Traffic Workload Envelope for Network Performance Guarantees with Multiplexing Gain

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

***Abstract*—Stochastic network calculus involves the use of a traffic bound or envelope to make admission control and resource allocation decisions for providing end-to-end quality-of-service guarantees. To apply network calculus in practice, the traffic envelope should: (i) be readily determined for an arbitrary traffic source, (ii) be enforceable by traffic regulation, and (iii) yield statistical multiplexing gain. Existing traffic envelopes typically satisfy at most two of these properties. A well-known traffic envelope based on the moment generating function (MGF) of the arrival process satisfies only the third property. We propose a new traffic envelope based on the MGF of the workload process obtained from offering the traffic to a constant service rate queue. We show that this traffic workload envelope can achieve all three properties and leads to a framework for a network service that provides stochastic delay guarantees. We demonstrate the performance of the traffic workload envelope with two bursty traffic models: Markov on-off fluid and Markov modulated Poisson Process (MMPP).**

***Index Terms*—Traffic envelope; quality-of-service; network calculus; statistical multiplexing; traffic regulator; admission control.**

I. Introduction

Provisioning quality-of-service (QoS) guarantees in a resource-efficient manner remains a challenging, yet important open problem for future networking. To this day, the Internet does not have the capability to provide end-to-end delay guarantees, which are desirable for time-critical multimedia applications. The packet-switching paradigm of the Internet enables very efficient network resource utilization due to statistical multiplexing, but it also makes provisioning for performance guarantees extremely challenging.

Stochastic network calculus is a theoretical framework in which traffic sources are characterized by a bounds or envelopes, and the traffic envelopes are used to obtain end-to-end QoS guarantees in the form of stochastic delay bounds. Application of stochastic network calculus to practical networks would require the following: 1) a means to characterize a given traffic source by a suitable traffic envelope; 2) a method to enforce conformance of a traffic flow to a given traffic envelope; 3) an admission control scheme based on traffic envelopes, which is capable of achieving statistical multiplexing gain. To our knowledge, such a traffic envelope satisfying all three of these requirements has been lacking in the networking literature.

Various traffic envelopes have been proposed in conjunction with stochastic network calculus. The well-known traffic arrival envelope [1], which we refer to as the A-envelope, imposes a bound on the moment generating function (MGF) of the traffic arrival process. An important feature of the A-envelope, in contrast to other traffic envelopes, such as stochastically bounded burstiness (SBB) and its variants [2], [3], is that statistical multiplexing gain can be achieved for a large number of independent flows. We note that SBB provides a bound on the tail probability of the arrival process. On the other hand, characterization of an arbitrary traffic source by an A-envelope is not straightforward since the sample paths of the arrival process increase without bound. For a similar reason, the A-envelope is not amenable to traffic regulation. Thus, the A-envelope fails as a practical traffic bound with respect to the first two requirements given above.

The main contribution of this paper is a new traffic envelope, which we refer to as the W-envelope, which satisfies the three stated requirements for practical application of stochastic network calculus to achieve end-to-end quality-of-service. The W-envelope is a bound on the MGF of the workload process that results from offering the traffic to a constant rate server. The W-envelope is closely related to the generalized stochastically bounded burstiness (gSBB) [4], which provides a bound on the tail probability of the workload process. The stationarity and ergodicity of the workload process facilitates characterization of an arbitrary traffic source in terms of a gSBB envelope [5]. The traffic regulator developed in [6] forces conformance of a traffic flow to a given gSBB envelope. Via a relationship between the W-envelope and the gSBB envelope, the traffic characterization method and traffic regulator extend to the W-envelope. An important property of the W-envelope is that a W-envelope for the superposition of flows characterized by W-envelopes can be obtained as the sum of their respective W-envelopes. This provides the basis for a computationally efficient admission control scheme that can achieve statistical multiplexing gain.

The remainder of the paper is organized as follows. Section II provides relevant background on traffic bounds and network calculus. In Section III, we introduce the W-envelope and establish its key properties. In Section IV, we outline a framework, based on the W-envelope, for providing delay

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

978-1-6654-3540-6/22/$31.00 © 2022 IEEE

guarantees. The W-envelopes for two widely used traffic models are obtained in Section V and numerical examples are presented in Section VI. The paper is concluded in Section VII.

II. Background and Motivation

We shall refer to a traffic process $A(t)$, which represents the amount of traffic arriving in the interval $(0, t]$ for $t > 0$, with the initialization $A(0) = 0$. Define $A(\tau, t) = A(t) - A(\tau)$, which represents the amount of traffic arriving in the interval $(\tau, t]$ for $0 \leq \tau < t$, with $A(t, t) = 0$ for all $t \geq 0$. The time parameter t is assumed continuous, but our results are also applicable to the discrete-time case.

