A Framework for Providing Stochastic Delay Guarantees in Communication Networks

Massieh Kordi Boroujeny and Brian L. Mark
Dept. of Electrical and Computer Engineering
George Mason University, Fairfax, VA, U.S.A.

***Abstract*—The provisioning of delay guarantees in packet-switched networks such as the Internet remains an important, yet challenging open problem. We propose and evaluate a framework, based on results from stochastic network calculus, for guaranteeing stochastic bounds on network delay at a statistical multiplexer. The framework consists of phase-type traffic bounds and moment generating function traffic envelopes, stochastic traffic regulators to enforce the traffic bounds, and an admission control scheme to ensure that a stochastic delay bound is maintained for a given set of flows. Through numerical examples, we show that a stochastic delay bound is maintained at the multiplexer, and contrast the proposed framework to an approach based on deterministic network calculus.**

***Index Terms*—Quality-of-service, delay guarantee, network calculus, traffic regulator, admission control.**

I. Introduction

Currently, the Internet does not provide end-to-end delay guarantees for traffic flows. Even if the path taken by a given traffic flow is fixed, e.g., via mechanisms such as software-defined networking (SDN) or multi-protocol label switching (MPLS), network congestion arising from other flows can result in highly variable delays. The variability and random nature of traffic flows in a packet-switched network make it very challenging to provide performance guarantees.

The standard approach to providing network performance guarantees consists of two basic elements:

1) *Admission control:* A new flow is admitted to the network only if sufficient resources are available to maintain a given performance guarantee.
2) *Traffic regulation:* Each traffic flow must be regulated to ensure that it does not use more resource than what was negotiated by the admission control scheme.

Admission control is challenging due to the random and bursty nature of traffic flows, which makes them difficult to characterize and regulate. Even when flows are modeled as random arrival processes, provisioning for end-to-end performance guarantees in a multi-hop network is generally intractable.

In his seminal work, Cruz [1], [2] proposed the so-called (σ, ρ) characterization of traffic, which imposes a deterministic bound on the burstiness of a traffic flow. By bounding traffic flows according to (σ, ρ) parameters, Cruz developed a network calculus which determined how these parameters propagate through network elements, from which end-to-end delay bounds could be derived. An important feature of the (σ, ρ) characterization is that it could be enforced by a traffic regulator. In practice, however, the deterministic (σ, ρ) characterization leads to very loose end-to-end delay bounds, which leads to very low utilization of the network resources. Nevertheless, the (σ, ρ) characterization was the basis for further research into stochastic bounds on traffic burstiness and stochastic network calculus to provide tighter, probabilistic end-to-end delay guarantees. Stochastic network calculus and associated performance bounds remains an active topic of research, with ongoing efforts aimed at improving the tightness of the stochastic delay bounds. To our knowledge, however, stochastic network calculus has not previously been applied within a practical framework to provide performance guarantees.

We present a practical framework for providing performance guarantees based on concepts from stochastic network calculus. Traffic flows are characterized by the phase-type traffic traffic descriptor proposed in [3] as well as a moment generating function (MGF) traffic envelope [4]. The phase-type traffic bound for each traffic flow is enforced by a stochastic traffic regulator, see [5]. An admission control scheme decides whether or not to admit a new traffic flow on the basis of both the phase-type descriptor and the MGF envelope to guarantee a stochastic delay bound for all admitted traffic flows. The main contribution of this paper is to demonstrate that stochastic delay guarantees can be achieved for admitted flows while maintaining relatively high traffic utilization in the network.

The remainder of the paper is organized as follows. In Section II, we discuss the phase-type traffic bound and its use as a traffic descriptor, as well as the MGF traffic envelope. In Section III, we discuss a scheme to enforce both a phase-type bound and an MGF envelope for a traffic process. In Section IV, we discuss an admission control scheme for a statistical multiplexer based on the phase-type bounds and results from stochastic network calculus. Numerical results, which demonstrate the proposed framework are presented in Section V. Concluding remarks are given in Section VI.

II. Stochastic Traffic Bounds

Let $A = \{A(s,t) : 0 \leq s \leq t\}$ denote a traffic arrival process, where $A(s,t)$ denotes the amount of traffic arriving in time interval $[s,t)$. For simplicity, we shall assume that the time parameters s and t are discrete unless otherwise specified, but the results that follow also carry over to the continuous-time case. Our proposed framework involves two types of

This work was supported in part by the U.S. National Science Foundation under Grant No. 1717033.

bounds on a traffic process: a phase-type traffic bound and an MGF traffic envelope.

