MGF-based SNC for Stationary Independent Markovian Processes with Localized Application of Martingales

Anne Bouillard[1*]

[1*]Huawei Technologies France, 18 quai du point du jour, Boulogne-Billancourt, 92100, France.

Corresponding author(s). E-mail(s): anne.bouillard@huawei.com;

Abstract

Stochastic Network Calculus is a probabilistic method to compute performance bounds in networks, such as end-to-end delays. It relies on the analysis of stochastic processes using formalism of (Deterministic) Network Calculus. However, unlike the deterministic theory, the computed bounds are usually very loose compared to the simulation. This is mainly due to the intensive use of the Boole's inequality. On the other hand, analyses based on martingales can achieve tight bounds, but until now, they have not been applied to sequences of servers. In this paper, we improve the accuracy of Stochastic Network Calculus by combining this martingale analysis with a recent Stochastic Network Calculus results based on the Pay-Multiplexing-Only-Once property, well-known from the Deterministic Network calculus. We exhibit a non-trivial class of networks that can benefit from this analysis and compare our bounds with simulation.

Keywords: Stochastic Network Calculus, Martingale, performance evaluation, generating functions

1 Introduction

New communication technologies aim at providing strong end-to-end delay guarantees. Some of these guarantees can be *deterministic*, in the sense that all packets must be transmitted in time less than some predefined value. However, these guarantees are in some cases too conservative, as the worst-case delay of a packet can be very large and rarely occurs. To this end, performances of type "the delay of a packet must be less than 10ms in 99.99% of the cases" is desired. Examples of applications are industrial networks, virtual reality, audio and video conferencing networks.

Network Calculus (NC) is a theory that aims at providing performance bounds, such as end-to-end delays or buffer occupancy for large classes of systems: general classes of arrival processes and service duration, different service policies, general topologies. From the seminal work of Cruz (1995), Deterministic Network Calculus (DNC, Chang (2000); Le Boudec and Thiran (2001); Bouillard et al (2018)) has succeeded to provide accurate performance bounds in many types of networks (aircraft, industrial network, time-sensitive networking...). Stochastic Network Calculus (SNC) takes its root in the works of Yaron and Sidi (1993) and Chang (1994), that respectively led to the tailbounded and moment-generating-function (MGF)-based SNC.

Several existing SNC analyses.

The *tailbounded* SNC, analyzed in depth in Liu and Jiang (2008), is based on the computation of the probability of elementary events, for instance, "the arrivals during a given interval of time exceeds b". The probability that the end-to-end delay satisfies some deadline is then obtained by the combination of these elementary events. The analysis of large networks requires a smart combination of these events, see e.g., Ciucu et al (2006). Bouillard and Nowak (2015) combine advanced results from DNC with this framework. As such, it can handle many scheduling policies and topologies. In particular when the stochastic processes involved are independent and have strong long-term properties (*e.g.*, the maximum backlog centric model), the performances scale linearly with the size of the network(Liu and Jiang, 2008, Sec 6.2.6).

The MGF-based SNC is based on the computation of the MGF of the arrival and service processes. When all these processes are independent, then their MGFs can easily be combined, making aggregation of flows more accurate than with the tail-bounded approach, as soon as several random processes are involved (Rizk and Fidler (2013)). The computation of end-to-end delays however still requires the combination of elementary events. When processes are independent, the MGF-based SNC leads to better bounds than the tailbounded SNC. In particular, Ciucu et al (2006) show that the performance bounds scale in $O(n \ln n)$ for a system with n servers in sequence, crossed by one-hop flows, whereas Fidler (2006) proved a $O(n)$ scale with MGF analysis. A precise comparison between tailbounded and MGF-based SNC has also been done in Rizk and Fidler (2013); Fidler and Rizk (2015).

When flows are not independent, the Hölder inequality is used to compute performance bounds with MGF-based SNC. Until recently, the analysis of tandem network also induced the application of some Hölder inequality to handle the dependency of processes after crossing a common server Nikolaus and Schmitt (2017). A recent work Bouillard et al (2022), based on the *Pay-Multiplexing-Only-Once* (PMOO) principle Bouillard et al (2008); Schmitt et al (2008), proposes to adapt a result from DNC in this framework and get rid of the Hölder inequalities for independent processes.

One shared drawback of these two methods is that they strongly rely on the application of the union bound (also known as Boole's inequality), which makes them highly inaccurate, even for small networks. Some improvements, like *flow prolongation* Nikolaus and Schmitt (2020), can be used to reduce the size of the network to analyze, hence improve the accuracy, but does not allow getting rid of these union bounds.

In order to avoid the use of the union bound, Poloczek and Ciucu (2014) use a martingale representation for the processes and rely on the Doob's inequality for super-martingales. The delay bounds computed that way are almost tight compared to the simulation. This method generalizes previous works by Duffield (1994) and Kingman (1964) from queuing theory to several service policies, to very general arrival processes (Ciucu and Poloczek (2019)) and to systems with replications (Ciucu et al (2021)). Most works focus on random arrivals with deterministic service. In contrast, Poloczek and Ciucu (2015) define a service-martingale to extend to stochastic services with applications to wireless connections and random access protocols. However, this method has been applied only to networks that have no server in sequence.

Contribution of the paper.

The contribution of this paper is to extend the martingale analysis from Poloczek and Ciucu (2014) to sequences of servers (called *tandem networks*). Since it does not seem possible to define a unique martingale representing the whole tandem network, we rather focus on the application of the martingale analysis at a server only and combine it with the recent results related to the MGF-based SNC of Bouillard et al (2022). We show on small examples the improvements that can be obtained, as well as the limits of our approach.

Note that another analysis of tandem networks can be founded in the literature Angrishi and Killat (2011), which is attempted to use sub-martingales for the analysis of tandem networks, and recently used for the analysis of URLLC (Ultra Reliable Low Latency Communications) networks Yu et al (2022). Unfortunately, this approach is not sound, and we explain why as another contribution of this paper.

Organization of the paper.

The paper is organized as follows: in Section 2, we introduce the necessary framework: MGF-based SNC, the class of stochastic models used in this paper, namely the Markov-modulated processes (MPP), and a more recent formalization of MGF-based SNC based on analytical combinatorics and corresponding results on PMOO analysis from Bouillard et al (2022). In Section 3, we present the main result of the paper. We start by generalizing Theorem 3 from Duffield (1994) and present a key theorem for the local application of the Doob's martingale inequality, and then use it for the analysis of tandem networks. For the sake of clarity, we explain our approach through a toy network before giving the general result. In Section 4, we explain why the analysis of Angrishi and Killat (2011) and Yu et al (2022) is not sound. Finally, in Section 5, we compare our performance bounds with the simulation and with the bounds of Bouillard et al (2022).

2 Stochastic Network Calculus framework

In this section we present the necessary framework for our analysis. First, we define the NC formalism. Second, we specialize it to the MGF-based SNC and to the class of MMPs. More details can be found in Chang (2000); Fidler (2006). Then we give a combinatorial presentation of the MGF-based SNC, that allow to present results from Bouillard et al (2022) on tandem networks.

