Sharp Per-Flow Delay Bounds for Bursty Arrivals: The Case of FIFO, SP, and EDF Scheduling

Florin Ciucu
University of Warwick

Felix Poloczek
University of Warwick / TU Berlin

Jens Schmitt
University of Kaiserslautern

***Abstract*—The practicality of the stochastic network calculus (SNC) is often questioned on grounds of potential looseness of its performance bounds. In this paper, it is uncovered that for bursty arrival processes (specifically Markov-Modulated On-Off (MMOO)), whose amenability to *per-flow* analysis is typically proclaimed as a highlight of SNC, the bounds can unfortunately be very loose (e.g., by several orders of magnitude off). In response to this uncovered weakness of SNC, the (Standard) per-flow bounds are herein improved by deriving a general sample-path bound, using martingale based techniques, which accommodates FIFO, SP, and EDF scheduling. The obtained (Martingale) bounds capture an extra exponential decay factor of $\mathcal{O}\left(e^{-\alpha n}\right)$ in the number of flows n. Moreover, numerical comparisons against simulations show that the Martingale bounds are not only remarkably accurate, but they also improve the Standard SNC bounds by factors as large as 100 or even 1000.**

I. Introduction

Several approaches to the classical queueing theory have emerged over the past decades. For instance, matrix analytic methods (MAM) not only provide a unified treatment for a large class of queueing systems, but they also lend themselves to practical numerical solutions; two key ideas are the proper accounting of the repetitive structure of underlying Markov processes, and the use of linear algebra rather than classic methods based on real analysis (see Neuts [31] and Lipsky [27]). Another unified approach targeting broad classes of queueing problems is the stochastic network calculus (SNC) (see Chang [7] and Jiang and Liu [20]), which can be regarded as a mixture between the deterministic network calculus conceived by Cruz [15] (see Le Boudec and Thiran [4]) and the effective bandwidth theory (see Kelly [22]). Because SNC solves queueing problems in terms of *bounds*, it is often regarded as an unconventional approach, especially by the queueing theory community.

MAM and SNC could be (slightly) compared by the way they apply to queues with fluid arrivals. In their simplest form, fluid arrival models were defined as Markov-Modulated On-Off (MMOO) processes by Anick, Mitra, and Sondhi [1], and were significantly extended thereafter, especially for the purpose of modelling the increasingly prevalent voice and video traffic in the Internet. By relating fluid models and Quasi-Birth-Death (QBD) processes, Ramaswami has argued that MAM can lend themselves to numerically more accurate solutions than spectral analysis methods [34]. In turn, SNC can produce alternative solutions with negligible numerical complexity, but these are arguably less relevant than exact solutions (simply because they are expressed as bounds). What does, therefore, justify more than two decades of research in SNC?

The answer lies in two key features of SNC: *scheduling abstraction* and *convolution-form networks* (see Ciucu and Schmitt [13]). The former expresses the ability of SNC to compute per-flow (or per-class) queueing metrics for a large class of scheduling algorithms, in a unified manner, by decoupling scheduling from queueing analysis. Concretely, given a flow A sharing a queueing system with other flows, the characteristics of the scheduling algorithm are first abstracted away in the so-called *service process*; thereafter, the derivation of queueing metrics for the flow A is scheduling independent. Furthermore, the per-flow results can be extended in a straightforward manner from a single queue to a large class of queueing networks (typically feed-forward), using convolution representations in a $(\min,+)$ algebra.

By relying on these two features, SNC could tackle several open queueing networks problems. The typical scenario involves the computation of end-to-end (e2e) non-asymptotic performance (e.g., delay) bounds of a single flow crossing a tandem network and sharing the single queues with some other flows. Such scenarios were solved for a large class of arrival processes (see, e.g., Ciucu *et al.* [10], [6] and Fidler [19] for MMOO processes, and Liebeherr *et al.* [25] for heavy-tailed and self-similar processes). Another important solution was given for the e2e delay distribution in a tandem (packet) network with Poisson arrival and exponential packet sizes, by circumventing the so-called Kleinrock's independence assumption (see Burchard *et al.* [5]). Other fundamentally difficult problems include the performance analysis of stochastic networks implementing network coding (see Yuan *et al.* [41]), the delay analysis of wireless channels under Markovian assumptions (see Zheng *et al.* [42]), the delay analysis of multi-hop fading channels (see Al-Zubaidy *et al.* [43]), bridging information theory and queueing theory by accounting for the stochastic nature and delay-sensitivity of real sources (see Lübben and Fidler [29]), or the computation of non-asymptotic per-flow capacity in ad-hoc networks (see Ciucu *et al.* [11]).

Based on its ability to solve some fundamentally hard queueing problems (in terms of bounds), SNC is justifiably proclaimed as a valuable alternative to the classical queueing theory (see Ciucu and Schmitt [13]). At the same time, SNC is also justifiably questioned on the tightness of its bounds. While the asymptotic tightness generally holds (see Chang [7], p. 291, and Ciucu *et al.* [10]), doubts on the bounds' numerical tightness shed skepticism on the practical relevance of SNC. This skepticism is supported by the fact that SNC largely employs the same probability methods as the effective bandwidth theory, which was argued to produce largely inaccurate results for non-Poisson arrival processes (see Choudhury *et al.* [8]).

In this paper, we reveal what is perhaps 'feared' by SNC proponents and expected by others: the bounds are very loose for the class of MMOO processes, which is very relevant as these can be tuned for various degrees of burstiness. In addition to providing numerical evidence for this fact (the bounds can be off by arbitrary orders of magnitude, e.g., by factors as large as 100 or even 1000), we also prove that the bounds are asymptotically loose in multiplexing regimes. Concretely, we (analytically) prove that the bounds are 'missing' an exponential decay factor of $\mathcal{O}\left(e^{-\alpha n}\right)$ in the number of flows n, where $\alpha > 0$; this missing factor was conjectured through numerical experiments in Choudhury *et al.* [8] in the context of effective bandwidth results (which scale identically as the SNC bounds).

While this paper convincingly uncovers a major weakness in the SNC literature, it also shows that the looseness of the bounds is generally not inherent in SNC but it is due to the 'temptatious' but 'poisonous' elementary tools from probability theory leveraged in its application. We point out that such methods have also been employed in the effective bandwidth literature dealing with scheduling; see Courcoubetis and Weber [14] for FIFO, Berger and Whitt [2] and Wischik [40] for SP, and Sivaraman and Chiussi [38] for EDF. Unlike the SNC results, which are given in terms of non-asymptotic bounds, the corresponding effective bandwidth results are typically given in large buffer asymptotics regimes; while exactly capturing the asymptotic decay rate, they fail to capture the extra $\mathcal{O}\left(e^{-\alpha n}\right)$ decay factor pointed out by Choudhury *et al.* [8] or by Botvich and Duffield [3].

To fix the weakness of existing SNC bounds, and also of existing effective bandwidth asymptotic results in scheduling scenarios, this paper leverages more advanced tools (i.e., martingale based techniques) and derives new (Martingale) bounds improving dramatically to the point of almost matching simulation results. We show the improvements for per-flow delay bounds in FIFO, SP, and EDF scheduling scenarios with MMOO flows, and in addition we prove the existence of the conjectured $\mathcal{O}\left(e^{-\alpha n}\right)$ decay factor. We point out that extensions to more general Markovian arrivals are immediate (see Appendix.A in the associated technical report [12]); due to their increased complexity, however, the generalized results do not easily lend themselves to visualizing the $\mathcal{O}\left(e^{-\alpha n}\right)$ decay factor uncovered herein for MMOO flows.