A. Stochastically Bounded Burstiness

In the exponentially bounded burstiness (EBB) concept of Yaron and Sidi [7], the tail distribution of the arrival process is bounded by an exponential function. The EBB envelope was later generalized to SBB by Starobinski and Sidi [2]:

Definition 1 (SBB)**.** *A traffic process $A(t)$ is said to be stochastically bounded bursty (SBB) with upper rate ρ and bounding function $f \in \mathcal{F}$ if, for all $0 \leq \tau \leq t$ and all $b \geq 0$,*

$$\mathsf{P}\left\{A(\tau, t) - \rho(t - \tau) \geq b\right\} \leq f(b), \tag{1}$$

$\mathcal{F}$ is the family of functions f such that for every $n, x \geq 0$, the n-fold integral $(\int_x^\infty \mathrm{d}u)^n f(u)$ is bounded.

A traffic process is EBB if it is SBB with a bounding function of the form $f(b) = \alpha e^{-\theta b}$, where $\alpha, \theta \geq 0$. The SBB envelope was extended in [3], [4] by the concept of *generalized* stochastically bounded burstiness (gSBB). Let

$$W_\rho(t) := \max_{0 \leq \tau \leq t} \left\{A(\tau, t) - \rho(t - \tau)\right\}, \tag{2}$$

denote the workload or backlog at time t of queue that has input traffic $A(t)$, a constant rate server with rate ρ, infinite buffer space, and is empty at time $t = 0$.

Definition 2 (gSBB)**.** *A traffic process $A(t)$ is generalized stochastically bounded bursty (gSSB) with upper rate ρ and bounding function $f \in \mathcal{BF}$ if, for all $t \geq 0$ and all $b \geq 0$,*

$$\mathsf{P}\left\{W_\rho(t) \geq b\right\} \leq f(b), \tag{3}$$

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

Significantly, gSBB is defined in terms of the workload process $W_\rho(t)$, rather than the arrival process $A(t)$. If the rate ρ exceeds the average arrival rate, the constant service rate queue will be stable and in this case, $W_\rho(t)$ will be a stationary and ergodic process. As noted in [4, Theorem 13], the process $W_\rho(t)$ exhibits stochastic ordering monotonicity[1]:

$$W_\rho(t_1) \leq_{\mathrm{st}} W_\rho(t_2) \leq_{\mathrm{st}} \cdots \leq_{\mathrm{st}} W_\rho(\infty), \tag{4}$$

for times $0 \leq t_1 \leq t_2 \leq \ldots$, where $W_\rho(\infty)$ denotes $W_\rho(t)$ in steady state, i.e., in the limit as $t \to \infty$. Thus, if (3) holds when $W_\rho(t)$ is in steady state, then it also holds for all $t \geq 0$ (see [5, Theorem 1]).

[1] $X \leq_{\mathrm{st}} Y$ means that $\mathsf{P}\{X > a\} \leq \mathsf{P}\{Y > a\}$ for any a.

A gSBB characterization for a superposition of independent gSBB traffic sources can be obtained by extending [4, Theorem 4] to N traffic sources.

Theorem 1. *Suppose $A_i(t)$ is gSBB with upper rate ρ_i and bounding function f_i, and $A_i(t)$, $1 \leq i \leq N$ are independent. Then the aggregate arrival process $A(t) = \sum_{i=1}^N A_i(t)$ is also gSBB with upper rate $\rho = \sum_{i=1}^N \rho_i$ and bounding function g defined as follows:*

$$g(x) = 1 - F_1 \bigstar F_2 \bigstar \cdots \bigstar F_N(x), \tag{5}$$

$$F_i(x) = 1 - f_i(x), \quad i = 1, \ldots, N, \tag{6}$$

where $\bigstar$ denotes Stieltjes convolution and is defined by

$$F_1 \bigstar F_2(x) := \int_0^x F_1(x - y) \mathrm{d}F_2(y). \tag{7}$$

B. Moment Generating Function Traffic Envelope

As the number of traffic sources N increases, computation of the N-fold Stieltjes convolution in (7) becomes impractical for real-time admission control. An alternative traffic bound based on the moment generating function (MGF) of the arrival process was proposed by Chang [1].

Definition 3. *A traffic process $A(t)$ has an MGF arrival envelope (A-envelope) given by $(\sigma(\theta), \rho(\theta))$ if for all $0 \leq \tau \leq t$,*

$$E\left[e^{\theta A(\tau, t)}\right] \leq e^{\theta[\rho(\theta)(t-\tau) + \sigma(\theta)]}, \tag{8}$$

where $\sigma(\theta)$ and $\rho(\theta)$ are nonnegative functions of $\theta \geq 0$.