A. Phase-type Traffic Bound

We consider stochastic bounds on the *burstiness* of a traffic flow, with respect to an *upper rate* ρ, which is chosen to be larger than or equal to the long term average traffic rate, i.e., $\rho \geq \lim_{t\to\infty} \frac{A(0,t)}{t}$, The concept of phase-type bounded traffic is defined as follows [3].

Definition 1. A traffic process A is characterized by a *phase-type traffic descriptor* $[\rho; (a, \boldsymbol{\pi}, \mathbf{Q}, T)]$ if

$$\mathsf{P}\left\{W_\rho(t; A) \geq \sigma\right\} \leq a\boldsymbol{\pi} e^{\mathbf{Q}\sigma}\mathbf{1}, \tag{1}$$

for all $t \geq 0$ and all $\sigma \in (0, T]$. Here, $\mathbf{1}$ is a column vector of all ones, $a \geq 0$, $T > 0$. The virtual workload of a constant rate queue with service rate ρ and input traffic A is defined by

$$W_\rho(t; A) := \max_{0\leq s\leq t}[A(s,t) - \rho(t-s)], \tag{2}$$

and $(\boldsymbol{\pi}, \mathbf{Q})$ denotes the parameter of a phase-type distribution [6].

When $T = \infty$, the phase-type traffic bound is a particular case of generalized stochastically bounded burstiness (gSBB), which was developed in [7], [8]. In [3], it was shown that the phase-type bound defined above is closed with respect to a stochastic network calculus based on the gSBB concept.

The concept of gSBB is closely related to the Stochastically Bounded Burstiness (SBB) concept introduced in [9], which in turn is a generalization of Exponentially Bounded Burstiness (EBB) [10]. A key feature gSBB vs. SBB is that it is based on the workload process $W_\rho(t; A)$, which can be reasonably assumed to be stationary and ergodic, rather than the arrival process A, which is neither stationary nor ergodic. Consequently, as discussed in Section III, a stochastic traffic regulator can be designed based on enforcement of a time-average approximation of the left-hand side of (1).

The problem of finding a phase-type traffic descriptor to fit a given traffic trace can be formulated as a semi-infinitely constrained optimization problem [11], which can be solved numerically for special phase-type distributions such as the hyperexponential distribution. In particular, the hyperexponential distribution provides a tight phase-type bound for a large class of traffic flows. Given the procedure developed in [11], we shall assume that each traffic flow that requests admission to the network has an associated phase-type traffic descriptor.

B. MGF Traffic Envelope

An alternative approach to characterizing a traffic process is to bound the moment generating function (MGF)

$$M_A(\theta; s, t) := E\left[e^{\theta A(s,t)}\right]$$

where $\theta > 0$ is a free parameter [4].

Definition 2. The MGF traffic envelope of traffic process A is defined by

$$E\left[e^{\theta A(s,t)}\right] \leq e^{\theta(\hat{\rho}(\theta)(t-s)+\hat{\sigma}(\theta))}, \tag{3}$$

where the parameters $\hat{\rho}(\theta) > 0$ and $\hat{\sigma}(\theta) \geq 0$ are functions of $\theta > 0$.

The MGF traffic envelope is analogous to the deterministic (σ, ρ) characterization in that it involves analogous parameters $\hat{\sigma}(\theta)$ and $\hat{\rho}(\theta)$ and it can be related to the EBB characterization via the Chernoff bound.

The MGF traffic envelope, however, has some advantages compared to the phase-type bound traffic descriptor, the most important being the following,

Theorem 1 (Sum of MGF envelopes). When n independent flows $A_1, \ldots, A_n$, with MGF envelope parameters $(\hat{\sigma}_1, \hat{\rho}_1), \ldots (\hat{\sigma}_n, \hat{\rho}_n)$, respectively, are superposed, the aggregate traffic process $A = A_1 + \ldots + A_n$ can be characterized by the MGF parameter $(\hat{\sigma}, \hat{\rho})$, where $\hat{\sigma} = \sum_{i=1}^n \hat{\sigma}_i$ and $\hat{\rho} = \sum_{i=1}^n \rho_i$.

This property of the MGF traffic envelope not only simplifies the computations involved in admission control, but more importantly, it captures the effect of statistical multiplexing gain. For this reason, our proposed framework uses *both* the phase-type bound traffic descriptor and the MGF traffic envelope. The problem of finding a MGF traffic envelope can be simplified by defining a finite set Θ of values to consider for the free parameter θ in (3). Then a set of MGF envelope parameters, $\{(\hat{\sigma}(\theta), \hat{\rho}(\theta)) : \theta \in \Theta\}$, could be determined using an approach similar to the procedure in [11] for fitting the phase-type traffic descriptor.

III. Stochastic Traffic Regulation

Next, we discuss methods for enforcing both a phase-type bound traffic descriptor and the MGF traffic envelope.