The sharp bounds obtained in this paper are the first in the conventional stochastic network calculus literature, i.e., involving service processes which decouple scheduling from the analysis. Their significance, relative to existing sharp bounds in the effective bandwidth literature (e.g., Duffield [17] and Chang [7], pp. 339-343, using martingale inequalities, or Liu *et al.* [28] by extending an approach of Kingman involving integral inequalities [23]), is that they apply at the *per-flow* level for various scheduling disciplines; in turn, existing sharp bounds only apply at the *aggregate* level. In other words, our sharp bounds generalize existing ones by accounting for FIFO, SP, and EDF scheduling.

The rest of the paper is structured as follows. In Section II, we identify, at an intuitive level, the elementary tool from probability theory which is 'responsible' for the very loose (Standard) bounds in SNC. In Section III, we describe the queueing model and provide some necessary SNC formalisms. The core of the paper is Section IV, which derives the new (Martingale) and reviews the existing (Standard) SNC per-flow delay bounds in multiplexing scenarios with MMOO flows; both analytical and numerical comparisons of the bounds (including uncovering the $\mathcal{O}\left(e^{-\alpha n}\right)$ decay factor) are further explored. Concluding remarks are presented in Section VI.

II. Three Bounding Steps in SNC and One Pitfall

This section overviews the SNC bounding approach to derive per-flow bounds for broad classes of arrivals and scheduling. In addition to highlighting the underlying bounding steps, an elementary example proves that *careless bounding* can lend itself to impractical results. For the sake of simplicity, the discussion is formulated in a discrete-time setting.

We consider a simplified queueing system in which a (cumulative) arrival process $A(t)$ shares a server with capacity C and infinite queue length with some other flows. We are particularly interested in the complementary distribution of $A(t)$'s backlog process $B(t) := A(t) - D(t)$ (where $D(t)$ is $A(t)$'s corresponding departure process), which is bounded in SNC for some $t, \sigma \geq 0$ according to

$$\mathbb{P}\left(B(t) > \sigma\right) \leq \mathbb{P}\left(\sup_{0 \leq s \leq t} \left\{A(s,t) - S(s,t)\right\} > \sigma\right) \ . \quad (1)$$

Here, $A(s,t) := A(t) - A(s)$ is the bivariate extension of $A(t)$, whereas $S(s,t)$ is another bivariate process, called a *service process*, encoding the information about the server, the scheduling, and the other arrival processes with which $A(t)$ shares the server. In the simplest setting with no other arrivals, $S(s,t) = C(t-s)$ and Eq. (1) (with equality) recovers Reich's equation. In another setting in which $A(t)$ receives the lowest priority, should the server implement a static priority (SP) scheduler, then $S(s,t) = C(t-s) - A_c(s,t)$, where $A_c(s,t)$ denotes the other (cross) arrivals at the server.

SNC typically continues with Eq. (1) by invoking the Union Bound, i.e.,

$$\text{Eq. (1)} \ldots \leq \sum_{s=0}^{t} \mathbb{P}\left(A(s,t) - S(s,t) > \sigma\right) \ . \quad (2)$$

The probability events can be further computed either by 1) convolving the distribution functions of $A(s,t)$ and $S(s,t)$, when available, and under appropriate independence assumptions, or by following a more elegant procedure using the Chernoff bound, i.e.,

$$\text{Eq. (2)} \ldots \leq \sum_{s=0}^{t} E\left[e^{\theta(A(s,t) - S(s,t))}\right] e^{-\theta\sigma} \ , \quad (3)$$

for some $\theta > 0$. The expectation can be split into a product of expectations, according to the statistical independence properties of $A(s,t)$ and $S(s,t)$, and the sum can be further reduced to some canonical form.

Eqs. (1)-(3) outline three major bounding steps. The first is 'proprietary' to SNC, in the sense that it involves the specific construction of a 'proprietary' service process $S(s,t)$ which decouples scheduling from analysis. The next two follow general purpose methods in probability theory, which are applied in the same form in the effective bandwidth theory, except that $S(s,t)$ is now a random process rather than a constant-rate function.

In particular, the second step reveals a convenient continuation of Eq. (1). The reason for this 'temptatious' step to be consistently invoked in SNC stems from the 'freedom' of seeking for bounds rather than exact results. As we will show over the rest of this section, and of the paper, this 'temptatious' step is also 'poisonous' in the sense that it can lead to very loose bounds.

As a simple and yet illustrative example, let us consider the stationary but non-ergodic process

$$A(s,t) = (t-s)X \ \forall 0 \le s \le t \ , \tag{4}$$

where X is a Bernoulli random variable taking values in $\{0,2\}$, each with probabilities $1-\varepsilon > .5$ and $\varepsilon > 0$. Assume also that $S(s,t) = t-s$. Clearly, for $\sigma > 0$ and for sufficiently large t, the backlog process satisfies

$$\mathbb{P}\left(B(t) > \sigma\right) = \varepsilon \ .$$

In turn, the application of the bound from Eq. (2) lends itself to a bogus bound, i.e.,

$$\mathbb{P}\left(B(t) > \sigma\right) \le \varepsilon t \ ,$$

for $\sigma < 1$ (for $\sigma \ge 1$, the bound diverges as well). The underlying reason behind this bogus result is that the Union Bound from Eq. (2) is agnostic to the statistical poperties of the increments of the arrival process $A(s,t)$.

The construction of $A(s,t)$ from Eq. (4) illustrates thus the poor performance of the Union Bound for arrivals with correlated increments, such as MMOO processes. Within the same class, another relevant arrival process is the fractional Brownian motion (fBM) which has long-range correlations; fBM was analyzed either by relying on approximations (e.g., Norros [32]) or by using the Union Bound (e.g., Rizk and Fidler [35]). The rest of the paper will unequivocally reveal that the Union Bound leads to very loose per-flow bounds in scheduled queueing scenarios with MMOO processes.

As a side remark, we point out that the Union Bound renders reasonably tight bounds when $X_s := A(s,t)$'s are rather uncorrelated (see Talagrand [39]). Shroff and Schwartz [37] argued that the effective bandwidth theory yields reasonable bounds only for Poisson processes. Along the same lines, Ciucu [9] provided numerical evidence that SNC itself renders reasonably tight bounds for Poisson arrivals.

III. Queueing Model

This section introduces the queueing model and necessary SNC formalisms. For the rest of the paper, the time model is continuous. Consider a stationary (bivariate) arrival process $A(s,t)$ defined as

$$A(s,t) := \int_{u=s}^{t} a(u)du \ \forall 0 \le s \le t \ , \quad A(t) := A(0,t) \ ,$$

where $a(s)\ \forall s \ge 0$ is the increment process.

According to Kolmogorov's extension theorem, the one-side (stationary) process $\{a(s) : 0 \le s < \infty\}$ can be extended to a two-side process $\{a(s) : -\infty < s < \infty\}$ with the same distribution. For convenience, we often work with the *reversed* cumulative arrival process $A^{\mathrm{r}}(s,t)$ defined as

$$A^{\mathrm{r}}(s,t) = \int_{u=s}^{t} a(-u)du \ \forall 0 \le s \le t \ , \quad A^{\mathrm{r}}(t) := A^{\mathrm{r}}(0,t) \ .$$

This definition is identical with that of $A(s,t)$, except that the time direction is reversed.

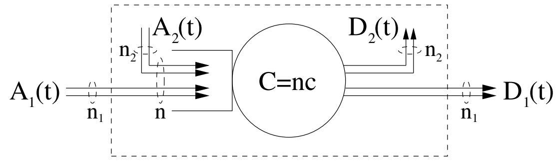


Fig. 1. A queueing system with two arrival processes $A_1(t)$ and $A_2(t)$, each containing n_1 and n_2 sub-flows. The server has a capacity $C = nc$, where $n = n_1 + n_2$. We are interested in the delay distribution of $A_1(t)$.