The defining equation (8) can be written as

$$E\left[e^{\theta[A(\tau, t) - \rho(t-\tau)]}\right] \leq e^{\theta\sigma}. \tag{9}$$

By applying the Chernoff bound, one can show (cf. [8]) that a source with A-envelope $(\sigma(\theta), \rho(\theta))$ is EBB/SBB with upper rate $\rho(\theta)$ and bounding function $f(b) = e^{\theta\sigma(\theta)} e^{-\theta b}$ for all $\theta \geq 0$. The superposition of a set of independent flows has an A-envelope given by the sum of the A-envelopes of the individual flows [9, Lemma 7.3.1].

Theorem 2. *If $A_i(t)$, $1 \leq i \leq N$, are independent and have A-envelopes $(\sigma_i(\theta), \rho_i(\theta))$, then $A(t) = \sum_{i=1}^N A_i(t)$ has A-envelope $(\sigma(\theta), \rho(\theta))$ where $\sigma(\theta) = \sum_{i=1}^N \sigma_i(\theta)$ and $\rho(\theta) = \sum_{i=1}^N \rho_i(\theta)$.*

C. Statistical Multiplexing and Admission Control

Suppose that a set of statistically independent and identically distributed traffic flows is offered as input to a constant service rate queue (i.e., a multiplexer) of capacity C. We consider an admission control scheme, which admits traffic flows subject to a quality-of-service (QoS) constraint:

$$P\{D(t) > d\} < \epsilon, \tag{10}$$

where $D(t)$ represents the delay through the multiplexer, d is a delay threshold, ϵ is a small positive number. Via Theorem 2, the A-envelope provides the basis for a computationally simple admission control scheme that can achieve statistical

multiplexing gain for a given QoS constraint. By contrast, admission control via Theorem 1 based on the gSBB envelope requires convolution of the bounding functions, which is not scalable to large numbers of flows.

III. Workload-based Traffic Envelope

We now develop a new traffic envelope based on the MGF of the workload process $W_\rho(t)$ defined in (2).

A. Definition and basic properties

Definition 4. *A traffic process $A(t)$ has an MGF workload envelope or W-envelope $(\sigma(\theta), \rho(\theta))$ if for all $t \geq 0$,*

$$E\left[e^{\theta W_{\rho(\theta)}(t)}\right] \leq e^{\theta\sigma(\theta)}, \tag{11}$$

where $\sigma(\theta)$ and $\rho(\theta)$ are nonnegative functions of $\theta \geq 0$.

Similar to the gSBB envelope, the W-envelope is more suitable for traffic characterization and traffic regulation than the A-envelope because it is based on $W_\rho(t)$, which has a steady-state distribution. The monotonicity property (4) implies that if (11) holds when $W_\rho(t)$ is in steady state, then it also holds for all $t \geq 0$.

Theorem 3. *If $A(t)$ has a W-envelope $(\sigma(\theta), \rho(\theta))$ then it is also characterized by an A-envelope with the same parameter $(\sigma(\theta), \rho(\theta))$.*

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

$$A(\tau,t)-\rho(t-\tau) \leq \max_{0\leq\tau\leq t}[A(\tau,t)-\rho(t-\tau)] = W_\rho(t). \tag{12}$$

Therefore,

$$E[e^{\theta[A(\tau,t)-\rho(t-\tau)]}] \leq E[e^{\theta W_\rho(t)}], \tag{13}$$

and the result follows immediately. □

Theorem 3 implies that a traffic regulator that enforces a W-envelope with parameter $(\sigma(\theta), \rho(\theta))$ also enforces an A-envelope with the same parameter. Consider N traffic processes $A_1(t), \ldots, A_N(t)$. Assume each process $A_i(t)$ is characterized by a W-envelope of the form (11) specified by parameters $(\sigma_i(\theta), \rho_i(\theta))$, i.e.,

$$E\left[e^{\theta W_{\rho_i}(t)}\right] \leq e^{\theta[\rho_i(\theta)(t-\tau)+\sigma_i(\theta)]}, \qquad i = 1, \ldots, N. \tag{14}$$

The following inequality is useful in obtaining a W-envelope for the aggregate traffic process $A(t) = \sum_{i=1}^N A_i(t)$.