A. (σ^*, ρ) Regulator

The deterministic (σ, ρ) regulator tends to provide a very loose bound on the traffic or to incur unnecessarily large delays on the traffic. To address these issues, a *stochastic* traffic regulator was proposed in [5], which enforces a probabilistic bound on a traffic process A:

$$\mathsf{P}\left\{W_\rho(t; A) \geq \gamma\right\} \leq f(\gamma), \quad \forall\gamma \in [0, T], \tag{4}$$

where $f(\gamma)$ is a non-increasing positive bounding function and T is a limit on the tail distribution of the workload. We refer to a regulator that enforces (4) as a stochastic (σ^*, ρ) regulator, where the burstiness parameter σ^* is variable.

Users specify their traffic flows with a descriptor $[\rho; (f(\gamma), T)]$ in terms of a bound of the form (4). In particular, for the phase-type bound the bounding function has the form $f(\gamma) = a\boldsymbol{\pi} e^{\mathbf{Q}\gamma}\mathbf{1}$ (cf. (1)). By applying results from stochastic network calculus, the admissibility of a given set of traffic flows with respect to a certain probabilistic end-to-end delay constraint can be determined. However, such an end-to-end delay guarantee can only be provided if the traffic flows conform to their negotiated traffic parameters. The (σ^*, ρ) regulator can be applied at the network edge to force compliance of each traffic flow to a negotiated phase-type bound parameter. Optionally, the regulator could be applied at internal nodes of the network to reshape traffic flows to their negotiated phase-type traffic bounds. This has the benefit of maintaining the negotiated traffic profile for each traffic flow over a multi-hop path, but requires the additional overhead of traffic regulation within the network.

B. MGF Traffic Envelope Regulator

According to Definition 2, the MGF envelope parameters $\hat{\rho}(\theta)$ and $\hat{\sigma}(\theta)$ satisfy (3). However, verification of (3), requires estimation of the MGF $E\left[e^{\theta A(s,t)}\right]$, which presents difficulties because the traffic process A is non-stationary and non-ergodic. Therefore, we introduce an alternative MGF envelope characterization.

Definition 3. The MGF workload envelope (or w-envelope) of traffic process A is defined by

$$E\left[e^{\theta W_{\hat{\rho}(\theta)}(t;A)}\right] \leq e^{\theta\hat{\sigma}(\theta)}, \tag{5}$$

where $\theta > 0$ is a free parameter, $\hat{\rho}(\theta) > 0$ and $\hat{\sigma}(\theta) \geq 0$, and $W_{\hat{\rho}}(A;t)$ is the workload defined in (2).

The MGF w-envelope provides an upper bound on the MGF traffic envelope in the following sense.

Theorem 2. Suppose a traffic process A has an MGF w-envelope with parameter $\{(\hat{\sigma}(\theta), \hat{\rho}(\theta)) : \theta \in \Theta\}$, i.e.,

$$E\left[e^{\theta W_{\hat{\rho}(\theta)}(t;A)}\right] \leq e^{\theta\hat{\sigma}(\theta)}, \qquad \theta \in \Theta. \tag{6}$$

Then it is also characterized by an MGF traffic envelope $\{(\hat{\sigma}(\theta), \hat{\rho}(\theta)) : \theta \in \Theta\}$, i.e.,

$$E\left[e^{\theta A(s,t)}\right] \leq e^{\theta[\sigma+\rho(t-s)]}, \qquad 0 \leq s \leq t, \;\; \theta \in \Theta. \tag{7}$$

Proof. For $0 \leq s \leq t$,

$$A(s,t) - \rho(t-s) \leq \max_{0\leq s\leq t}[A(s,t) - \rho(t-s)] = W_\rho(t;A).$$

Therefore,

$$E[e^{\theta[A(s,t)-\rho(t-s)]}] \leq E[e^{\theta W_\rho(t;A)}],$$

and the result follows immediately. □

Theorem 2 implies that a traffic regulator which enforces an MGF w-envelope with parameter $(\hat{\sigma}, \rho)$ also enforces an MGF traffic envelope with the same parameter. To enforce an MGF traffic envelope for a traffic process A, we can estimate the left-hand side of (6), for each value of $\theta \in \Theta$, via a time-average and regulate it to ensure that the inequality is maintained. This can be accomplished by designing a stochastic regulator along the lines of the (σ^*, ρ) regulator in [5].

IV. Admission Control

We develop an admission control scheme based on a stochastic delay bound derived from the phase-type traffic bound and MGF traffic envelope. The phase-type bound provides a tighter delay bound when a small to moderate number of flows is considered. When the number of flows becomes larger, the MGF envelope can yield a tighter bound due to the statistical multiplexing effect. Therefore, we propose a *hybrid* admission control scheme which uses both types of traffic bounds.