Working with time reversed processes is particularly convenient in that the steady-state queueing process (say in a queueing system with constant-rate capacity C fed by the one-side increment process $a(s)$) can be represented by Reich's equation

$$Q = \sup_{t \ge 0} \left\{A^{\mathrm{r}}(t) - Ct\right\} \ .$$

The evaluation of Q needs an additional *stability* condition, e.g., $\limsup_{t\to\infty} \frac{A^{\mathrm{r}}(t)}{t} < C$ a.s. (see Chang [7], pp. 293-294); this condition is fulfilled by the (stronger) Loynes' condition, i.e., $a(s)$ is also ergodic and $\lim_{t\to\infty} \frac{A(t)}{t} = E\left[a(1)\right] < C$ a.s.

In this paper, we mostly consider the queueing system depicted in Figure 1. Two cumulative arrival processes $A_1(t)$ and $A_2(t)$, each containing n_1 and n_2 sub-flows, are served by a server with constant-rate $C = nc$, where $n = n_1 + n_2$. The parameter c is referred to as the *per-(sub)flow capacity*, and will be needed in the context of asymptotic analysis. For clarity, $A_1(t)$ and $A_2(t)$ are suggestively referred to as the *through* and *cross* (aggregate) flows, respectively. The data units are infinitesimally small and are referred to as *bits*. The queue has an infinite size capacity, and is assumed to be stable.

The performance measure of interest is the *virtual* delay process for the (through) flow $A_1(t)$, defined as

$$W_1(t) := \inf\left\{d \ge 0 : A_1(t-d) \le D_1(t)\right\} \ \forall t \ge 0 \ ,$$

where $D_1(t)$ is the corresponding departure process of $A_1(t)$ (see Figure 1). The attribute *virtual* expresses the fact that $W_1(t)$ models the delay experienced by a *virtual* bit departing at time t. Note that $W_1(t)$ is the horizontal distance between the curves $A_1(t)$ and $D_1(t)$, starting backwards from the point $(t, D_1(t))$ in the Euclidean space.

In SNC, queueing performance metrics (e.g., bounds on the distribution of the delay process $W_1(t)$) are derived by constructing *service processes*, which relate the departure and arrival processes by a $(\min,+)$ convolution. For instance, in the case of $A_1(t)$ and $D_1(t)$, the corresponding service process is a stochastic process $S_1(s,t)$ satisfying

$$D_1(t) \ge A_1 * S_1(t) \ \forall t \ge 0 \ , \tag{5}$$

where '$*$' is the $(\min,+)$ convolution operator defined for all sample-paths as $A_1 * S_1(t) := \inf_{0\le s\le t}\left\{A_1(s) + S_1(s,t)\right\}$.

The service process $S_1(s,t)$ encodes the information about the cross flow $A_2(t)$ and the scheduling algorithm; other information such as the packet size distribution is omitted herein in accordance to the infinitesimal data units assumption. Conceptually, the service process representation from Eq. (5) encodes $A_1(t)$'s own service view, as if it was alone at the network node (i.e., not competing for the service capacity C with other flows). Although the representation is not exact due to the inequality from Eq. (5), it suffices for the purpose of deriving upper bounds on the distribution of $W_1(t)$. The driving key property is that Eq. (5) holds for *all* arrival processes $A_1(t)$. Due to this property, the service representation in SNC is *somewhat analogous* with the impulse-response representation of signals in linear and time invariant (LTI) systems (see Ciucu and Schmitt [13] for a recent discussion on this analogy).

In this paper, we will compute the distribution of the through flow's delay process $W_1(t)$ for three scheduling algorithms: First-In-First-Out (FIFO), Static Priority (SP), and Earliest-Deadline-First (EDF). The corresponding service processes $S_1(s,t)$, enabling the derivations of the delay bounds, will be presented in Section IV-A.

IV. SNC Bounds for MMOO Processes

Consider the queueing scenario from Figure 1, in which all the sub-flows comprising $A_1(t)$ and $A_2(t)$ are Markov-Modulated On-Off (MMOO) processes. Being tunable for various degrees of burstiness, they are particularly relevant for testing the tightness of related delay bounds. Moreover, due to their apparent simplicity, the MMOO processes allow the explicit derivation of the conjectured $\mathcal{O}\left(e^{-\alpha n}\right)$ decay factor mentioned in the Introduction.

After defining the MMOO processes, we derive Martingale bounds for the distribution of $W_1(t)$ for FIFO, SP, and EDF scheduling. Then we overview the corresponding Standard bounds in SNC. Lastly, we compare these bounds both asymptotically, as well as against simulations.

Each MMOO sub-flow is modulated by a continuous time Markov process $Z(t)$ with two states denoted by 0 and 1, and transition rates μ and λ as depicted in Figure 2 with $n=1$. The cumulative arrival process for each sub-flow is defined as

$$A'(s,t) := \int_{u=s}^{t} Z(u)P du \; \forall 0 \le s \le t \ , \ A'(t) := A'(0,t) \ , \tag{6}$$

where $P>0$ is the peak rate. In other words, $A'(t)$ models a data source transmitting with rates 0 and P while $Z(t)$ delves in the 0 and 1 states, respectively. The steady-state 'On' probability is $p := \frac{\mu}{\lambda+\mu}$ and the average rate is pP.

When n such statistically independent sources are multiplexed together then the corresponding modulating Markov process, denoted with abuse of notation as $Z(t)$ as well, has the states $\{0,1,\ldots,n\}$ and the transition rates as depicted in Figure 2. The cumulative arrival process for the aggregate flow is defined identically as for each sub-flow, i.e.,

$$A(s,t) := \int_{u=s}^{t} Z(u)P du \; \forall 0 \le s \le t \ , \ A(t) := A(0,t) \ .$$

Note that, by definition, $A(s,t)$ is continuous.

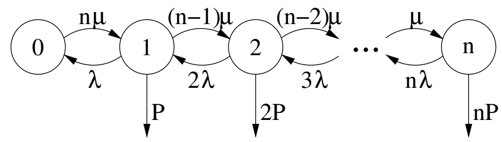


Fig. 2. A Markov-modulated process for the aggregation of n homogeneous MMOOs.

A. Martingale Bounds

Recall our main goal of deriving bounds on the distribution of the through flow's delay process $W_1(t)$ for the FIFO, SP, and EDF scheduling scenarios in the network model from Figure 1. We start this section with a technically abstract but instrumental result which will enable the analysis of all three scheduling scenarios.

Theorem 1: (Martingale Sample-Path Bound) Consider the single-node queueing scenario from Figure 1, in which n sub-flows are statistically independent MMOO processes with transition rates μ and λ, peak rate P, and all starting in the steady-state. The arrival processes (flows) are $A_1(t)$ and $A_2(t)$, each being modulated by the (stationary) Markov processes $Z_1(t)$ and $Z_2(t)$ with n_1 and n_2 states, respectively, with $n_1+n_2=n$. Assume that the utilization factor $\rho := \frac{pP}{c}$ satisfies $\rho<1$ for stability, where p is the steady-state 'On' probability; assume also that $P>c$ to avoid a trivial scenario with zero delay. Then the following sample-path bound holds for all $0 \le u \le t$ and σ

$$\mathbb{P}\left(\sup_{0\le s<t-u}\left\{A_1(s,t-u)+A_2(s,t)-C(t-s)\right\}>\sigma\right) \le K^n e^{-\gamma(C_1 u+\sigma)} \ , \tag{7}$$

where $C_1 = n_1 c$, $K = \rho\left(\frac{\rho-p}{1-p}\right)^{\frac{p}{\rho}-1}$, and $\gamma = \frac{(\lambda+\mu)(1-\rho)}{P-c}$.