Lemma 1.

$$W_\rho(t) \leq \sum_{i=1}^N W_{\rho_i}(t). \tag{15}$$

Proof.

$$W_\rho(t) = \max_{0\leq\tau\leq t}[A(\tau,t) - \rho(t-\tau)]. \tag{16}$$

Let

$$\tau^* = \text{argmax}_{0\leq\tau\leq t}[A(\tau,t) - \rho(t-\tau)], \tag{17}$$

such that

$$W_\rho(t) = A(\tau^*, t) - \rho(t-\tau^*). \tag{18}$$

Then, clearly,

$$A_i(\tau^*, t) - \rho_i(t-\tau^*) \leq \max_{0\leq\tau\leq t}[A_i(\tau,t) - \rho(t-\tau)], \tag{19}$$

for $i = 1, \ldots, N$, Summing both sides of (19) for $i = 1, \ldots, N$, we obtain

$$A(\tau^*, t)-\rho(t-\tau^*) \leq \sum_{i=1}^N \max_{0\leq\tau\leq t}[A_i(\tau,t)-\rho_i(t-\tau)]. \tag{20}$$

Applying (18) and the definition of $W_{\rho_i}(t)$ in the above inequality, we obtain (22). □

Theorem 4. *The aggregate traffic $A(t)$ has W-envelope $(\sigma(\theta), \rho(\theta))$ where $\sigma(\theta) = \sum_{i=1}^N \sigma_i(\theta)$ and $\rho(\theta) = \sum_{i=1}^N \rho_i(\theta)$.*

Proof of Theorem 4. Fix $\theta \geq 0$ and let $W_\rho(t)$ denote the workload at a multiplexer with input traffic $A(t)$ and constant service rate ρ. Then

$$E\left[e^{\theta W_\rho(t)}\right] = E\left[e^{\theta \max_{0\leq\tau\leq t}[A(\tau,t)-\rho(t-\tau)]}\right]. \tag{21}$$

Applying Lemma 1, we have

$$E\left[e^{\theta W_\rho(t)}\right] \leq E\left[e^{\theta\sum_{i=1}^N W_{\rho_i}(t)}\right] = E\left[\prod_{i=1}^N e^{\theta W_{\rho_i}(t)}\right]. \tag{22}$$

Next, we use induction to show that

$$E\left[\prod_{i=1}^N e^{\theta W_{\rho_i}(t)}\right] \leq \prod_{i=1}^N e^{\theta\sigma_i} = e^{\theta\sum_{i=1}^N \sigma_i}. \tag{23}$$

Note that (23) holds trivially when $N = 1$. Suppose that (23) holds for $N = k \geq 1$. Let $Y(t) = \sum_{i=1}^k W_{\rho_i(t)}$. Applying the Cauchy-Schwarz inequality for random variables, we have

$$\begin{aligned} &E\left[e^{\theta Y(t)} \cdot e^{\theta W_{\rho_{k+1}}(t)}\right] \\ &\leq \left(E\left[e^{2\theta Y(t)}\right]\right)^{\frac{1}{2}} \cdot \left(E\left[e^{2\theta W_{\rho_{k+1}}(t)}\right]\right)^{\frac{1}{2}}. \end{aligned} \tag{24}$$

We have

$$E\left[e^{\theta W_{\rho_{k+1}}(t)}\right] \leq e^{\theta\sigma_{k+1}} \tag{25}$$

for all $\theta > 0$. By the induction hypothesis.

$$E\left[e^{\theta Y(t)}\right] = E\left[\prod_{i=1}^k e^{\theta W_{\rho_i}(t)}\right] \leq e^{\theta\sum_{i=1}^k \sigma_i}. \tag{26}$$

Applying (25) and (26) into (24), we obtain

$$E\left[\prod_{i=1}^{k+1} e^{\theta W_{\rho_i}(t)}\right] \leq e^{\theta\sum_{i=1}^k \sigma_i} \cdot e^{\theta\sigma_{k+1}} = e^{\theta\sum_{i=1}^{k+1} \sigma_i}. \tag{27}$$

This establishes (23) by the principle of induction. Combining (22) and (23), we obtain (11). □

Unlike Theorems 1 and 2, statistical independence of the set of flows $\{A_i(t) : 1 \leq i \leq N\}$ is not required. Applying

Theorem 4, we see that the A-envelope $(\sigma(\theta), \rho(\theta))$ of $A(t)$ satisfies

$$E\left[e^{\theta[A(\tau,t)-\rho(t-\tau)]}\right] \leq E\left[e^{\theta W_\rho(t)}\right] \leq e^{\theta\sigma}. \quad (28)$$

Thus, a traffic regulator that enforces conformance of $A_i(t)$ according to the W-envelope $(\sigma_i(\theta), \rho_i(\theta))$ for $i = 1, \ldots, N$, enforces *both* the A-envelope and W-envelope of the aggregate traffic $A(t)$ according to the parameter $(\sigma(\theta), \rho(\theta))$, where $\sigma(\theta) = \sum_{i=1}^{N} \sigma_i(\theta)$ and $\rho(\theta) = \sum_{i=1}^{N} \rho_i(\theta)$.

IV. Framework for Providing Delay Guarantees

Next, we outline a framework for providing stochastic delay guarantees using the W-envelope.

