Stochastic Upper and Lower Bounds for General Markov Fluids

Florin Ciucu
University of Warwick

Felix Poloczek
University of Warwick / TU Berlin

Jens Schmitt
University of Kaiserslautern

***Abstract*—Promising perspectives of a hypothetical 'Tactile Internet', or 'Internet at the speed of light', whereby network latencies become imperceptible to users, have (again) triggered a broad interest to understand and mitigate Internet latencies. In this paper we revisit the queueing analysis of the versatile Markov Fluid traffic model, which was mainly investigated in the 1980-90s, yet with low accuracy. We derive upper bounds on the tail distribution of the queue size, which improve state-of-the-art results by an exponential factor $\mathcal{O}(\kappa^n)$ in a special case, where $0 < \kappa < 1$ and n is the number of multiplexed sources; additionally, we provide the first lower bounds. The underlying results are quite general in that they can be easily adapted to derive the delay distribution for SP, FIFO, and EDF scheduling. Our overall results rely on a powerful martingale methodology which was recently shown to be highly accurate.**

I. Introduction

Latency/delay is a fundamental metric of communication and human perception. Not surprisingly, several recent studies reported a strong correlation between Internet latencies and the revenue of major online service providers, e.g., Google, Bing, or Amazon [39], [26], [42]; a typical cited argument is that an additional 100ms in latency would cost Amazon 1% of sales. Moreover, innovations in electronics, systems, and protocols, together with the apparent abundance of computing resources and network bandwidth, have recently sparked the hope for a major paradigm shift called 'Tactile Internet' [22], or 'Internet at the speed of light' [41]. A potential major benefit of consistently achieving negligible network latencies (e.g., sub-100ms or sub-10ms) would be the immediate opportunity for major innovations in exciting areas such as health-care (e.g., remote surgery), transportation (e.g., fully automatic driving), or entertaining.

The increasing and broadly recognized importance of network latency has recently materialized, at a large scale, into a dedicated workshop [2] and the funding of a major European Union FP-7 project [1]; a relevant outcome of such concerted efforts was a comprehensive survey on a taxonomy of latency sources and mitigating techniques [8]. While there has been much work on system solutions such as traffic engineering and replication strategies to reduce latencies [19], there has also been shown that achieved improvements reached a point of diminishing returns [31], particularly due to the dominating queueing delays (i.e., the time spent by packets in buffers, waiting to be forwarded).

As the development of system solutions to mitigate Internet latencies can strongly benefit from the fundamental/theoretical understanding of network delays, there is a need for the parallel development of related analytical tools. A similar moment occurred in the 1980-90s, when novel analytical/queueing tools emerged to especially cope with the case of non-renewal arrivals. To some extent, this goal was driven by the increasing prevalence of audio and video content, which, when viewed as stochastic processes, are subject to some form of statistical correlations. In fact, such a new characteristic of Internet traffic determined the questioning of the common and analytically convenient assumption of Poisson arrivals (at the packet level), which was convincingly shown to be largely misleading when improperly used [36]. Besides accounting for non-renewal arrivals, an additional goal of emerging analytical theories was to permit the queueing/delay analysis over a network path; such a problem is in itself extremely challenging in the case of non-Poisson arrivals, as attested by the state-of-the-art in the classical queueing theory.

A general queueing tool which can deal with non-renewal arrivals over a network path is the network calculus [7], [10]. Initially conceived by Cruz as a deterministic queueing theory [17], its main and radically new conceptual characteristic of deterministically bounding arrivals was extended in a probabilistic framework to account for not necessarily bounded arrivals. Given its wide modelling scope and also the ability to render (deterministic) worst-case measures for network queueing metrics such as end-to-end delays, the deterministic branch of network calculus has some notable practical applications, e.g., towards the certification of the Airbus A380 AFDX backbone [5]. The broad applicability of (deterministic) network calculus comes however at the price of providing bounds (i.e., not exact results), whose practical tightness in terms of efficient numerical algorithms remains an open issue in the network case [6] despite significant recent progress [3]. In turn, the stochastic branch of network calculus suffers from similar low accuracy issues, including the single-node case whereby the stochastic bounds can be loose by orders of magnitude in the case of non-renewal arrivals [14]; this drawback is exacerbated in the network case, especially under heavy-tailed arrivals [32].

In a stochastic setting, an alternative queueing tool is the theory of effective bandwidth [27] which can address a broad class of (non-renewal) arrival processes in a unified and conceivably quite elegant manner. Unlike stochastic network calculus, which yields results in terms of non-asymptotic bounds, effective bandwidth provides exact results but only

in the so-called large-buffer or many-sources asymptotic regimes [34]. Unfortunately, when fitted to practical (finite) regimes, such results are however largely inaccurate and thus practically questionable [12]; a convincing explanation is provided in [40], i.e., for more variable sources than Poisson such as most Markov-modulated processes, the underlying *additivity property* reveals both the elegant nature of effective bandwidth but also the conceivably conservative nature of the exact results in finite regimes.

Motivated by the need for a unified queueing tool to address non-renewal arrivals, and more importantly in terms of practically accurate results, this paper revisits the queueing analysis of the versatile class of Markov Fluid traffic models. Informally, a source is a Markov Fluid if the associated arrival process is a function of a (continuous-time) Markov process on a finite state space [28]. The seminal model is the so-called Anick/Mitra/Sondhi On-Off model, in which a source produces data (as continuous 'fluid') either at some positive constant or zero rates, depending on the state of a modulating Markov process [4]; for related generalizations, which are able to model larger classes of traffic patterns (e.g., video), see [43], [28]. It is worth pointing out that fluid models are the continuous approximation of their discrete counterpart, and are essentially motivated by computational reasons such as dealing with (discrete) *state space explosion* or various underlying *granularity and sizes* (e.g., modelling a large number of small packets in a short time interval) [24].

In this paper we generalize existing stochastic upper bounds, available for the specific On-Off process [14], to the case of general Markov fluids. The obtained stochastic (upper) bound on the steady-state queue size distribution is shown to be tighter than the existing state-of-the-art result from [35] by an exponential factor $\mathcal{O}\left(\kappa^n\right)$, in the case of a superposition of n On-Off sources, where κ is an explicit constant with $0 < \kappa < 1$. Additionally, we provide for the first time the matching lower bounds, which have the same analytical structure as the upper bounds. Furthermore, by leveraging the framework of the stochastic network calculus, we provide (per-flow) delay bounds in a scheduling scenario with either FIFO (first-in-first-out), SP (static priority), or EDF (earliest deadline first) scheduling, when the flows are Markov fluids as well.