In the whole paper, we assume time and space are discrete. We deal with bivariate functions, and always assume that their definition domain is $\mathbb{N}^2_{\leq} = \{(t,u) \in \mathbb{N}^2 \mid t \leq u\}$ and that they are in the set $\mathcal{F} = \{f : \mathbb{N}^2_{\leq} \to \mathbb{N}_+ \mid \forall t \geq 0,\ f(t,t) = 0\}$. The main notations, defined below, are summarized in Table 1.

2.1 Network Calculus formalism

Arrival processes.

A bivariate process $A \in \mathcal{F}$ of a flow represents the amount of data of that flow arrived in the network during any interval of time: let $a_t \in \mathbb{N}$ be the amount of data

t, u, v	time variables
Flows and data processes	
i, m	index of flows, number of flows
a_t	arrivals at time slot t
A, A_i	bivariate arrival processes
$F_A(\theta, z)$	arrival bounding generating function
f_i, ℓ_i	first and last servers crossed by flow i
D, D_i	bivariate departures processes
Service processes	
j, n	index of a server, number of servers
s_t	service at timeslot t
S, S_j	bivariate service processes
$F_S(\theta, z)$	service bounding generating function
Performance bounds	
$q(t)$	backlog at time t
$d(t)$	delat at time t
$F_d(\theta, z)$	delay bounding generating function
Generating functions	
$F(z) = \sum_{k\in\mathbb{N}} f_k z^k$	Generating function associated with sequence $(f_k)_{k\in\mathbb{N}}$
r_F	dominant singularity (or radius of convergence) of F
Markov-modulated processes	
$\mathcal{X}$, P, π	state space, transition matrix, and stationary distribution
$\varphi_x(\theta)$	MGF associated to state x
$\psi(\theta)$	exponential transition matrix,
$\lambda(\theta)$, $\nu(\theta)$	its largest eigenvalue and associated eigenvector
$(\sigma(\theta), \rho(\theta))$	MGF-based SNC characterization of a processes
$M(\theta, u, v)$	martingale representation of a process

Table 1 Table of notations

arriving during the t-th time slot. We define for all $t \leq u$, $A(t, u) = \sum_{v=t}^{u-1} a_v$, with the convention $A(t, t) = 0$. Consequently, A is additive: $A(t, u) + A(u, v) = A(t, v)$.

S-servers.

Let $S \in \mathcal{F}$ be a bivariate function. A server is a dynamic S-server if the relation between its bivariate arrival and departure processes $A \in \mathcal{F}$ and $D \in \mathcal{F}$ satisfies for all $t \geq 0$, $A(0, t) \geq D(0, t) \geq \min_{0\leq u\leq v} A(0, u) + S(u, t)$.

This notion of dynamic S-server is often too weak to perform network analysis, and we need the notion of work-conserving S-servers: define s_t as the amount of service offered by the server during time slot t. Then the service offered by this server is $S(t, u) = \sum_{v=t}^{u-1} s_v$, which defines S as an additive bivariate function of $\mathcal{F}$. If during

$(t, u]$ the server is never empty (for all $v \in (t, u]$, $A(0, v) > D(0, v)$), then for all $t \leq v \leq u$, $D(v, u) = S(v, u)$).

In general, a server is crossed by several flows, with respective arrival and departure processes A_i and D_i, $i \in \{1, \ldots, m\}$ in case of m flows. We say that a server is a dynamic S-server (resp. a work-conserving S-server) if it is for the *aggregated* arrival and departure processes $A = \sum_{i=1}^m A_i$ and $D = \sum_{i=1}^m D_i$. Moreover, we assume that the system is causal, and we also have for all $i \in \{1, \ldots, m\}$ and all $t \geq 0$, $A_i(0, t) \geq D_i(0, t)$.

Performance bounds.

Consider a dynamic S-server and A and D its respective arrival and departure bivariate processes. The backlog at time t is $q(t) = A(0, t) - D(0, t)$ and the virtual delay at time t is $d(t) = \inf\{T \geq 0 \mid A(0, t) \leq D(0, t + T)\}$.

Theorem 1 (Performance bounds Chang (2000); Fidler (2006))**.** *Let A be a bivariate process crossing an S-dynamic server. Then*

- $q(t) \leq \sup_{0 \leq s \leq t} A(s, t) - S(s, t)$*;*
- *for all* $T \in \mathbb{N}$, $d(t) \geq T \implies \exists u \leq t,\ A(u, t) > S(u, t + T - 1)$*.*

We call the delay *virtual delay* because this is the delay when data exit the system in their arrival order. This is not necessarily the case when several flows cross the same system.

2.2 Markov modulated processes and SNC

SNC is the study of systems described above when A and S are described by stochastic processes, and we want to upper-bound the *violation probability* (v.p.) of some backlog or delay, more precisely, $\mathbf{P}(q(t) \geq b)$ and $\mathbf{P}(d(t) \geq T)$.

The $(\sigma(\theta), \rho(\theta))$ representations is used to bound these quantities, using MGF-based SNC. In this paragraph, we first present this $(\sigma(\theta), \rho(\theta))$ characterization, then define the MMPs, that have a $(\sigma(\theta), \rho(\theta))$ representation and will be used in this paper.

2.2.1 MGF-based SNC and $(\sigma(\theta), \rho(\theta))$-representations

MGF-based SNC mainly uses two probabilistic inequalities: the union bound (or Boole's inequality, denoted **(UB)** in the equations) and Chernoff bounds ($\forall\theta > 0$, $\mathbf{P}(X \geq x) \leq \frac{\mathbf{E}[e^{\theta X}]}{e^{\theta x}}$, denoted **(CB)**), that require computing MGF of a process. For example, the backlog v.p. would be for all $\theta > 0$,

$$\mathbf{P}(q(t) \geq b) \leq \mathbf{P}(\exists u \leq t,\ A(u,t) - S(u,t) \geq b) \overset{\textbf{(UB)}}{\leq} \sum_{u \leq t} \mathbf{P}(A(u,t) - S(u,t) \geq b)$$
$$\overset{\textbf{(CB)}}{\leq} \sum_{u \leq t} \mathbf{E}[e^{\theta(A(u,t) - S(u,t))}] e^{-\theta b}.$$

Independence between the arrival and service processes $(\perp)$ is a common assumption in MGF-based SNC we make. Hence, we now need to bound the MGFs of $A(u,t)$ and $S(u,t)$, using the $(\sigma(\theta), \rho(\theta))$ representations.

- An arrival process A is $(\sigma_A(\theta), \rho_A(\theta))$-constrained if for all $(t,u) \in \mathbb{N}^2_{\leq}$, $\mathbf{E}[e^{\theta A(t,u)}] \leq e^{\theta(\sigma_A(\theta) + \rho_A(\theta)(u-t))}$.
- A service process is $(\sigma_S(\theta), \rho_S(\theta))$-constrained if for all $(t,u) \in \mathbb{N}^2_{\leq}$, $\mathbf{E}[e^{-\theta S(t,u)}] \leq e^{\theta(\sigma_S(\theta) - \rho_S(\theta)(u-t))}$.

