_e$.

Finally, by Eqn. (2) and $\mathcal{H}_C^{1,\varepsilon}(\tau)$ obtained, a GSE of an aggregate of homogeneously regulated flows can be given by

$$\mathcal{H}_C^{\varepsilon}(\tau) = \min\{P_h\tau + M_h, \rho_h\tau + \sigma_h, N\rho_0\tau + N\sigma_0\}. \quad (8)$$

Note that, to apply Lemma 1 to obtaining a GSE, it's needed to make sure $\mathcal{H}_C^{\varepsilon,l}(l) \geq A_C^*(l)$. Eqn. (7) and the expressions of P_e , ρ_e and σ_e imply that, given ε , the smallest l leads to the lowest $\mathcal{G}_C^{\varepsilon_g}$ and $\mathcal{H}_C^{\varepsilon}$. Hence, for a fixed ε , l is chosen such that

$$\mathcal{G}_C^{\varepsilon_g}(\gamma l + a) = A_C^*(l). \quad (9)$$

C. Discussion

Being *General* envelopes, GSE, LEE and GEE are all available for bounding probabilistically the aggregated traffic of regulated flows. However, when they are used to solve statistical bounds on service measures, next section will show that GSE needs neither the assumption on Gaussian traffic process (as LEE) nor the estimation of busy period (as GEE). Consequently, GSE can be applied to statistical service analysis in more realistic networks.

The formulation of GSE covers the existing *EB* envelopes. The ones presented in [9] [10] [11] are similar to GSE and GEE in terms of being defined from sample path point of view. For any choice of σ , each of them always corresponds to a GSE, and the reverse conversion exists only when $\mathcal{H}^{\varepsilon}(\tau)$ is specialized with *EB* type. In contrast, other *EB* envelopes [8] [12] [13] hold merely for single τ . Theorem 2 in [10] has shown how to convert these pointwise envelopes to the one of sample path view, so is a GSE.

When flow traffic can be expressed by an *EB* envelope, the formulations of sample path view impose the least constraint on the violation probability function $\varepsilon = f(\sigma)$. That is, $f(x)$ needs merely to be a tail distribution [9] [10], e.g., a power function as $x^{-\alpha}$, or even the ones decaying slowly. This feature makes GSE and the *EB* envelopes in [9] [10] [11] be available for characterizing non-Gaussian heavy-tailed traffic processes in realistic networks, e.g., α -stable self-similar processes.

III. STATISTICAL SERVICE BOUNDS

Based on the notion of GSE, we now derive the statistical bounds on two service measures and the output traffic of a flow in a service curve system. Let $A(t)$ and $D(t)$ denote the traffic arrival and departure of a flow at time t , $A(t_1, t_2)$ and $D(t_1, t_2)$ are particularly used for the ones during the time period $[t_1, t_2]$. By convention, $A(t) = D(t) = 0$ for any $t \leq 0$, $A(t_2, t_1) = -A(t_1, t_2)$, and $D(t_2, t_1) = -D(t_1, t_2)$. One measure is virtual delay $W(t)$, the delay experienced by a bit which arrives at time t . For a lossless system, $W(t) = \inf\{d \geq 0 : A(t) \leq D(t + d)\}$. Another one is backlog at time t , i.e., $B(t) = A(t) - D(t)$.

For a flow $A(t)$, the system is said to serve this flow with a service curve $S(t)$ if and only if its traffic departure $D(t) \geq A(t) \otimes S(t)$. Here $\otimes$ means min-plus convolution [5], i.e., $f(t) \otimes g(t) = \inf_{u \in [0, t]} [f(u) + g(t - u)]$. The concatenation of multiple service curve nodes is also a service curve system. Hence, the results in this section are available for statistical service analysis both at a single node and in a network.

Before the derivation of statistical service bounds, we first present a lemma as below.

Lemma 2: Given an arrival process $A(t)$ and its global statistical envelope $\mathcal{H}^{\varepsilon}(\tau)$, the following probabilistic inequality holds for any nonnegative function $g(t)$ and any $t \geq 0$.

$$Pr\{\forall \tau \geq 0 : A(t + \tau) - A(t) \otimes g(t) \leq \mathcal{H}^{\varepsilon}(\tau) \otimes g(\tau)\} \geq 1 - \varepsilon \quad (10)$$

Here $\otimes$ means min-plus de-convolution [5], i.e., $f(t) \otimes g(t) = \sup_{u \geq 0} [f(t + u) - g(u)]$.