Unlike most of alternative results from the effective bandwidth (1980-90s), e.g., [28], [16] and stochastic network calculus literature (1990s-), e.g., [10], [13], [23], our overall results employ a powerful martingale approach which was proposed by Kingman to derive bounds in GI/G/1 (renewal) queues [29]. This methodology was adopted in [20], [35], [10] to the case of non-renewal processes. The numerical tightness of the obtained upper bounds on the queue-size distribution, and their noticeable superiority to alternative effective bandwidth/stochastic network results, was recently exposed in [14], [37], who further provided per-flow queueing results under several scheduling policies with similar high accuracy. An additional practical benefit of the martingale approach is that it provides the most critical component of latency, i.e., its tail (e.g., the 95^{th}-percentile) and not its average; in fact, the importance of the tail to the development of the much desired and anticipated "latency tail-tolerant" systems has been convincingly exposed in a recent Google study [19].

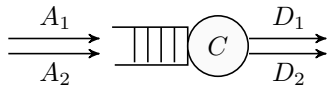


Fig. 1. A queueing system with two arrival processes

The remainder of the paper is organized as follows: In § II we present a general tool to analyze Markov fluids in an infinite buffer queueing model. This tool is subsequently applied to derive upper and lower bounds on the queue size distribution (§ III), and to the (per-flow) delay bounds in a scheduled system (§ IV). In § V we provide a collateral but important result, i.e., the first analytical construction for the space-parameter in the classical effective bandwidth approximation. Finally in § VI we summarize the paper.

II. A General Queueing Result

In this section we first derive a general queueing result which can be instantiated to several scenarios, e.g., the queue size distribution for a single process or the delay distribution in the case of scheduling.

The general queueing model is depicted in Figure 1. Given a continuous time model, $A_1(t)$ and $A_2(t)$ are two (cumulative) Markov fluid processes served at some constant rate $C > 0$ with corresponding departure processes $D_1(t)$ and $D_2(t)$, respectively. Each process $A_k(t)$, $k = 1, 2$, is modulated by a reversible Markov process $Z_k(t)$ with $n_k + 1$ states, generator $\mathbf{Q}_k = (q_{k,i,j})_{i,j=0,\ldots,n_k}$, equilibrium distribution $\pi_k = (\pi_{k,0}, \ldots, \pi_{k,n_k})$, arrival rates $\mathbf{r}_k = (r_{k,0}, \ldots, r_{k,n_k})$, and increment process $a_k(s) = r_{k,Z_k(s)}$ $\forall s \geq 0$. We remark that if $Z_k(t)$ was not reversible, then one could consider as input the corresponding reversed process. The reversibility property enables expressing Reich's equation for the steady-state queue size as

$$Q = \sup_{t \geq 0} \left\{ A_1(t) + A_2(t) - Ct \right\} \ .$$

For each arrival process $A_k(t)$ we consider the generalized eigenvalue problem

$$\mathbf{Q}_k \mathbf{h}_k = -\gamma_k \mathbf{u}_k \mathbf{h}_k, \ k = 1, 2 \ , \tag{1}$$

where $\mathbf{Q}_k$ are the underlying generators, γ_k are the eigenvalues, $\mathbf{h}_k$ are the right eigenvectors, $\mathbf{u}_k$ are diagonal matrices with $(u_{k,0}, u_{k,1}, \ldots, u_{k,n_k})$ on the diagonal, and where

$$u_{k,j} = r_{k,j} - C_k$$

are the instantaneous queueing drifts for $j = 0, 1, \ldots, n_k$. Here, C_1 and C_2 are positive values such that $C_1 + C_2 = C$, and can be regarded as the per-class/flow allocated capacity.

Assuming the per-class stability conditions

$$\sum_{j=0}^{n_k} \pi_{k,j} u_{k,j} < 0, \ k = 1, 2 \ , \tag{2}$$

Lemma 5.1 from [35] guarantees the existence of real generalized eigenvalues $-\gamma_k$ (as the ones with the biggest negative real parts) and also of the generalized eigenvectors

$$\mathbf{h}_k = (h_{k,0}, h_{k,1}, \ldots, h_{k,n_k})^{\mathrm{T}}$$

with positive coordinates. Thus, $\gamma_k > 0$ for $k = 1, 2$.

We now present a general result which will be used throughout the paper.

Lemma 1. *Consider the single-node queueing scenario from Figure 1 and the solutions for the generalized eigenvalue problems from Eq. (1). Then for all $0 \leq u \leq t$ and $\sigma \geq 0$*

$$\begin{aligned}\mathbb{P}\left(\sup_{0\leq s<t-u}\{A_1(s,t-u)+A_2(s,t)-C(t-s)\}>\sigma\right)\\ \leq \inf_{0\leq\gamma\leq\min\gamma_k}\inf_{C_1+C_2=C}\kappa e^{-\gamma(C_1u+\sigma)}\ ,\end{aligned} \tag{3}$$

where

$$\kappa = \frac{\sum_{i,j}\pi_{1,i}\pi_{2,j}h_{1,i}^{\frac{\gamma}{\gamma_1}}h_{2,j}^{\frac{\gamma}{\gamma_2}}}{\min_{u_{1,i}+u_{2,j}\geq 0}h_{1,i}^{\frac{\gamma}{\gamma_1}}h_{2,j}^{\frac{\gamma}{\gamma_2}}}\ ,$$

whereas the condition $C_1 + C_2 = C$ is subject to the stability conditions from Eq. (2).

Let us make several observations about the two infimum operators. The parameter γ in the outer infimum reconciles the different burstiness of the two not necessarily homogeneous flows $A_1(t)$ and $A_2(t)$, loosely expressed through the exponential decay factor. The extreme optimal value $\gamma = \min\{\gamma_1, \gamma_2\}$ is attained when $\sigma \to \infty$; in turn, a numerical optimization after γ is necessary in finite regimes of σ. In turn, due to the implicit expression of κ in terms of the (generalized) eigenvectors from Eq. (1), which depend on C_1 and C_2, the values for the inner infimum are subject to further numerical optimizations.

Lemma 1 generalizes a result from [14] (see Theorem 1 therein) to the case of general and not necessarily homogeneous Markov fluid processes. The theorem also generalizes a result from [35] (see Proposition 5.1 therein), restricted to $A_2(t) = 0$, and also the seminal result from Kingman [29] to the non-renewal case. We point out that the key benefit of our generalization result is that it can lend itself to per-flow delay bounds in a scheduling scenario (see § IV).

Proof. Fix $u \geq 0$ and $\sigma \geq 0$. Since the two arrival processes are reversible, we can rewrite the probability from Eq. (3), by shifting the time origin, as

$$\begin{aligned}&\mathbb{P}\left(\sup_{t>u}\{A_1(u,t)+A_2(t)-Ct\}\geq\sigma\right)\\ &\quad=\mathbb{P}\Bigg(\sup_{t>u}\{A_1(u,t)+A_2(u,t)-C(t-u)\}\\ &\qquad+A_2(u)-C_2u>C_1u+\sigma\Bigg)\ .\end{aligned} \tag{4}$$

Let the following stopping time

$$\begin{aligned}T := \inf\Bigg\{t>u : A_1(u,t)+A_2(u,t)-C(t-u)\\ +A_2(u)-C_2u>C_1u+\sigma\Bigg\}\ .\end{aligned} \tag{5}$$

In the rest of the proof we shall bound $\mathbb{P}(T < \infty)$, which is exactly the target probability from Eq. (4).

