Tail-Limited Phase-Type Burstiness Bounds for Network Traffic

Massieh Kordi Boroujeny, Brian L. Mark, and Yariv Ephraim
Dept. of Electrical and Computer Engineering
George Mason University
Fairfax, Virginia, U.S.A.
mkordibo@gmu.edu, bmark@gmu.edu, yephraim@gmu.edu

Abstract—**The bursty nature of network traffic makes it difficult to characterize accurately, and may give rise to heavy-tailed queue distributions within the network. Building on prior work in stochastic network calculus, we propose traffic burstiness bounds based on the class of phase-type distributions and develop an approach to estimate the parameter of such bounds using the expectation-maximization (EM) algorithm. By limiting the tail of the burstiness bound, our approach achieves a better fit of the phase-type distribution to the empirical data from heavy-tailed traffic. The proposed tail-limited phase-type burstiness bounds fall within the framework for stochastic network calculus based on generalized stochastically bounded burstiness.We demonstrate the effectiveness of the proposed methodology with a numerical example involving a heavy-tailed M/G/1 queue.**[1]

Index Terms—**communication networks, stochastic network calculus, traffic burstiness, phase-type distribution, EM algorithm, heavy-tailed queue.**

I. Introduction

Providing performance guarantees in communication networks is a challenging research problem due to the bursty nature of variable bit rate traffic streams. To provide a guarantee, for example, on the end-to-end delay of a given traffic stream, sufficient network resources need to be allocated. An admission control scheme is also needed to ensure that the resource requirements of a new traffic stream can be accommodated without compromising those of the existing traffic streams in the network. Overallocation of network resources to provide performance guarantees can lead to very poor network utilization. One approach to this issue is to characterize network traffic by sophisticated stochastic models and to derive end-to-end network performance metrics based on such models. Unfortunately, network traffic characterization is difficult due to the bursty nature of the traffic and analytical end-to-end network performance results are known only under very simple assumptions such as Poisson traffic and independence among nodes in the network. Another approach, pioneered by Cruz and others (see, e.g., [6], [7]) is to characterize the traffic in terms of mathematically simple bounds and then to compute bounds on end-to-end network delay. The original work of Cruz was based on a two-parameter (σ, ρ) deterministic bound on a traffic stream and an associated network calculus to derive deterministic end-to-end delay bounds. However, the determistic network calculus was found to provide bounds that were too loose in practice since they capitalize on the worst case scenario. Attention then turned to stochastic bounds on traffic burstiness and an associated *stochastic* network calculus to derive end-to-end bounds that are tighter with respect to network resource allocation, but probabilistic rather than deterministic.

In this paper, we develop traffic burstiness bounds based on the concept of "generalized" stochastically bounded burstiness (gSBB) proposed in [10], [19], which in turn is closely related to the stochastically bounded burstiness (SBB) concept proposed earlier in [14]. The SBB concept is a generalization of exponentially bounded burstiness (EBB), which was originally proposed in [17], [18]. A closely related traffic burstiness bound based on moment generating functions was developed in [4], [5]. Stochastic network calculus [5], [9] seeks to derive end-to-end network performance bounds from such traffic burstiness bounds. In earlier work [11], we proposed the use of the phase-type distribution to obtain specialized SBB-type bounds, referred to as phase-type bounded burstiness (PHBB), which can give rise to performance bounds significantly tighter than those obtained via EBB. This was demonstrated using numerical examples involving the Markov modulated Poisson Process (MMPP) fed as input traffic to $\cdot/M/1$ and $\cdot/E_2/1$ queues by leveraging results from [12].

We make several contributions in this work. We further refine the notion of PHBB from [11] by specializing the gSBB concept using phase-type distributions. We refer to the corresponding traffic burstiness bounds as gPHBB (generalized PHBB). In addition, we propose to bound the tail distribution of traffic burstiness up to a specified limit. In particular, this allows us to apply gSBB-type traffic burstiness bounds to heavy-tailed traffic, which cannot be bounded mathematically by a phase-type bound, which inherently has an exponentially decaying tail. We refer to this characterization of traffic as *tail-limited* gPHBB. Much of the research on stochastic network calculus has focused on the derivation of stochastic network delay bounds from the traffic burstiness bounds. Relatively little attention has been devoted to deriving or estimating the parameter of a traffic burstiness bound. We develop an EM (expectation-maximization) algorithm to estimate the tail-limited gPHBB parameter for an arbitrary traffic source from an empirical traffic trace. Our approach can be used to

[1] This work was supported in part by the U.S. National Science Foundation under Grant CCF-1717033.

characterize network traffic via stochastic bounds within the SBB/gSBB framework and applied to provide stochastic end-to-end delay guarantees. We provide a numerical example to demonstrate the effectiveness of the tail-limited gPHBB characterization of heavy-tailed traffic.

The remainder of the paper is organized as follows. In Section II, we briefly review the concepts of SBB/gSBB and the associated stochastic network calculus framework. In Section III, we define the concept of tail-limited gPHBB by specializing gSBB using the phase-type distribution and imposing a limit on the tail distribution. In Section IV, we develop an EM algorithm to estimate the gPHBB parameter of a given traffic source. In Section V, we provide a numerical example involving the application of tail-limited gPHBB to an M/G/1 heavy-tailed queue. Concluding remarks are provided in Section VI.