A. Traffic Characterization

Consider a traffic process $A(t)$ that is gSBB with upper rate ρ and bounding function f. The following theorem[2], provides a characterization of the W-envelope of $A(t)$.

Theorem 5. *Let traffic process $A(t)$ be gSBB with upper rate $\rho > 0$ and bounding function f and assume the workload process $W_\rho(t)$ is upper bounded by $\sigma_{\max}$. Then $A(t)$ has W-envelope $(\sigma(\theta), \rho)$ given by*

$$\sigma(\theta) = \frac{1}{\theta} \ln\left[1 + \theta\eta(\theta)\right], \quad (29)$$

where $\theta \geq 0$ and

$$\eta(\theta) = \int_0^{\sigma_{\max}} f(y) e^{\theta y} \mathrm{d}y. \quad (30)$$

The upper bound $\sigma_{\max}$ can be chosen based on the physical buffer capacities of the network elements and it can be enforced using the deterministic (σ, ρ) regulator of Cruz [10]. Given a gSBB envelope, a W-envelope characterization can be obtained using Theorem 5.

B. Traffic Regulation

Using the gSBB traffic regulator developed in [6] and Theorem 5, the network can regulate a traffic flow according to a negotiated W-envelope parameter $(\sigma(\theta), \rho(\theta))$, as defined in (11). Through the relationship between the gSBB bounding function $f(\cdot)$ and the desired $\sigma(\theta)$ obtained using Theorem 5, we find the required $f(\cdot)$ for the gSBB characterization. Using the traffic regulator in [6], the traffic flow can be forced to conform to a gSBB envelope with bounding function $f(\cdot)$, which in turn guarantees conformance to the W-envelope parameter $(\sigma(\theta), \rho(\theta))$.

C. Admission Control

We consider an admission control scheme for a multiplexer of capacity C. A set of N traffic flows $\{A_i(t) : i = 1, \ldots, N\}$ is admissible if offering the aggregate traffic $A(t) = \sum_{i=1}^{N} A_i(t)$ as input to the multiplexer results in delay $D(t)$ satisfying the following quality-of-service (QoS) constraint:

$$P\{D(t) > d\} < \epsilon. \quad (31)$$

Assume that $A_i(t)$ has W-envelope $(\sigma_i(\theta), \rho_i(\theta))$. By Theorem 4, $A(t)$ has W-envelope $(\sigma(\theta), \rho(\theta))$, where $\sigma(\theta) = \sum_{i=1}^{N} \sigma_i(\theta)$ and $\rho(\theta) = \sum_{i=1}^{N} \rho_i(\theta)$.

Clearly, (31) cannot be satisfied if $\rho > C$. Assuming $\rho \leq C$,

$$W_C(t) \leq W_\rho(t), \quad (32)$$

holds almost surely, where $W_C(t)$ denotes the workload at the multiplexer. For a FCFS server, $D(t) = W_C(t)/C$; hence,

$$\mathsf{P}\{D(t) \geq d\} = \mathsf{P}\{W_C(t) \geq d\ C\} \overset{(a)}{\leq} \mathsf{P}\{W_\rho(t) \geq d\ C\} \overset{(b)}{\leq} \mathsf{E}\left[e^{\theta W_\rho(t)}\right] e^{-\theta dC} \overset{(c)}{\leq} e^{\theta(\sigma(\theta)-Cd)}, \quad (33)$$

where (a) follows from (32), (b) follows from the Chernoff bound, and (c) follows from Theorem 4. The free parameter θ on the right-hand side (RHS) of (33), can be optimized to obtain the tightest upper bound on $\mathsf{P}\{D(t) \geq d\}$. The QoS criterion (31) will be satisfied if the RHS of (33) is less than ϵ. This admission control scheme could, in principle, be extended to accommodate flows traversing multi-hop paths.

V. Workload Envelope for Two Traffic Models

We now consider the workload envelopes for Markov fluid models and Markov modulated Poisson Processes (MMPPs).

A. Markov Fluid Model

We consider an Markov on-off fluid model, which consists of an underlying Markov chain with two states: 0 (*off*) and 1 (*on*) [11]. In the *on* state, the source generates fluid at a constant rate of one unit of information per unit time, while in the *off* state, no fluid is generated. The sojourn time in each *on* state is exponentially distributed with mean one, while that in each *off* state is exponentially distributed with mean λ^{-1}.