Let $\mathbb{P}_{i,j}$ denote the underlying probability measure conditioned on $Z_1(u) = i$ and $Z_2(0) = j$, for $0 \leq i \leq n_1$ and $0 \leq j \leq n_2$. Next we define the following two processes

$$\begin{aligned}\widetilde{M}_{1,t} &:= \frac{h_{1,Z_1(t)}}{h_{1,i}}e^{-\int_u^t\frac{(\mathbf{Q}_1\mathbf{h}_1)_{Z_1(s)}}{h_{1,Z_1(s)}}ds}\ \forall t\geq u \text{ and}\\ \widetilde{M}_{2,t} &:= \frac{h_{2,Z_2(t)}}{h_{2,j}}e^{-\int_0^t\frac{(\mathbf{Q}_2\mathbf{h}_2)_{Z_2(s)}}{h_{2,Z_2(s)}}ds}\ \forall t\geq 0\ .\end{aligned}$$

$M_1(t)$ and $M_2(t)$ are martingales with respect to (wrt) $\mathbb{P}_{i,j}$ and the natural filtration (see [21], p. 175). Considering the solution of the generalized eigenvalue problem from Eq. (1), we can rewrite

$$\begin{aligned}\widetilde{M}_{1,t} &= \frac{h_{1,Z_1(t)}}{h_{1,i}}e^{\gamma_1\int_u^t u_{1,Z_1(s)}ds}\ \forall t\geq u \text{ and}\\ \widetilde{M}_{2,t} &= \frac{h_{2,Z_2(t)}}{h_{2,j}}e^{\gamma_2\int_0^t u_{2,Z_2(s)}ds}\ \forall t\geq 0\ .\end{aligned}$$

For $0 \leq \gamma \leq \min\{\gamma_1, \gamma_2\}$ we consider the transformations

$$M_{k,t} = \widetilde{M}_{k,t}^{\frac{\gamma}{\gamma_k}}\ k = 1, 2\ .$$

Denoting by $\mathcal{F}_{k,s}$ the natural filtrations of $M_{k,t}$ we can write for $0 \leq s \leq t$

$$\begin{aligned}E[M_{k,t}\mid\mathcal{F}_{k,s}] &= E\left[\widetilde{M}_{k,t}^{\frac{\gamma}{\gamma_k}}\,\middle|\,\mathcal{F}_{k,s}\right]\leq E\left[\widetilde{M}_{k,t}\,\middle|\,\mathcal{F}_{k,s}\right]^{\frac{\gamma}{\gamma_k}}\\ &\leq \widetilde{M}_{k,s}^{\frac{\gamma}{\gamma_k}} = M_{k,s}\ ,\end{aligned}$$

where the first line is due to Jensen's inequality (applied to the concave function $x \mapsto x^{\frac{\gamma}{\gamma_k}}$ for $x \geq 0$) and the second inequality is due to the martingale property of $\widetilde{M}_{k,t}$. Therefore, the new processes $M_{k,t}$ are also martingales; we point out that their construction is motivated by the need of having the same decay rate, i.e., γ, in the corresponding exponentials.

Next we invoke a result from [11] (stating that the product of two independent martingales is also a martingale) and the Optional Switching Theorem ([25], p. 488), and obtain that the 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 definition). It can be explicitly written as

$$M_t = \begin{cases}\left(\frac{h_{2,Z_2(t)}}{h_{2,j}}\right)^{\frac{\gamma}{\gamma_2}}e^{\gamma(A_2(t)-C_2t)}, t\leq u\\ \left(\frac{h_{1,Z_1(t)}}{h_{1,i}}\right)^{\frac{\gamma}{\gamma_1}}\left(\frac{h_{2,Z_2(t)}}{h_{2,j}}\right)^{\frac{\gamma}{\gamma_2}}e^{\gamma(A_1(u,t)+A_2(t)-Ct)}, t>u\end{cases}$$

Referring now to the stopping time T from Eq. (5), which may be unbounded, we construct the bounded stopping times $T \wedge v$ for all $v \in \mathbb{N}$. For these times, the Optional Sampling Theorem (see, e.g., [25], p. 489) yields

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

for all $v \in N$, where the expectations are taken wrt the underlying probability measures $\mathbb{P}_{i,j}$. Moreover, from the definition of T as an infimum over a set, it holds for $v \geq 0$ that

$$\left(u_{1,Z_1(T)} + u_{2,Z_2(T)}\right) I_{\{T\leq v\}} \geq 0 \ , \tag{6}$$

where $I_{\{.\}}$ denotes the indicator function. Using now that $E_{i,j}\left[M_0\right] = 1$ we obtain for $v > u$

$$\begin{aligned} 1 &\geq E_{i,j}\left[M_{T\wedge v} I_{\{T\leq v\}}\right] \\ &\geq \min\left(\frac{h_{1,l_1}}{h_{1,i}}\right)^{\frac{\gamma}{\gamma_1}} \left(\frac{h_{2,l_2}}{h_{2,j}}\right)^{\frac{\gamma}{\gamma_2}} e^{\gamma(C_1 u+\sigma)} \mathbb{P}_{i,j}\left(T \leq v\right) \ , \end{aligned}$$

where the 'min' operator is taken over the set $\{(l_1, l_2) : u_{1,l_1} + u_{2,l_2} \geq 0\}$ according to Eq. (6).

Finally, by deconditioning on i and j (recall that $Z_1(u)$ and $Z_2(0)$ are in steady-state by construction) we obtain

$$\mathbb{P}\left(T \leq v\right) \leq \kappa e^{-\gamma(C_1 u+\sigma)} \ .$$

Letting $v \to \infty$ completes the proof. □

III. Queue Size Distribution: Upper and Lower Bounds

Here we apply Lemma 1 to derive upper bounds on the steady-state queue size distribution, and then provide the corresponding lower bounds.