II. Stochastically Bounded Burstiness

The concept of stochastically bounded burstiness is defined in [14] as follows.

Definition 1 (SBB). A continuous-time traffic process $R = \{R(t) : t \geq 0\}$ is said to have stochastically bounded burstiness (SBB) with upper rate ρ and bounding function $f(\sigma) \in \mathcal{F}$ if, for all $t, s \geq 0$ and all $\sigma \geq 0$,

$$\mathsf{P}\left\{\int_s^t R(\tau)\,\mathrm{d}\tau - \rho(t-s) \geq \sigma\right\} \leq f(\sigma), \tag{1}$$

where $\mathcal{F}$ is defined as the family of functions such that for every $n, \sigma \geq 0$, the n-fold integral $(\int_\sigma^\infty \mathrm{d}u)^n f(u)$ is bounded. Let $R^{s,t} := \int_s^t R(\tau)\,\mathrm{d}\tau$ denote the amount of traffic arriving in the interval $[s,t)$. For a discrete-time traffic process, essentially the same definition of SBB applies, except that s and t are nonnegative integers, $R(t)$ represents the amount of traffic arriving during time-slot t, and $R^{s,t} := \sum_{u=s+1}^t R(u)$. In this paper, we will mostly work in continuous-time, although the results generally carry over to the discrete-time case.

The SBB concept was motivated as a generalization of exponentially bounded burstiness (EBB), originally proposed in [17].

Definition 2 (EBB). A traffic process R is EBB if it is SBB with a bounding function of the form $f(\sigma) = Ae^{-\alpha\sigma}$ where $A, \alpha \geq 0$.

In [11], we proposed a bounding function based on the phase-type distribution, which is a large class of probability distributions including exponentials, mixtures of exponentials, and convolutions of mixtures of exponentials.

Definition 3 (PHBB). A traffic process R has phase-type bounded burstiness (PHBB) if it is SBB with a bounding function of the form $f(\sigma) = A\boldsymbol{\pi}e^{\mathbf{Q}\sigma}\mathbf{1}$ where $\mathbf{1}$ is a column vector of all ones, $(\boldsymbol{\pi}, \mathbf{Q})$ represents the parameter of a phase-type distribution, and $A \geq 0$.

Due to the greater modeling fidelity of the phase-type distribution compared to the exponential distribution, tighter bounds on traffic burstiness can potentially be achieved with PHBB compared to EBB at the expense of a more complicated parameter.

The idea of SBB was further developed in [10], [19] with the concept of *generalized* stochastically bounded burstiness. Let

$$W(t) := \max_{0 \leq s \leq t}\left\{R^{s,t} - \rho(t-s)\right\}, \tag{2}$$

Definition 4 (gSBB). A traffic process R is said to have generalized stochastically bounded burstiness (gSSB) with upper rate ρ and bounding function $f(\sigma) \in \mathcal{BF}$ if, for all $t \geq 0$ and all $\sigma \geq 0$,

$$\mathsf{P}\left\{W(t) \geq \sigma\right\} \leq f(\sigma), \tag{3}$$

where $\mathcal{BF}$ is defined as the family of positive, non-increasing functions.

Comparing Eqs. (1) and (3), we note that the gSBB characterization is more restrictive than that of SBB in the following sense: For a given bounding function, if a traffic process is gSBB then it is also SBB, but the converse may not hold. The gSBB concept has several advantages over SBB. The class of bounding functions, $\mathcal{BF}$, for gSBB is less restrictive than the class $\mathcal{F}$ appearing in the definition of SBB. In the definition of gSBB, the process $W(t)$ can be interpreted as the virtual workload of a constant rate queue with service rate ρ and input traffic R. This property is useful in establishing stochastic network calculus results, and as we shall see in Section IV, central to our approach for estimating the parameter of the gPHBB traffic burstiness bound discussed next in Section III.

III. Generalized Phase-Type Traffic Bounds

In this section, we develop phase-type traffic bounds as a useful specialization of the gSBB bounds in [10], [19] and introduce a further refinement by limiting the tail of the bounding function.

A. Phase-type Distribution

The phase-type distribution is defined in terms of a Markov chain $X = \{X(t) : t \geq 0\}$ with state space $E = \{1, 2, \ldots, n, n+1\}$, where states $1, 2, \ldots, n$ are transient states and $n+1$ is an absorbing state. The generator of X has the form [2]

$$\begin{pmatrix} \mathbf{Q} & \mathbf{q} \\ \mathbf{0} & 0 \end{pmatrix}, \tag{4}$$