Remark 1. • $\mathbf{E}[e^{\theta A(t,u)}]$ *is increasing in* θ*, but might not be finite from some value, in which case we set* $\sigma(\theta) = \rho(\theta) = +\infty$*;*

- $-\theta$ *is used in the MGF of the service processes instead of* θ*. This is to upper bound the arrival processes, and lower bound on the service processes.*

We can now complete the computation of the backlog v.p., as we obtain a geometric sum (GS):

$$
\begin{aligned}
\mathbf{P}(q(t) \geq b) &\leq \sum_{u \leq t} \mathbf{E}[e^{\theta A(u,t)}]\mathbf{E}[e^{-\theta S(u,t)}]e^{-\theta b}. && (\perp) \\
&\leq \sum_{u \leq t} e^{\theta(\sigma_A(\theta+\rho_A(\theta)(u-t))} e^{\theta(\sigma_S(\theta)-\rho_S(\theta)(u-t))} e^{-\theta b} && (\sigma,\rho) \\
&\leq \frac{e^{\theta(\sigma_A(\theta)+\sigma_S(\theta)-b)}}{1-e^{\theta(\rho_A(\theta)-\rho_S(\theta))}}. && \text{(GS)} \qquad (1)
\end{aligned}
$$

The sum is finite if and only if $\rho_S(\theta) > \rho_A(\theta)$, and in that case, the bound is valid for all $t \in \mathbb{N}$. A system is said *stable* if for all t, $\mathbf{E}[q(t)] < \infty$. Here, a sufficient condition is the existence of a positive $\theta > 0$ such that $\rho_S(\theta) > \rho_A(\theta)$.

V.p. for the delay can be obtained similarly or from (Bouillard et al, 2022, Lemma 2), and

$$
\mathbf{P}(d(t) \geq T) \leq \frac{e^{\theta(\sigma_A(\theta)+\sigma_S(\theta)+\rho_A(\theta)-\rho_S(\theta)T)}}{1-e^{\theta(\rho_A(\theta)-\rho_S(\theta)}}. \qquad (2)
$$

2.2.2 Markov modulated processes

Many stochastic processes have $(\sigma(\theta), \rho(\theta))$ representations. In this paper, we focus on the family of MMPs detailed in (Chang, 2000, Chapter 7).

Let us denote by y_t the amount of arrival or service at time slot t (y_t is either a_t or s_t depending on the context). In short, in an MMP, the amount of data y_t at each time slot t follows the distribution that depends on a state described by a homogeneous discrete-time Markov chain (MC).

Example 1 (Markov-modulated On-Off process (MMOO))**.** *An MMOO process: the MC has two states,* `On` *and* `Off`*. When in the* `Off` *state, there is no arrival, and the amount of arrival when in the* `On` *state is generated according to a distribution.*

Conditionally to being in the `On` *state, the amount of data generated is independent of the past evolution of the process.*

More generally, let us consider $(X(t))_{t\in\mathbb{N}}$ an ergodic MC on a finite space $\mathcal{X}$, with transition matrix P and stationary distribution π. Define for all $x \in \mathcal{X}$, $(Y_x(t))_{t\in\mathbb{N}}$ an i.i.d sequence of random variables (rv) with MGF $\varphi_x : \theta \mapsto \mathbf{E}[e^{\theta Y_x(0)}]$. The quantity $Y_x(t)$ represents the number of data generated at time slot t if the MC is in state x. We assume that $(X(t))_{t\in\mathbb{N}}$ and $(Y_x(t))_{t\in\mathbb{N}}$, $x \in \mathcal{X}$ are mutually independent.

Example 2 (MMOO process (continued))**.** *The Markov chain* $(X(t))_{t\in\mathbb{N}}$ *has two states:* $\mathcal{X} = \{\mathtt{Off}, \mathtt{On}\}$. *Its transition matrix and stationary distribution are given by*

$$P = \begin{pmatrix} 1-p & p \\ q & 1-q \end{pmatrix} \quad \textit{and} \quad \pi = \left(\frac{q}{p+q}, \frac{p}{p+q}\right).$$

There is no data generated in state `Off`*, so* $\varphi_{\mathtt{Off}}(\theta) = 1$ *for all* $\theta \in \mathbb{R}$*. If for example data are generated according to a Poisson process of intensity* μ *in state* `On`*, then* $\varphi_{\mathtt{On}}(\theta) = e^{\mu(e^\theta - 1)}$ *for all* $\theta \in \mathbb{R}$*.*

The sequence $(y_t)_{t\in\mathbb{N}}$, with $y_t = Y_{X(t)}(t)$ for all $t \in \mathbb{N}$ is an MMP.

In this paper, we assume that the process is stationary: $X(0)$ is distributed according to the stationary distribution π and then, for all $t \in \mathbb{N}$, $X(t)$ is also distributed according to π: $\pi P = \pi$.

As a consequence, for all $T \in \mathbb{N}$, the time-reversed process $(X(T-t))_{t\leq T}$ is a MC. By Kolmogorov's extension one can extend this process to $(X(T-t))_{t\in\mathbb{N}}$. This process is also an ergodic MC with stationary probability π, and transition matrix $P^r = (\frac{\pi(j)}{\pi(i)} P_{j,i})_{i,j\in\mathcal{X}}$.

MMOO process: A two-state ergodic MC is time-reversible and that $P^r = P$.

The *exponential transition matrix* associated to the MMP is $\psi(\theta) = (P^r_{i,j}\varphi_j(\theta))_{i,j\in\mathcal{X}}$. For all values of θ such that $\psi(\theta)$ is finite, this matrix is primitive, and from the Perron-Frobenius theorem,

- $\psi(\theta)$ has an eigenvalue $\lambda(\theta)$ that is strictly positive, and strictly larger in modulus than any other eigenvalue of $\psi(\theta)$. This eigenvalue is simple (its associated eigenspace has dimension 1);
- the unique right-eigenvector $\nu(\theta)$ associated to $\lambda(\theta)$ and satisfying $\langle \nu(\theta), \pi \rangle = \sum_{x \in \mathcal{X}} \pi_x \nu(\theta)_x = 1$ is strictly positive (all its coefficients are strictly positive).

MMOO process: For the MMOO process, we have

$$\psi(\theta) = \begin{pmatrix} 1-p & pe^{\mu(e^\theta - 1)} \\ q & (1-q)e^{\mu(e^\theta - 1)} \end{pmatrix}.$$

The spectral analysis is left to the interested reader, that can refer to (Chang, 2000, Ex. 7.2.7).

We say that an arrival process A is generated by an MMP if there exists an MMP $(a_t)_{t \in \mathbb{N}}$ such that for all $(t, u) \in \mathbb{N}^2_{\leq}$, $A(t, u) = \sum_{i=t}^{u-1} a_i$, and similarly for a work-conserving S-server.

In the following we will consider several arrival processes A_i and servers S_j. The quantities defined above will be indexed by A_i or S_j accordingly.