Proof: For any $t \geq 0$,

$$\begin{aligned} & Pr\{\forall \tau \geq 0 : A(t + \tau) - A(t) \otimes g(t) \leq \mathcal{H}^{\varepsilon}(\tau) \otimes g(\tau)\} \\ & \geq Pr\{\forall \tau \geq 0 : \sup_{u \in [0, t]} [A(t - u, t + \tau) - g(u)] \\ & \leq \sup_{u \in [0, t]} [\mathcal{H}^{\varepsilon}(u + \tau) - g(u)]\} \\ & \geq Pr\{\forall \tau \geq 0, u \in [0, t] : A(t - u, t + \tau) - g(u) \\ & \leq \mathcal{H}^{\varepsilon}(u + \tau) - g(u)\} \\ & \geq 1 - \varepsilon. \end{aligned} \quad (11)$$

Eqn. (11) holds since $\{\forall \tau \geq 0, u \in [0, t] : A(t - u, t + \tau) - g(u) \leq \mathcal{H}^{\varepsilon}(u + \tau) - g(u)\} \subseteq \{\forall \tau \geq 0 : \sup_{u \in [0, t]} [A(t - u, t + \tau) - g(u)] \leq \sup_{u \in [0, t]} [\mathcal{H}^{\varepsilon}(u + \tau) - g(u)]\}$, and Eqn. (12) results from the definition of GSE. ■

Replacing $g(t)$ in (10) with a service curve $S(t)$ shows that

$$Pr\{\forall \tau \geq 0 : A(t + \tau) - A(t) \otimes S(t) \leq \mathcal{H}^{\varepsilon}(\tau) \otimes S(\tau)\} \geq 1 - \varepsilon. \quad (13)$$

The statistical bounds on virtual delay, backlog and traffic departure of a flow served by a service curve system are provided in the following theorem.

Theorem 1 (Statistical Service Bounds): Consider a flow with GSE $\mathcal{H}^{\varepsilon}(\tau)$ (the maximum violation probability is ε), served by a system with service curve $S(t)$. For any $t \geq 0$, there exists the following statistical bounds on virtual delay $W(t)$, backlog $B(t)$ and traffic departure $D(t)$ of this flow.

$$d_s = \inf\{d \geq 0 \mid \forall t \geq 0 : \mathcal{H}^{\varepsilon}(t - d) \leq S(t)\}, \quad (14)$$

$$b_s = (\mathcal{H}^e \otimes S)(0), \quad (15)$$

$$\mathcal{H}_d^e(\tau) = \mathcal{H}^e(\tau) \otimes S(\tau), \quad (16)$$

respectively hold in the senses that $\Pr\{W(t) \leq d_s\} \geq 1 - \varepsilon$, $\Pr\{B(t) \leq b_s\} \geq 1 - \varepsilon$, $\Pr\{\forall \tau \geq 0 : D(t, t + \tau) \leq \mathcal{H}_d^e(\tau)\} \geq 1 - \varepsilon$.

Proof: For any $t \geq 0$, there exists

$$\begin{aligned} & \Pr\{D(t) \geq A(t - d_s)\} \\ & \geq \Pr\{A(t) \otimes S(t) \geq A(t - d_s)\} \end{aligned} \quad (17)$$

$$\geq \Pr\{A(t) \otimes \mathcal{H}^e(t - d_s) \geq A(t - d_s)\} \quad (18)$$

$$\begin{aligned} & = \Pr\left\{\inf_{u \in [0, t]} [A(u) + \mathcal{H}^e(t - d_s - u)] \geq A(t - d_s)\right\} \\ & \geq \Pr\left\{\inf_{u \in [0, t]} [\mathcal{H}^e(t - d_s - u) - A(u, t - d_s)] \geq 0\right\} \end{aligned} \quad (19)$$

$$\geq 1 - \varepsilon. \quad (20)$$

Note that, in the above formulas, Eqn. (17) holds as $D(t) \geq A(t) \otimes S(t)$ implies $\{A(t) \otimes S(t) \geq A(t - d_s)\} \subseteq \{D(t) \geq A(t - d_s)\}$, Eqn. (18) is true since Eqn. (14) indicates $S(t) \geq \mathcal{H}^e(t - d_s)$, Eqn. (19) holds because $A(t - d_s)$ isn't concerned with the inf computation over u , and Eqn. (20) comes from the definition of GSE. This finishes the proof of statistical delay bound d_s expressed by Eqn. (14).