where $\mathbf{Q} = [q_{ij} : i, j = 1, \ldots, n]$ is an $n \times n$ matrix such that q_{ij} is the transition rate from state i to state j and $\mathbf{q} = -\mathbf{Q}\mathbf{1}$ is an $n \times 1$ column vector. Define $\pi_i = \mathsf{P}(X(0) = i)$ for $i = 1, \ldots, n+1$ and the vector $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_n)$. Hence, the initial distribution of X is given by $(\boldsymbol{\pi}, \pi_{n+1})$, where π_{n+1} is the probability that the chain starts in the absorbing state. Let $\tau := \inf\{t \geq 0 : X(t) = n+1\}$ be the time until absorption of the Markov process X. The random variable τ is said to be phase-type with parameter $(\boldsymbol{\pi}, \mathbf{Q})$. In this case, the cumulative distribution function and survival function of τ are given, respectively, by

$$F_\tau(t) = 1 - \boldsymbol{\pi} e^{\mathbf{Q}t}\mathbf{1}, \tag{5}$$
$$S_\tau(t) = \mathsf{P}(\tau > t) = 1 - F_\tau(t) = \boldsymbol{\pi} e^{\mathbf{Q}t}\mathbf{1}, \tag{6}$$

for $t \geq 0$. The class of phase-type distributions has the important property of being dense in the family of distributions of nonnegative random variables; i.e., the distribution of any random variable taking values in $[0, \infty)$ can be approximated arbitrarily closely by a phase-type distribution [16, Theorem 5.2]. In addition, phase-type distributions are mathematically tractable and form a closed set with respect to operations such as convolutions or mixtures.

B. Tail-Limited Generalized Phase-Type Bounded Burstiness

Now we specialize the gSBB concept to bounds based on phase-type distributions and restrict the tail of the bound to a limit $T > 0$.

Definition 5 (Tail-limited gPHBB). A traffic process $R(t)$ has tail-limited generalized phase-type bounded burstiness (gPHBB) with upper rate ρ and bounding parameter $(A, \boldsymbol{\pi}, \mathbf{Q}, T)$ if

$$\mathsf{P}\left\{W(t) \geq \sigma\right\} \leq A\boldsymbol{\pi} e^{\mathbf{Q}\sigma}\mathbf{1}, \tag{7}$$

for all $t \geq 0$ and all $\sigma \in [0, T]$. Here, $A \geq 0$, $T > 0$, $W(t)$ is given by (2), and $(\boldsymbol{\pi}, \mathbf{Q})$ represents the parameter of a phase-type distribution. When no tail limit is imposed, i.e., $T = \infty$, the traffic process is referred to simply as gPHBB.

Tail-limited gPHBB traffic processes inherit the stochastic network calculus of gSBB processes developed in [10], [19], which are analogous to the stochastic network calculus for SBB and EBB processes [14], [17], respectively. Using properties of the phase-type distribution, results for a stochastic network calculus based on tail-limited gPHBB can be derived.

The following *Characterization* theorem follows directly from Definition 5.

Theorem 1 (*Characterization*). Consider a work-conserving system that transmits at a constant rate of ρ and is fed with a single traffic stream with rate process $R(t)$ and $W_q(t)$ is the queue workload at time t. Then $R(t)$ is tail-limited gPHBB with upper rate ρ and bounding parameter $(A, \boldsymbol{\pi}, \mathbf{Q}, T)$ if and only if

$$\mathsf{P}\{W_q(t) \geq \sigma\} \leq A\boldsymbol{\pi} e^{\mathbf{Q}\sigma}\mathbf{1}, \tag{8}$$

for all $t \geq 0$ and all $T \geq \sigma \geq 0$.

The following *Sum* and *Input-Output* theorems for tail-limited gPHBB are useful for deriving stochastic bounds on end-to-end network delay.

Theorem 2 (*Sum*). Let $R_1(t)$ and $R_2(t)$ be tail-limited gPHBB traffic processes with upper rates ρ_1 and ρ_2 respectively, and bounding parameters $(A, \boldsymbol{\alpha}, \mathbf{G}, T_1)$ and $(B, \boldsymbol{\beta}, \mathbf{H}, T_2)$, respectively. Then $R_1(t) + R_2(t)$ is tail-limited gPHBB with upper rate $\rho = \rho_1 + \rho_2$ and bounding parameter $(C, \boldsymbol{\pi}, \mathbf{Q}, T)$ where $T = \min(T_1, T_2)$, $C = A + B$,

$$\boldsymbol{\pi} = \left[\frac{A\boldsymbol{\alpha}}{A+B}, \frac{B\boldsymbol{\beta}}{A+B}\right], \quad \mathbf{Q} = \begin{pmatrix} p\mathbf{G} & \mathbf{0} \\ \mathbf{0} & (1-p)\mathbf{H} \end{pmatrix}, \tag{9}$$

and p is a real number such that $0 < p < 1$.

