Phase-Type Bounds on Network Performance

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

***Abstract*—Evaluation of end-to-end network performance using realistic traffic models is a challenging problem in networking. The classical theory of queueing networks is feasible only under rather restrictive assumptions on the input traffic models and network elements. An alternative approach, first proposed in the late 1980s, is to impose deterministic bounds on the input traffic that can be used as a basis for a network calculus to compute end-to-end network delay bounds. Such deterministic bounds are inherently loose as they must accommodate worst case scenarios. Since the early 1990s, efforts have shifted to development of a stochastic network calculus to provide probabilistic end-to-end performance bounds. In this paper, we capitalize on the approach of stochastically bounded burstiness (SBB) which was developed for a general class of bounding functions, and was demonstrated for a bound that is based on a mixture distribution. We specialize the SBB bounds to bounds based on the class of phase-type distributions, which includes mixture distributions as a particular case. We develop the phase-type bounds and demonstrate their performance.**[1]

***Index Terms*—network performance, network calculus, phase-type distribution, delay bounds.**

I. Introduction

A challenging problem in the field of networking is how to provide end-to-end performance guarantees while still reaping the benefits of statistical multiplexing gain enabled by the packet switching paradigm. Queueing theory and teletraffic theory, which are by now considered classical branches of applied probability, can yield analytical performance results for a single node or network element under fairly general input traffic and service processes. The theory of queueing networks extends queueing theory to a network of nodes, but only under rather restrictive assumptions on the input traffic processes and service processes. In particular, the theory of queueing networks cannot generally accommodate the bursty traffic models that characterize modern networks. An alternative approach is to impose bounds on the input traffic that can be used as a basis for a network calculus to compute end-to-end network delay bounds. In his pioneering papers [1], [2], Cruz introduced the so-called (σ, ρ)-characterization of traffic and developed an associated deterministic network calculus to compute end-to-end delay bounds. Since such deterministic end-to-end delay bounds must accommodate worst-case scenarios, they tend to be very loose. Hence, interest has turned towards the development of a stochastic network calculus to provide probabilistic end-to-end performance bounds. The influential book by Chang [3] summarizes the state-of-the-art on deterministic and stochastic network calculus prior to 2000. Since 2000, stochastic network calculus has continued to be viewed as a challenging research topic of active interest in the networking community.

Of interest here is the class of traffic processes having exponentially bounded burstiness (EBB) proposed by Yaron and Sidi in [4], which in turn was heavily influenced by the (σ, ρ)-traffic characterization and network calculus of Cruz. The EBB concept has been applied to the analysis of end-to-end network delay, e.g., in [5]. The basic idea of EBB is to stochastically bound the input traffic to a network element using an exponential function. This in turn leads to an exponential bound on the tail distribution of the queue workload distribution, as well as a probabilistic bound on the output traffic. The input-output relation is the basis for a stochastic network calculus to compute end-to-end network performance bounds. In a later work, Starobinski and Sidi [6] extended the EBB concept to the so-called stochastically bounded burstiness (SBB), whereby input traffic is bounded by a function that satisfies fairly general conditions, of which the exponential function is a special case. In this paper we specialize the SBB bound of [6] to functions that stem from phase-type distributions. The family of phase-type distributions is closed under mixture and convolution operations, and hence includes the exponential mixture bound used in [6] as a particular case. We develop the phase-type based bound and demonstrate its performance in a preliminary numerical study.

The remainder of the paper is organized as follows. In Section II, we provide a brief summary of the concepts of EBB and SBB and the associated stochastic network calculus. In Section III, we review properties of the phase-type distribution and specialize the SBB bounds to phase-type bounds. In Section IV, we provide a numerical example for a queue with Markov modulated Poisson process input traffic model. The paper is concluded with further comments in Section V.

II. Stochastic Network Delay Bounds

In this section, we review the EBB concept in [4] and its generalization to SBB in [6].[2] These approaches to bounding a traffic process provide the basis of a stochastic network calculus for deriving end-to-end network delay bounds.

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

[2]As in the original papers, we shall, with minor abuse of grammar, use the abbreviations EBB and SBB both as nouns and adjectives depending the context.