Let $\tau = 0$ in Eqn. (13). That is, for any $t \geq 0$,

$$\Pr\{A(t) - A(t) \otimes S(t) \leq (\mathcal{H}^e \otimes S)(0)\} \geq 1 - \varepsilon. \quad (21)$$

The definition of service curve implies $B(t) = A(t) - D(t) \leq A(t) - A(t) \otimes S(t)$, which together with Eqn. (21) completes the proof of Eqn. (15).

Note that $D(t + \tau) \leq A(t + \tau)$ and $D(t) \geq A(t) \otimes S(t)$. Then, for any $\tau \geq 0$, $D(t, t + \tau) = D(t + \tau) - D(t) \leq A(t + \tau) - A(t) \otimes S(t)$. Referring to Eqn. (13) completes the proof of Eqn. (16). ■

Given the GSE of flow arrival, Theorem 1 provides general results of statistical bounds on delay, backlog and traffic departure of a flow in a service curve system. Moreover, Theorem 1 also provides uniform expressions for both statistical and deterministic service bounds. In fact, $\varepsilon = 0$ in Theorem 1 implies Theorem 1.4.1, 1.4.2 and 1.4.3 in [5], the basic conclusions of network calculus.

IV. AGGREGATE OF REGULATED FLOWS IN A RATE-LATENCY SERVER NETWORK

In this section, we will go further to apply the general results provided by Theorem 1 to computing end-to-end statistical delay bound of an aggregate of regulated flows in a network of rate-latency servers [16]. The service curve of a rate-latency server with service rate R and latency T can be expressed with

$$S(t) = R(t - T)^+, \quad (22)$$

where the upper label $+$ means $m^+ = \max(m, 0)$. By virtue of generality and simplicity, rate-latency server has been suggested as the standard model for an Internet router. A concatenation of multiple rate-latency servers is also of the same type, whose service rate and latency can be readily obtained by the ones of each node.

The rate-latency server is fed with an aggregate of independent flows, which have been homogeneously regulated by

peak rate constrained leaky buckets. The GSE of such an aggregate has been provided by Eqn. (8). To compute end-to-end statistical delay bound, one can consider the equivalent rate-latency server of multiple nodes. This implies that the assumptions on the independence and regulation are needed only at the ingress node. Therefore, the above assumptions do not restrict the results being applied in a multi-node scenario.

Theorem 1 indicates that statistical delay bound with violation probability ε can be given by the maximum horizontal deviation between the two curves $\mathcal{H}_C^e(t) - t$ and $S(t) - t$. $\mathcal{H}_C^e(t)$ shown in Eqn. (8) is concave and piecewise linear. It has three turning points at 0, $\tau_h = \frac{\tau_0 - a}{\gamma}$, and l , where $\mathcal{H}_C^e(\tau) = M_h$, $P_h \tau_h + M_h$ (equal to $P_e \tau_0$) and $N(\rho_0 l + \sigma_0)$ respectively. $S_C(t)$ given by Eqn. (22) is convex and piecewise linear, and has a turning point at T . Thus, a statistical delay bound of the flow aggregate can be offered by

$$d_s = T + \frac{\max\{M_h, (P_e \tau_0 - R \tau_h)^+, (N \rho_0 l + N \sigma_0 - R l)^+\}}{R}, \quad (23)$$

and the violation probability is ε , the same as $\mathcal{H}_C^e(\tau)$. The max operator in Eqn. (23) indicates that d_s could be $d_{s0} = T + M_h/R$, $d_{s1} = T + (P_e \tau_0 - R \tau_h)^+/R$, or $d_{s2} = T + (N \rho_0 l + N \sigma_0 - R l)^+/R$, respectively corresponding to $P_h \leq R$, $\rho_h \leq R$ and $\rho_h > R$.