$(\sigma(\theta), \rho(\theta))$***-characterization of a process***

If an arrival process A is generated by an MMP with exponential transition matrix $\psi_A(\theta)$, then from (Chang, 2000, Ex. 7.2.7),

- $\rho_A(\theta) = \frac{1}{\theta} \ln \lambda_A(\theta)$ and
- a simple adaptation of (Beck, 2016, Chapter 10) to the special case of stationary processes shows that $\sigma_A(\theta) = \frac{1}{\theta} \ln \big(\frac{1}{\min_{x \in \mathcal{X}} \nu_A(\theta)_x} \big)$.

Similarly, if a service process S is generated by an MMP with exponential transition matrix $\psi_S(\theta)$, then $\rho_S(\theta) = -\frac{\ln \lambda_S(-\theta)}{\theta}$ and $\sigma_S(\theta) = \frac{1}{\theta} \ln \big(\frac{1}{\min_{x \in \mathcal{X}} \nu_S(-\theta)_x} \big)$.

2.3 Analytic combinatorics and (σ, ρ)-representation

In this paragraph, we use generating functions (GFs) from analytic combinatorics (Sedgewick and Flajolet (1996)) also used in Bouillard et al (2022). GFs allow more compact and general characterizations than the traditional $(\sigma(\theta), \rho(\theta))$ representations, in particular for representing servers in tandem. We first give an example explaining the use of a more general representation than $(\sigma(\theta), \rho(\theta))$. Then we provide a generalization that uses GFs, and the computation of performance bounds.

2.3.1 Processes without a $(\sigma(\theta), \rho(\theta))$ representation

Let us anticipate on Theorem 4 (cf. Paragraph 2.4), and consider two work-conserving servers in tandem. The end-to-end dynamic server of this tandem is characterized by a product of geometric series: assume S_1 and S_2 the respective service processes of the two work-conserving servers having the same $(\sigma(\theta), \rho(\theta)$ representation. The end-to-end service is then $\forall(u,t) \in \mathbb{N}_\leq$, $S_{e2e}(u,t) \geq \inf_{u \leq v \leq t} S_1(u,v) + S_2(v,t)$. We use the the union bound to upper-bound $\mathbf{E}[e^{-\theta S_{e2e}(u,t)}]$:

$$\begin{aligned}
\mathbf{E}\big[e^{-\theta S_{e2e}(u,t)}\big] &\leq \mathbf{E}\big[e^{-\theta(\inf_{u \leq v \leq t} S_1(u,v)+S_2(v,t))}\big] = \mathbf{E}\big[\sup_{u \leq v \leq t} e^{-\theta(S_1(u,v)+S_2(v,t))}\big] \\
&\leq \sum_{u \leq v \leq t} \mathbf{E}\big[e^{-\theta(S_1(u,v)+S_2(v,t))}\big] && \textbf{(UB)} \\
&\leq \sum_{u \leq v \leq t} e^{\theta(2\sigma(\theta)-\rho(\theta)[v-u+t-v])} && (\perp),(\sigma,\rho) \\
&= \sum_{u \leq v \leq t} e^{\theta(2\sigma(\theta)-\rho(\theta)(t-u))} = (t-u+1)e^{\theta(2\sigma(\theta)-\rho(\theta)(t-u))}. && (3)
\end{aligned}$$

The bound given in Eq. (3) for the S_{e2e}-dynamic server is not a $(\sigma(\theta), \rho(\theta))$-representation. However, similar to the $(\sigma(\theta), \rho(\theta))$-framework, $\mathbf{E}[e^{-\theta S_{e2e}(u,t)}]$ depends only on the time variables by their difference:

$$\mathbf{E}[e^{-\theta S_{e2e}(t,t+k)}] \leq (k+1)e^{\theta(2\sigma(\theta)-\rho(\theta)k} = [z^k]\frac{e^{2\theta\sigma(\theta)}}{\left(1-e^{-\theta\rho_S(\theta)}z\right)^2},$$

where $[z^k]G(z)$ is the k-th term of the GF G.

2.3.2 Bounding generating functions

Generating functions

Let $f = (f_k)_{k\in\mathbb{N}}$ be a non-negative sequence. The GF associated to f is $F(z) = \sum_{k\in\mathbb{N}} f_k z^k$, and $f_k = [z^k]F(z)$ denotes the k-th term of the sequence. An important example is the geometric sequence: for all $k \in \mathbb{N}$, $f_k = r^k$ and $F(z) = \sum_{k\in\mathbb{N}}(rz)^k = (1-rz)^{-1}$.

Analytic combinatorics is a branch of combinatorics. For a collection of combinatorial objects, f_k corresponds to the number of objects of size k. The idea is to represent this collection by the function $F(z) = \sum_{k\in\mathbb{N}} f_k z^k$, the GF. One advantage of GF is the close relation between the asymptotic behavior of the sequence and the singularities of its GF. The radius of convergence (or dominant singularity) of F is defined by $r_F = \sup\{z \geq 0 \mid \sum_{k\in\mathbb{N}} f_k z^k < \infty\}$. One important result is that if F has a dominant singularity of multiplicity 1, then $f_k \sim_{k\to\infty} c r_F^{-k}$ for some constant c (Sedgewick and Flajolet, 1996, Ch. 5). This can be checked for the example of the geometric series $F(z) = (1-rz)^{-1}$, whose unique (hence dominant) singularity is r^{-1}.

Bounding generating functions for processes

An arrival process has the arrival bounding generating function (arrival bgf) $F_A(\theta, z)$ if for all $k \in \mathbb{N}$, for all $t \in \mathbb{N}$, $\mathbf{E}[e^{\theta A(t,t+k)}] \leq [z^k]F_A(\theta, z)$. When A has a $(\sigma_A(\theta), \rho_A(\theta))$ representation, then $F_A(\theta, z) = \frac{e^{\theta\sigma_A(\theta)}}{1-e^{\theta\rho_A(\theta)}z}$ is an arrival bgf of A. Its radius of convergence is $r_A(\theta) = e^{-\theta\rho_A(\theta)}$.

Similarly, a service process has the service bounding generating function (service bgf) $F_S(\theta, z)$ if for all $k \in \mathbb{N}$, for all $t \in \mathbb{N}$, $\mathbf{E}[e^{-\theta S(t,t+k)}] \leq [z^k]F_S(\theta, z)$. If S is $(\sigma_S(\theta), \rho_S(\theta))$-characterized, then , $F_S(\theta, z) = \frac{e^{\theta\sigma_S(\theta)}}{1-e^{-\theta\rho_S(\theta)}z}$. Its radius of convergence is $r_S(\theta) = e^{\theta\rho_S(\theta)}$.