An A-envelope for a Markov on-off fluid can be obtained from the *minimum envelope rate* defined in [1]. We shall use the A-envelope given by $(\sigma = 0, \rho = a^*(\theta))$, where $a^*(\theta)$ is the minimum envelope rate given by [1, Eq. (46)] (with $\mu = \nu = 1$):

$$a^*(\theta) = \left[\theta - 1 - \lambda + \sqrt{(\theta - 1 + \lambda)^2 + 4\lambda)}\right] / 2\theta. \quad (34)$$

We now derive the W-envelope for a Markov on-off fluid source. Applying [11, Eq. (46)] with $C = \rho < 1$ and $N = 1$,

$$\mathsf{P}\{W_\rho(t) > x\} = -a_0(\mathbf{1}'\boldsymbol{\phi}_0)e^{z_0 x}, \quad (35)$$

where $W_\rho(t)$ is assumed to be in steady-state, $\mathbf{1}$ is a column vector of all ones, $'$ denotes transpose,

$$z_0 = \frac{\lambda(1-\rho)-\rho}{\rho(1-\rho)},\ \boldsymbol{\phi}_0 = \left[\frac{1-\rho}{\rho}, 1\right]',\ a_0 = \frac{-\lambda}{1+\lambda}. \quad (36)$$

Note that the traffic utilization for a single Markov on-off fluid source fed to a constant rate server with rate ρ is $U := p_{\text{on}}/\rho = \lambda/(\rho(1+\lambda))$. The traffic utilization should

[2]The proof is omitted due to space constraints.

be less than 1; otherwise, the workload process will grow without bound. We also pick $\rho < 1$ since otherwise, the workload would be almost surely 0. Therefore, $\frac{\lambda}{1+\lambda} < \rho < 1$. Applying (35) and (36), for a single source fed into a constant rate server with rate ρ, we get

$$\mathsf{P}\{W_\rho(t) > x\} = \frac{\lambda}{\rho(1+\lambda)} e^{z_0 x}. \tag{37}$$

Using (37), we can derive

$$E\left[e^{\theta W_\rho(t)}\right] = 1 - \frac{\lambda}{\rho(1+\lambda)} + \frac{\lambda z_0}{\rho(1+\lambda)(\theta+z_0)}, \tag{38}$$

for $\theta \in (0, -z_0)$. Then the W-envelope parameter $\sigma(\theta)$ can be obtained by equating $e^{\theta\sigma(\theta)}$ and $E\left[e^{\theta W_\rho(t)}\right]$, which yields

$$\sigma(\theta) = \frac{1}{\theta} \log E\left[e^{\theta W_\rho(t)}\right]. \tag{39}$$

B. MMPP Traffic Model

We next consider a model of bursty traffic generated by a three-state MMPP with parameter values as in [12, p. 79], which are derived from matching the arrival process of I, P and B frames in an MPEG-4 encoded video stream to the three states of the MMPP. The packet sizes are modeled according to a special phase-type distribution referred to as "G3" in [12, Table 1] with probability density function (pdf)

$$f_L(l) = p\ \mathrm{Er}(r_1, 1/\mu_1) + (1-p)\ \mathrm{Er}(r_2, 1/\mu_2), \tag{40}$$

where $\mathrm{Er}(r, 1/\mu)$ denotes the pdf of an $r-$stage Erlang distribution with mean r/μ. This phase-type distribution is a mixture of Erlangs, which closely approximates the empirical distribution of measured Internet packet sizes obtained in [13]. The Laplace transform (LT) of the packet length pdf is

$$\tilde{H}(s) = p\left(1 + \frac{s}{\mu_1}\right)^{-r_1} + (1-p)\left(1 + \frac{s}{\mu_2}\right)^{-r_2}. \tag{41}$$

Suppose N independent MMPP traffic sources are offered as input to a multiplexer with capacity C. Let $\mathbf{\Lambda}_i$ and $\boldsymbol{R}_i$ denote the rate matrix and generator matrix of the ith source, $i = 1, \ldots, N$. Then the input traffic is an MMPP with arrival matrix and rate matrix given by [14]

$$\mathbf{\Lambda} = \mathbf{\Lambda}_1 \oplus \ldots \oplus \mathbf{\Lambda}_N \text{ and } \boldsymbol{R} = \boldsymbol{R}_1 \oplus \ldots \oplus \boldsymbol{R}_N,$$

respectively, where $\oplus$ represents the Kronecker sum. When the service times are independent and generally distributed, the resulting queue is denoted by MMPP/G/1.