Eqn. (8) and (9) show that there exist infinite number of combinations $(\varepsilon_g, \gamma, a, l)$ corresponding to each ε required. Accordingly, the solution to $\mathcal{H}_C^e(\tau)$ is not unique. Because the quality of d_s provided by Eqn. (23) depends on ε_g , γ , a and l , we need to optimize d_s over ε_g , γ , a and l .

The first step of optimization is to reduce the number of arbitrary parameters. For simplicity, let $z = \rho_e/N\rho_0$, and $f(z, \gamma) = \varepsilon e^{2N(z-1)^2} \frac{\sqrt{\gamma}-1}{\sqrt{\gamma}+1}$. It's obvious that $z > 1$, $N' = N(z-1)$, $\varepsilon_g = e^{-2N(z-1)^2}$. Combining Eqn. (7) with (9) shows that a and l can be expressed with γ and z as follows.

$$a = \frac{\sigma_0}{\rho_0} \cdot \frac{2 - z}{z + (z\gamma - 1)f(z, \gamma)} \quad (24)$$

$$l = \frac{\sigma_0}{\rho_0} \cdot \frac{2 - z}{(z\gamma - 1) + \frac{z}{f(z, \gamma)}} \quad (25)$$

Replacing a and l in Eqn. (23) shows there are now only two variables, γ and z , to be optimized over.

The second step is to compute the optimal d_{s0} , d_{s1} and d_{s2} , and determine their implications on the choice of γ and z . The expressions of P_h and ρ_h show that $P_h \leq R$, $\rho_h \leq R$ and $\rho_h > R$ respectively correspond to $\eta\gamma[(z-1)\frac{P_0}{\rho_0} + 1] \leq 1$, $\eta z\gamma \leq 1$ and $\eta z\gamma > 1$, where $\eta = N\rho_0/R$ means the utilization. Then the optimal d_{s0} , d_{s1} and d_{s2} can be detailed individually with

$$d_0 = \min\left\{\eta\gamma[(z-1)\frac{P_0}{\rho_0} + 1] \leq 1 : T + \frac{\sigma_0}{\rho_0} \cdot \frac{2 - z}{\gamma[z + (z\gamma - 1)f(z, \gamma)]}\right\}. \quad (26)$$

$$\begin{aligned} d_1 &= \min\left\{\eta z\gamma \leq 1 : T + \eta\tau_0\left[1 + \frac{P_0}{\rho_0}(z-1) - \frac{1}{\eta\gamma}\right] \right. \\ &\quad \left. + \frac{\sigma_0}{\rho_0} \cdot \frac{2 - z}{\gamma[z + (z\gamma - 1)f(z, \gamma)]}\right\}, \end{aligned} \quad (27)$$

$$d_2 = \min\{\eta z \gamma > 1 : T + \frac{N\sigma_0}{R} - (1-\eta)\frac{\sigma_0}{\rho_0} \cdot \frac{2-z}{(z\gamma-1) + \frac{z}{f(z,\gamma)}}\}. \quad (28)$$

Note that there are four basic constraints for the above optimizations, including (a) $\gamma > 1$, (b) $1 + \sqrt{-\log \varepsilon/(2N)} < z < 2$, (c) $a < \tau_0$, and (d) $l > \tau_0$. Among these constraints, (a) is necessary, (b) is required as $\varepsilon_g = e^{-2N(z-1)^2} > \varepsilon$ and $a, l > 0$, and (c) together with (d) assures that Eqn. (23) is effective.

The last step is to obtain the optimal d_s . Let $x = \sqrt{-(N \log \varepsilon)/2}$. By $\gamma > 1$ and $\varepsilon_g < \varepsilon$, it's readily to verify that, (1) if $x < \frac{(1-\eta)R}{P_0}$, $P_h \leq R$, $\rho_h \leq R$ and $\rho_h > R$ are all possible, (2) if $\frac{(1-\eta)R}{P_0} \leq x < \frac{(1-\eta)R}{\rho_0}$, either $\rho_h \leq R$ or $\rho_h > R$ holds, and $P_h \leq R$ is impossible, and (3) if $x \geq \frac{(1-\eta)R}{\rho_0}$, only $\rho_h > R$ is possible, neither $P_h \leq R$ nor $\rho_h \leq R$ holds any more. Considering $P_h \leq R$, $\rho_h \leq R$ and $\rho_h > R$ are the conditions of respectively taking d_{s0} , d_{s1} and d_{s2} as d_s , we certainly conclude that the optimal d_s can be formulated with