Proof: As $R_1(t)$ and $R_2(t)$ are gSBB, we can apply the Sum theorem for gSBB [19, Theorem 3]. In this case, a bounding function of the aggregated traffic is given by $g(\sigma) = f_1(p\sigma) + f_2((1-p)\sigma)$, where

$$f_1(\sigma) = A\boldsymbol{\alpha} e^{\mathbf{G}\sigma}\mathbf{1}, \quad \text{for} \quad T_1 > \sigma > 0$$
$$f_2(\sigma) = B\boldsymbol{\beta} e^{\mathbf{H}\sigma}\mathbf{1}, \quad \text{for} \quad T_2 > \sigma > 0.$$

We have

$$\begin{aligned} g(\sigma) &= A\boldsymbol{\alpha} e^{p\mathbf{G}\sigma}\mathbf{1} + B\boldsymbol{\beta} e^{(1-p)\mathbf{H}\sigma}\mathbf{1} \\ &= (A+B)\left[\frac{A\boldsymbol{\alpha}}{A+B}, \frac{B\boldsymbol{\beta}}{A+B}\right]\begin{pmatrix} e^{p\mathbf{G}} & \mathbf{0} \\ \mathbf{0} & e^{(1-p)\mathbf{H}} \end{pmatrix}\mathbf{1}, \end{aligned}$$

for $T = \min(T_1, T_2) \geq \sigma > 0$. By setting $T = \min(T_1, T_2)$, $g(\sigma)$ is well-defined. ∎

Theorem 3 (*Input-Output Relation*). Let $R_{\mathrm{i}}(t)$ be the input traffic rate process to a work-conserving element, which transmits at rate C. Suppose that $R_{\mathrm{i}}(t)$ is tail-limited gPHBB with upper rate $\rho < C$ and bounding parameter $(A, \boldsymbol{\pi}, \mathbf{Q}, T)$. Let $R_{\mathrm{o}}(t)$ denote the output traffic rate process. Then the following hold:

1) $R_{\mathrm{o}}(t)$ is less bursty than $R_{\mathrm{i}}(t)$, almost surely; i.e.,
$$\max_{0 \leq s \leq t}\left\{R_{\mathrm{o}}^{s,t} - \rho(t-s)\right\} \leq \max_{0 \leq s \leq t}\left\{R_{\mathrm{i}}^{s,t} - \rho(t-s)\right\}, \text{ a.s.}$$
2) $R_{\mathrm{o}}(t)$ is tail-limited gPHBB with upper rate ρ and the same bounding parameter $(A, \boldsymbol{\pi}, \mathbf{Q}, T)$.

Proof:

1) This relation follows directly from [19, Theorem 5] and does not depend on the bounding function.
2) Since $R_{\mathrm{i}}(t)$ is tail-limited gPHBB with upper rate ρ and bounding parameter $(A, \boldsymbol{\pi}, \mathbf{Q}, T)$, which is a special case of gSBB, from [19, Corollary(Input-Output Relation)] it follows that $R_{\mathrm{o}}(t)$ is tail-limited gPHBB with upper rate ρ and the same bounding parameter $(A, \boldsymbol{\pi}, \mathbf{Q}, T)$.

∎

IV. Parameter Estimation via EM Algorithm

We develop a method for estimating the parameter of a tail-limited gPHBB bound for a traffic source based on an EM algorithm. Our approach leverages the interpretation of $W(t)$ in Definition 4 (gSBB) as the virtual workload in a constant rate server queue, as well as the phase-type form of the bounding function in Definition 5 (gPHBB). For a given upper rate ρ, the traffic is, in effect, offered to a queue with a constant rate server of rate ρ. Samples of the virtual workload of the queue are used to estimate, via the EM algorithm, the parameter $(\boldsymbol{\pi}, \mathbf{Q})$ of a phase-type distribution that would

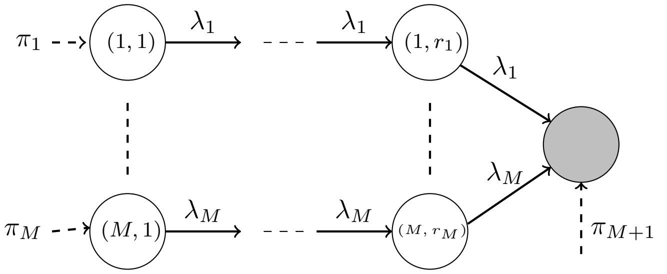

Fig. 1. Hyper-Erlang form of a phase-type random variable (r_1 may differ from r_M).

satisfy (7) at equality when $A = 1$ and $T = \infty$. The left-hand side of (7) is known either in theoretical form (as in the example of Section V) or can be approximated empirically from the observation samples. The value of the tail-limit parameter T is assumed to be specified in advance, depending on the performance requirements of the traffic stream. The EM parameter estimate $(\hat{\boldsymbol{\pi}}, \hat{\mathbf{Q}})$ is then applied to derive a tight tail-limited gPHBB bound of the form in (7) by adjusting the value of A. Note that when $A = 1$, the right-hand side of (7) using the parameter estimate for $(\boldsymbol{\pi}, \mathbf{Q})$ may not be a true tail-limited upper bound or, if it is, possibly a tighter bound could be obtained with a smaller value of A. Thus, we set A to the smallest value that ensures the upper bound property of (7).