978-1-5386-0579-0/18/$31.00 ©2018 IEEE

A. Stochastically Bounded Burstiness

A stochastic process $W = \{W(t) : t \geq 0\}$ is called exponentially bounded (EB) if there exist $\alpha \geq 0$ and $A \in [0, 1]$ such that

$$\mathsf{P}\{W(t) \geq \sigma\} \leq Ae^{-\alpha\sigma}, \tag{1}$$

for all $t \geq 0$ and all $\sigma \geq 0$, A traffic process with instantaneous rate process $R = \{R(t) : t \geq 0\}$ is said to be EBB with upper rate ρ if there exist $\alpha \geq 0$ and $A \in [0, 1]$ such that

$$\mathsf{P}\left\{\int_s^t R(t)\,\mathrm{d}t \geq \rho(t-s) + \sigma\right\} \leq Ae^{-\alpha\sigma}, \tag{2}$$

for all $t \geq s \geq 0$ and all $\sigma \geq 0$. For a discrete-time traffic process $\{R_k : k = 0, 1, \ldots\}$, essentially the same definition of EBB applies, except that s and t are nonnegative integers and the integral is replaced by a summation:

$$\mathsf{P}\left\{\sum_{u=s+1}^{t} R_u \geq \rho(t-s) + \sigma\right\} \leq Ae^{-\alpha\sigma}. \tag{3}$$

In this paper, we shall focus on the continuous-time case. Analogous results could be developed for the discrete-time case.

For some traffic models, the exponential bound of EBB can be quite loose. This bound was extended in [6] to employ a general bounding function. A stochastic process $W(t)$ is called stochastically bounded (SB) if, for all $t \geq 0$ and all $\sigma \geq 0$

$$\mathsf{P}\{W(t) \ \geq \sigma\} \leq f(\sigma), \tag{4}$$

where $f(\sigma) \in \mathcal{F}$, and $\mathcal{F}$ 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. For example, if $f(\sigma) = e^{-\alpha\sigma}$, then $(\int_\sigma^\infty \mathrm{d}u)^n f(u) = (1/\alpha)^n e^{-\alpha\sigma}$, which will be bounded if $\alpha > 0$. A traffic process with instantaneous rate process $R(t)$ is called SBB with upper rate ρ and bounding function $f(\sigma)$ if, for all $t \geq s \geq 0$ and all $\sigma \geq 0$,

$$\mathsf{P}\left\{\int_s^t R(t)\,\mathrm{d}t \geq \rho(t-s) + \sigma\right\} \leq f(\sigma). \tag{5}$$

B. Stochastic Network Calculus

We consider a network model that starts at $t = 0$ and all the network queues are empty at that time. Buffers are assumed to be infinite. The network is assumed to be a work-conserving system, which means that in every element of the network, no work is created or destroyed, and the server of the element never idles in the presence of a non-empty queue. In [6], several important results are established that can be used to develop a network calculus for assessing network delays using probabilistic bounds.

- The *SBB Characterization Theorem* [6, Theorem 1] considers a work-conserving system that transmits at a rate of ρ, fed with a traffic stream with rate process $R(t)$. If $W(t)$, the queue workload at time t, is SB with bounding function $f(\sigma)$ then the input traffic stream will be SBB with the same bounding function $f(\sigma)$ and upper rate ρ.
- The *SBB Sum Theorem* [6, Theorem 2] states that when two SBB traffic streams $R_1(t)$ and $R_2(t)$ with bounding functions $f_1(\sigma)$ and $f_2(\sigma)$ and upper rates ρ_1 and ρ_2 are fed into a network element with constant service rate, the aggregate traffic rate process $R_1(t) + R_2(t)$ will also be SBB with upper rate $\rho_1 + \rho_2$ and bounding function $g(\sigma) = f_1(p\sigma) + f_2((1-p)\sigma)$, where p is any value such that $0 < p < 1$.
- The *SBB Input-Output Relation Theorem* [6, Theorem 3] considers a traffic rate process $R_{\mathrm{in}}(t)$ fed as input to a work-conserving network element that transmits at rate C. If $R_{\mathrm{in}}(t)$ is SBB with upper rate $\rho < C$ and bounding function $f(\sigma)$, then the queue workload process $W(t)$ and the output rate process $R_{\mathrm{out}}(t)$ have the following properties:
  i) $R_{\mathrm{out}}(t)$ is SBB with upper rate ρ and bounding function
  $$g(\sigma) = f(\sigma) + \frac{1}{C-\rho}\int_\sigma^\infty f(u)\,\mathrm{d}u, \tag{6}$$
  ii) $W(t)$ is SB with bounding function
  $$g(\sigma) = f(\sigma) + \frac{1}{C-\rho}\int_\sigma^\infty f(u)\,\mathrm{d}u. \tag{7}$$

By the Sum Theorem, if the individual inputs to different nodes of a network are SBB, their aggregated input stream will also be SBB. Then by the Input-Output Relation Theorem, $W(t)$ and $R_{\mathrm{out}}(t)$ of these nodes will be SB and SBB, respectively. Following the same steps, we can extend this further to other nodes, and eventually to the entire network. Thus, if the input traffic streams to the network can be characterized as being SBB, then the traffic streams in all links of the network and the queue workloads at all network elements can be characterized as being SBB and SB, respectively.