$$d_{\text{sopt}} = \begin{cases} \min\{d_0, d_1, d_2\}, & x < \frac{(1-\eta)R}{P_0} \\ \min\{d_1, d_2\}, & \frac{(1-\eta)R}{P_0} \leq x < \frac{(1-\eta)R}{\rho_0} \\ d_2, & x \geq \frac{(1-\eta)R}{\rho_0} \end{cases} \quad (29)$$

where d_0 , d_1 , and d_2 are shown in Eqn. (26), (27) and (28).

V. NUMERICAL EVALUATION

In this section, we will evaluate the statistical delay bound provided in Section IV by numerical experiments. There are two types of flow in the experiments. Type 1 is specified with $P_1 = 1.5\text{Mbps}$, $\rho_1 = 0.15\text{Mbps}$ and $\sigma_1 = 95.4\text{kbits}$, while type 2 is expressed with $P_2 = 6.0\text{Mbps}$, $\rho_2 = 0.15\text{Mbps}$ and $\sigma_2 = 10.345\text{kbits}$. Each rate-latency server node serves a flow aggregate with service rate $R = 50\text{Mbps}$ and latency $T = 0.08\text{ms}$. As for a concatenation of H nodes, the equivalent rate-latency server are expressed with $R = 50\text{Mbps}$ and latency $T = H \times 0.08\text{ms}$.

End-to-end statistical delay bounds are computed based on Eqn. (29), and simultaneously by the approach introduced in [6]. The former is labeled with *GSE* in the figures, while the latter is tagged with *SUMDfB* by the approach of *Sum of Delay-from-Backlog bounds*. In particular, *SUMDfB* curves are attained by Eqn. (5), (16), (17) and (18) in [6]. Here the delay jitter bound is taken as the deterministic delay bound from the ingress node to the concerned core node. All numerical results are normalized with respect to the deterministic delay bound d_d , which is computed based on Proposition 1.4.1 in [5].

Fig. 1 shows how d_s/d_d varies with violation probability ε , hop number H and utilization η . The factors which influence d_s/d_d include ε and flow type. Higher ε and flow type with larger burstiness both lead to smaller d_s/d_d , i.e., better improvement upon deterministic bounds. For example, type 1 is more bursty than type 2, so d_s/d_d with an aggregate of type 1 flows is obviously better than the one with type 2. Both Fig. 1(c) and 1(d) indicate that the increase of H nearly has no impact on d_s/d_d , which is natural due to the merit of service curve model. It's observed in Fig. 1(e) and 1(f) that higher η has merely limited influence on $\frac{d_s}{d_d}$ though both d_s and d_d simultaneously increase with η .

The comparison between *GSE* and *SUMDfB* curves shows that, at most of time, Eqn. (29) offers significantly tighter statistical delay bounds. This indicates that independence at each hop is a worst-case assumption for the distribution of end-to-end delay. The exceptions exist in the cases of single node and some high violation probabilities, in which the bounds by Eqn. (29) are slightly looser. Another observation from the comparison is that, the variances of violation probability, number of hops, and utilization have obviously smaller impacts on $\frac{d_s}{d_d}$ by Eqn. (29) than the ones based on the approach in [6].

VI. CONCLUSION

In this paper, we compute statistical service bounds in a network fed with an aggregate of regulated flows. Based on the notion of global statistical envelope, statistical bounds on delay, backlog and traffic departure can be readily derived, as network calculus has done in the deterministic sense. Although the notion of global statistical envelope is presented originally to characterize the aggregated traffic of regulated flows, it actually covers a large variety of traffic source that existing studies have done. Thus our theoretical results are general rather than merely for an aggregate of regulated flows. The effectiveness of our theoretical results is verified by numerical experiments. The numerical results confirm that sample path view of stochastic flow traffic leads to not only concise forms of formulation, but also tight statistical performance bounds.