The distribution of the virtual waiting time $V(t)$ in steady-state can be obtained from results in [15] for a MAP/G/1 queue, since an MMPP is a special case of a Markovian Arrival Process (MAP). Assuming that the service rate of the queue is normalized to one, the LT of the steady-state virtual waiting time pdf of an MMPP/G/1 queue is given by [15]

$$\tilde{V}(s) = s(1-\gamma)\mathbf{g}[sI + \mathbf{D}(\tilde{H}(s))]^{-1}\mathbf{1}, \tag{42}$$

where $\gamma = \lambda_{\text{avg}}/\mu_{\text{avg}}$ is the utilization of the queue, λ_{avg} is the average packet arrival rate, μ_{avg}^{-1} is the mean packet length, and $\tilde{H}(s)$ is given in (41). The matrix function $\mathbf{D}(z)$ is given by $\mathbf{D}(z) = \mathbf{D}_0 + \mathbf{D}_1 z$, where $\mathbf{D}_0 = \boldsymbol{R} - \mathbf{\Lambda}$ and $\mathbf{D}_1 = \mathbf{\Lambda}$. The row vector $\mathbf{g}$ is the invariant probability vector associated with the stochastic matrix $\mathbf{G}$ defined in [15, Eq. (22)], i.e., $\mathbf{g}$ is the solution to

$$\mathbf{gG} = \mathbf{g}, \quad \mathbf{g1} = 1. \tag{43}$$

By relating the virtual waiting time for the MAP/G/1 system considered in [15] to the workload of the desired system, with server rate C, we obtain the LT $\tilde{W}(s)$ of the steady-state workload pdf as follows:

$$\tilde{W}(s) = sC(1-\gamma/C)\mathbf{g}[sCI + \mathbf{D}(\tilde{H}(s))]^{-1}\mathbf{1}. \tag{44}$$

Using (44), with $C = \rho$ and the MMPP parameter, we can then derive the W-envelope parameter $\sigma(\theta)$ as follows:

$$\sigma(\theta) = \frac{1}{\theta} \log E\left[e^{\theta W_\rho(t)}\right] = \frac{1}{\theta} \log \tilde{W}(-\theta). \tag{45}$$

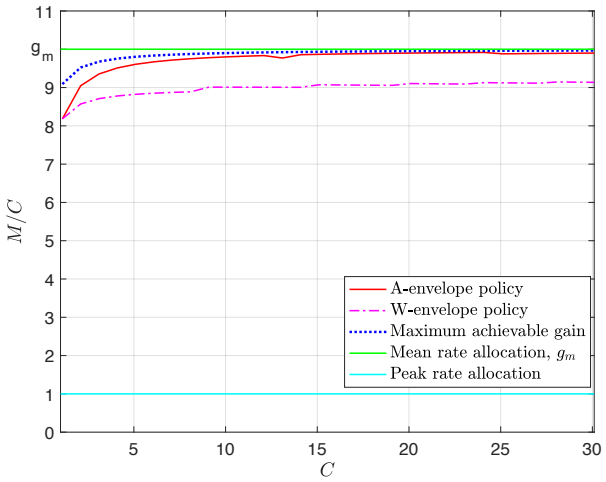

Fig. 1: Multiplexing gain vs. C with Markov on-off fluid models: $p_{\text{on}} = 0.1$, $d = 100$, $\epsilon = 10^{-3}$.

VI. Numerical Examples

Using the analytical results from Section V, we numerically investigate the performance of the W-envelope for Markov fluid and MMPP traffic sources.

1) Markov Fluid Model: Let M denote the maximum number of Markov on-off fluid sources that can be supported at a multiplexer with capacity C. The statistical multiplexing gain can be quantified by the ratio $g = M/C$. Under *peak rate* allocation, the number of sources that can be supported is $M_p = C$ and in this case, the multiplexing gain is $g_p = M_p/C = 1$. Under *mean rate* allocation, the workload at the multiplexer will grow without bound, but the number $M_m = C/p_{\text{on}} = C(1+\lambda^{-1})$ provides an upper bound on the number of sources that can be supported under any admission control policy. We define $g_m := M_m/C = 1+\lambda^{-1}$. In general, $g \in [g_p, g_m) = \left[1,\ 1+\lambda^{-1}\right)$.

Figure 1 shows the multiplexing gain under the A-envelope and W-envelope admission control policies (Theorems 2 and 4,

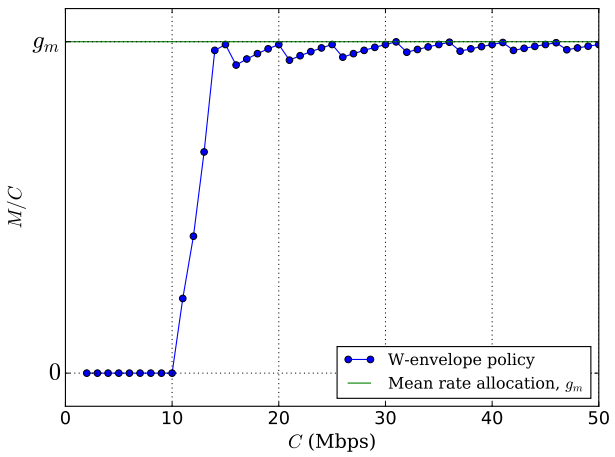

Fig. 2: Multiplexing gain vs. C with MMPP bursty traffic: $d = 4$ ms, $\epsilon = 10^{-3}$.

respectively) when the QoS constraint (31) is specified by $\epsilon = 10^{-3}$ and $d = 100$. The traffic sources are independent and identically distributed with $p_{\text{on}} = 0.1$. The achievable gain curve was obtained using the analytical result in [11, Eq. (46)]. Figure 1 shows that both policies achieve high multiplexing gains even for moderate values of the capacity C. It is important to note that the A-envelope policy relies on independence of the traffic flows, whereas the W-envelope policy does not.