A. *Hyper-Erlang Model*

Various EM algorithms for fitting data to a phase-type distribution have been developed in the literature, notably the algorithm of Asmussen [1]. In this section, we adopt an EM algorithm developed by Thummler *et al.* [15] for estimating the parameter of a hyper-Erlang distribution. Although the hyper-Erlang distribution is a special case of a phase-type distribution, the class of hyper-Erlang distributions is also dense in the family of distributions with nonnegative support [15]. When a phase-type parameter $(\boldsymbol{\pi}, \mathbf{Q})$ is specialized to the form of a hyper-Erlang distribution, the number of nonzero components in $\mathbf{Q}$ is significantly fewer than in the general case. Hence, fitting with the hyper-Erlang distribution is computationally simpler and less prone to overfitting.

The hyper-Erlang distribution may be viewed as a mixture of Erlang distributions. Consider a hyper-Erlang model consisting of a mixture of M Erlang distributions, where the orders of the Erlang distributions are given by $\mathbf{r} = (r_1, \ldots, r_M)$ and the mixture probabilities are given by $\tilde{\boldsymbol{\pi}} = (\pi_1, \ldots, \pi_M)$. The ith component of the mixture is an Erlang distribution of order r_i parameterized by λ_i, with probability density function

$$p_i(x; \lambda_i) = \frac{(\lambda_i x)^{r_i - 1}}{(r_i - 1)!} \lambda_i e^{-\lambda_i x}, \quad x \geq 0, \tag{10}$$

for $i = 1, \ldots, M$. The parameter of the hyper-Erlang distribution is given by $\boldsymbol{\Theta} = (\tilde{\boldsymbol{\pi}}, \mathbf{r}, \boldsymbol{\lambda})$, where $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_M)$.

A hyper-Erlang random variable can be represented in terms of a phase-type random variable parameterized by $(\boldsymbol{\pi}, \mathbf{Q})$. The corresponding Markov chain for the hyper-Erlang random variable is shown in Fig. 1, where the absorbing state is shaded. In this case,

$$\mathbf{Q} = \mathrm{diag}\{\mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_M\}, \tag{11}$$

where

$$\mathbf{q}_i = \begin{pmatrix} -\lambda_i & \lambda_i & 0 & \ldots & 0 \\ 0 & -\lambda_i & \lambda_i & \ldots & 0 \\ \vdots & \ddots & \ddots & \vdots & \vdots \\ 0 & 0 & \ldots & -\lambda_i & \lambda_i \\ 0 & 0 & \ldots & 0 & -\lambda_i \end{pmatrix}_{r_i \times r_i} . \tag{12}$$

The initial probability vector $\boldsymbol{\pi}$ is given by

$$\boldsymbol{\pi} = (\pi_1, \underbrace{0, \ldots, 0}_{r_1 - 1}, \pi_2, \underbrace{0, \ldots, 0}_{r_2 - 1}, \ldots, \pi_M, \underbrace{0, \ldots, 0}_{r_M - 1}, \pi_{M+1}) \tag{13}$$

In this case, the probability density function of τ is given by

$$f_\tau(t) = \sum_{i=1}^{M} \pi_i \frac{(\lambda_i t)^{r_i - 1}}{(r_i - 1)!} \lambda_i e^{-\lambda_i t}, \quad t \geq 0. \tag{14}$$

Note that the vector $\tilde{\boldsymbol{\pi}}$ of mixture probabilities in the hyper-Erlang model is a subvector of $\boldsymbol{\pi}$ in the phase-type representation.

The hyper-Erlang distribution is a particular case of a phase-type distribution consisting of $3M$ parameter values consisting of the components of the vectors $\mathbf{r} \in \mathbb{N}_+^M$, $\tilde{\boldsymbol{\pi}} \in \mathbb{R}_+^M$, and $\boldsymbol{\lambda}^M$, where $\mathbb{N}_+$ denotes the set of nonnegative integers and $\mathbb{R}_+$ denotes the set of nonnegative reals. When traffic is fed to a constant rate server, the probability that the queue workload is empty is given by $\mathsf{P}\{W(t) = 0\} = 1 - \rho$, where ρ is the utilization factor. To capture this effect, we introduce an additional $(M+1)$st branch to the hyper-Erlang model with corresponding branch probability denoted by π_{M+1}. In this case, the hyper-Erlang probability density function becomes

$$p(x; \boldsymbol{\Theta}) = \sum_{\substack{i=1 \\ x_k \neq 0}}^{M} \pi_i p_i(x_k; \lambda_i) + \pi_{M+1} \mathbf{1}_{\{x=0\}}, \tag{15}$$

for $i = 1, \ldots, M$, where $\mathbf{1}_{\mathcal{A}}(\cdot)$ represents the indicator function on the set $\mathcal{A}$. We shall assume that the vector $\mathbf{r}$ of Erlang orders for the hyper-Erlang distribution is constant. Accordingly, the parameter of the (extended) hyper-Erlang model is given by $\boldsymbol{\Theta} = (\tilde{\boldsymbol{\pi}}, \pi_{M+1}, \boldsymbol{\lambda})$. The hyper-Erlang parameter can then be mapped to a phase-type parameter $(\boldsymbol{\pi}, \mathbf{Q})$ to derive the gPHBB bound given in Definition 5. In addition, the parameter A must be chosen to ensure that the workload survival function $\mathsf{P}\{W(t) \geq \sigma\}$ is upper-bounded in accordance with (7).