While K and γ correspond to the multiplicative factor and the exponential decay rate, respectively, we point out that the crucial element in the sample-path bound from Eq. (7) is the *parameter* u, which can be explicitly *tuned* depending on the scheduling algorithm for the bits of $A_1(t)$ and $A_2(t)$. From a conceptual point of view, the parameter u encodes the information about the underlying scheduling, whereas the theorem further enables the *per-flow* delay analysis for several common scheduling algorithms: FIFO, SP, and EDF (see Subsections IV-A1–IV-A3).

The delay bounds obtained from Theorem 1 generalize the delay bounds obtained by Palmowski and Rolski [33], by further accounting for several scheduling algorithms[1]. The bounds from [33] can be recovered by applying Theorem 1 with $A_2(t) := 0$ (i.e., no cross traffic and thus no scheduling being considered) and $u := 0$. The key to the proof of Theorem 1 is a subtle martingale construction, accounting for the time shifting parameter u, followed by a standard application of the Optional Sampling theorem. For relevant martingale notions and results we refer to [12] (Appendix.B). Also, for the generalization of Theorem 1 to general Markov fluid processes we refer to [12] (Appendix.A); as mentioned in

[1]More exactly, [33] gives backlog bounds at the aggregate level which can be immediately translated into delay bounds, given the fixed server capacity.

the Introduction, the generalized result does not lend itself to visualizing the conjectured $\mathcal{O}\left(e^{-\alpha n}\right)$ decay factor, for which reason this paper *deliberately* focuses on MMOO processes.

Proof: Fix $u \geq 0$ and σ. For convenience, let us bound the probability from Eq. (7) by shifting the time origin and using the time-reversed representation of arrival processes described in Section III, i.e.,

$$\mathbb{P}\Big(\sup_{t>u}\left\{A_1^{\mathrm{r}}(u,t)+A_2^{\mathrm{r}}(u,t)-C(t-u)\right\}+A_2^{\mathrm{r}}(u)-C_2u > C_1u+\sigma\Big) , \qquad (8)$$

where $C_2 = n_2c$. This representation is possible because the underlying Markov modulating processes of $A_1(t)$ and $A_2(t)$, i.e., $Z_1(t)$ and $Z_2(t)$, respectively, are *time-reversible* processes (see, e.g., Mandjes [30], p. 57); the reversibility is a consequence of the fact that $Z_1(t)$ and $Z_2(t)$ are stationary birth-death processes (see Kelly [21], pp. 10-11). Denote by $Z_1^{\mathrm{r}}(t)$ and $Z_2^{\mathrm{r}}(t)$ the time-reversed versions of $Z_1(t)$ and $Z_2(t)$, respectively.

Let us define the stopping time

$$T := \inf\Big\{t>u : A_1^{\mathrm{r}}(u,t)+A_2^{\mathrm{r}}(u,t)-C(t-u) +A_2^{\mathrm{r}}(u)-C_2u > C_1u+\sigma\Big\} . \qquad (9)$$

This construction is motivated by the fact that $\mathbb{P}(T<\infty)$ is exactly the probability from Eq. (8). The goal of the rest of the proof is to bound $\mathbb{P}(T<\infty)$.

Let $\mathbb{P}_{i,j}$ denote the underlying probability measure conditioned on $Z_1^{\mathrm{r}}(u)=i$ and $Z_2^{\mathrm{r}}(0)=j$, for $0\leq i\leq n_1$ and $0\leq j\leq n_2$. Denote also the stationary probability vectors of $Z_1^{\mathrm{r}}(u)$ and $Z_2^{\mathrm{r}}(u)$ by $(\pi_{1,0},\ldots,\pi_{1,n_1})$ and $(\pi_{2,0},\ldots,\pi_{2,n_2})$, respectively.

Define the following two processes

$$\begin{aligned} M_1(t) &:= e^{-\theta(Z_1^{\mathrm{r}}(t)-i)}e^{\gamma\int_u^t(PZ_1^{\mathrm{r}}(s)-C_1)ds}\ \forall t\geq u \text{ and}\\ M_2(t) &:= e^{-\theta(Z_2^{\mathrm{r}}(t)-j)}e^{\gamma\int_0^t(PZ_2^{\mathrm{r}}(s)-C_2)ds}\ \forall t\geq 0 , \end{aligned}$$

where $\theta := \log\frac{\mu}{\lambda}\frac{P-c}{c}$. Note that $\theta<0$ due to the stability condition $\rho<1$.

According to Palmowski and Rolski [33], both $M_1(t)$ and $M_2(t)$ are martingales with respect to (wrt) $\mathbb{P}_{i,j}$ and the natural filtration (for the original result see Ethier and Kurtz [18], p. 175). Moreover, according to Lemmas 1 and 2 from the Appendix in [12], the following process

$$M_t := \begin{cases} M_2(t) & , \quad t\leq u\\ M_1(t)M_2(t) & , \quad t>u \end{cases}$$

is also a martingale (note that $M_1(u)=1$, by construction).

Because T may be unbounded, we need to construct the bounded stopping times $T\wedge k := \min\{T,k\}$ for all $k\in\mathbb{N}$. For these times, the Optional Sampling theorem (see Theorem 3 in Appendix.B of [12]) yields

$$E_{i,j}\left[M_0\right] = E_{i,j}\left[M_{T\wedge k}\right] ,$$

for all $k\in N$, where the expectations are taken wrt $\mathbb{P}_{i,j}$. Using $E_{i,j}\left[M_0\right]=1$ and according to the construction of $M_2(t)$ we further obtain for $k>u$

$$\begin{aligned} 1 &\geq E_{i,j}\left[M_{T\wedge k}I_{\{T\leq k\}}\right]\\ &\geq e^{-\theta\left(\frac{C_1+C_2}{P}-(i+j)\right)}e^{\gamma(C_1u+\sigma)}\mathbb{P}_{i,j}\left(T\leq k\right) , \end{aligned}$$

where $I_{\{\cdot\}}$ denotes the indicator function. The first term in the product follows from $\theta<0$ and

$$\left(Z_1^{\mathrm{r}}(T)+Z_2^{\mathrm{r}}(T)\right)P \geq C_1+C_2 ,$$

according to the construction of T from Eq. (9) and the continuity property of the arrival processes. The second term follows from $\gamma>0$ and the construction of T.

By deconditioning on i and j (note that $Z_1^{\mathrm{r}}(u)$ and $Z_2^{\mathrm{r}}(0)$ are in steady-state by construction) we obtain

$$\mathbb{P}\left(T\leq k\right) \leq \sum_{i,j}\pi_{1,i}\pi_{2,j}e^{\theta\left(\frac{C_1+C_2}{P}-(i+j)\right)}e^{-\gamma(C_1u+\sigma)} .$$

Using the identities

$$\sum_{i=0}^{n_1}\pi_{1,i}e^{\theta\left(\frac{C_1}{P}-i\right)} = K^{n_1} \text{ and } \sum_{j=0}^{n_2}\pi_{2,j}e^{\theta\left(\frac{C_2}{P}-j\right)} = K^{n_2}$$

(see [33]) and taking $k\to\infty$ we finally obtain that

$$\mathbb{P}\left(T<\infty\right) \leq K^n e^{-\gamma(C_1u+\sigma)} ,$$

which completes the proof. $\square$

In the following we fix $0\leq d\leq t$ and derive bounds on $\mathbb{P}\left(W_1(t)>d\right)$ for FIFO, SP, and EDF scheduling; the derivations follow more or less directly by instantiating the parameters of Theorem 1 for each scheduling case.

1) FIFO: The FIFO server schedules the bits of $A_1(t)$ and $A_2(t)$ in the order of their arrival times.