A. Upper Bounds

Consider a server with constant rate $C > 0$ serving a Markov fluid process $A(t)$ which is modulated by a reversible Markov process $Z(t)$ with $n+1$ states, generator $\mathbf{Q} = (q_{i,j})_{i,j=0,\ldots,n}$, equilibrium distribution $\pi = (\pi_0, \ldots, \pi_n)$, arrival rates $\mathbf{r} = (r_0, \ldots, r_n)$, and increment process $a(s) = r_{Z(s)}\ \forall s \geq 0$. The steady-state queue size is

$$Q = \sup_{t\geq 0}\left\{A_1(t) + A_2(t) - Ct\right\} \ ,$$

and the steady-state virtual delay is defined via

$$\begin{aligned} \{W(t) \geq d\} &= \{A_1(t-d) + A_2(t-d) \geq D_1(t)\} \\ &\subseteq \left\{\sup_{t\geq 0}\left(A_1(t) + A_2(t) - Ct\right) \geq Cd\right\} \ . \end{aligned}$$

Let the generalized eigenvalue problem

$$\mathbf{Q}\mathbf{h} = -\gamma \mathbf{u}\mathbf{h} \ , \tag{7}$$

where $\mathbf{u}$ is a diagonal matrix with $(u_0, u_1, \ldots, u_n)$ on the diagonal, and where

$$u_j = r_j - C$$

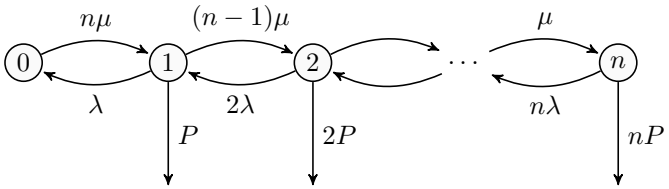


Fig. 2. A superposition of n On-Off processes

are the instantaneous queueing drifts for $j = 0, 1, \ldots, n$. Assume further the stability condition

$$\sum_{j=0}^{n} \pi_j u_j < 0 \ .$$

As mentioned earlier, there exists a real generalized eigenvalue $-\gamma < 0$, and the corresponding eigenvector

$$\mathbf{h} = (h_0, h_1, \ldots, h_n)^{\mathrm{T}}$$

with positive coordinates.

An immediate consequence of Lemma 1, instantiated with $A_2(t) := 0$ and $u := 0$, is the following:

Corollary 2. *(Queue Size Distribution: Upper Bound) Consider the previous queueing system with a single Markov fluid $A(t)$ served at rate C. Then the stationary queue size distribution Q and waiting time distribution $W(t)$ satisfy for all $\sigma, t \geq 0$*

$$\mathbb{P}\left(Q \geq \sigma\right) \leq \frac{\sum_i \pi_i h_i}{\min_{u_i\geq 0} h_i} e^{-\gamma\sigma} \ , \tag{8}$$

$$\mathbb{P}\left(W(t) \geq d\right) \leq \frac{\sum_i \pi_i h_i}{\min_{u_i\geq 0} h_i} e^{-\gamma C d} \ , \tag{9}$$

where γ and $\mathbf{h} = (h_0, h_1, \ldots, h_n)^T$ are the solutions of the generalized eigenvalue problem from Eq. (7).

Let us next compare this bound with the state-of-the-art bound from [35], i.e.,

$$\mathbb{P}\left(Q \geq \sigma\right) \leq \frac{\sum_i \pi_i h_i}{\min_i h_i} e^{-\gamma\sigma} \ . \tag{10}$$

$$\mathbb{P}\left(W(t) \geq d\right) \leq \frac{\sum_i \pi_i h_i}{\min_i h_i} e^{-\gamma C d} \ . \tag{11}$$

Clearly, our bound is tighter due to the additional constraint on the 'min' operator from Eq. (8).

To give an explicit order of the improvement, consider the classical scenario when the process $A(t)$ is a superposition of n Markov-modulated On-Off processes (see Figure 2). Each sub-process is modulated by a Markov process with two states, denoted by 'On' and 'Off', and which communicate at rates λ and μ. While in the 'On' state, each sub-process generates 'fluid' at some constant rate P. To avoid trivial situations we assume that $nP > C$ and that the utilization factor $\rho = \frac{n\frac{\mu}{\lambda+\mu}P}{C}$ satisfies the stability condition $\rho < 1$.

A key advantage of the chosen multiplexed On-Off model is that it lends itself to an *explicit* solution for the generalized eigenvalue problem from Eq. (7), i.e.,

$$\left\{ \begin{array}{l} \gamma = \frac{n(\lambda+\mu)(1-\rho)}{nP-C} \\ h_i = e^{-\theta i}, i = 0, 1, \ldots, n \end{array} \right. , \tag{12}$$

where $\theta = \log \frac{\mu}{\lambda} \frac{nP-C}{C}$ (see also [35]). Thus, an explicit bound in Corollary 2 is

$$\mathbb{P}(Q \geq \sigma) \leq \kappa^n e^{-\gamma\sigma} , \tag{13}$$

where $\kappa = \rho \left(\frac{\rho-p}{1-p} \right)^{\frac{p}{\rho}-1}$ and $p = \frac{\mu}{\lambda+\mu}$ is the steady-state probability for an On-Off process to be in the 'On' state. As it was shown in [14] that $0 < \kappa < 1$, whereas the prefactor from Eqs. (10) and (11) is clearly greater than 1, it follows that the improvement of our bound from Corollary 2 relative to the one from Eq. (10) and (11) is of the order $\mathcal{O}(\kappa^n)$ (in the specific case of a superposition of On-Off processes).

The technical explanation for this drastic improvement can be found in the proof of Lemma 1. More specifically, the key observation resides in Eq. (6): in the current case of a single fluid $A(t)$, the equation 'says' that there must be a non-negative drift $a(Z(T)) - C$ when the stopping time T is attained; although seemingly elementary, this observation does have the reported drastic impact on the bounds. We point out that this observation is reminiscent of the works in [9], [35].

B. Lower Bounds

We now provide the matching lower bounds for the upper bounds from Corollary 2. We consider the same scenario from § III-A with a single Markov fluid $A(t)$ served at rate $C > 0$.

Corollary 3. *(*QUEUE SIZE DISTRIBUTION: LOWER BOUND*) The stationary queue size distribution Q and waiting time distribution $W(t)$ satisfy for all $\sigma, t \geq 0$*

$$\mathbb{P}(Q \geq \sigma) \geq \frac{\sum_i \pi_i h_i}{\max_{u_i \geq 0} h_i} e^{-\gamma\sigma} , \tag{14}$$

$$\mathbb{P}(W(t) \geq d) \geq \frac{\sum_i \pi_i h_i}{\max_{u_i \geq 0} h_i} e^{-\gamma C d} , \tag{15}$$

where γ and $\mathbf{h} = (h_0, h_1, \ldots, h_n)^T$ are the solutions of the generalized eigenvalue problem from Eq. (7).