III. Phase-type Network Delay Bounds

In this section, we develop phase-type network delay bounds based on the SBB calculus in [6]. The phase-type bounds provide a useful specialization of the SBB bounds in [6]. 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 [7, Theorem 4.2] [8, 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. We use properties of phase-type random variables to relate bounds on the input traffic to a network element to bounds on the queue workload as well as bounds on the output traffic, as is usually done in network calculus [3], [4], [6].

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 [9]

$$\begin{pmatrix} \mathbf{Q} & \mathbf{q} \\ \mathbf{0} & 0 \end{pmatrix}, \tag{8}$$

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 $\mathbf{q} = \mathrm{col}(q_1,\ldots,q_n)$ such that q_i is the transition rate from transient state i to the absorbing state $n+1$. The submatrix $\mathbf{Q}$ is invertible and the vector $\mathbf{q}$ is related to $\mathbf{Q}$ as follows:

$$\mathbf{q} = -\mathbf{Q}\mathbf{1}, \tag{9}$$

where $\mathbf{1}$ denotes a column vector of ones of the appropriate dimension, which is n in this case. 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 phase-type with parameter $(\boldsymbol{\pi}, \mathbf{Q})$:

$$\tau \sim \mathrm{PH}_n(\boldsymbol{\pi}, \mathbf{Q}). \tag{10}$$

In this case, the cumulative distribution function and survival function of τ are given, respectively, by

$$F_\tau(t) = 1 - \boldsymbol{\pi} e^{\mathbf{Q}t}\mathbf{1}, \quad t \geq 0 \tag{11}$$

$$S_\tau(t) = \mathsf{P}(\tau > t) = 1 - F_\tau(t) = \boldsymbol{\pi} e^{\mathbf{Q}t}\mathbf{1}, \quad t \geq 0. \tag{12}$$

The Laplace transform of τ is given by

$$M_\tau(s) := \mathsf{E}\left\{e^{-s\tau}\right\} = \pi_{n+1} + \boldsymbol{\pi}[sI - \mathbf{Q}]^{-1}\mathbf{q}, \tag{13}$$

where I denotes an identity matrix of appropriate dimension, in this case $n \times n$. The expected value of the phase-type random variable τ is given by

$$\mathsf{E}\{\tau\} = -\boldsymbol{\pi}\mathbf{Q}^{-1}\mathbf{1}. \tag{14}$$

The transition probabilities among the transient states of X are given by

$$\mathsf{P}(X(t) = j, \tau > t \mid X(0) = i) = \left[e^{\mathbf{Q}t}\right]_{ij}, \tag{15}$$

where $i,j \in \{1,2,\ldots,n\}$. As the states $1,2,\ldots,n$ are transient, we have

$$\lim_{t\to\infty} \left[e^{\mathbf{Q}t}\right]_{ij} = 0, \tag{16}$$

The family of phase-type distributions is closed under convolution and mixture operations (see [9], Theorems 3.1.26 and 3.1.27, respectively). Suppose, for example, that $\tau_1 \sim \mathrm{PH}_n(\boldsymbol{\alpha}, \mathbf{G})$ and $\tau_2 \sim \mathrm{PH}_m(\boldsymbol{\beta}, \mathbf{H})$ and τ_1 and τ_2 are independent. Then $\tau_{\mathrm{sum}} = \tau_1 + \tau_2$ is a phase-type random variable with $n+m$ transient states such that

$$\tau_{\mathrm{sum}} \sim \mathrm{PH}_{m+n}\left((\boldsymbol{\alpha}, \mathbf{0}), \begin{pmatrix} \mathbf{G} & \mathbf{g}\boldsymbol{\beta} \\ \mathbf{0} & \mathbf{H} \end{pmatrix}\right), \tag{17}$$

where $\mathbf{g} = -\mathbf{G}\mathbf{1}$. Thus, if $X_1, X_2, \ldots, X_n$ are independent exponential random variables with $X_i \sim \exp(\lambda_i)$, $i = 1,\ldots,n$, then the distribution of the sum $\tau = X_1 + X_2 + \ldots + X_n$ is given by $\tau \sim \mathrm{PH}_n(\boldsymbol{\pi}, \mathbf{Q})$ where

$$\mathbf{Q} = \begin{pmatrix} -\lambda_1 & \lambda_1 & 0 & 0 & \cdots & 0 \\ 0 & -\lambda_2 & \lambda_2 & 0 & \cdots & 0 \\ 0 & 0 & -\lambda_3 & \lambda_3 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & \lambda_{n-1} \\ 0 & 0 & 0 & 0 & \cdots & -\lambda_n \end{pmatrix}, \tag{18}$$

$\boldsymbol{\pi} = (1, 0, 0, \cdots, 0)$.