In fact, this paper considers only statistical multiplexing gain of aggregated flows, but no statistical resource sharing among differentiated flow aggregates. The latter is the aim of our ongoing study to relax some constraints in related works. Moreover, in our future work, we are intended to carry out simulations with more realistic traffic and practical scheduling disciplines to evaluate the theoretical results in this paper.

REFERENCES

- [1] S. Blake, F. Baker, and D. Black, "An architecture for differentiated service," RFC2475, June 1998.
- [2] D. Wrege, E. W. Knightly, H. Zhang, and J. Liebeherr, "Deterministic delay bounds for vbr video in packet-switching networks: Fundamental limits and practical tradeoffs," *IEEE/ACM Trans. Networking*, vol. 4, pp. 352–362, June 1996.
- [3] R. Boorstyn, A. Burchard, J. Liebeherr, and C. Oottamakorn, "Statistical services assurances for traffic scheduling algorithms," *IEEE J. Select. Areas Commun.*, vol. 18, pp. 2651–2664, Dec. 2000.
- [4] J. Liebeherr, "A statistical network calculus for computer networks," Feb. 2005. [Online]. Available: <http://www.cs.virginia.edu/~jorg/archive/talks/statcalc-Feb05-vt.ppt>
- [5] J. Y. L. Boudec and P. Thiran, *Network Calculus*. Springer-Verlag.
- [6] M. Vojnovic and J. L. Boudec, "Stochastic analysis of some expedited forwarding networks," in *Proc. IEEE INFOCOM 2002*, New York, June 2002.
- [7] J. C. R. Bennett, K. Benson, A. Charny, W. F. Courtney, and J. L. Boudec, "Delay jitter bounds and packet scale rate guarantee for expedited forwarding," *IEEE/ACM Trans. Networking*, vol. 10, p. 10, Aug. 2002.
- [8] F. Ciucu, A. Burchard, and J. Liebeherr, "A network service curve approach for the stochastic analysis of networks," in *Proc. ACM Sigmetrics'05*, Alberta, Canada, June 2005.
- [9] Y. Jiang and P. J. Emsted, "Analysis of stochastic service guarantees in communication networks: A traffic model," in *Proc. 19th International Teletraffic Congress (ITC19)*, Aug. 2005.

Fig. 1. Statistical delay bound, violation probability, number of hops, and utilization: (a) flow type 1, $H = 5$, $\eta = 0.2, 0.8$; (b) the same as (a) excepts with flow type 2; (c) flow type 1, $\varepsilon = 10^{-9}$, $\eta = 0.2, 0.8$; (d) the same as (c) excepts with flow type 2; (e) flow type 1, $H = 5$, $\varepsilon = 10^{-9}, 10^{-6}$; (f) the same as (e) excepts with flow type 2

- [10] Q. Yin, Y. Jiang, and P. Y. Kong, "Analysis on generalized stochastically bounded bursty traffic for communication networks," in *Proc. IEEE Local Computer Networks 2002*, Nov. 2002.
- [11] S. Ayyorgun and W. Feng, "A probabilistic definition of burstiness characterization: A system approach," Los Alamos National Laboratory, Tech. Rep. LA-UR-03-3668, May 2003.
- [12] D. Starobinski and M. Sidi, "Stochastically bounded burstiness for communication networks," *IEEE Trans. Inform. Theory*, vol. 46, pp. 206–212, Jan. 2000.
- [13] O. Yaron and M. Sidi, "Performance and stability of communication networks via robust exponential bounds," *IEEE/ACM Trans. Networking*, vol. 1, pp. 372–385, Mar. 1993.
- [14] C. Li, A. Burchard, and J. Liebeherr, "A calculus with effective bandwidth," Department of Computer Science, University of Virginia, Tech. Rep. CS-2003-20, Nov. 2003.
- [15] J. Liebeherr, S. D. Patek, and A. Burchard, "Statistical per-flow service bounds in a network with aggregate provisioning," in *Proc. IEEE INFOCOM 2003*, San Francisco, Mar. 2003.
- [16] D. Stiliadis and A. Varma, "Latency-rate servers: A general model for analysis of traffic scheduling algorithms," *IEEE/ACM Trans. Networking*, vol. 6, pp. 611–624, Oct. 1998.
- [17] W. Hoeffding, "Probability inequalities for sums of bounded random variables," *American Statistical Association Journal*, pp. 13–30, Mar. 1963.