Remarkably, the only difference to the corresponding upper bound is that the factor $\min_{u_i \geq 0} h_i$ from Eq. (8) is now replaced by $\max_{u_i \geq 0} h_i$. In particular, in a scenario where the set $\{u_i \mid u_i \geq 0\}$ only consists of a single element, upper and lower bounds coincide and hence, Eqs. (8) and (14) (as well as Eqs. (9) and (15)) provide exact results. The proof of Corollary 3 is similar to that of Lemma 1, but it uses an additional stopping time following an idea from [38] to derive bounds in the GI/G/1 (renewal) queue:

Proof. Fix $\sigma \geq 0$. The queue size distribution can be written as

$$\mathbb{P}(Q \geq \sigma) = \mathbb{P}\left(\sup_{t \geq 0} (A(t) - Ct) \geq \sigma \right) . \tag{16}$$

As in the proof of Lemma 1, let $\mathbb{P}_i$ denote the underlying probability measure conditioned on $Z(0) = i$ for $0 \leq i \leq n$ (recall that $Z(t)$ is the underlying modulating Markov process of $A(t)$). Again, the process

$$M_t \quad := \quad \frac{h_{Z(t)}}{h_i} e^{-\int_0^t \frac{(\mathbf{Q}\mathbf{h})_{Z(s)}}{h_{Z(s)}} ds} \ \forall t \geq 0$$

is a martingale wrt $\mathbb{P}_i$ and the natural filtration. Considering the solution of the generalized eigenvalue problem from Eq. (7), we have $\forall t \geq 0$

$$\begin{aligned} M_t &= \frac{h_{Z(t)}}{h_i} e^{\gamma \int_0^t u_{Z(s)} ds} \\ &= \frac{h_{Z(t)}}{h_i} e^{\gamma(A(t)-Ct)} . \end{aligned}$$

Let us now define the stopping time

$$T = \inf \{ t \geq 0 \mid A(t) - Ct \geq \sigma \}$$

which is the equivalent of the one from Eq. (5), in the current context with a single fluid. Define further a second stopping time for some $y > 0$:

$$T_y = \min \{ T, \inf \{ t \geq 0 \mid A(t) - Ct \leq -y \} \} , \tag{17}$$

as the first hitting time of the boundary of the interval $[-y, \sigma]$. Since T_y is a finite stopping time, relative to the natural filtration $\mathcal{F}_t$, it follows from the Optional Stopping Theorem that the process $(M_{T_y \wedge v})_v$ is a martingale, which is bounded and hence uniformly integrable. Since $M_{T_y \wedge v} \to M_{T_y}$ a.s. and in L^1, we can further write

$$\begin{aligned} &E_i[M_0] = E_i\left[M_{T_y \wedge 0}\right] = E_i\left[M_{T_y}\right] \\ &= E_i\left[M_{T_y} \mid A(T_y) \geq CT_y + \sigma\right] \mathbb{P}_i(A(T_y) \geq CT_y + \sigma) \\ &\quad + E_i\left[M_{T_y} \mid A(T_y) \leq CT_y - y\right] \mathbb{P}_i(A(T_y) \leq CT_y - y) , \end{aligned} \tag{18}$$

where the expectations are taken wrt $\mathbb{P}_i$.

To deal with the first term we first observe that

$$\{A(T_y) \geq CT_y + \sigma\} \quad \Rightarrow \quad \{T_y = T\} .$$

Because the process $A(t)$ is continuous it follows that the 'hitting' condition from Eq. (17) is attained with equality, i.e.,

$$A(T_y) = CT_y + \sigma .$$

Moreover, since $\gamma > 0$, we can bound the conditional expectations from Eq. (18) as follows

$$\begin{aligned} 1 &= E_i[M_0] \\ &\leq \frac{\sup_{l, u_l \geq 0} h_l}{h_i} e^{\gamma\sigma} \mathbb{P}_i(A(T_y) \geq CT_y + \sigma) \\ &\quad + \frac{\sup_{l, u_l \geq 0} h_l}{h_i} e^{-\gamma y} . \end{aligned}$$

Letting $y \to \infty$ the second term vanishes and thus

$$1 \leq \frac{\sup_{l, u_l \geq 0} h_l}{h_i} e^{\gamma\sigma} \mathbb{P}_i(T < \infty) .$$

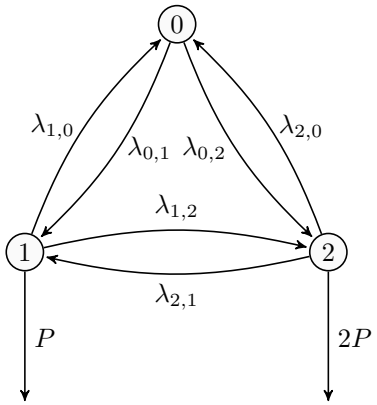


Fig. 3. 3-state Markov process.

By recalling that $\{T < \infty\}$ is the same (a.s.) with the event from Eq. (16), and finally deconditioning on i (i.e., $\mathbb{P}(T < \infty) = \sum_i \pi_i \mathbb{P}_i(T < \infty)$), the proof for the queue size Q is complete.

The proof for the waiting time $W(t)$ is entirely analogous. □

For an explicit lower bound, consider again the case when $A(t)$ is the superposed process from Figure 2. Denoting $p = \frac{\mu}{\lambda+\mu}$, the equilibrium distribution π of $Z(t)$ has the components $\pi_i = \binom{n}{i} p^i (1-p)^{n-1}$ for $i = 0, 1, \ldots, n$. Recalling the explicit solution of the underlying generalization eigenvalue problem from Eq. (12), the prefactor of the exponential bound from Eq. (14) becomes after elementary calculations

$$\begin{aligned} \sum_{i=0}^{n} \binom{n}{i} p^i (1-p)^{n-i} e^{\theta(n-i)} &= e^{\theta n} \left(p e^{-\theta} + 1 - p\right)^n \\ &= \rho^n \ , \end{aligned}$$

where $\rho = \frac{n\frac{\mu}{\lambda+\mu}P}{C}$ is the utilization factor. The explicit lower bound is thus

$$\mathbb{P}(Q \geq \sigma) \geq \rho^n e^{-\gamma\sigma} \ , \tag{19}$$

which has the same analytical structure as the corresponding upper bound from Eq. (13), except for the constant in the prefactor (i.e., ρ vs. κ, both belonging to $(0, 1)$).

C. Evaluation

In order to numerically validate the bounds given in Corollaries 2 and 3, we consider the simulation scenario as in Figure 3: A 3-state Markov process alternates with rates $\lambda_{i,j}$ between one inactive and two active states. In the active states, fluid is generated with rates P and $2P$, respectively; in the inactive state no fluid is generated.

Figure 4 shows the CCDF of the aggregate delay distribution for a scenario with $n = 2, 5, 10$ such Markov processes being multiplexed. The parameters are $\lambda_{i,j} = 1$ (for all i, j), $P = 1$, and $C > 0$ is scaled such that for the link utilization holds $\rho = 0.75$ (Figure 4a) and $\rho = 0.9$ (Figure 4b), respectively. By the symmetry of the transition rates, the Markov process is reversible. The analytical results are shown as coloured areas between the upper and lower bounds from Eqs. (9) and (15), the simulations are displayed as single points.