Next consider a mixture of phase-type distributions defined by

$$\tau_{\mathrm{mix}} = \begin{cases} \tau_1 & \text{with probability } p, \\ \tau_2 & \text{with probability } 1-p, \end{cases}$$

where $p \in [0,1]$. Then

$$\tau_{\mathrm{mix}} \sim \mathrm{PH}_{n+m}\left((p\boldsymbol{\alpha}, (1-p)\boldsymbol{\beta}), \begin{pmatrix} \mathbf{G} & \mathbf{0} \\ \mathbf{0} & \mathbf{H} \end{pmatrix}\right). \tag{19}$$

In particular, if τ is a random variable such that with probability π_i, τ is exponentially distributed with parameter λ_i for $i = 1,\ldots,n$, then $\tau \sim \mathrm{PH}_n(\boldsymbol{\pi}, \mathbf{Q})$, where

$$\mathbf{Q} = \mathrm{diag}\{-\lambda_1, -\lambda_2, \ldots, -\lambda_n\}, \tag{20}$$

$$\boldsymbol{\pi} = (\pi_1, \pi_2, \cdots, \pi_n). \tag{21}$$

B. Phase-type Bounded Burstiness

Now we specialize the general SBB concept to bounds based on phase-type distributions. As mentioned earlier, phase-type distributions are dense in the set of densities with non-negative support and every density function in this set can be approximated arbitrarily well by a phase-type distribution [7, Theorem 4.2] [8, Theorem 5.2].

Definition 1. A stochastic process $W(t)$ is phase-type bounded (PHB) with bounding parameter $(\boldsymbol{\pi}, \mathbf{Q}, A)$ if $A \in [0,1]$ and $(\boldsymbol{\pi}, \mathbf{Q})$ is the parameter of a phase-type random variable such that

$$\mathsf{P}\{W(t) \ \geq \sigma\} \leq A\boldsymbol{\pi} e^{\mathbf{Q}\sigma}\mathbf{1}, \tag{22}$$

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

Definition 2. A continuous-time traffic stream with traffic rate process $R(t)$ has phase-type bounded burstiness (PHBB) with upper rate ρ and bounding parameter $(\boldsymbol{\pi}, \mathbf{Q}, A)$ if

$$\mathsf{P}\left\{\int_s^t R(t)\,\mathrm{d}t \geq \rho(t-s) + \sigma\right\} \leq A\boldsymbol{\pi} e^{\mathbf{Q}\sigma}\mathbf{1}, \tag{23}$$

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

Next, we show that phase-type bounding functions belong to the family of functions $\mathcal{F}$ defined immediately after (4).

Theorem 1. Let $(\boldsymbol{\pi}, \mathbf{Q}, A)$ be a phase-type bounding parameter. Then $f(\sigma) = A\boldsymbol{\pi} e^{\mathbf{Q}\sigma}\mathbf{1}$ is monotonically decreasing and $f \in \mathcal{F}$.

Proof. Since $(\boldsymbol{\pi}, \mathbf{Q})$ is the parameter of a phase-type distribution, the function $S(\sigma) = \boldsymbol{\pi} e^{\mathbf{Q}\sigma}\mathbf{1}$ is the associated survival function, which by definition is monotonically decreasing.

Therefore, $f(\sigma)$ is a monotonically decreasing function. To show that $f \in \mathcal{F}$, we need to show that $(\int_\sigma^\infty \mathrm{d}u)^n f(u)$ is bounded. For the phase-type bounding function we have

$$\int_\sigma^\infty A\boldsymbol{\pi} e^{\mathbf{Q}u}\mathbf{1}\mathrm{d}u = A\boldsymbol{\pi}\int_\sigma^\infty e^{\mathbf{Q}u}\,\mathrm{d}u\;\mathbf{1} \tag{24}$$

$$= A\boldsymbol{\pi}\left[\mathbf{Q}^{-1}e^{\mathbf{Q}u}\right]_\sigma^\infty\;\mathbf{1} = -A\boldsymbol{\pi}\mathbf{Q}^{-1}e^{\mathbf{Q}\sigma}\mathbf{1}. \tag{25}$$

From (16), $\lim_{u\to\infty} e^{\mathbf{Q}u} = \mathbf{0}$. Hence, the right-hand side of (25) is bounded. Repeating this argument $n-1$ more times shows that $(\int_\sigma^\infty \mathrm{d}u)^n f(u)$ is bounded. □

Theorem 2 (*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 $R(t)$. Let $W(t)$ be the workload in the system at time t. If $W(t)$ is PHB with parameter $(\boldsymbol{\pi}, \mathbf{Q}, A)$ then $R(t)$ is PHBB with upper rate ρ and bounding parameter $(\boldsymbol{\pi}, \mathbf{Q}, A)$.