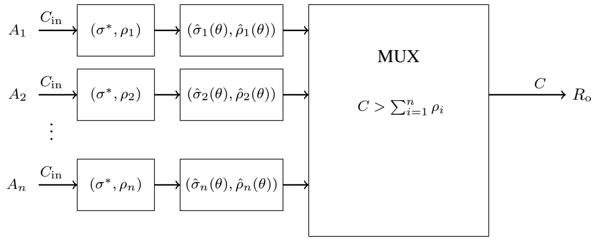
A_1
C_{in}
(σ^*, ρ_1)
$(\hat{\sigma}_1(\theta), \hat{\rho}_1(\theta))$
A_2
C_{in}
(σ^*, ρ_2)
$(\hat{\sigma}_2(\theta), \hat{\rho}_2(\theta))$
$\vdots$
A_n
C_{in}
(σ^*, ρ_n)
$(\hat{\sigma}_n(\theta), \hat{\rho}_n(\theta))$
MUX
$C > \sum_{i=1}^n \rho_i$
C
R_{o}

Fig. 1. Multiplexer with n independent traffic flows.

A. Admission Control via Phase-Type Bound

Consider a multiplexer of capacity C with a set of n independent traffic flows, $\mathcal{A} = \{1, \ldots, n\}$, as inputs characterized by phase-type traffic descriptors $\Delta_i = [\rho_i, (a_i, \boldsymbol{\pi}_i, \mathbf{Q}, T_i)]$, $i = 1, \ldots, n$. The essential task of the admission controller is to determine whether or not a stochastic delay bound of the following form can be satisfied:

$$\mathsf{P}\{D \geq d\} < \epsilon, \tag{8}$$

where D represents the delay experienced by a packet in the multiplexer, $\epsilon > 0$ is a small number, e.g., $\epsilon = 10^{-3}$, and d represents a "maximum" tolerable delay for a packet from any of the admitted flows. Clearly, a necessary condition for (8) to be satisfied is $\sum_{i=1}^n \rho_i < C$.

A phase-type traffic bound for the aggregate traffic input to the multiplexer can be determined by repeated application of the following theorem.

Theorem 3 (*Independent Sum*). Let A_1 and A_2 be independent traffic processes characterized by phase-type traffic descriptors $\Delta_1 = [\rho_1, (a, \boldsymbol{\alpha}, \mathbf{G}, T_1)]$, and $\Delta_2 = [\rho_2, (b, \boldsymbol{\beta}, \mathbf{H}, T_2)]$, respectively. The aggregate process $A = A_1 + A_2$ is bounded by the phase-type traffic descriptor $\Delta = [\rho, (c, \boldsymbol{\pi}, \mathbf{Q}, T)]$. where $\rho = \rho_1 + \rho_2$, $T = \min(T_1, T_2)$, $c = a + b - ab$,

$$\boldsymbol{\pi} = \left[\frac{a(1-b)}{c}\boldsymbol{\alpha}, \frac{b(1-a)}{c}\boldsymbol{\beta}, \frac{ab}{c}\boldsymbol{\alpha}, \mathbf{0}\right], \tag{9}$$

$$\mathbf{Q} = \begin{pmatrix} \mathbf{G} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{H} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{G} & \mathbf{g}\boldsymbol{\beta} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{H} \end{pmatrix},$$

where $\mathbf{g} = -\mathbf{G}\mathbf{1}$.

This theorem can be derived from [7, Theorem 4] for gSBB flows by applying properties of the phase-type distribution. In Theorem 3, if the number of phases represented by the phase-type traffic descriptors Δ_1 and Δ_2 equals m, then the number of phases in Δ will be $4m$. To avoid this expansion in the size of the phase-type traffic descriptor, we can apply the numerical method developed in [11] to determine a phase-type traffic descriptor $\tilde{\Delta}$, which approximates the true aggregate descriptor Δ, but only has m phases. Thus, we obtain a practical procedure for obtaining a phase-type traffic descriptor for the superposition of an arbitrary set of independent flows characterized by phase-type traffic descriptors.

Given the above procedure for determining a phase-type traffic descriptor for an aggregate traffic flow A at the input to a multiplexer, a relationship between ϵ and d in (8) can be derived from following theorem:

Theorem 4. Let A be a traffic process with phase-type descriptor $[\rho; (a, \boldsymbol{\pi}, \mathbf{Q}, T)]$ that is input to a FIFO system with constant transmission rate $C > \rho$. Then the steady-state delay D through the system can be bounded as follows:

$$\mathsf{P}\{D \geq d\} \leq a\boldsymbol{\pi} e^{(C - \frac{\rho}{d})\mathbf{Q}d}\mathbf{1}, \tag{10}$$

for all $t \geq 0$ and all $\frac{T+\rho}{C} \geq d \geq \frac{\rho}{C}$.