To gain insight into the impact of multiplexing, we refer to Figure 5: For different violation probabilites $\varepsilon = 10^{-2}, 10^{-4}, 10^{-6}$, the delay is given in dependency of the number of multiplexed sources n. Again, the link utilization is $\rho = 0.75$ (Figure 5a) and $\rho = 0.9$ (Figure 5b) and the bounds from Eqs. (9) and (15) are shown as coloured areas; in addition, the state-of-the-art (upper) bounds from Eqs. 10 and (11) are displayed as lines. One immediately sees that the beneficial effect of multiplexing (multiplexing gain) manifests itself in an exponential decay of the corresponding delay. Although this effect is also captured by the state-of-the-art bounds, they are too large by a factor of at least three times the width of the newly obtained interval, revealing their large inaccuracy.

IV. Per-Flow Delay Distribution under Scheduling

Clearly, the formulation from Lemma 1 is cumbersome, especially due to the apparently obscure parameter u. It has however the key advantage that it can be broadly applied to obtain per-flow/class bounds in the general queueing scenario from Figure 1, besides the previous results on the queue size distribution. We will derive in particular bounds on the virtual delay $W_1(t)$ of the *tagged* flow $A_1(t)$, under three scheduling policies: FIFO, SP, and EDF. While the derivations resemble those from [14], we present the generalized results (i.e., holding for general Markov fluids) for the sake of completeness.

First, we describe a common step to the derivations of all the three cases. In a scheduled scenario, the departure process $D_1(t)$ of the tagged flow has the representation

$$D_1(t) \geq A_1 * S_1(t) := \inf_{0 \leq s \leq t} \{A_1(s) + S_1(s,t)\} \ , \tag{20}$$

where $S_1(t)$ is the *service process*, encoding information about the cross flow $A_2(t)$ and the specific scheduling policy, and where '$*$' is a $(\min, +)$ convolution operator [10].

Using the equivalence of events

$$\{W_1(t) \geq d\} \quad = \quad \{A_1(t-d) \geq D_1(t)\} \ ,$$

we can bound the distribution of $W_1(t)$ as

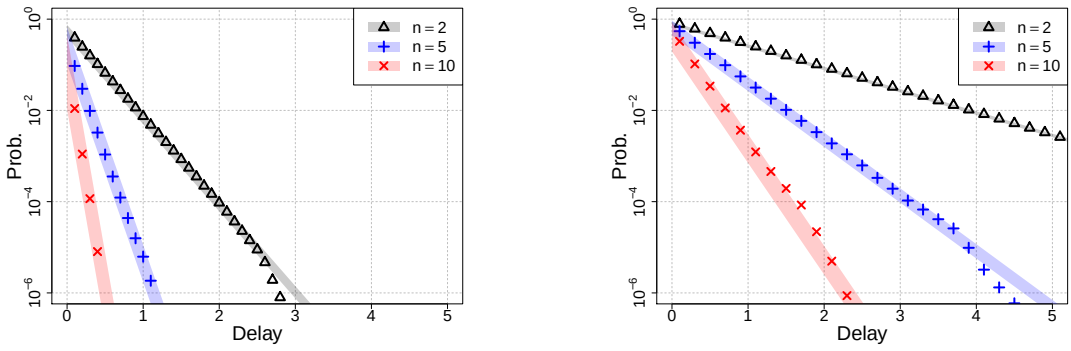
$$\mathbb{P}(W_1(t) \geq d) \leq \mathbb{P}(A_1(t-d) \geq A_1 * S_1(t)) \ . \tag{21}$$

The three delay bounds (i.e., for FIFO, SP, and EDF) can be then obtained by plugging in the service process for the corresponding scheduling policy, and by choosing a suitable value for the parameter u in Lemma 1.

A. FIFO

Under this policy, the fluid from the sources $A_1(t)$ and $A_2(t)$ is processed in the order of the respective arrival times. The service process of the tagged flow is [18]

$$S_1(s,t) = [C(t-s) - A_2(s, t-x)]_+ I_{\{t-s>x\}} \ , \tag{22}$$



(a) $\rho = 0.75$ (b) $\rho = 0.9$

Fig. 4. CCDF and upper/lower bounds for the delay of a 3-state Markov fluid with unit rates, $n = 2, 5, 10$

for some fixed $x \geq 0$, independent of s and t. Eq. (21) continues as follows

$$\mathbb{P}\left(W_1(t) \geq d\right) \leq \mathbb{P}\Big(\sup_{0 \leq s < t-d} \big\{ A_1(s, t-d) - \left[C(t-s) - A_2(s, t-x)\right]_+ I_{\{t-s>x\}} \big\} \geq 0 \Big) .$$

Note that we could restrict the range of s from $[0, t]$ to $[0, t-d)$, using the positivity of the '$[\cdot]_+$' operator and the monotonicity of $A_1(s,t)$. By making the choice $x := d$, it follows that

$$\mathbb{P}\left(W_1(t) \geq d\right) \leq \mathbb{P}\Big(\sup\nolimits_{0 \leq s < t-d} \big\{ A_1(s, t-d) + A_2(s, t-d) - C(t-s) \big\} \geq 0 \Big) .$$

Further applying Lemma 1 with $u := 0$ and $\sigma := Cd$ yields:

Corollary 4. *(Per-Flow Delay Distribution: FIFO) Under FIFO scheduling, the delay of flow $A_1(t)$ satisfies for all $d \geq 0$*

$$\mathbb{P}\Big(W_1(t) \geq d\Big) \leq \kappa e^{-\gamma C d} , \tag{23}$$

where κ and γ are given in Lemma 1.

B. SP

Under this policy, fluid from $A_1(t)$ is only served when there is no unprocessed fluid from $A_2(t)$. A service process for the low-priority tagged flow is given by [10]

$$S_1(s,t) = C(t-s) - A_2(s,t) ,$$

such that Eq. (21) continues as follows:

$$\mathbb{P}\left(W_1(t) \geq d\right) \leq \mathbb{P}\left(\sup_{0 \leq s < t-d} \left\{ A_1(s, t-d) + A_2(s,t) - C(t-s) \right\} \geq 0 \right) .$$

Recalling the arbitrary split $C_1 + C_2 = C$, Lemma 1 yields (with $u := d$ and $\sigma := 0$):

Corollary 5. *(Per-Flow Delay Distribution: SP) Under SP scheduling, the delay of flow $A_1(t)$ satisfies for all $d \geq 0$*

$$\mathbb{P}\Big(W_1(t) \geq d\Big) \leq \kappa e^{-\gamma C_1 d} , \tag{24}$$

where κ and γ are given in Lemma 1.

C. EDF

An EDF server associates the relative deadlines d_1^* and d_2^* to the fluids of $A_1(t)$ and $A_2(t)$, respectively. All fluids are served in the order of their remaining deadlines, even when they are negative. A service process for the tagged flow $A_1(t)$ is given by [33]

$$S_1(s,t) = \left[C(t-s) - A_2(s, t-x+\min\{x,y\})\right]_+ I_{\{t-s>x\}} , \tag{25}$$

for some $x > 0$ and where $y := d_1^* - d_2^*$.