2) MMPP Bursty Traffic Model: To model a bursty traffic source, the MMPP parameter values are chosen according to [12, p. 79], with arrival matrix $\mathbf{\Lambda} = \text{diag}\{116, 274, 931\}$ in [packets/s] and rate matrix

$$\boldsymbol{R} = \begin{bmatrix} -0.12594 & 0.12594 & 0 \\ 0.25 & -2.22 & 1.97 \\ 0 & 2 & -2 \end{bmatrix} \tag{46}$$

in $[\text{s}^{-1}]$. The packet length parameters in (40) are set as $p = 0.54$, $r_1 = r_2 = 5$, $\mu_1^{-1} = 5.2$ bytes, $\mu_2^{-1} = 191.2$ bytes. The Erlang mixture distribution closely approximates the empirical distribution of measured Internet packet sizes obtained in [13]. The average packet length is 454 bytes, which yields an average bit rate of 1.24 Mbps.

Under a mean rate allocation scheme, the number $M_m = C/\gamma = C\mu_{\text{avg}}/\lambda_{\text{avg}}$ provides an upper bound on the number of sources that can be supported under any admission control policy, and we define $g_m := M_m/C = \mu_{\text{avg}}/\lambda_{\text{avg}}$. The gain g of a general policy satisfies $g \leq g_m$. Figure 2 shows the multiplexing gain under the W-envelope policy when the QoS constraint (31) is specified by $\epsilon = 10^{-3}$ and $d = 4$ ms. Observe that the W-envelope policy achieves a gain close to the upper bound g_m even for moderate values of C. Although the gain g can decrease slightly with C, the number of admitted flows is always monotonically increasing.

VII. Conclusion

Motivated by the desire to provide performance guarantees for time-critical applications in future networks, we proposed the W-envelope, a traffic bound on the MGF of the workload process resulting from offering a traffic source to a constant rate server. In contrast to the MGF traffic arrival envelope (A-envelope), the W-envelope is amenable to traffic regulation [6] and traffic fitting [5]. Admission control based on the W-envelope does not require independence of the traffic flows, yet it can achieve statistical multiplexing gain.

Although our numerical results were based on analytical W-envelope expressions for Markov fluid and MMPP traffic models, an empirical W-envelope can be obtained for an arbitrary traffic trace using the method in [5]. By leveraging network softwarization and virtualization, a service for time-critical applications requiring delay guarantees can be realized using the proposed W-envelope framework.

References

[1] 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.

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

[3] Q. Yin, Y. Jiang, S. Jiang, and P. Kong, "Analysis on generalized stochastically bounded bursty traffic for communication networks," in *IEEE Conf. on Local Computer Nets. (LCN)*, Nov. 2002, pp. 141–149.

[4] 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.

[5] M. Kordi Boroujeny, B. L. Mark, and Y. Ephraim, "Fitting network traffic to phase-type bounds," in *54rd Conf. on Info. Sciences and Systems (CISS)*, Princeton, NJ, March 2020.

[6] M. Kordi Boroujeny and B. L. Mark, "Design of a stochastic traffic regulator for end-to-end network delay guarantees," *IEEE/ACM Trans. Netw.*, pp. 1–13, 2022, doi:10.1109/TNET.2022.3181858.

[7] 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.

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

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

[10] R. Cruz, "A calculus for network delay. I. Network elements in isolation," *IEEE Trans. Inf. Theory*, vol. 37, no. 1, pp. 114–131, 1991.

[11] D. Anick, D. Mitra, and M. M. Sondhi, "Stochastic theory of a data-handling system with multiple sources," *The Bell System Technical Journal*, vol. 61, no. 58, pp. 1871–1894, Oct. 1982.

[12] G. Dán, V. Fodor, and G. Karlsson, "Packet size distribution: An aside?" in *Quality of Service in Multiservice IP Networks*, M. Ajmone Marsan, G. Bianchi, M. Listanti, and M. Meo, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2005, pp. 75–87.

[13] C. Fraleigh, S. Moon, B. Lyles, C. Cotton, M. Khan, D. Moll, R. Rockell, T. Seely, and S. Diot, "Packet-level traffic measurements from the Sprint IP backbone," *IEEE Network*, vol. 17, no. 6, pp. 6–16, 2003.

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

[15] D. 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.