To derive a bound on the distribution of the through flow's (virtual) delay process $W_1(t)$, we rely on a service process construction for FIFO scheduling, as mentioned in Section III. We use the service process from Cruz [16] extended to bivariate stochastic processes, i.e.,

$$S_1(s,t) = \left[C(t-s)-A_2(s,t-x)\right]_+ I_{\{t-s>x\}} , \qquad (10)$$

for some fixed *parameter* $x\geq 0$ and independent of s and t (for a proof, in the slightly simpler case of univariate processes, see Le Boudec and Thiran [4], pp. 177-178; for a more recent and general proof see Liebeherr *et al.* [26]). By notation, $[y]_+ := \max\{y,0\}$ for some real number y.

Before proceeding further, we ought to point out that readers unfamiliar with network calculus may find the expression of the service process $S_1(s,t)$ rather difficult to grasp. Note that the meaning of $S_1(s,t)$ becomes more intuitive when setting the parameter $x:=0$; the resulting service process holds also in the case of SP scheduling, when the bits of A_1 have the lowest priority, but is conceivably weak in the case of FIFO. The role of the parameter x is simply to strengthen the SP service process; moreover, x can be optimized (e.g., when computing a bound on the delay distribution).

Resuming the derivation of the delay bounds, note first the equivalence of events

$$W_1(t)>d \Leftrightarrow A_1(t-d)>D_1(t) .$$

By using the definition of the service process from Eq. (5), we can next bound the distribution of $W_1(t)$ as follows

$$\begin{aligned}&\mathbb{P}\left(W_1(t) > d\right)\\ &\quad \le \mathbb{P}\left(A_1(t-d) > A_1 * S_1(t)\right)\\ &\quad = \mathbb{P}\Big(\sup_{0\le s<t-d}\Big\{A_1(s,t-d) - \left[C(t-s) - A_2(s,t-x)\right]_+\\ &\qquad\qquad \cdot I_{\{t-s>x\}}\Big\} > 0\Big) \ . \quad (11)\end{aligned}$$

Here we restricted the range of s from $[0,t]$ to $[0,t-d)$, using the positivity of the '$[\cdot]_+$' operator and the monotonicity of $A_1(s,t)$.

Because x is a free parameter in the FIFO service process construction from Eq. (10), let us choose now $x := d$. With this choice it follows from above that

$$\begin{aligned}&\mathbb{P}\left(W_1(t) > d\right)\\ &\le \mathbb{P}\left(\sup_{0\le s<t-d}\left\{A_1(s,t-d) + A_2(s,t-d) - C(t-s)\right\} > 0\right) .\end{aligned}$$

Finally, by applying Theorem 1 with $u := 0$ and $\sigma := Cd$ (recall in particular that the parameter u encodes the information about scheduling), we get the following
Martingale Delay Bound (FIFO):

$$\mathbb{P}\Big(W_1(t) > d\Big) \le K^n e^{-\gamma C d} \ , \quad (12)$$

where K and γ are given in Theorem 1. Note that the bound is invariant to the number of sub-flows n_1, which is a property characteristic to a *virtual* delay process (for FIFO); such a dependence will be established by changing the measure from a virtual delay process to a packet delay process (see Section V).

2) SP: Here we consider an SP server giving higher priority to the bits of the cross flow $A_2(t)$. We are interested in the delay distribution of the lower priority flow; the case of the higher priority flow is a consequence of the previous FIFO result.

We follow the same procedure of first encoding $A_1(t)$'s service view in a service process, e.g., (see Fidler [19]),

$$S_1(s,t) = C(t-s) - A_2(s,t) \ . \quad (13)$$

now in the case of SP scheduling; recall the previous side remark that $S_1(s,t)$ is also a loose service process for FIFO.

To bound the distribution of $W_1(t)$ we continue the first two lines of Eq. (11) as follows

$$\begin{aligned}&\mathbb{P}\left(W_1(t) > d\right)\\ &\quad \le \mathbb{P}\left(\sup_{0\le s<t-d}\left\{A_1(s,t-d) + A_2(s,t) - C(t-s)\right\} > 0\right) .\end{aligned}$$

By applying Theorem 1 with $u := d$ and $\sigma := 0$, we get the following
Martingale Delay Bound (SP):

$$\mathbb{P}\Big(W_1(t) > d\Big) \le K^n e^{-\gamma C_1 d} \ , \quad (14)$$

where K and γ are given in Theorem 1. Note that, as expected, the SP delay bound recovers the FIFO delay bound from Eq. (12) when there is no cross flow, i.e., in the case when $C_1 = C$.

3) EDF: An EDF server associates the relative deadlines d_1^* and d_2^* with the bits of $A_1(t)$ and $A_2(t)$, respectively. Furthermore, all bits are served in the order of their remaining deadlines, even when they are negative (we do not consider bit losses).

A service process for $A_1(t)$ is for some $x > 0$

$$S_1(s,t) = \left[C(t-s) - A_2(s, t-x+\min\{x,y\})\right]_+ I_{\{t-s>x\}} \ , \quad (15)$$

where $y := d_1^* - d_2^*$ (see Liebeherr *et al.* [26]). This service process generalizes the FIFO one from Eq. (10) (which holds when $y = 0$, i.e., the associated deadlines to the flows are equal), and it also generalizes a previous EDF service process by Li *et al.* [24] (which is restricted to $x = 0$).

To derive the delay bound let us first choose $x := d$, as we did for FIFO. Next we distinguish two cases depending on the sign of y.

If $y \ge 0$ then the continuation of Eq. (11) is

$$\begin{aligned}\mathbb{P}\left(W_1(t) > d\right) &\le \mathbb{P}\Big(\sup_{0\le s<t-d}\Big\{A_1(s,t-d)\\ &\quad + A_2(s, t-d+\min\{d,y\}) - C(t-s)\Big\} > 0\Big) .\end{aligned}$$

By changing the variable $t \leftarrow t + d - \min\{d,y\}$ we get

$$\begin{aligned}\mathbb{P}\left(W_1(t) > d\right) &\le \mathbb{P}\Big(\sup_{0\le s<t-\min\{d,y\}}\Big\{A_1(s, t-\min\{d,y\})\\ &\quad + A_2(s,t) - C(t-s+d-\min\{d,y\})\Big\} > 0\Big) .\end{aligned}$$

(We point out that as we are looking for the steady-state distribution of $W_1(t)$, we can omit the technical details of writing $W_1(t+d-\min\{d,y\})$ above.) We can now apply Theorem 1 with $u := \min\{d,y\}$ (note that both d and y are positive) and $\sigma := C(d - \min\{d,y\})$, and get the following
Martingale Delay Bound (EDF) ($d_1^* \ge d_2^*$ Case):

$$\mathbb{P}\Big(W_1(t) > d\Big) \le K^n e^{\gamma C_2 \min\{d_1^* - d_2^*, d\}} e^{-\gamma C d} \ , \quad (16)$$

where K and γ are given in Theorem 1.