This theorem can be derived from [7, Theorem 8] for gSBB flows by applying properties of the phase-type distribution.

B. Admission Control via MGF Envelope

In conjunction with Theorem 1, the following result can be used to perform admission control based on MGF traffic envelopes [4].

Theorem 5. Suppose a traffic process A with MGF envelope $(\hat{\sigma}(\theta), \hat{\rho}(\theta))$, $\theta \in \Theta$, is offered as input to a constant rate server of capacity $C > \rho$. Then the steady-state system delay D can be bounded as follows:

$$P\{D \geq d\} \leq \frac{e^{\theta\hat{\sigma}(\theta)}}{1 - e^{-\theta(C - \hat{\rho}(\theta))}}\ e^{-\theta d}, \qquad \theta \in \Theta. \tag{11}$$

The parameter θ can be optimized to minimize the right-hand side of (11) .

C. Hybrid Admission Control Scheme

We proposed a hybrid admission control scheme that combines the phase-type traffic descriptor and MGF traffic envelope characterizations of the input traffic flows. The basic setup is depicted in Fig. 1. Each flow passes through a (σ^*, ρ) stochastic regulator (see Section III-A), which enforces a phase-type traffic descriptor negotiated between the network and the traffic flow. Similarly, an MGF traffic w-envelope for each flow is enforced by an MGF traffic regulator (see Section III-B).

Given a set of flows $\mathcal{A} = \{1, \ldots, n\}$, the hybrid admission control scheme checks two admission criteria with respect to the stochastic delay constraint (8):

1) Using the procedure based on the phase-type traffic descriptors outlined in Section IV-A, determine whether or not $\mathcal{A}$ is admissible.
2) Using the procedure based on MGF envelope parameters outlined in Section IV-B, determine whether or not $\mathcal{A}$ is admissible.

If $\mathcal{A}$ is admissible under criterion 1 *or* criterion 2, then $\mathcal{A}$ is considered admissible.

V. Numerical Study

In this section, we demonstrate key aspects of the proposed framework using traffic flows modeled as MMPPs and discrete-time Markov on-off fluid processes.

A. Markov Modulated Poisson Process

The MMPP is a popular continuous-time model for traffic flows possessing a high degree of burstiness [12]. An m-state MMPP is a doubly-stochastic Poisson point process $N(t)$ parameterized by a diagonal arrival matrix $\mathbf{\Lambda} = \text{diag}\{\lambda_1, \ldots, \lambda_m\}$, where $\lambda_i \geq 0$ is the Poisson arrival rate when the underlying Markov chain is in state i and a rate matrix $\mathbf{R} = [r_{ij}]$, $1 \leq i, j \leq m$, where $r_{ij} \geq 0$ is the departure rate of the Markov chain from state i to $j \neq i$. For $1 \leq i \leq m$, $r_{ii} < 0$ and $-r_{ii}$ is the departure rate of the Markov chain from state i. The rate matrix $\mathbf{R}$ is the generator matrix of the modulating Markov chain. For example, a 2-state MMPP is parameterized by arrival and rate matrices given, respectively, by

$$\mathbf{\Lambda} = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix}, \quad \mathbf{R} = \begin{bmatrix} -r_1 & r_1 \\ r_2 & -r_2 \end{bmatrix}. \tag{12}$$

The superposition of n independent MMPPs is again an MMPP. The rate matrix $\mathbf{R}$ and the arrival rate matrix $\mathbf{\Lambda}$ of the aggregated process are given, respectively, by

$$\mathbf{R} = \mathbf{R}_1 \oplus \ldots \oplus \mathbf{R}_n, \quad \mathbf{\Lambda} = \mathbf{\Lambda}_1 \oplus \ldots \oplus \mathbf{\Lambda}_n, \tag{13}$$

where $\oplus$ denotes the Kronecker-sum [12]. For example, the superposition of two independent, identically distributed 2-state MMPPs results in a 3-state MMPP with arrival matrix

$$\mathbf{\Lambda} = \begin{bmatrix} 2\lambda_2 & 0 & 0 \\ 0 & \lambda_1 + \lambda_2 & 0 \\ 0 & 0 & 2\lambda_1 \end{bmatrix} \tag{14}$$

and rate matrix

$$\mathbf{R} = \begin{bmatrix} -2r_2 & 2r_2 & 0 \\ r_1 & -r_1 - r_2 & r_2 \\ 0 & 2r_1 & -2r_1 \end{bmatrix}. \tag{15}$$

Assume that the packet lengths times are independent and generally distributed. Then the MMPP $N(t)$ together with the packet lengths specifies a continuous-time traffic arrival process A. When the process A is fed as input to a multiplexer with constant service rate, the system can be modeled as an MMPP/$G/1$ queue. A closed-form expression for the Laplace transform of the virtual waiting time of a MMPP/$G/1$ queue is given in [13].