B. *EM Algorithm*

Given an observation sequence of samples of the queue workload process, $\mathbf{x} = (x_1, \ldots, x_K)$, the log-likehood of the data is given by

$$\log L(\mathbf{x}; \boldsymbol{\Theta}) = \log p(\mathbf{x}; \boldsymbol{\Theta}) = \log \prod_{k=1}^{K} p(x_k; \boldsymbol{\Theta}), \tag{16}$$

where the last equality assumes independence of the observed samples. We follow the approach of [15], in which the Erlang order vector $\mathbf{r}$ is chosen from a set $\mathcal{R} = \{\mathbf{r} \geq \mathbf{0} : \mathbf{r1} = n\}$, where $\mathbf{0}$ is a row vector of all zeros and n is a fixed positive number chosen in advance. The number of Erlang components, M, in the vectors $\mathbf{r} \in \mathcal{R}$ ranges from 1 to n. The phase-type parameter $\mathbf{\Theta}$ corresponding to each $\mathbf{r} \in \mathcal{R}$ is estimated and then the estimate with the highest incomplete data log-likelihood, given by (16), is chosen.

In [15], the unobserved data y_k, representing the Erlang branch from which the sample x_k was drawn, is introduced to derive an EM algorithm based on complete data. Let $\mathbf{y} = (y_1, \ldots, y_k)$ represent the unobserved data sequence. The EM algorithm developed in [15] in effect maximizes the complete log-likelihood function $\log L(\mathbf{x}, \mathbf{y}; \mathbf{\Theta})$. To accommodate the positive probability mass at $x_k = 0$, this EM algorithm requires a slight modification. We omit the details here, but provide the key re-estimation formulas of the EM algorithm as follows:

$$\hat{\pi}_i = \begin{cases} \dfrac{1}{K} \displaystyle\sum_{\substack{k=1 \\ x_k \neq 0}}^{K} q(i \mid x_k; \hat{\mathbf{\Theta}}), & i = 1, 2, \ldots, M, \\ \frac{K_0}{K}, & i = M+1, \end{cases} \tag{17}$$

and

$$\hat{\lambda}_i = r_i \cdot \frac{\sum_{\substack{k=1 \\ x_k \neq 0}}^{K} q(i \mid x_k, \hat{\mathbf{\Theta}})}{\sum_{\substack{k=1 \\ x_k \neq 0}}^{K} x_k q(i \mid x_k; \hat{\mathbf{\Theta}})} \tag{18}$$

for $i = 1, \ldots, M$. In the above equations, $q(y_k \mid x_k; \hat{\mathbf{\Theta}})$ represents the posterior probability mass function of the unobserved sample y_k given the observed data sample x_k and is given by

$$q(y_k \mid x_k; \hat{\mathbf{\Theta}}) = \frac{\hat{\pi}_{y_k} \cdot p_{y_k}(x_k; \hat{\lambda}_{y_k})}{\sum_{i=1}^{M} \hat{\pi}_i \cdot p_i(x_k; \hat{\lambda}_i)} \tag{19}$$

for $y_k \in 1, 2, \ldots, M$, $x_k \neq 0$, and

$$q(M+1 \mid x_k; \hat{\mathbf{\Theta}}) = \begin{cases} 1, & x_k = 0, \\ 0, & x_k \neq 0. \end{cases} \tag{20}$$

To initialize the EM algorithm, we set $\pi_{M+1} = \frac{K_0}{K}$, as π_{M+1} will be fixed through all iterations of the algorithm.

V. Case Study

A. M/G/1 Heavy-Tailed Queue

In this section, we adopt the model of the M/G/1 queue in [3]. In this model, the service time, denoted by $\tau_{\boldsymbol{\theta}}$, depends on a gamma-distributed random variable $\boldsymbol{\theta}$. The conditional probability density function of $\tau_{\boldsymbol{\theta}}$ given $\boldsymbol{\theta} = \theta$ is given by

$$\mathsf{P}\{\tau_{\boldsymbol{\theta}} < t \mid \boldsymbol{\theta} = \theta\} = 1 - \delta \left(\frac{\theta}{\theta + t} \right)^{v}, \tag{21}$$

where $1 < v < 2$, $0 < \delta \leq 1$, and the density of $\boldsymbol{\theta}$ is given by

$$f_{\boldsymbol{\theta}}(\theta) = \frac{s^{2-v}}{\Gamma(2-v)} \theta^{1-v} e^{-s\theta}, \tag{22}$$

where $s > 0$ is a constant and $\Gamma(\cdot)$ is the gamma function. For the particular case $v = 3/2$, the cumulative distribution function of $\tau_{\boldsymbol{\theta}}$ is shown in [3] to have the form

$$\mathsf{P}\{\tau_{\boldsymbol{\theta}} \leq t\} = 1 + \delta \left[\frac{2\sqrt{st}}{\sqrt{\pi}} - (1 + 2st) e^{st} \operatorname{erfc}(\sqrt{st}) \right], \tag{23}$$

where the complementary error function is defined by

$$\operatorname{erfc}(x) = \frac{2}{\sqrt{\pi}} \int_{x}^{\infty} e^{-u^2} \mathrm{d}u. \tag{24}$$