The second case, i.e., $y < 0$, is slightly more complicated. The reason is that $\min\{d,y\} = y$ (see Eq. (15)) such that the continuation of Eq. (11) becomes

$$\begin{aligned}&\mathbb{P}\Big(\sup_{0\le s<t-d}\Big\{A_1(s,t-d) - \left[C(t-s) - A_2(s,t-d+y)\right]_+\\ &\qquad\qquad \cdot I_{\{t-s>d\}}\Big\} > 0\Big) \ . \quad (17)\end{aligned}$$

Note that when $s \in [t-d+y, t-d)$, then one must consider $A_2(s,t-d+y) := 0$ according to the conventions from [26]. Therefore, one can perform the splitting $[0,t-d) = [0,t-d+y) \cup [t-d+y, t-d)$; thereafter, by changing the variable $t \leftarrow t+d$, the continuation of Eq. (17) is

$$\begin{aligned}&\le \mathbb{P}\Big(\max\Big\{\sup_{0\le s<t+y}\left\{A_2(s,t+y) + A_1(s,t) - C(t-s)\right\}\\ &\qquad , \sup_{t+y\le s<t}\left\{A_1(s,t) - C(t-s)\right\}\Big\} > Cd\Big)\\ &\le \mathbb{P}\left(\sup_{0\le s<t+y}\left\{A_2(s,t+y) + A_1(s,t) - C(t-s)\right\} > Cd\right)\\ &\qquad + \mathbb{P}\left(\sup_{0\le s<t}\left\{A_1(s,t) - C(t-s)\right\} > Cd\right)\end{aligned}$$

In the third line we applied the Union Bound [*sic*], which is conceivably tight because the two elements in the '$\max$' are rather uncorrelated. Moreover, we extended the left margin in the last supremum (in the fourth line), as we are looking for upper bounds, whereas the Martingale argument from Theorem 1 is insensitive to where the left margin starts.

The last two probabilities can be directly evaluated with Theorem 1. For the first one we set $u := -y$ (note that y is now negative) and $\sigma := Cd$. For the second one we set $u := 0$, $n_2 := 0$, $\sigma := Cd$, and we properly rescale the per-flow capacity c and utilization factor ρ (see below). In this way we get the following

Martingale Delay Bound (EDF) (**$d_1^* < d_2^*$ Case**):

$$\mathbb{P}\Big(W_1(t) > d\Big) \leq K^n e^{\gamma C_2(d_1^*-d_2^*)} e^{-\gamma Cd} + K'^n e^{-\gamma' Cd} , \quad (18)$$

with the same K and γ from Theorem 1, whereas K' and γ' are obtained alike K and γ, but after rescaling $c' := \frac{n_1+n_2}{n_1} c$ and $\rho' := \frac{n_1}{n_1+n_2}\rho$.

Note that the first EDF bound from Eq. (16) recovers the FIFO bound when the associated deadlines are equal, i.e., when $d_1^* = d_2^*$. In turn, the second EDF bound from Eq. (18) would also recover the FIFO bound, but only by dispensing with the unnecessary splitting of the interval $[0, t-d)$ since $y = 0$.

B. Standard Bounds

Here we list the standard bounds on $A_1(t)$'s virtual delay for FIFO, SP, and EDF (for derivations see Section IV.B in [12]). These bounds will be compared, both analytically and numerically, against the previous Martingale bounds.

$$\textbf{FIFO} \quad : \quad \ldots \leq \inf_{\{\theta : c > r_\theta\}} L e^{-\theta Cd} \quad (19)$$

$$\textbf{SP} \quad : \quad \ldots \leq \inf_{\{\theta : c > r_\theta\}} L e^{-\theta(C - n_2 r_\theta)d} \quad (20)$$

$$\textbf{EDF}^1 \quad : \quad \ldots \leq \inf_{\{\theta : c > r_\theta\}} L e^{\theta n_2 r_\theta \min\{d_1^*-d_2^*, d\}} e^{-\theta Cd} \quad (21)$$

$$\textbf{EDF}^2 \quad : \quad \ldots \leq \inf_{\{\theta : c > r_\theta\}} L e^{\theta(C - n_1 r_\theta)(d_1^*-d_2^*)} e^{-\theta Cd} + \inf_{\{\theta : c' > r_\theta\}} L' e^{-\theta Cd} \quad (22)$$

where $L = \frac{ce}{c - r_\theta}$, $r_\theta = \frac{-b+\sqrt{\Delta}}{2\theta}$, $b = \lambda + \mu - \theta P$, and $\Delta = b^2 + 4\mu\theta P$; for the expression of L' see [12].

C. Many-Sources Asymptotics Comparison

Here we prove that, unlike the Standard delay bounds, the new Martingale bounds capture the conjectured $\mathcal{O}\left(e^{-\alpha n}\right)$ decay factor. The underlying scaling regime is: the total number of flows n is scaled up, whereas the rest of the parameters (the utilization factor ρ, the per-flow rate $r_0 = pP$, and the per-flow capacity c) remain unchanged.

Let us first observe that the factors K (defined in Theorem 1) and L (defined for the Standard bounds after Eq. (22)) from the two sets of bounds satisfy

$$K < 1 \text{ and } L > 1 \ .$$

The second property is immediate. In turn, for the factor K, we first note that the function $h(x) := \rho\left(\frac{\rho - x}{1 - x}\right)^{\frac{x-\rho}{\rho}}$ is decreasing for $0 < x < \rho$ (this can be shown immediately by differentiating $\log h(x)$ and using the inequality $\log(1+x) \geq \frac{x}{1+x}$ $\forall x > -1$. Since $\lim_{x \downarrow 0} h(x) = \rho^{-1}$ and $K = h(p)$ it then follows that $K < 1$.

Delay Bounds / Scheduling	Martingale	Standard
FIFO, SP, EDF	$\mathcal{O}\left(e^{-\alpha n} e^{-\eta d n}\right)$	$\mathcal{O}\left(e^{-\eta d n}\right)$

TABLE I. SCALING LAWS FOR THE MARTINGALE AND STANDARD BOUNDS ON $\mathbb{P}\left(W_1(t) > d\right)$ IN THE TOTAL NUMBER OF FLOWS $n = n_1 + n_2$; α AND η ARE INVARIANT TO n

Table I illustrates the scaling laws of the Martingale and Standard delay bounds for the three scheduling algorithms. The factors $\alpha > 0$ and $\eta > 0$ are invariant to n and can be fitted for each individual case; e.g., in the case of FIFO, $\alpha = -\log K$ and $\eta = \gamma c$. We remark that all pairs of bounds have the same *asymptotic decay rate* η. The critical observation is that, unlike the Standards bounds, the Martingale bounds have an additional factor $e^{-\alpha n}$ decaying exponentially with n. This scaling behavior was conjectured by Choudhury *et al.* [8] through numerical evaluations. We point out that [8] further conjectured an additional factor $\beta > 0$, invariant to n, which is, however not, captured by the Martingale bounds.

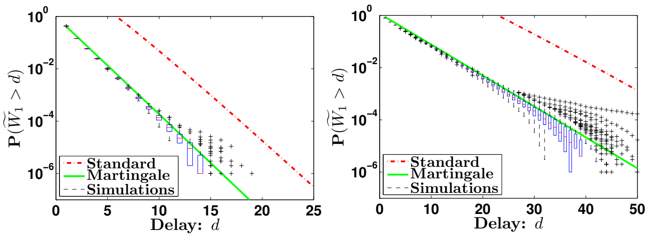
V. Numerical Evaluations

In this section, we compare the Martingale and Standard bounds against simulation results. The parameters of a single MMOO sub-flow are $\lambda = 0.5$, $\mu = 0.1$, and $P = 1$ (the average 'Off' period is five fold the average 'On' period).

We consider two utilization levels ($\rho = 0.75$ and $\rho = 0.9$), and a multiplexing regime with $n_1 = n_2 = 10$. The packet sizes in a packet-level simulator are set to 1; fractional packet sizes are additionally set when the dwell times in the states of the Markov process from Figure 2 are not integers. The simulator measures the delays of the through flow's first 10^7 packets, and it discards the first 10^6. For numerical confidence, 100 independent simulations are being run and the results are presented as box-plots.