Proof. The result follows from [6, Theorem 1], where the bounding function is given by $f(\sigma) = A\boldsymbol{\pi}e^{\mathbf{Q}\sigma}\mathbf{1}$. □

Theorem 3 (*Sum*). Let $R_1(t)$ be PHBB with upper rate ρ_1 and bounding parameter $(\boldsymbol{\alpha}, \mathbf{G}, A_1)$, and $R_2(t)$ be PHBB with upper rate ρ_2 and bounding parameter $(\boldsymbol{\beta}, \mathbf{H}, A_2)$. Then $R_1(t) + R_2(t)$ is PHBB with upper rate $\rho = \rho_1 + \rho_2$ and bounding parameter $(\boldsymbol{\pi}, \mathbf{Q}, A)$ where $A = A_1 + A_2$,

$$\boldsymbol{\pi} = \left[\frac{A_1}{A}\boldsymbol{\alpha}, \frac{A_2}{A}\boldsymbol{\beta}\right], \quad \mathbf{Q} = \begin{pmatrix} p\mathbf{G} & \mathbf{0} \\ \mathbf{0} & (1-p)\mathbf{H}\end{pmatrix}, \tag{26}$$

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

Proof. As $R_1(t)$ and $R_2(t)$ are PHBB, a special case of SBB, we can apply the Sum theorem for SBB [6, Theorem 2]. 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

$$f_1(\sigma) = A_1\boldsymbol{\alpha}e^{\mathbf{G}\sigma}\mathbf{1}, \quad f_2(\sigma) = A_2\boldsymbol{\beta}e^{\mathbf{H}\sigma}\mathbf{1}. \tag{27}$$

We have

$$g(\sigma) = A_1\boldsymbol{\alpha}e^{p\mathbf{G}\sigma}\mathbf{1} + A_2\boldsymbol{\beta}e^{(1-p)\mathbf{H}\sigma}\mathbf{1} \tag{28}$$

$$= \left[A_1\boldsymbol{\alpha}, A_2\boldsymbol{\beta}\right]\begin{pmatrix} e^{p\mathbf{G}} & \mathbf{0} \\ \mathbf{0} & e^{(1-p)\mathbf{H}}\end{pmatrix}\mathbf{1}, \tag{29}$$

from which the result follows. □

Theorem 4 (*Input-Output Relation*). Let $R_{\mathrm{in}}(t)$ be the input traffic rate process to a work-conserving element, which transmits at rate C. Suppose that $R_{\mathrm{in}}(t)$ is PHBB with upper rate $\rho < C$ and bounding parameter $(\boldsymbol{\pi}, \mathbf{Q}, A)$. Let $W(t)$ denote the queue workload process and let $R_{\mathrm{out}}(t)$ denote the output traffic rate process. Then the following hold:

1) $W(t)$ is PHB with bounding parameter

$$\left(\frac{\boldsymbol{\pi}(C-\rho) - \boldsymbol{\pi}\mathbf{Q}^{-1}}{\mathsf{E}\{\tau\} + C - \rho}, \mathbf{Q}, \frac{A(C-\rho+\mathsf{E}\{\tau\})}{C-\rho}\right), \tag{30}$$

where $\mathsf{E}\{\tau\} = -\boldsymbol{\pi}\mathbf{Q}^{-1}\mathbf{1}$ is the mean of phase-type random variable $\tau \sim \mathrm{PH}(\boldsymbol{\pi}, \mathbf{Q})$.

2) $R_{\mathrm{out}}(t)$ is PHBB with upper rate ρ and bounding parameter as given in (30).

Proof. 1) Since $R_{\mathrm{in}}(t)$ is PHBB with upper rate $\rho < C$, we can apply the general SBB input-output relation theorem given in [6, Theorem 3]. In this case, $W(t)$ will be bounded with bounding function $g(\sigma) = f(\sigma) + \frac{1}{C-\rho}\int_\sigma^\infty f(u)\,\mathrm{d}u$, where $f(\sigma) = A\boldsymbol{\pi}e^{\mathbf{Q}\sigma}\mathbf{1}$. We have

$$g(\sigma) = A\boldsymbol{\pi}e^{\mathbf{Q}\sigma}\mathbf{1} - \frac{A\boldsymbol{\pi}\mathbf{Q}^{-1}e^{\mathbf{Q}\sigma}\mathbf{1}}{C-\rho} \tag{31}$$

$$= A\left[\boldsymbol{\pi} - \frac{\boldsymbol{\pi}\mathbf{Q}^{-1}}{C-\rho}\right]e^{\mathbf{Q}\sigma}\mathbf{1} \tag{32}$$

$$= \frac{A(C-\rho+\mathsf{E}\{\tau\})}{C-\rho}\left[\frac{\boldsymbol{\pi}(C-\rho) - \boldsymbol{\pi}\mathbf{Q}^{-1}}{\mathsf{E}\{\tau\} + C - \rho}\right]e^{\mathbf{Q}\sigma}\mathbf{1}. \tag{33}$$

The factor in square brackets represents a probability distribution since

$$\frac{\boldsymbol{\pi}(C-\rho) - \boldsymbol{\pi}\mathbf{Q}^{-1}}{E\{\tau\} + C - \rho}\cdot\mathbf{1} = \frac{C - \rho - \boldsymbol{\pi}\mathbf{Q}^{-1}\mathbf{1}}{\mathsf{E}\{\tau\} + C - \rho} = 1,$$

where we have used (14). Therefore, $g(\sigma)$ is a phase-type bounding function for the output traffic rate process.