Cauchy product of geometric series

In Eq. (3), we recognize the Cauchy product of two sequences (the MGFs of the service processes): the Cauchy product of $(f_k)_{k\in\mathbb{N}}$ and $(g_k)_{k\in\mathbb{N}}$ is $(h_k)_{k\in\mathbb{N}}$ with $h_k = \sum_{k'=0}^{k} f_{k'}g_{k-k'}$ for all $k \in \mathbb{N}$. An important result from analytical combinatorics is that the GF of the Cauchy product is the product of the GFs: $H(z) = F(z)G(z)$. Applying this to the product of n geometric series translates into the following equalities: for all $\alpha_1, \ldots, \alpha_n \in (0, 1)$,

$$\sum_{\substack{k_1,\ldots,k_n\geq 0\\ k_1+\cdots+k_n=k}} \prod_{j=1}^{n} \alpha_j^{k_j} = [z^k] \prod_{j=1}^{n} \frac{1}{1 - z\alpha_j} \text{ and } \sum_{k_1,\ldots,k_n\geq 0} \prod_{j=1}^{n} \alpha_j^{k_j} = \prod_{j=1}^{n} \frac{1}{1 - \alpha_j}. \tag{4}$$

2.3.3 Bounding performance using generating functions

Defining a delay bounding generating function (delay bgf) of a process as a generating function $F_d(z)$ such that for all $T \geq 0$, $\mathbf{P}(d(t) \geq T) \leq [z^T]F_d(z)$ also allows more compact representations. Since the backlog v.p. $\mathbf{P}(q(t) \geq b)$ has a simpler expression, we do not define one for the backlog.

We can state the theorem for performance computation.

Theorem 2 ((Bouillard et al, 2022, Corollary 3 and Lemma 2))**.** *Consider a dynamic S-server offering a service bounded by the service bgf $F_S(\theta, z)$ and crossed by a flow with bivariate process A that is (σ_A, ρ_A)-constrained. Assume independence between the arrival and service processes. For all θ such that $r_A(\theta)r_S(\theta) > 1$,*

1. *Backlog v.p. bound:* $\mathbf{P}(q(t) \geq b) \leq e^{-\theta b}e^{\theta\sigma_A(\theta)}F_S(\theta, e^{\theta\rho_A(\theta)})$.
2. *Delay bgf:* $F_d(\theta, z) = e^{\theta\sigma_A(\theta)} \frac{e^{\theta\rho_A(\theta)}F_S(\theta, e^{\theta\rho_A(\theta)}) - zF_S(\theta, z)}{1 - ze^{-\theta\rho_A(\theta)}}$.

Remind that $r_A(\theta) = e^{-\theta\rho_A(\theta)}$ and $r_S(\theta) = e^{\theta\rho_S(\theta)}$. It is shown in (Duffield, 1994, Lemma 1) that in the stationary regime $f : \theta \mapsto \ln \mathbf{E}[e^{\theta(a_1-s_1)}] = -\ln(r_A(\theta)r_S(\theta))$ is convex on its definition set, and $f(0) = 0$. Consequently (Duffield, 1994, Lemma 2), there exists $\theta > 0$ satisfying $r_A(\theta)r_S(\theta) > 1$ (*i.e.*, f is decreasing on some interval $[0, x]$) if and only if $f'(0) < 0$, that is $\mathbf{E}[a_1] - \mathbf{E}[s_1] < 0$: there are in average fewer arrivals than the service offered, and the system is said *stable*.

We set $\theta^* = \sup\{\theta \geq 0 \mid r_A(\theta)r_S(\theta) > 1\}$. This value can be $+\infty$ in case the arrivals are almost surely less than the services at each time step.

In the particular case of a $(\sigma_S(\theta), \rho_S(\theta))$-characterized dynamic server, $F_S(\theta, z) = \frac{e^{\theta\sigma_S(\theta)}}{1-e^{-\theta\rho_S(\theta)}z}$, and the condition $r_A(\theta)r_S(\theta) > 1$ also reads $\rho_A(\theta) - \rho_S(\theta) < 0$ or $\rho_S(\theta) > \rho_A(\theta)$. For all $\theta < \theta^*$, the delay bgf is

$$F_d(\theta, z) = \frac{e^{\theta(\sigma_A(\theta)+\sigma_S(\theta)+\rho_A(\theta))}}{1 - e^{-\theta(\rho_S(\theta)-\rho_A(\theta))}} \cdot \frac{1}{1 - ze^{-\theta\rho_S(\theta)}},$$

and the backlog and delay v.p. are exactly those in Equations (1) and (2).

We can observe that the v.p depends on the choice of θ that needs to be optimized. For a fixed value of θ, the v.p. of the backlog decreases exponentially fast with rate θ. The higher the value of θ, the higher the decay rate, but the denominator that decreases to 0 when θ approaches θ^*.

2.4 Tandem network model

In this paragraph, we describe our model of tandem network. Examples are given in Figure 1.

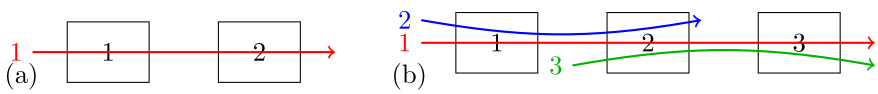


Fig. 1 Examples of network: (a) two-server tandem; (b) interleaved tandem.

Assumption 1 (Topology)**.** *The notations will be used throughout the paper.*

(H_1) *The network is composed of n servers, numbered from 1 to n. Each server j is a S_j-work-conserving server. We assume FIFO per-flow and arbitrary multiplexing: there is no assumption about the service order of data, except that data from the same flow are served in their arrival order.*

(H_2) *There are m flows circulating in the network, numbered from 1 to m. For all flow i, there exist $f_i \leq \ell_i \in \mathbb{N}_n$, such that flow i crosses servers $f_i, f_i + 1, \dots, \ell_i$. We denote by A_i the bivariate arrival process of flow i. We write $i \in \mathrm{Fl}(j)$ if flow i crosses server j.*

(H_3) *Flow 1, for which we are going to compute performance bounds and is then also called the flow of interest, crosses the whole tandem network ($f_1 = 1$ and $\ell_1 = n$).*

PMOO's technique consists in computing an end-to-end dynamic server for flow 1 in order to guarantees its performances. The formula, intuitively obtained by removing the cross traffic when it interferes with flow 1, is given in next theorem.

Theorem 3 (End-to-end dynamic server (Bouillard et al, 2022, Theorem 6))**.** *With the notations and assumptions (H_1)–(H_3), the end-to-end service offered to flow 1 is a dynamic S_{e2e}-server with $\forall 0 \leq t_1 \leq t_{n+1}$ and $[\cdot]_+ = \max(\cdot, 0)$,*

$$S_{e2e}(t_1, t_{n+1}) = \Big[\inf_{\substack{t_j \leq t_{j+1}, \\ 1 \leq j \leq n}} \sum_{j=1}^{n} [S_j(t_j, t_{j+1})] - \sum_{i=2}^{m} A_i(t_{f_i}, t_{\ell_i + 1}) \Big]_+ .$$

Example 3. *Dynamic end-to-end servers for the networks (a) and (b) of Figure 1 are:*

- $S^{(a)}_{e2e}(t_1, t_3) = [\inf_{t_1 \leq t_2 \leq t_3} S_1(t_1, t_2) + S_2(t_2, t_3)]_+$ *and*
- $S^{(b)}_{e2e}(t_1, t_4) = [\inf_{t_1 \leq t_2 \leq t_3 \leq t_4} S_1(t_1, t_2) + S_2(t_2, t_3) + S_3(t_3, t_4) - A_2(t_1, t_3) - A_3(t_2, t_4)]_+$.