The distribution of the stationary waiting time W for the M/G/1 queue is given by [3]:

$$\begin{aligned} \mathsf{P}\{W \leq t\} = 1 - &\frac{1+\sqrt{\rho}}{2} \sqrt{\rho} e^{(1-\sqrt{\rho})^2 st} \cdot \operatorname{erfc}\left[(1-\sqrt{\rho}) \sqrt{st} \right] \\ + &\frac{1-\sqrt{\rho}}{2} \sqrt{\rho} e^{(1-\sqrt{\rho})^2 st} \cdot \operatorname{erfc}\left[(1+\sqrt{\rho}) \sqrt{st} \right]. \end{aligned} \tag{25}$$

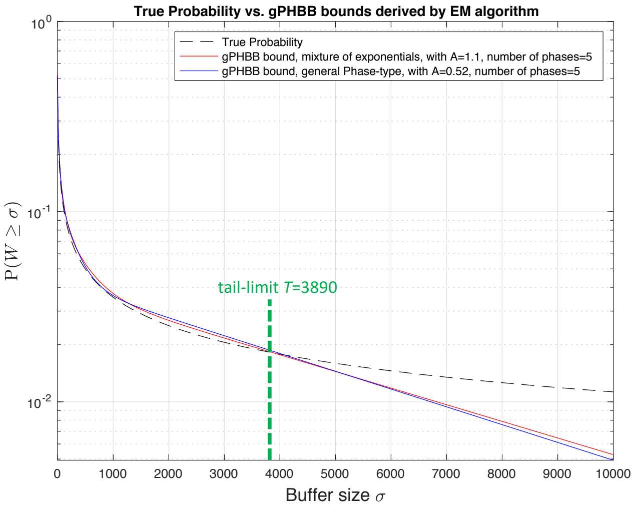

Fig. 2. Estimated gPHBB bound and true tail probability for a heavy-tailed M/G/1 queue.

B. Numerical Example

We consider a heavy-tailed M/G/1 queue as described above. We set $s = \delta = 1$ and $\lambda = 0.5$. The stationary waiting time distribution of the queue is given by (25). For this example, we consider the queue workload, which is simply related to the waiting time by a constant factor. The probability that the queue is empty is given by $\mathsf{P}\{\sigma = 0\} = 1 - \rho = 1 - \lambda\beta = 0.5$ The survival function of this stationary waiting time is shown in Fig. 2 as the *true probability* curve.

Since the workload distribution is heavy-tailed, the survival function cannot be bounded by a phase-type survival function. We demonstrate that the workload distribution can be bounded by a tail-limited gPHBB. In this particular case, we have set the tail limit parameter to $T = 3890$, in units representing the workload, e.g., bytes of data. The tail limit is explicitly shown in Fig. 2, and represents where the tail of the gPHBB curve is to be cut off. In other words, the gPHBB curve is only claimed to bound the true workload survival function up to the tail limit T.

To estimate the gPHBB parameter, we apply the EM algorithm for the hyper-Erlang model with the total number of phases set to $n = 5$. We generated $N = 10^6$ random samples drawn from the heavy-tailed workload distribution given in (25). For this example, the hyper-Erlang parameter estimate turns out to be a mixture of exponentials, such that $M = n = 5$ and $\boldsymbol{\pi} = \tilde{\boldsymbol{\pi}}$. This is in agreement with an observation in [8] that when the probability density function of the queue workload is completely monotone, as in this example, it can be well approximated by a mixture of exponentials. The following gPHBB parameter values were obtained: $A = 1.1$,

$$\boldsymbol{\pi} = [0.037, 0.059, 0.12, 0.14, 0.14], \tag{26}$$

$$\boldsymbol{\lambda} = [2.0\mathrm{e}^{-4}, 0.36\mathrm{e}^{-3}, 1.5\mathrm{e}^{-2}, 7.6\mathrm{e}^{-2}, 0.38], \tag{27}$$

where $\mathrm{e}^d := 10^d$. The phase-type matrix $\mathbf{Q}$ was obtained from $\boldsymbol{\lambda}$ using (11).

We have also computed a gPHBB bound using the EM algorithm for the general phase-type distribution described in [1]. In this case, we have a phase-type bound with 5 phases. Estimates of the parameters A and $\boldsymbol{\pi}$ were obtained as, respectively, $A = 0.52$ and

$$\boldsymbol{\pi} = [4.7\mathrm{e}^{-2}, 4.4\mathrm{e}^{-9}, 0.95, 8.1\mathrm{e}^{-9}, 5.7\mathrm{e}^{-7}].$$