2) Since $R_{\mathrm{in}}(t)$ is PHBB, following the same argument as above we can establish that $R_{\mathrm{out}}(t)$ is bounded with upper rate ρ and bounding function $g(\sigma) = f(\sigma) + \frac{1}{C-\rho}\int_\sigma^\infty f(u)\,\mathrm{d}u$, where $f(\sigma) =$ is $\mathrm{PH}_n(\boldsymbol{\pi}, \mathbf{T}, A)$. The proof relies on the facts that $\int_\sigma^\infty f(u)\,\mathrm{d}u \in \mathcal{F}$ and $f(\sigma)$ is a decreasing function of σ, which are established in Theorem 1.

□

IV. Case Study

In this section, we apply the PHBB bounds to a network element fed by a Markov modulated Poisson process (MMPP) as traffic input and compare with the EBB bounds.

A. MMPP/G/1 queue

The MMPP is a common model for traffic with a high degree of burstiness [10]. A 2-state MMPP is parameterized by an arrival matrix

$$\boldsymbol{\Lambda} = \begin{bmatrix}\lambda_0 & 0 \\ 0 & \lambda_1\end{bmatrix} \tag{34}$$

and a rate matrix

$$\mathbf{R} = \begin{bmatrix}-r_0 & r_0 \\ r_1 & -r_1\end{bmatrix}, \tag{35}$$

which is the generator of the modulating Markov chain. When the service times are independent and generally distributed, the resulting queue is denoted as MMPP/$G/1$. A relatively simple form for the Laplace transform of the virtual waiting time of a two-state MMPP/$G/1$ queue is given in [11] in terms of a transition probability matrix

$$\mathbf{D} = [d_{ij} : i,j = 0,1] = \begin{bmatrix}1-d_0 & d_0 \\ d_1 & 1-d_1\end{bmatrix}, \tag{36}$$

where d_{ij} is the probability that a busy period starting in the underlying state i ends in underlying state j. The virtual

waiting time in an MMPP/G/1 queue is equivalent to the queue workload or buffer content in a system with a constant rate server where packets arrives according to an MMPP with identically distributed, independent packet lengths drawn from a common distribution. Let $H(s)$ denote the Laplace transform of the packet length, which is equivalent to the packet service time, assuming that the service rate is one unit (e.g., bit) per unit time (e.g., seconds). The matrix $\mathbf{D}$ is determined by numerically solving the following equations [11, Eqs. (86), (87)]:

$$d_0 + d_1 = 1 - H(r_0 + r_1 + \lambda_0 d_0 + \lambda_1 d_1) \quad (37)$$

$$d_0(r_1 + \lambda_1 d_1) = d_1(r_0 + \lambda_0 d_0) \quad (38)$$

Let $\mathbf{f} = (f_0, f_1)$ denote the steady-state distribution vector associated with $\mathbf{D}$, satisfying

$$\mathbf{fD} = \mathbf{f}, \quad \mathbf{f1} = 1. \quad (39)$$

The Laplace transform of the queue workload is given as follows [11]:

$$W(s) = \frac{N(s)}{D(s)}, \quad (40)$$

where

$$N(s) = s(1-\rho)[s - r_0 - r_1 + (H(s)-1)(f_0\lambda_1 + f_1\lambda_0)] \quad (41)$$

$$D(s) = s^2 + [(H(s)-1)(\lambda_0 + \lambda_1) - (r_0 + r_1)]s + (H(s)-1)[(H(s)-1)\lambda_0\lambda_1 - r_0\lambda_1 - r_1\lambda_0]. \quad (42)$$

B. Numerical Examples

We shall consider queues with MMPP input traffic model and two service time distributions: exponential and Erlang-2. The corresponding queues are denoted by MMPP/M/1 and MMPP/E_2/1, respectively. In [6], a numerical example based on a discrete-time queue with Markov modulated Bernoulli process input traffic model was considered. In that model, the queue workload distribution was a mixture of exponential distributions.