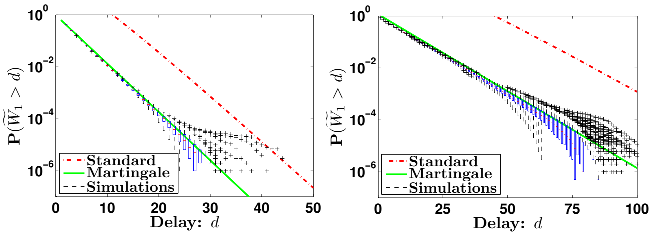
For the soundness of the comparisons against simulations, it is important to remark that the delay analysis so far concerned the *virtual delay process* $W_1(t)$, which corresponds to the delay of a through flow's infinitesimal unit, should it depart, or equivalently arrive, at time t; more concretely, we note that the bounds computed with SNC on virtual delays are identical, should they concern a virtual arrival or departure unit. In the packet level simulator, however, it is the *packet delay process* which is being measured, and which is denoted here by $\widetilde{W}_1(n)$ (the index 'n' corresponds to the packet number for the through flow). Therefore, one has to properly perform a suitable change of probability measures in order to provide meaningful numerical comparisons.

We follow a Palm calculus argument and relate the measure of the virtual delay process to that of the packet delay process (see Shakkottai and Srikant [36]). For convenience, we work in reversed time and focus on time 0 where steady-state is assumed to be reached. Denoting $W_1 := W_1(0)$, we can write by conditioning



(a) $\rho = 0.75$ ($n_1 = n_2 = 10$) (b) $\rho = 0.90$ ($n_1 = n_2 = 10$)

Fig. 3. FIFO delay bounds



(a) $\rho = 0.75$ ($n_1 = n_2 = 10$) (b) $\rho = 0.90$ ($n_1 = n_2 = 10$)

Fig. 4. SP delay bounds

$$
\begin{aligned}
\mathbb{P}(W_1 > d) &= \mathbb{P}(W_1 > d \mid a_1(0) > 0)\,\mathbb{P}(a_1(0) > 0) \\
&+ \mathbb{P}(W_1 > d \mid a_1(0) = 0)\,\mathbb{P}(a_1(0) = 0) \\
&\geq \mathbb{P}(W_1 > d \mid a_1(0) > 0)\,\mathbb{P}(a_1(0) > 0) \\
&= \mathbb{P}\left(\widetilde{W_1} > d\right)\mathbb{P}(a_1(0) > 0) \ , \qquad (23)
\end{aligned}
$$

where $a_1(0)$ denotes the instantaneous arrivals of the through flow at time 0, and $\widetilde{W_1}$ denotes the steady-state packet delay process of the through flow. Note that for the inequality we eliminated the second term in the sum above.

Therefore,

$$
\mathbb{P}\left(\widetilde{W_1} > d\right) \leq \frac{1}{1-(1-p)^n}\mathbb{P}(W_1 > d) \ . \qquad (24)
$$

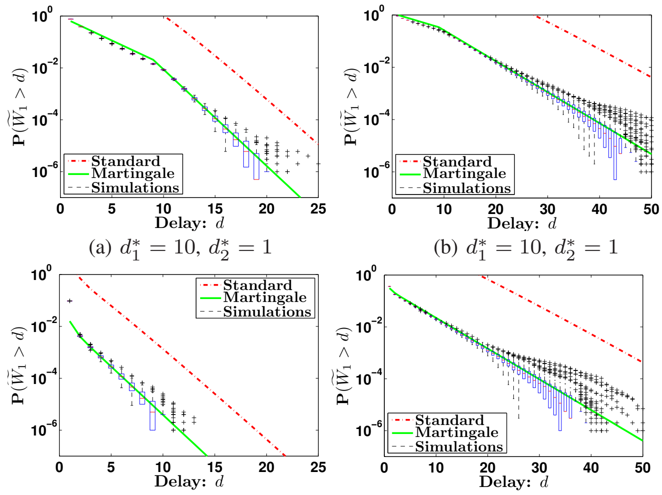
(Recall that p is the steady-state 'On' probability of the MMOO process from Figure 2.)

Next we compare the distribution of $\widetilde{W_1}$ against the one of the measured (simulated) delay process. All the Martingale and Standard bounds which we compute for FIFO, SP, and EDF scheduling are scaled up by the additional prefactor from Eq. (24) needed for the change of measure. We note that this scaling is conservative because of the inequality from Eq. (23).

Figure 3 illustrates the comparisons for FIFO scheduling (recall Eqs. (12) and (19) for the Martingale and Standard delay bounds, which are scaled as in Eq. (24)); the y-axis uses a log scale. The irregular tail behavior (including the presence of many outliers[2,3]at $\rho = 90\%$) of the box-plots is due to the restriction of the simulation runs to 10^7 packets. Both scenarios, corresponding to two utilization levels, clearly indicate that the Standard bounds are very loose, as they overestimate the simulation results by a factor of roughly 10^2 at 75% utilization (see (a)), and even 10^3 at 90% utilization (see (b)). In turn, the Martingale bounds are remarkably accurate even at a 90% utilization level.

[2]Outliers are depicted in the box-plots with the '+' symbol; on each box, the central mark is the median, and the edges of the box are the 25th and 75th percentiles.

[3]The long stretch of the box-plots and the presence of many outliers is caused by the choice of 10^7 arrivals, in order to illustrate the need for very long simulation runs (e.g., 10^8 arrivals, in which case the box-plots would significantly shrink and most outliers would disappear).



(a) $d_1^* = 10$, $d_2^* = 1$ (b) $d_1^* = 10$, $d_2^* = 1$

(c) $d_1^* = 1$, $d_2^* = 10$ (d) $d_1^* = 1$, $d_2^* = 10$

Fig. 5. EDF delay bounds ($n_1 = n_2 = 10$, $\rho = 75\%$ in (a) and (c), and $\rho = 90\%$ in (b) and (d))

The same observations hold for SP scheduling, as indicated by Figure 4; recall the Martingale and Standard delay bounds from Eq. (14) and (20), respectively, which are again scaled as in Eq. (24).

The tightness of the Martingale bounds, in contrast to the looseness of Standard bounds, further holds in the case of EDF scheduling, for both cases (i.e., $d_1^* > d_2^*$ and $d_1^* < d_2^*$), as illustrated in Figure 5; recall the Martingale and Standard delay bounds from Eqs. (16)-(18) and Eqs. (21)-(22), respectively. Note that the bending of the curves, e.g., in (a), is due to the choice of d_1^* and d_2^*: the bounds behave like the SP bounds for $d \leq d_1^* - d_2^*$, and asymptotically like the FIFO bounds thereafter.

The presented numerical results provide thus concrete evidence on the (quite severe) looseness of existing (Standard) SNC probabilistic bounds, and, more importantly, that the new Martingale bounds are remarkably accurate, even in a deliberately low multiplexing regime with only $n_1 = 10$ flows.

VI. Conclusions

In this paper, we have put our finger in a wound of the stochastic network calculus: the lingering issue of the tightness of the SNC bounds. To some degree, this issue has been evaded by the SNC literature for some time although it is a, if not the, crucial one. In fact, we demonstrated that the typical (Standard SNC) way of calculating performance bounds results in loose delay bounds for several scheduling disciplines, i.e., FIFO, SP, and EDF. This becomes particularly obvious when comparing the (Standard SNC) analytical results to simulation results, where discrepancies up to many orders of magnitude can be observed. So, we strongly confirm the often rumored conjecture about SNC's looseness; the same looseness is characteristic to the majority of effective bandwidth results as well.

Yet, the paper does not stop at these bad news, but in an attempt to understand the problems of Standard SNC, which mainly lie in not properly accounting for the correlation structure of the arrival processes (by coarse usage of the Union bound), we find a new way to calculate performance bounds using the SNC framework based on martingale techniques. Here, SNC still serves as the "master method", yet the Union bound is substituted by the usage of martingale inequalities, to make a long story short. Comparing the new Martingale SNC bounds to the simulation results shows that they are remarkably close in most cases, which rehabilitates the SNC as a general framework for performance analysis. So, the SNC can arguably still be regarded as a valuable methodology with the caveat that it has to be used with the right probabilistic techniques in order not to arrive at practically irrelevant results.