In the latter example, the end-to-end service considers the service of server j on the time interval $[t_j, t_{j+1}]$. The arrival process of flow 2 is taken into account on the time *interval* $[t_1, t_3]$*, which corresponds to the periods concerning servers 1 and 2, which are precisely the servers on the path of flow 2.*

Assumption 2 (Processes)**.** *We make the following assumptions on the processes:*

(H_4) *arrival processes* $(A_i)_{i=1}^m$ *and service processes* $(S_j)_{j=1}^n$ *are stationary MMPs, with notations explained in Paragraph 2.2;*

(H_5) *arrival processes* $(A_i)_{i=1}^m$ *and service processes* $(S_j)_{j=1}^n$ *are mutually independent.*

We now can give a bound of the end-to-end service using a service bgf.

Theorem 4 (End-to-end service bgf (Bouillard et al, 2022, Theorem 8))**.** *With notations and assumptions* (H_1)*–*(H_5)*, a service bgf for flow 1 is*

$$F_{S_{e2e}}(\theta, z) = e^{\theta(\sum_{i=2}^m \sigma_{A_i}(\theta) + \sum_{j=1}^n \sigma_{S_j}(\theta))} \prod_{j=1}^n \frac{1}{1 - e^{-\theta(\rho_{S_j}(\theta) - \sum_{i \in \mathrm{Fl}(j)\setminus\{1\}} \rho_{A_i}(\theta))} z}.$$

Theorem 4 is obtained from Theorem 3 by bounding the MGF of the end-to-end service $\mathbf{E}[e^{-\theta S_{e2e}(t,t+k)}]$ for each interval length k. MGF-based SNC uses the union bound to replace the expectation of a maximum (the infimum transforms into a supremum because of the "$-\theta$" MGF of a service process) as a sum of the expectations. The sum of products of expectations then translate into Cauchy products of series, and the product of the GFs (see Eq. (4)).

Example 4. *Service bgf for the end-to-end servers of networks (a) and (b) of Figure 1 are (dependencies in* θ *are omitted in the second example for the sake of concision):*

- $F_{S_{e2e}^{(a)}}(\theta, z) = F_{S_1}(\theta, z) \cdot F_{S_2}(\theta, z) = \frac{e^{\theta(\sigma_{S_1}(\theta) + \sigma_{S_2}(\theta))}}{\left(1 - e^{-\theta\rho_{S_1}(\theta)} z\right)\left(1 - e^{-\theta\rho_{S_2}(\theta)} z\right)}$ *is the product of the two service bgfs of the work-conserving servers and*
- $F_{S_{e2e}^{(b)}}(\theta, z) = \frac{e^{\theta(\sigma_{S_1} + \sigma_{S_2} + \sigma_{S_3} + \sigma_{A_2} + \sigma_{A_3})}}{\left(1 - e^{-\theta(\rho_{S_1} - \rho_{A_2})} z\right)\left(1 - e^{-\theta(\rho_{S_2} - \rho_{A_2} - \rho_{A_3})} z\right)\left(1 - e^{-\theta(\rho_{S_3} - \rho_{A_3})} z\right)}$ *can be interpreted as the product of the three* residual servers*, when the arrival processes of the cross traffic is removed from the service.*

Stability and bottlenecks

Consider a tandem network described as above. This network is stable if the backlog of flow 1 is bounded in expectation. In our setting, this corresponds to the existence of $\theta > 0$ such that $\forall b \in \mathbb{N}$, $\mathbf{P}(q(t) > b) \leq C(\theta)e^{-\theta b}$ for some constant $C(\theta) \in \mathbb{R}_+$.

When Theorem 4 is combined with the arrival curve for flow 1, the backlog v.p. bound is

$$\mathbf{P}(q(t) \geq b) \leq \frac{e^{-\theta b} e^{\theta(\sum_{i=1}^{m} \sigma_{A_i}(\theta) + \sum_{j=1}^{n} \sigma_{S_j}(\theta))}}{\prod_{j=1}^{n} \left(1 - e^{-\theta(\rho_{S_j}(\theta) - \sum_{i \in \mathrm{Fl}(j)} \rho_{A_i}(\theta))}\right)},$$

and is defined for all θ such that $r_A(\theta) r_{S_{e2e}}(\theta) > 1$. Here, $r_A(\theta) = e^{-\theta \rho_A(\theta)}$ and the dominant singularity of $F_{e2e}(\theta, z)$ is

$$r_{S_{e2e}}(\theta) = \min_{j \in \{1,\dots,n\}} e^{\theta(\rho_{S_j}(\theta) - \sum_{i \in \mathrm{Fl}(j) \setminus \{1\}} \rho_{A_i}(\theta))}.$$

In other words, for all server j, one must have $\rho_{S_j}(\theta) > \sum_{i \in \mathrm{Fl}(j)} \rho_{A_i(\theta)}$. Let us denote for each server j,

$$\theta_j^* = \sup\left\{\theta \geq 0 \mid \forall j \in \{1, \dots, n\},\ \rho_{S_j}(\theta) > \sum_{i \in \mathrm{Fl}(j)} \rho_{A_i}(\theta)\right\}.$$

The formula for the backlog v.p. is then valid for all $\theta < \theta^* = \min_{j=1}^{n} \theta_j^*$. We call the *bottleneck(s)* the server(s) at which this minimum is reached.

3 Tandem analysis with localized use of martingales

In this section, we describe the main contribution of the paper. We extend to the case of multiple servers the use of martingale in the network calculus framework. As stated in the introduction, the use of the martingale is localized at one server. Nevertheless, performance bounds are improved. In this analysis, we rather follow the approach of Duffield (1994). First, in Paragraph 3.1, we generalize their Theorem 3 to Theorem 5 that is key for the analysis. Then in Paragraph 3.2, we detail the analysis for the two-server tandem of Figure 1(a), that is representative enough of the approach. Finally, we will state the general result, Theorem 6 in Paragraph 3.3.

3.1 Martingales for arrival and service processes

Consider an arrival process A generated by a MMP. Using the notation of Paragraph 2.2, we can define for all $(u, v) \in \mathbb{N}^2_\leq$,

$$M_A(\theta, u, v) = e^{\theta A(u,v) - \theta \rho_A(\theta)(v-u)} \nu_A(\theta)_{X_A(u)}.$$

Let $\mathcal{F}_A(u, v)$ be the σ-algebra generated by $(X_A(u), \ldots, X_A(v), a_u, \ldots, a_{v-1})$.

Lemma 1. *For all $\theta \in \mathbb{R}$ such that $\psi_A(\theta)$ is defined, for all $(u, v) \in \mathbb{N}^2_\leq$, $\{M_A(\theta, u - \tau, v)\}_{\tau \in \mathbb{N}}$ is a martingale with respect to the filtration $(\mathcal{F}_A(u - \tau, v))_{\tau \in \mathbb{N}}$.*

This result is almost a rewriting of Lemma 1 of Duffield (1994) to the specific case of MMP and is omitted for the sake pf space. However, the complete proof can be found in Bouillard (2024). The main difference is on that we use bivariate functions for the service, and then make explicit that the martingale is used in reversed-time. If A has i.i.d increments, then $M_A(\theta, u, v) = e^{\theta A(u,v) - \theta \rho_A(\theta)(v-u)}$.