For the sake of brevity, we only consider the case $y \geq 0$. Setting $x := d$ as for the FIFO case, Eq. (21) continues to

$$\mathbb{P}\left(W_1(t) \geq d\right) \leq \mathbb{P}\left(\sup_{0 \leq s < t-d} \left\{ A_1(s, t-d) + A_2(s, t-d + \min\{d,y\}) - C(t-s) \right\} \geq 0 \right) .$$

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

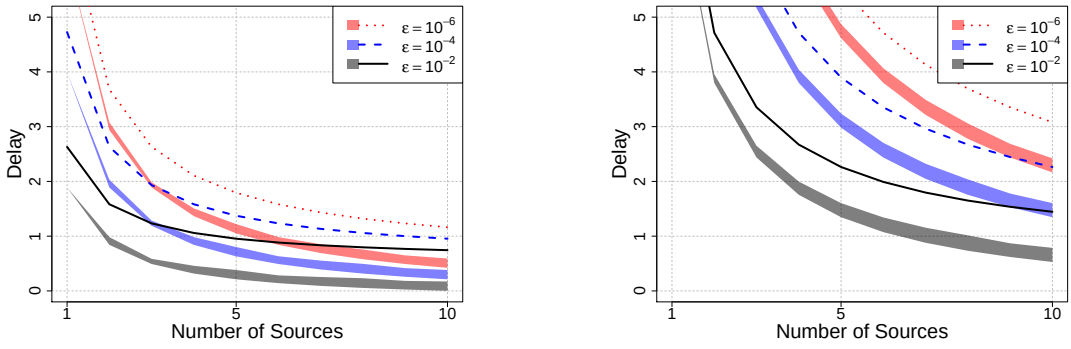
$$\mathbb{P}\left(W_1(t) \geq d\right) \leq \mathbb{P}\left(\sup_{0 \leq s < t-\min\{d,y\}} \left\{ A_1(s, t-\min\{d,y\}) + A_2(s,t) - C(t-s+d-\min\{d,y\}) \right\} \geq 0 \right) .$$

We can now apply Lemma 1 with $u := \min\{d,y\}$ (note that both d and y are positive) and $\sigma := C(d - \min\{d,y\})$, and finally obtain:

Corollary 6. *(Per-Flow Delay Distribution: EDF) Under EDF scheduling with $y := d_1^* - d_2^* \geq 0$, the delay of flow $A_1(t)$ satisfies for all $d \geq 0$*

$$\mathbb{P}\Big(W_1(t) \geq d\Big) \leq \kappa e^{\gamma C_2 \min\{d_1^* - d_2^*, d\}} e^{-\gamma C d} ,$$

where κ and γ are given in Lemma 1.



(a) $\rho = 0.75$ (b) $\rho = 0.9$

Fig. 5. Comparison of upper/lower bounds from Eqs. (9), (15) (depicted as shaded areas) on the ε-quantiles of the delay with state-of-the-art upper bound from Eq. (11) (depicted as lines), for $\varepsilon = 10^{-2}, 10^{-4}, 10^{-6}$.

V. An Analytical Construction for the Space-Parameter in the Effective Bandwidth Approximation

As a collateral result, we now give the first analytical construction for the *space-parameter* in the *effective bandwidth approximation*, which was proposed in the $1980 - 90$s as a fundamental technique for traffic engineering. For an arrival process $A(t)$, the effective bandwidth of $A(t)$ is defined for some $\theta > 0$ as [27]

$$\alpha(\theta, t) := \frac{1}{t\theta} \log E\left[e^{\theta A(t)}\right] \ .$$

Let

$$\alpha_\theta := \lim_{t\to\infty} \alpha(\theta, t)$$

and θ^* be the solution of

$$\alpha_\theta = C \ ,$$

where C is the rate at which $A(t)$ is served at a stable queue. The effective bandwidth approximation states that the steady-state queue size distribution satisfies

$$\mathbb{P}(Q > \sigma) \sim \kappa e^{-\theta^*\sigma} \ , \tag{26}$$

where κ is the *asymptotic constant*, θ^* is the *asymptotic decay rate*, and $f(x) \sim g(x)$ means that $f(x)/g(x) \to 1$ as $x \to \infty$ (see [12], [15]).

The space-parameter of the effective bandwidth approximation is generally the parameter θ in the expression of $\alpha(\theta, t)$, and more particularly θ^*, which was shown to play a fundamental role in traffic engineering; unfortunately, the construction of θ^* is based on either numerical search or simulations [15]. To the best of our knowledge, the next result provides the first analytical construction of θ^* (in the case of Markov fluids).

Lemma 7. *(Space-Parameter Construction) For the previous queueing scenario it holds*

$$\theta^* = \gamma \ , \tag{27}$$

where γ was defined in Corollary 2 (as the solution of the generalized eigenvalue problem from Eq. (7)).

Proof. From the construction of γ we have that

$$(\mathbf{Q} + \gamma\mathbf{u})\,\mathbf{h} = 0 \ . \tag{28}$$

Let the diagonal matrix $\mathbf{V}$ with $(r_0\gamma, r_1\gamma, \ldots, r_n\gamma)$ on the diagonal, and construct the matrix

$$\mathbf{Q}_\gamma := \mathbf{Q} + \mathbf{V} \ .$$

Then it holds that

$$\mathbf{Q}_\gamma \mathbf{x} = \alpha_\gamma\gamma\mathbf{I}\mathbf{x} \tag{29}$$

where $\alpha_\gamma\gamma$ is the spectral radius of $\mathbf{Q}_\gamma$ and $\mathbf{x}$ is the corresponding (positive) eigenvector (see [28]).

Let us now observe that

$$\mathbf{Q} + \gamma\mathbf{u} = \mathbf{Q}_\delta - C\gamma\mathbf{I} \ .$$

Combining with Eqs. (28) and (29) we obtain that

$$0 = (\mathbf{Q}_\gamma - \alpha_\gamma\gamma\mathbf{I})\,\mathbf{x} = (\mathbf{Q}_\gamma - C\gamma\mathbf{I})\,\mathbf{h} \ .$$

Therefore, $C\gamma$ is an eigenvalue for the eigenvalue problem from Eq. (29) and thus

$$\alpha_\gamma \geq C \ ,$$

since by construction $\alpha_\gamma\gamma$ is the corresponding spectral radius.