B. Admission Control via Phase-Type Bounds

We consider the scenario shown in Fig. 1, in which five statistically independent traffic flows A_i, $i = 1, \cdots, 5$ arrive on input links with capacity C_{in} to a multiplexer with constant service rate C. All flows are identically distributed 2-state MMPPs characterized by Poisson arrival rates $\lambda_1 = 0.75$ and

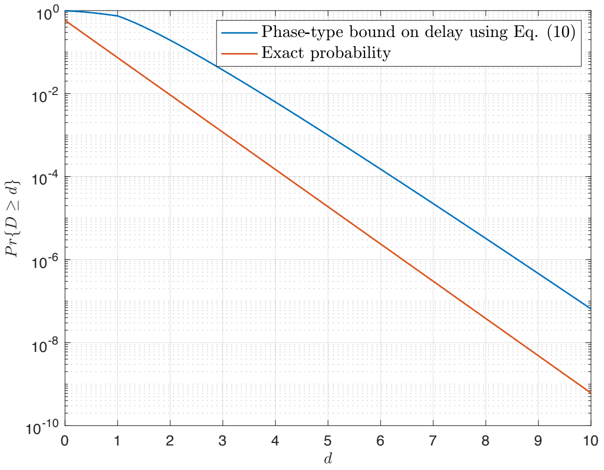

Fig. 2. Stochastic delay bound via phase-type traffic bounds.

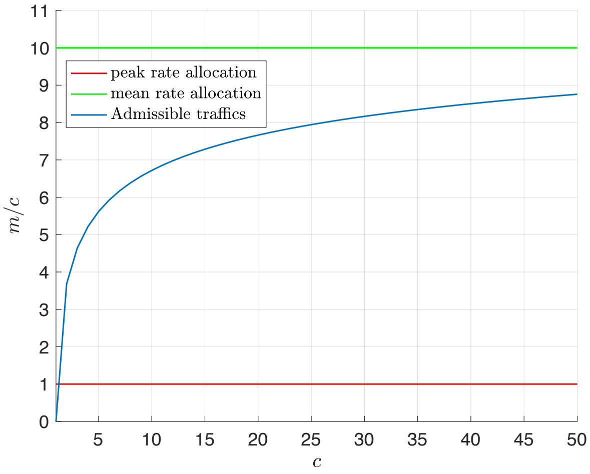

Fig. 3. Statistical multiplexing gain via MGF traffic envelopes.

$\lambda_2 = 0.5$ and rate matrix given by $r_1 = 2$ and $r_2 = 1$. Hence, the average arrival rate of each flow is 0.583 packets/unit of time. All packet lengths are assumed to be exponentially distributed with mean $\mu^{-1} = 1$. The input link capacity is set as $C_{\text{in}} = 5$ and the multiplexer has a constant rate server of rate $C = 5$.

The superposition of traffic flows A_i, $i = 1, \ldots, 5$, is a 6-state MMPP whose parameter can be determined using (13). With the 6-state MMPP as the input traffic, the multiplexer can be modeled as an MMPP/M/1 queue. Therefore, a closed-form solution for the virtual waiting time distribution at the multiplexer can be determined using results from [13]. The orange curve in Fig. 2 shows the tail distribution of the incurred delay in this case. Using definition (1), a phase-type bound can be obtained for each input traffic stream by considering the virtual waiting time distribution of a 2-state MMPP/M/1 queue. Using results from [13], this distribution has the form of an hyperexponential distribution. For this scenario, the procedure for fitting a traffic flow to a phase-type traffic descriptor (see Section II-A) can be bypassed. We shall set the parameter ρ equal to mean rate of the MMPP, i.e., 0.583. In this case, the phase-type bounding parameters of A_i, for $i = 1, \ldots, 5$ can be chosen to exactly match the tail of the workload distribution. Thus, we can assume $T = \infty$, and we obtain $a = 0.583$,

$$\boldsymbol{\pi} = [0.0.9982, 0.0018], \quad \mathbf{Q} = \begin{bmatrix} -0.413 & 0 \\ 0 & -0.858 \end{bmatrix}. \quad (16)$$

Since the traffic flows are MMPPs, they automatically satisfy the derived phase-type bounds and hence do not need to be regulated, although (σ^*, ρ) regulators are shown in Fig. 1 for the general case. Using the phase-type descriptor in (16), and Theorem 3, the phase-type descriptor of the aggregate arrival traffic can be derived. In this case, the aggregate traffic is characterized by a phase-type descriptor with a 92-state phase-type parameter. In this example, the approximation procedure described in Section IV-A to limit the number of phases in the phase-type descriptor was not performed. Using this phase-type descriptor for the aggregate traffic, a bound on the delay, as shown in Fig. 2, can be derived via Theorem 4. The blue curve in Fig. 2 shows the bound on the delay. From Fig. 2 the phase-type bounds can be used to provide the following stochastic delay guarantee: $\mathsf{P}\{D \geq 5\} < 10^{-3}$. The output link utilization in this case is $5(0.583)/C = 0.583$.