Similarly, if S is the service process of a work-conserving server generated by a MMP, we can define for all $(u, v) \in \mathbb{N}_\leq$,

$$M_S(\theta, u, v) = e^{-\theta S(u,v) + \theta \rho_S(\theta)(v-u)} \nu_S(-\theta)_{X_S(u)}.$$

If $\mathcal{F}_S(u, v)$ is the σ-algebra generated by $(X_S(u), \ldots, X_S(v), s_u, \ldots, s_{v-1})$, then for all $\theta \in \mathbb{R}$ such that $\psi_S(-\theta)$ is defined, for all $(u, v) \in \mathbb{N}^2_\leq$, $\{M_S(\theta, u - \tau, v)\}_{\tau \in \mathbb{N}}$ is a martingale with respect to the filtration $(\mathcal{F}_S(u - \tau, v))_{\tau \in \mathbb{N}}$.

Theorem 5. *Consider mutually independent arrival processes* $(A_i)_{i=1}^m$ *and service process* S *satisfying Assumptions* $(H_5) - (H_6)$*, and* $(u_i, v_i)_{i=0}^m \in (\mathbb{N}^2_{\leq})^{m+1}$ *and* Y *a random variable independent of* $(A_i)_{i=1}^m$ *and* S*. Define for all* $\tau \in \mathbb{N}$*,*

$$W_\tau = \sum_{i=1}^m A_i(u_i - \tau, v_i) - S(u_0 - \tau, v_0).$$

For all $\theta \in \mathbb{R}_+$ *satisfying* $\sum_{i=1}^m \rho_{A_i}(\theta) - \rho_S(\theta) \leq 0$*, there exists a constant* $\xi_{(A_i),S}(\theta)$ *independent of* $(u_i, v_i)_{i=0}^m$ *such that*

$$\mathbf{P}(\sup_{\tau \geq 0} W_\tau \geq Y) \leq \xi_{(A_i),S}(\theta)\mathbf{E}[e^{-\theta Y}]e^{\theta(\sum_{i=1}^m \rho_{A_i}(\theta)(v_i - u_i) - \rho_S(\theta)(v_0 - u_0)))}.$$

Compared to Theorem 3 of Duffield (1994), we clearly separate the arrival processes and the service processes. This enables us to consider them from different end point: while in Duffield (1994), the starting time is 0 and backward processes are implicitly used, the equivalent would be to set $u_i = v_i = 0$ for all $i \in \{0, \ldots, m\}$. So here, we introduce more flexibility in the definition of the process $(W_\tau)_{\tau \in \mathbb{N}}$. Nevertheless, the proof of Theorem 5 follows the lines of Theorem 3 of Duffield (1994) and is omitted here for the sake of space. However, it is detailed in Bouillard (2024).

The constant $\xi_{(A_i),S}(\theta)$ depends on the MMP of the arrival and service processes. It can be expressed as $\xi_{(A_i),S}(\theta) = (\inf\{\nu(\theta)_x \mid x \in \mathcal{P}\})^{-1}$, where $\nu(\theta)$ is the tensor product of the $\nu_X(\theta)$'s, $X \in \{A_1, \ldots, A_m, S\}$ and $\mathcal{P}$ is the set of states that have a positive probability to receive more arrivals than the amount of service offered. When all processes have i.i.d. increments, then $\xi_{(A_i),S}(\theta) = 1$. More generally, with the choice of $\sigma_X(\theta)$ in Paragraph 2.2, we have $\xi_{(A_i),S}(\theta) \leq e^{\theta(\sigma_S(\theta) + \sum_{i=1}^m \sigma_{A_i}(\theta))}$, with equality if there can be more arrivals than service in all states.

3.2 Analysis of a two-server tandem network

We now compute new bounds of the v.p. for the backlog and delay for tandem networks on the small, yet representative example of Figure 1(a).

3.2.1 Backlog violation probability

Consider the network of Figure 1(a). From Theorems 1 and 3, the v.p. for backlog b is

$$\mathbf{P}(q(t_3) \geq b) \leq \mathbf{P}(\exists t_1 \leq t_2 \leq t_3,\ A_1(t_1,t_3) - S_1(t_1,t_2) - S_2(t_2,t_3) \geq b).$$

The next computation is done in several steps: first in (5) we partially use the union bound and sum on all t_2, then in (6), we apply Theorem 5 with $Y = b + S_2(t_2,t_3)$, $u_0 = u_1 = t_2$, $v_0 = t_2$ and $v_1 = t_3$. This is valid for all $\theta \in [0, \theta_1^*]$. In (7), we use the $(\sigma(\theta), \rho(\theta))$ representations to bound the expectation, and finally, in (8), we sum all the terms. The sum is finite for all $\theta \in [0, \theta_2^*)$. For all $\theta \in [0, \theta_1^*] \cap [0, \theta_2^*)$,

$$\begin{aligned}
\mathbf{P}(q(t_3) \geq b) &\leq \mathbf{P}\big(\sup_{t_1 \leq t_2 \leq t_3} A_1(t_1,t_3) - S_1(t_1,t_2) - S_2(t_2,t_3) \geq b\big) \\
\textbf{(UB)} &\leq \sum_{t_2 \leq t_3} \mathbf{P}\big(\sup_{t_1 \leq t_2} A_1(t_1,t_3) - S_1(t_1,t_2) \geq b + S_2(t_2,t_3)\big) \quad (5)\\
&\leq \sum_{t_2 \leq t_3} \mathbf{P}\big(\sup_{\tau \geq 0} A_1(t_2-\tau,t_3) - S_1(t_2-\tau,t_2) \geq b + S_2(t_2,t_3)\big) \\
\textbf{(Th. 5)} &\leq \sum_{t_2 \leq t_3} \xi_{A_1,S_1}(\theta) e^{\theta \rho_{A_1}(\theta)(t_3-t_2)} \mathbf{E}\big[e^{-\theta(b+S_2(t_2,t_3))}\big] \quad (6)\\
(\sigma,\rho) &\leq \sum_{t_2 \leq t_3} \xi_{A_1,S_1}(\theta) e^{\theta \rho_{A_1}(\theta)(t_3-t_2) - \rho_{S_2}(\theta)(t_3-t_2) + \sigma_{S_2}(\theta) - b} \quad (7)\\
\textbf{(GS)} &\leq \frac{\xi_{A_1,S_1}(\theta) e^{\theta(\sigma_{S_2}(\theta)-b)}}{1 - e^{-\theta(\rho_{S_2}(\theta) - \rho_{A_1}(\theta))}}. \quad (8)
\end{aligned}$$

One can first remark that with slightly modified constant terms, this formula is similar to the backlog v.p. for the network made only of server 2 and flow 1 (see Eq. (1)).