The estimate of the $\mathbf{Q}$ matrix was as follows:

$$\begin{bmatrix} -4.9\mathrm{e}^{-2} & 1.7\mathrm{e}^{-5} & 3.3\mathrm{e}^{-2} & 1.8\mathrm{e}^{-5} & 1.2\mathrm{e}^{-2} \\ 3.1\mathrm{e}^{-6} & -1.5\mathrm{e}^{-2} & 2.6\mathrm{e}^{-7} & 1.4\mathrm{e}^{-2} & 1.2\mathrm{e}^{-3} \\ 1.3\mathrm{e}^{-1} & 1.0\mathrm{e}^{-6} & -2.6\mathrm{e}^{-2} & 1.3\mathrm{e}^{-7} & 2.4\mathrm{e}^{-4} \\ 1.1\mathrm{e}^{-6} & 3.3\mathrm{e}10^{-3} & 7.5\mathrm{e}^{-9} & -3.4\mathrm{e}^{-3} & 1.4\mathrm{e}^{-4} \\ 4.3\mathrm{e}^{-3} & 8.7\mathrm{e}^{-4} & 3.5\mathrm{e}^{-6} & 5.1\mathrm{e}^{-4} & -5.7\mathrm{e}^{-3} \end{bmatrix}.$$

From Fig. 2, the gPHBB bound appears to be slightly looser than the one obtained using the hyper-Erlang model. This can be explained by overfitting of the more general phase-type model compared to the hyper-Erlang model.

VI. Conclusion

We proposed the use of phase-type distributions to specialize the general bounding function in the gSBB traffic burstiness bounding framework [10], [19]. We established key properties of the proposed tail-limited gPHBB bounds. We developed an approach to estimate a tail-limited gPHBB bound for a given traffic source based on the EM algorithm. A numerical example was provided to demonstrate the gPHBB bound for an M/G/1 queue with heavy-tailed service using results from [3]. We showed that the notoriously difficult case of a heavy-tailed input traffic can be bounded meaningfully using phase-type bounds over a finite time horizon.

The proposed approach for obtaining gPHBB bounds could be applied to variable bit rate (VBR) traffic sources with tight delay constraints, for example, in multimedia streaming applications. For real-time traffic, training data could be used to obtain an initial estimate of the gPHBB bounds via the EM-based approach. An online algorithm could be developed to adapt the gPHBB bounds to the time-varying characteristics of a real-time traffic stream. Such an approach presumes that the network is capable of renegotiating the parameter of a traffic stream in real-time [13]. This is a topic of ongoing investigation.

References

[1] S. Asmussen, O. Nerman, and M. Olsson, “Fitting phase-type distributions via the EM algorithm,” *Scandinavian Journal of Statistics*, vol. 23, no. 4, pp. 419–441, 1996.

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

[3] O. J. Boxma and J. W. Cohen, “The M/G/1 queue with heavy-tailed service time distribution,” *IEEE J. Sel. Areas Commun.*, vol. 16, no. 5, pp. 749–763, Jun. 1998.

[4] C. S. Chang, “Stability, queue length and delay of deterministic and stochastic queueing networks,” *IEEE Trans. Autom. Control*, vol. 39, no. 5, pp. 913–931, May 1994.

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

[6] R. L. Cruz, “A calculus for network delay. II. Network analysis,” *IEEE Trans. Inf. Theory*, vol. 37, no. 1, pp. 132–141, Jan. 1991.

[7] ——, “A calculus for network delay. I. Network elements in isolation,” *IEEE Trans. Inf. Theory*, vol. 37, no. 1, pp. 114–131, Jan. 1991.

[8] A. Feldmann and W. Whitt, “Fitting mixtures of exponentials to long-tail distributions to analyze network performance models,” in *Proceedings of INFOCOM ’97*, vol. 3, Apr. 1997, pp. 1096–1104.

[9] Y. Jiang and Y. Liu, *Stochastic Network Calculus*. Springer-Verlag, 2008.

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

[11] M. Kordi Boroujeny, Y. Ephraim, and B. L. Mark, “Phase-type bounds on network performance,” in *52nd Annual Conference on Information Sciences and Systems (CISS)*, Princeton, NJ, Mar. 2018, pp. 1–6.

[12] D. M. Lucantoni, “New results on the single server queue with a batch markovian arrival process,” *Communications in Statistics. Stochastic Models*, vol. 7, no. 1, pp. 1–46, 1991.

[13] B. L. Mark and G. Ramamurthy, “Real-time estimation and dynamic renegotiation of UPC parameters for arbitrary traffic sources in ATM networks,” *IEEE/ACM Trans. Netw.*, vol. 6, no. 6, pp. 811–827, Dec. 1998.

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

[15] A. Thummler, P. Buchholz, and M. Telek, “A novel approach for phase-type fitting with the EM algorithm,” *IEEE Trans. Depend. Sec. Comput.*, vol. 3, no. 3, pp. 245–258, Jul. 2006.

[16] R. W. Wolff, *Stochastic Modeling and the Theory of Queues*. New Jersey: Prentice-Hall, 1989.

[17] 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.

[18] ——, “Generalized processor sharing networks with exponentially bounded burstiness arrivals,” in *Proceedings of INFOCOM ’94 Conference on Computer Communications*, Jun. 1994, pp. 628–634 vol.2.

[19] Q. Yin, Y. Jiang, S. Jiang, and P. Y. Kong, “Analysis on generalized stochastically bounded bursty traffic for communication networks,” in *27th Annual IEEE Conference on Local Computer Networks (LCN)*, Nov. 2002, pp. 141–149.