1) MMPP/M/1 *Queue:* In the case of exponential service, the Laplace transform of the packet length distribution is given by

$$H(s) = \frac{\mu}{s + \mu}. \quad (43)$$

Let λ_{avg} denote the average arrival rate to the queue. Then the queue utilization is given by

$$\rho = \frac{\lambda_{\text{avg}}}{\mu} = \frac{r_1}{\mu(r_0 + r_1)}\lambda_0 + \frac{r_0}{\mu(r_0 + r_1)}\lambda_1. \quad (44)$$

For our numerical example, we set the parameters as follows: $r_0 = 2$, $r_1 = 10^{-2}$, $\lambda_0 = 12$, $\lambda_1 = 3$, $\mu = 4$. Using (44), we compute $\rho = 0.7612$. Applying the result (40) and inverting the Laplace transform, the density of the queue workload process is obtained as follows:

$$f_W(\sigma) = 0.56e^{-1.07\sigma} + 0.159e^{-0.669\sigma} + 0.239\delta(\sigma), \quad \sigma \geq 0, \quad (45)$$

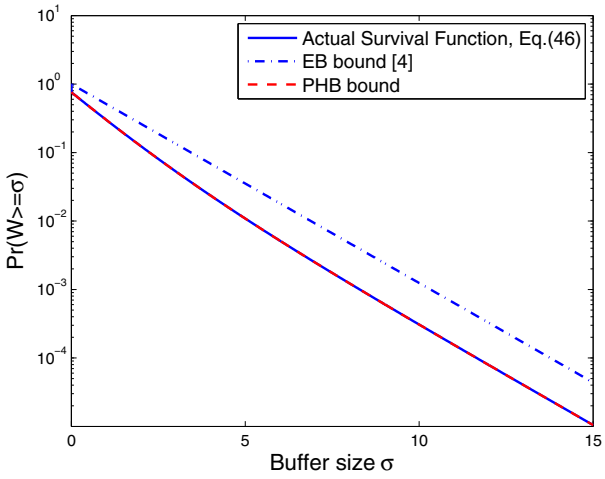

Fig. 1. EB and PHB bounds for MMPP/M/1 queue.

where $\delta(\sigma)$ denotes the Dirac delta function. The delta function in the density function implies a discontinuity in the distribution and survival functions at $\sigma = 0$. Since we are interested in tail probabilities, we shall only consider the case $\sigma > 0$.

The survival function for the queue workload process $W(t)$ is then given by

$$\mathsf{P}\{W(t) \geq \sigma\} = 0.5234e^{-1.07\sigma} + 0.2377e^{-0.669\sigma}, \quad (46)$$

for $t \geq 0$, $\sigma > 0$. Based on the (46), we obtain an EB bound as follows:

$$\mathsf{P}\{W(t) \geq \sigma\} \leq e^{-0.669\sigma}, \quad t \geq 0, \sigma > 0. \quad (47)$$

Since a mixture of exponential distributions is a special case of the phase-type distribution, the right-hand side of (46) can be used directly to obtain the phase-type bound. In this case, the PHB bound can be chosen to match (46), i.e., $W(t)$ is PHB with bounding parameters $\boldsymbol{\pi} = (0.6877, 0.3123)$,

$$\mathbf{Q} = \begin{bmatrix} -1.07 & 0 \\ 0 & -0.669 \end{bmatrix}, \quad (48)$$

and $A = 0.7611$. The MMPP/M/1 model is analogous to the discrete-time queueing model studied in [6], which resulted in an SBB bound in the form of a mixture of exponentials. The EB and the PHB (equal to the exact result) are shown in semi-log scale in Figure 1.

2) MMPP/E_2/1 *Queue:* When the packet length has an Erlang-2 distribution, we have

$$H(s) = \left(\frac{2\mu}{s + 2\mu}\right)^2, \quad (49)$$

where $\frac{1}{2\mu}$ is the mean of each one of two independent exponential random variables, which are added together to form the Erlang-2 random variable. The queue utilization in this case is also given by (44). The parameters are set as follows: $r_0 = 2$, $r_1 = 10^{-3}$, $\lambda_0 = 20$, $\lambda_1 = 3$, $\mu = 4$.

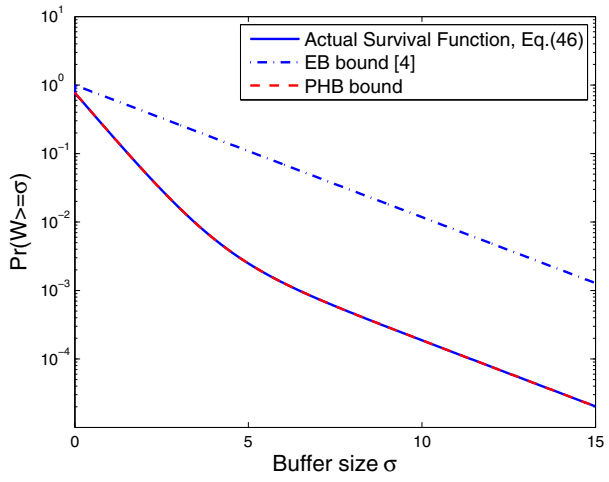

Fig. 2. EB bound, PH bound, and true tail probability for MMPP/$E_2/1$ queue.