To show the converse, i.e., $\alpha\gamma \leq C$, consider the exact asymptotic decay of the distribution of the queue occupancy (of $A(t)$ when fed at a queue with capacity C), i.e.,

$$\lim_{\sigma\to\infty} \frac{1}{\sigma} \log \mathbb{P}(Q \geq \sigma) = -\theta^* \ , \tag{30}$$

where $\alpha_{\theta^*} = C$ (see [27]). In other words, θ^* is the exact asymptotic decay rate. As Corollary 2 predicts γ as a decay rate, in terms of an upper bound, it follows that $\gamma \leq \theta^*$. Finally, since α_θ is increasing in θ (see [10], p. 241), it follows that

$$\alpha_\gamma \leq \alpha_{\theta^*} = C \ ,$$

completing the proof. Alternatively, one can invoke the lower bound from Corollary 3. Note that a more direct proof can be

given by invoking the exact result from Eq. (30) along with the upper/lower bounds from Eqs. (10) and (14), respectively. □

VI. Conclusions

In this paper we have advanced the queueing analysis of the versatile class of general Markov fluids. In particular, we have improved the state-of-the-art upper bounds on the queue size distribution by an exponential factor in the special case of Markov modulated On-Off sources. We have further provided the first matching lower bounds, and also the first bounds on the per-flow delay distribution under FIFO, SP, and EDF scheduling.

An attractive feature of our stochastic bounds is that they are obtained using a powerful martingale methodology, which essentially invokes Kolmogorov-Doob inequality arguments, and which was shown to render accurate stochastic bounds. An alternative and arguably more powerful technique based on integral equations was provided by Kingman [30] in the case of renewals and a discrete-time setting; its extension to the non-renewal case can in principle provide clues for further improving the current upper/lower bounds.

References

[1] http://riteproject.eu/.

[2] [Online] Workshop on Reducing Internet Latency. Sept. 2013. http://www.internetsociety.org/latency2013.

[3] Fast symbolic computation of the worst-case delay in tandem networks and applications. *Performance Evaluation*, 91:270 – 285, 2015.

[4] 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.

[5] H. Bauer, J. Scharbarg, and C. Fraboul. Worst-case end-to-end delay analysis of an avionics AFDX network. In *Design, Automation Test in Europe Conference Exhibition (DATE)*, pages 1220–1224, March 2010.

[6] S. Bondorf and J. B. Schmitt. hould network calculus relocate? an assessment of current algebraic and optimization-based analyses. In *QEST*, 2016.

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

[8] B. Briscoe, A. Brunstrom, A. Petlund, D. Hayes, D. Ros, I.-J. Tsang, S. Gjessing, G. Fairhurst, C. Griwodz, and M. Welzl. Reducing internet latency: A survey of techniques and their merits. *IEEE Communications Surveys Tutorials [to appear]*. https://riteproject.files.wordpress.com/2014/07/latency_preprint-312.pdf.

[9] E. Buffet and N. G. Duffield. Exponential upper bounds via martingales for multiplexers with Markovian arrivals. *Journal of Applied Probability*, 31(4):1049–1060, Dec. 1994.

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

[11] A. Cherny. Some particular problems of martingale theory. In Y. Kabanov, R. Liptser, and J. Stoyanov, editors, *From Stochastic Calculus to Mathematical Finance*, pages 109–124. Springer, 2006.

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

[13] F. Ciucu, A. Burchard, and J. Liebeherr. A network service curve approach for the stochastic analysis of networks. In *ACM Sigmetrics*, pages 279–290, 2005.

[14] F. Ciucu, F. Poloczek, and J. Schmitt. Sharp per-flow delay bounds for bursty arrivals: The case of FIFO, SP, and EDF scheduling. In *IEEE Infocom*, pages 1896–1904, Apr. 2014.

[15] C. Courcoubetis, V. A. Siris, and G. D. Stamoulis. Application and evaluation of large deviation techniques for traffic engineering in broadband networks. In *ACM Sigmetrics*, pages 212–221, June 1998.

[16] C. Courcoubetis and R. Weber. Buffer overflow asymptotics for a buffer handling many traffic sources. *Journal of Applied Probability*, 33(3):886–903, Sept. 1996.

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

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

[19] J. Dean and L. A. Barroso. The tail at scale. *Communications of the ACM*, 56(2):74–80, Feb. 2013.

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

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

[22] G. Fettweis. The Tactile Internet: Applications and challenges. *IEEE Vehicular Technology Magazine*, 9(1):64–70, Mar. 2014.

[23] Y. Ghiassi-Farrokhfal and F. Ciucu. On the impact of finite buffers on per-flow delays in FIFO queues. In *24th International Teletraffic Congress (ITC)*, 2012.

[24] M. Gribaudo and M. Telek. Fluid models in performance analysis. In M. Bernardo and J. Hillston, editors, *Formal Methods for Performance Evaluation*, volume 4486 of *Lecture Notes in Computer Science*, pages 271–317. Springer, 2007.

[25] G. Grimmett and D. Stirzaker. *Probability and Random Processes*. Oxford University Press, 2001.

[26] T. Hoff. [Online] Latency is everywhere and it costs you sales - how to crush it. July 2009. http://highscalability.com/blog/2009/7/25/latency-is-everywhere-and-it-costs-you-sales-how-to-crush-it.html.

[27] 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.

[28] G. Kesidis, J. Walrand, and C. Chang. Effective bandwidths for multiclass Markov fluids and other ATM sources. *IEEE/ACM Transactions on Networking*, 1(4):424–428, Aug. 1993.

[29] J. F. C. Kingman. A martingale inequality in the theory of queues. *Cambridge Philosophical Society*, 60(2):359–361, Apr. 1964.

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

[31] R. Krishnan, H. V. Madhyastha, S. Srinivasan, S. Jain, A. Krishnamurthy, T. Anderson, and J. Gao. Moving beyond end-to-end path information to optimize cdn performance. In *Internet Measurement Conference (IMC)*, pages 190–201, 2009.

[32] 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.

[33] 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.

[34] R. R. Mazumdar. *Performance Modeling, Loss Networks, and Statistical Multiplexing*. Synthesis Lectures on Communication Networks. Morgan & Claypool Publishers, 2009.

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

[36] V. Paxson and S. Floyd. Wide-area traffic: The failure of Poisson modelling. *IEEE/ACM Transactions on Networking*, 3(3):226–244, June 1995.

[37] F. Poloczek and F. Ciucu. Scheduling analysis with martingales. *Performance Evaluation*, 79:56 – 72, Sept. 2014. Special Issue: Performance 2014.

[38] S. M. Ross. Bounds on the delay distribution in GI/G/1 queues. *Journal of Applied Probability*, 11(2):417–421, June 1974.

[39] E. Schurman and J. Brutlag. The user and business impact of server delays, additional bytes and HTTP chunking in web search. *O'Reilly Velocity Web Performance and Operations Conference*, June 2009.

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

[41] A. Singla, B. Chandrasekaran, P. B. Godfrey, and B. Maggs. The Internet at the speed of light. In *ACM Workshop on Hot Topics in Networks (HotNets)*, pages 1–7, 2014.

[42] S. Souders. [Online] Velocity and the bottom line. July 2009. http://radar.oreilly.com/2009/07/velocity-making-your-site-fast.html.

[43] T. E. Stern and A. I. Elwalid. Analysis of separable Markov-modulated rate models for information-handling systems. *Advances in Applied Probability*, 23(1):105–139, Mar. 1991.