To compare the stochastic delay guarantee with a deterministic guarantee, such as that provided by the (σ, ρ) characterization of Cruz [1], we can increase the value of C such that $\mathsf{P}\{D \geq 5\}$ is close to zero, say 10^{-10}. As shown in Fig. 2, we can derive the exact tail probability $\mathsf{P}\{D \geq 5\}$ for every C. By increasing the value of C, we have that $\mathsf{P}\{D \geq 5\} \leq 10^{-10}$ when $C > 8.5$. Therefore, the link utilization that can be achieved when a deterministic guarantee is provided can be most $5(0.583)/8.5 \approx 0.34$.

C. Admission Control via MGF Traffic Envelope

As mentioned in Section II-B, an advantage of the MGF traffic envelope representation is that it can capture statistical multiplexing gain. Here, we shall consider a discrete-time Markov on-off fluid flow as a model for traffic flows. Such a process consists of an underlying discrete-time 2-state Markov process. In state 1 (On-state) the source generates a constant fluid flow of packets at rate r and in state 2 (Off-state), the source does not generate packets. The underlying Markov process has transition probability matrix $\mathbf{P}$ given by

$$\mathbf{P} = \begin{bmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{bmatrix}, \quad (17)$$

where p_{ij} for $i, j = 1, 2$ are the transition probabilities from state i to state j. The steady-state probability of states On and Off are given, respectively, by

$$p_{\text{on}} = \frac{p_{12}}{p_{12} + p_{21}}, \quad p_{\text{off}} = \frac{p_{21}}{p_{12} + p_{21}}. \quad (18)$$

The mean arrival rate of the process is $p_{\text{on}} r$. This process is also characterized by a burst parameter $\beta = 1/p_{12} + 1/p_{21}$.

According to [4], the MGF traffic envelope of the Markov On-Off process is given by $\hat{\sigma}(\theta) = 0$ and $\hat{\rho}(\theta) =$

$$\frac{1}{\theta}\ln\!\left(\frac{p_{11}+p_{22}e^{\theta r}+\sqrt{(p_{11}+p_{22}e^{\theta r})^2-4(p_{11}+p_{22}-1)e^{\theta r}}}{2}\right), \tag{19}$$

for $\theta > 0$. For the special case of a *memoryless* On-Off process, we have $p_{11} = p_{21}$ and $p_{22} = p_{12}$. In this case, $p_{\mathrm{on}} = p_{22}$ and the MGF traffic envelope simplifies to the form $\hat{\sigma}(\theta) = 0$ and

$$\hat{\rho}(\theta) = \frac{1}{\theta}\ \ln\left(p_{\mathrm{on}}e^{\theta r} + 1 - p_{\mathrm{on}}\right), \qquad \theta > 0. \tag{20}$$

Consider a multiplexer with constant rate C. Suppose a maximum of M identically distributed and statistically independent input Markov on-off fluid flows can be supported at the multiplexer while satisfying (8) for some specific d and ϵ. We are interested in evaluating the number of admissible flows per unit capacity, given by M/C, as C increases. As an example, we shall assume that each Markov on-off fluid flow, A_i, for $1 \le i \le M$, has average rate $p_{\mathrm{on}}r = 0.1$, peak rate $r = 1$, and burst parameter $\beta = 300$. Each Markov on-off fluid flow can be characterized by an MGF traffic envelope $(\hat{\sigma}(\theta), \hat{\rho}(\theta))$, which can be enforced by a regulator, as discussed in Section II-B. In this case, $\hat{\rho}_i(\theta)$ for $1 \le i \le M$, is given by (19), where $\theta > 0$ is a free parameter. Also as mentioned before, $\hat{\sigma}_i(\theta) = 0$ for $1 \le i \le M$.

According to Theorem 1, the aggregate traffic process, $A = A_1 + A_2 + \ldots + A_M$ can be characterized by the MGF parameter $\rho = M\hat{\rho}_1(\theta)$ and $\hat{\sigma} = 0$. The admission control scheme imposes a stochastic delay constraint of the form (8) with $d = 100$ and $\epsilon = 10^{-3}$. For each value of M and C, by using Theorem 5, and by optimizing the free parameter $\theta > 0$ we can derive a statistical bound on the delay for $d = 100$. If the derived statistical bound on the delay is less than ϵ, then such a choice of M and C is acceptable. For each value of C, we try to find the maximum value of acceptable M such that the desired statistical bound on the delay is satisfied.