As a final remark, the paper advocates a conceptual shift in applying the SNC, by 1) coupling it with the mainstream queueing literature, in particular by "getting a firm grip on arrivals", and 2) carefully leveraging the main two features of SNC (i.e., dealing with scheduling and multi-node) in order to obtain sharp bounds. The next immediate and fundamental challenge is to derive sharp end-to-end delay bounds.

References

[1] D. Anick, D. Mitra, and M. Sondhi. Stochastic theory of a data-handling system with multiple sources. *Bell Systems Technical Journal*, 61(8):1871–1894, Oct. 1982.

[2] A. W. Berger and W. Whitt. Effective bandwidths with priorities. *IEEE/ACM Transactions on Networking*, 6(4):447–460, Aug. 1998.

[3] D. Botvich and N. Duffield. Large deviations, economies of scale, and the shape of the loss curve in large multiplexers. *Queueing Systems*, 20(3-4):293–320, Sept. 1995.

[4] J.-Y. Le Boudec and P. Thiran. *Network Calculus*. Springer Verlag, Lecture Notes in Computer Science, LNCS 2050, 2001.

[5] A. Burchard, J. Liebeherr, and F. Ciucu. On superlinear scaling of network delays. *IEEE/ACM Transactions on Networking*, 19(4):1043–1056, Aug. 2011.

[6] A. Burchard, J. Liebeherr, and S. D. Patek. A min-plus calculus for end-to-end statistical service guarantees. *IEEE Transactions on Information Theory*, 52(9):4105–4114, Sept. 2006.

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

[8] G. Choudhury, D. Lucantoni, and W. Whitt. Squeezing the most out of ATM. *IEEE Transactions on Communications*, 44(2):203–217, Feb. 1996.

[9] F. Ciucu. Network calculus delay bounds in queueing networks with exact solutions. In *20th International Teletraffic Congress (ITC)*, pages 495–506, 2007.

[10] F. Ciucu, A. Burchard, and J. Liebeherr. Scaling properties of statistical end-to-end bounds in the network calculus. *IEEE Transactions on Information Theory*, 52(6):2300–2312, June 2006.

[11] F. Ciucu, R. Khalili, Y. Jiang, L. Yang, and Y. Cui. Towards a system theoretic approach to wireless network capacity in finite time and space. In *IEEE Infocom*, 2014.

[12] F. Ciucu, F. Poloczek, and J. Schmitt. Sharp bounds in stochastic network calculus. *CoRR*, abs/1303.4114, 2013.

[13] F. Ciucu and J. Schmitt. Perspectives on network calculus - No free lunch but still good value. In *ACM Sigcomm*, 2012.

[14] C. Courcoubetis and R. Weber. Effective bandwidths for stationary sources. *Probability in Engineering and Informational Sciences*, 9(2):285–294, Apr. 1995.

[15] R. Cruz. A calculus for network delay, parts I and II. *IEEE Transactions on Information Theory*, 37(1):114–141, Jan. 1991.

[16] R. L. Cruz. SCED+: Efficient management of quality of service guarantees. In *IEEE Infocom*, pages 625–634, 1998.

[17] N. G. Duffield. Exponential bounds for queues with markovian arrivals. *Queueing Systems*, 17(3-4):413–430, Sept. 1994.

[18] S. N. Ethier and T. G. Kurtz. *Markov processes – characterization and convergence*. John Wiley & Sons Inc., 1986.

[19] M. Fidler. An end-to-end probabilistic network calculus with moment generating functions. In *IEEE International Workshop on Quality of Service (IWQoS)*, pages 261–270, 2006.

[20] Y. Jiang and Y. Liu. *Stochastic Network Calculus*. Springer, 2008.

[21] F. P. Kelly. *Reversibility and Stochastic Networks*. Wiley, Chichester, 1979.

[22] F. P. Kelly. Notes on effective bandwidths. In *Stochastic Networks: Theory and Applications. (Editors: F.P. Kelly, S. Zachary and I.B. Ziedins) Royal Statistical Society Lecture Notes Series, 4*, pages 141–168. Oxford University Press, 1996.

[23] J. F. C. Kingman. Inequalities in the theory of queues. *Journal of the Royal Statistical Society, Series B*, 32(1):102–110, 1970.

[24] C. Li, A. Burchard, and J. Liebeherr. A network calculus with effective bandwidth. *IEEE/ACM Transactions on Networking*, 15(6):1442–1453, Dec. 2007.

[25] J. Liebeherr, A. Burchard, and F. Ciucu. Delay bounds in communication networks with heavy-tailed and self-similar traffic. *IEEE Transactions on Information Theory*, 58(2):1010–1024, Feb. 2012.

[26] 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*, 29(5):1009–1020, May 2011.

[27] L. Lipsky. *Queueing Theory: A Linear Algebraic Approach*. Springer Publishing Company, Incorporated, 2nd edition, 2008.

[28] Z. Liu, P. Nain, and D. Towsley. Exponential bounds with applications to call admission. *Journal of the ACM*, 44(3):366–394, May 1997.

[29] R. Lübben and M. Fidler. Non-equilibrium information envelopes and the capacity-delay-error-tradeoff of source coding. *CoRR*, abs/1107.3087, 2011.

[30] M. Mandjes. *Large Deviations for Gaussian Queues: Modelling Communication Networks*. Wiley, 2007.

[31] M. F. Neuts. *Matrix-Geometric Solutions in Stochastic Models: An Algorithmic Approach*. Dover, 1981.

[32] I. Norros. On the use of fractional Brownian motion in the theory of connectionless networks. *IEEE Journal on Selected Areas in Communications*, 13(6):953–962, Aug. 1995.

[33] Z. Palmowski and T. Rolski. A note on martingale inequalities for fluid models. *Statistics & Probability Letters*, 31(1):13–21, Dec. 1996.

[34] V. Ramaswami. Matrix analytic methods for stochastic fluid flows. In *16th International Teletraffic Congress (ITC)*, pages 1019–1030, 1999.

[35] A. Rizk and M. Fidler. Non-asymptotic end-to-end performance bounds for networks with long range dependent fBm cross traffic. *Computer Networks*, 56(1):127–141, Jan. 2012.

[36] S. Shakkottai and R. Srikant. Delay asymptotics for a priority queueing system. In *ACM Sigmetrics*, pages 188–195, 2000.

[37] N. B. Shroff and M. Schwartz. Improved loss calculations at an ATM multiplexer. *IEEE/ACM Transactions on Networking*, 6(4):411–421, Aug. 1998.

[38] V. Sivaraman and F. M. Chiussi. Statistical analysis of delay bound violations at an earliest deadline first scheduler. *Performance Evaluation*, 36(1):457–470, Aug. 1999.

[39] M. Talagrand. Majorizing measures: The generic chaining. *Annals of Probability*, 24(3):1049–1103, July 1996.

[40] D. Wischik. Sample path large deviations for queues with many inputs. *Annals of Applied Probability*, 11(2):379–404, May 2001.

[41] Y. Yuan, K. Wu, W. Jia, and Y. Jiang. Performance of acyclic stochastic networks with network coding. *IEEE Transactions on Parallel and Distributed Systems*, 22(7), July 2011.

[42] K. Zheng, F. Liu, L. Lei, C. Lin, and Y. Jiang. Stochastic performance analysis of a wireless finite-state Markov channel. *IEEE Transactions on Wireless Communications*, 12(2):782–793, Feb. 2013.

[43] H. Al-Zubaidy, J. Liebeherr, and A. Burchard. A (min, $\times$) network calculus for multi-hop fading channels. In *IEEE Infocom*, 2013.