Applying (40), we obtain the following density function for the queue workload:

$$f_W(\sigma) = 1.05e^{-1.38\sigma} - 0.318e^{-11.6\sigma} - 4.65 \cdot 10^{-5}e^{-14\sigma} + 0.007e^{-0.444\sigma} + 0.248\delta(\sigma), \tag{50}$$

for $\sigma \geq 0$. The survival function of the queue workload for $\sigma > 0$ is then given by

$$\mathsf{P}\{W(t) \geq \sigma\} = 0.7609e^{-1.38\sigma} - 0.0274e^{-11.6\sigma} - 0.3321 \cdot 10^{-5}e^{-14\sigma} + 0.0158e^{-0.444\sigma}. \tag{51}$$

Note that $W(t)$ is not a mixture of exponentials, nor a phase-type distribution due to the negative coefficients on the right-hand side of (51). Nevertheless $W(t)$ has a matrix exponential distribution [9] and can be bounded using EB and phase-type bounds. For the EB bound we have

$$\mathsf{P}\{W(t) \geq \sigma\} \leq e^{-0.444\sigma}, \quad t \geq 0, \ \sigma > 0. \tag{52}$$

The phase-type bound can be obtained by simply dropping the negative terms on the right-hand side of (51), i.e.,

$$\mathsf{P}\{W(t) \geq \sigma\} \leq 0.7609e^{-1.38\sigma} + 0.0158e^{-0.444\sigma}. \tag{53}$$

This bound is a mixture of exponentials, which is equivalent to the SBB bound for the example of a discrete-time queue considered in [6]. In this case $W(t)$ is PHB with bounding parameter $(\boldsymbol{\pi}, \mathbf{Q}, A)$ given by $\boldsymbol{\pi} = (0.9806, 0.0194)$, and

$$\mathbf{Q} = \begin{bmatrix} -1.38 & 0 \\ 0 & -0.444 \end{bmatrix}, \tag{54}$$

and $A = 0.7760$. In Figure 2 the EB and PHB curves are shown along with the true tail probability curve. The PHB curve lines up closely with the true tail probability, whereas the EB provides a loose bound. We remark that an example of a queueing model different from the MMPP/$E_2/1$, for example, a heavy-tailed queue as in [12], would be needed to highlight the potential benefit of the PHB bound vs. the mixture of exponentials bound considered in [6].

V. Conclusion

We proposed the use of phase-type distributions to specialize the general bounding function in the SBB traffic burstiness bounding framework [6]. Phase-type distributions generalize an earlier proposed bounding function in the form of a mixture distribution [6]. The use of phase-type distribution can potentially lead to tighter bounds. We are currently studying this aspect of the proposed bound and examples to realize this potential. In this paper we proved that the bounding function that is based on phase-type distributions is admissible and we demonstrated the bounds for an example of an MMPP/$G/1$ queue using results from [11].

References

[1] R. L. Cruz, "A calculus for network delay, Part I: Network Elements in isolation," *IEEE Transactions on Information Theory*, vol. 37, no. 1, pp. 132–141, 1991.

[2] R. L. Cruz, "A calculus for network delay, Parts II: Network Analysis," *IEEE Transactions on Information Theory*, vol. 37, no. 1, pp. 132–141, 1991.

[3] C. S. Chang, *Performance Guarantees in Communication Networks*. Springer, 2000.

[4] O. Yaron and M. Sidi, "Performance and Stability of Communication Networks via Robust Exponential Bounds," *IEEE/ACM Transactions on Networking*, vol. 1, no. 3, pp. 372–385, 1993.

[5] J. Liebeherr, Y. Ghiassi-Farrokhfal, and A. Burchard, "On the impact of link scheduling on end-to-end delays in large networks," *IEEE Journal on Selected Areas in Communications, Special Issue on Trading Rate for Delay at the Transport and Application Layers*, vol. 29, pp. 1009–1020, May 2011.

[6] D. Starobinski and M. Sidi, "Stochastically bounded burstiness for communication networks," *IEEE Transactions on Information Theory*, vol. 46, no. 1, pp. 206–212, 2000.

[7] S. Asmussen, *Applied Probability and Queues*. New York: Springer-Verlag, 2nd ed., 2003.

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

[9] M. Bladt and B. Nielsen, *Matrix-Exponential Distributions in Applied Probability*. Springer, 2017.

[10] W. Fischer and K. Meier-Hellstern, "The Markov-modulated Poisson process (MMPP) cookbook," *Performance Evaluation*, vol. 18, no. 2, pp. 149–171, 1993.

[11] 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.

[12] O. J. Boxma and J. W. Cohen, "The M/G/1 queue with heavy-tailed service time distribution," *IEEE Journal on Selected Areas in Communications*, vol. 16, pp. 749–763, June 1998.