With *mean rate* allocation, $1/0.1 = 10$ flows can be supported per unit capacity, whereas with *peak rate* allocation, only one flow can be supported per unit capacity. By performing admission control according to the MGF bound parameters, statistical multiplexing gain is achieved as C increases, as shown in Fig. 3. In particular, as C increases, the number of admissible flows per unit capacity, M/C increases and approaches the mean rate allocation of 10 flows per unit capacity. This shows that statistical multiplexing gain is achieved.

VI. Conclusion

We presented a practical framework for providing stochastic delay guarantees based on results from stochastic network calculus. Key elements of our approach are the phase-type traffic bound [3], the MGF traffic envelope [4], a method for fitting a traffic flow to a phase-type bound [11], stochastic traffic regulators to enforce compliance of a traffic flow to a negotiated traffic descriptors [5] and an admission control scheme. Each flow characterizes its traffic process by a phase-type traffic descriptor, which can be determined using the procedure developed in [11]. Similarly, an MGF traffic envelope can be determined for each traffic flow. Both types of traffic descriptors are enforced by stochastic regulators and are used in the proposed admission control scheme.

Our numerical study showed that much higher traffic utilization can be achieved compared to the deterministic (σ, ρ) framework, while providing a stochastic delay guarantee. Moreover, even higher utilization can be achieved by taking into account statistical multiplexing gain. The main contribution of this work is to show how results from stochastic network calculus can be applied in a practical framework to provide performance guarantees. In ongoing work, we are extending the proposed framework to multi-hop networking scenarios.

References

[1] R. L. Cruz, “A calculus for network delay. I. Network elements in isolation,” *IEEE Trans. Inf. Theory*, vol. 37, pp. 114–131, Jan. 1991.

[2] ——, “A calculus for network delay. II. Network analysis,” *IEEE Trans. Inf. Theory*, vol. 37, pp. 132–141, Jan. 1991.

[3] M. Kordi Boroujeny, B. L. Mark, and Y. Ephraim, “Tail-limited phase-type burstiness bounds for network traffic,” in *Proc. Inf. Sci. Sys. (CISS)*, Mar. 2019, pp. 1–6.

[4] M. Fidler and A. Rizk, “A guide to the stochastic network calculus,” *IEEE Commun. Surveys Tuts.*, vol. 17, no. 1, pp. 92–104, 2015.

[5] M. Kordi Boroujeny, B. L. Mark, and Y. Ephraim, “Stochastic traffic regulator for end-to-end network delay guarantees,” in *IEEE Int. Conference on Communications (ICC 2020)*, Dublin, Ireland, June 2020.

[6] M. Bladt and B. Nielsen, *Matrix-Exponential Distributions in Applied Probability*. New York, NY: Springer, 2017.

[7] Q. Yin, Y. Jiang, S. Jiang, and P. Y. Kong, “Analysis on generalized stochastically bounded bursty traffic for communication networks,” in *Proc. IEEE Local Comput. Netw. (LCN)*, Nov. 2002, pp. 141–149.

[8] Y. Jiang, Q. Yin, Y. Liu, and S. Jiang, “Fundamental calculus on generalized stochastically bounded bursty traffic for communication networks,” *Comput. Netw.*, vol. 53, no. 12, pp. 2011 – 2021, Aug. 2009.

[9] D. Starobinski and M. Sidi, “Stochastically bounded burstiness for communication networks,” *IEEE Trans. Inf. Theory*, vol. 46, no. 1, pp. 206–212, Jan. 2000.

[10] O. Yaron and M. Sidi, “Performance and stability of communication networks via robust exponential bounds,” *IEEE/ACM Trans. Netw.*, vol. 1, no. 3, pp. 372–385, Jun. 1993.

[11] M. Kordi Boroujeny, B. L. Mark, and Y. Ephraim, “Fitting network traffic to phase-type bounds,” in *54rd Annual Conference on Information Sciences and Systems (CISS 2020)*, Princeton, NJ, Mar. 2020.

[12] W. Fischer and K. Meier-Hellstern, “The Markov-modulated Poisson process (MMPP) cookbook,” *Perform. Eval.*, vol. 18, no. 2, pp. 149 – 171, 1993.

[13] D. M. Lucantoni, “New results on the single server queue with a batch Markovian arrival process,” *Stoch. Models*, vol. 7, no. 1, pp. 1–46, 1991.