This formula can also be compared with the one obtained with the PMOO analysis from Bouillard et al (2022): $\forall\theta \in [0, \min(\theta_1^*, \theta_2^*))$,

$$\mathbf{P}(q(t) \geq b) \leq \frac{e^{\theta(\sigma_{S_2}(\theta)+\sigma_{S_1}(\theta)+\sigma_{A_1}(\theta)-b)}}{\big(1-e^{-\theta(\rho_{S_1}(\theta)-\rho_{A_1}(\theta))}\big)\big(1-e^{-\theta(\rho_{S_2}(\theta)-\rho_{A_1}(\theta))}\big)}.$$

Intuitively, if $\theta_1^* < \theta_2^*$ (server 1 is the bottleneck), the optimal value taken for the PMOO formula is around θ_1^*, and the value of $1-e^{-\theta(\rho_{S_1}(\theta)-\rho_A(\theta))}$ be very small. The localized use of the martingale analysis then drastically improves the bound. On the contrary, if $\theta_2^* \leq \theta_1^*$, the gain is more limited as the optimal value for θ will approach θ_2^*, and in the two expressions, the factor $(1-e^{-\theta(\rho_{S_2}(\theta)-\rho_A(\theta))})$ is small resulting in a large prefactor in both analyses. The gain for the partial use of the martingale is then approximately $(1-e^{-\theta(\rho_{S_1}(\theta)-\rho_{A_1}(\theta))})^{-1}$ only.

The gain would then be much larger if the application of the martingale analysis could be localized at server 2. Unfortunately, this seems to be a much more difficult problem. Indeed, a first attempt for this would be to compute

$$\mathbf{P}(q(t_3) \geq b) \leq \sum_{t_1 \leq t_3} \mathbf{P}\big(\sup_{t_1 \leq t_2 \leq t_3} A_1(t_1,t_3) - S_1(t_1,t_2) - S_2(t_2,t_3) \geq b\big).$$

In this latter expression, the process $(S_1(t_1,t_2)+S_2(t_2,t_3))_{t_2\in[t_1,t_3]}$ is not a martingale, and the presented approach cannot be applied.

The issue can be solved when the first server is a constant-rate server: there exists a constant C_1 such that for all $(t_1,t_2) \in \mathbb{N}^2_{\leq}$, $S_1(t_1,t_2) = C_1(t_2-t_1)$. Then instead of fixing t_2 for applying Theorem 5 at server 2, one can fix $k = t_2 - t_1$. In line (9), t_1 is replaced by $t_2 - k$. In line (10), use the constant rate service property: $S_1(t_2-k,t_2) = C_1 k$ does not depend on t_2. Theorem 5 is then applied in line (11) with $u_1 = t_3 - k$, and $u_0 = v_0 = v_1 = t_3$.

$$
\begin{aligned}
\mathbf{P}(q(t_3) \geq b) &\leq \mathbf{P}\big(\sup_{t_1 \leq t_2 \leq t_3} A_1(t_1,t_3) - S_1(t_1,t_2) - S_2(t_2,t_3) \geq b\big) \\
\textbf{(UB)} &\leq \sum_{k \geq 0} \mathbf{P}\big(\sup_{t_2 \leq t_3} A_1(t_2-k,t_3) - S_1(t_2-k,t_2) - S_2(t_2,t_3) \geq b\big) \qquad (9) \\
&\leq \sum_{k \geq 0} \mathbf{P}\big(\sup_{t_2 \leq t_3} A_1(t_2-k,t_3) - S_2(t_2,t_3) \geq b + C_1 k\big) \qquad (10) \\
&\leq \sum_{k \geq 0} \mathbf{P}\big(\sup_{\tau \geq 0} A_1(t_3-k-\tau,t_3) - S_2(t_3-\tau,t_3) \geq b + C_1 k\big) \\
\textbf{(Th. 5)} &\leq \sum_{k \geq 0} \xi_{A_1,S_2}(\theta) e^{-\theta(b-\rho_{A_1}(\theta)k)} e^{-\theta C_1 k} \qquad (11) \\
\textbf{(GS)} &\leq \frac{\xi_{A_1,S_2}(\theta) e^{-\theta b}}{1 - e^{-\theta(\rho_{C_1} - \rho_{A_1}(\theta))}}.
\end{aligned}
$$

Remark that if S_2 is also a constant rate-server, and end-to-end server is $S_1 \wedge S_2$, and the network can be analyzed as a single-server network.

3.2.2 Delay violation probability

Let us now focus on the computation of the v.p. of the delay, again with the network of Figure 1(a). Recall that we have, from Theorems 1 and 3,

$$
\begin{aligned}
d(t_3-T+1) \geq T &\Rightarrow \exists t_1 \leq t_3 - T,\ A_1(t_1,t_3-T+1) > \inf_{t_1 \leq t_2 \leq t_3} S_1(t_1,t_2) + S_2(t_2,t_3) \\
&\Rightarrow \exists t_2 \leq t_3, \quad \sup_{t_1 \leq t_2 \wedge t_3 - T} A_1(t_1,t_3-T+1) - S_1(t_1,t_2) > S_2(t_2,t_3).
\end{aligned}
$$

One can then write, using the union bound,

$$
\begin{aligned}
\mathbf{P}(d(t_3-T+1) \geq T) &\leq \sum_{t_2 \leq t_3} \mathbf{P}\big(\sup_{t_1 \leq t_2 \wedge t_3 - T} A_1(t_1,t_3-T+1) - S_1(t_1,t_2) > S_2(t_2,t_3)\big) \\
&= \sum_{t_2 \leq t_3 - T} \mathbf{P}\big(\sup_{t_1 \leq t_2} A_1(t_1,t_3-T+1) - S_1(t_1,t_2) > S_2(t_2,t_3)\big) \\
&\quad + \sum_{t_3 - T < t_2 \leq t_3} \mathbf{P}\big(\sup_{t_1 \leq t_3 - T} A_1(t_1,t_3-T+1) - S_1(t_1,t_2) > S_2(t_2,t_3)\big).
\end{aligned}
$$

In the last equality, we distinguish two cases, depending on how t_2 and $t_3 - T$ compare. We will deal with them separately.

In the first sum sign, that we denote P_1, one can apply Theorem 5 with $Y = S_2(t_2, t_3)$, $u_0 = u_1 = v_0 = t_2$ and $v_1 = t_3 - T + 1$. For all $\theta_1 \in [0, \theta_1^*] \cap [0, \theta_2^*)$,

$$
\begin{aligned}
P_1 &\leq \sum_{t_2 \leq t_3 - T} \xi_{A_1,S_1}(\theta_1) e^{\theta_1 \sigma_{S_2}(\theta_1)} e^{-\theta_1(\rho_{S_2}(\theta_1)(t_3-t_2) - \rho_{A_1}(\theta_1)(t_3-T+1-t_2))} \\
\textbf{(GS)} &\leq \frac{\xi_{A_1,S_1}(\theta_1) e^{\theta_1(\sigma_{S_2}(\theta_1) + \rho_{A_1}(\theta_1) - \rho_{S_2}(\theta_1)T)}}{1 - e^{-\theta_1(\rho_{S_2}(\theta_1) - \rho_{A_1}(\theta_1))}}.
\end{aligned}
$$

With slightly modified constant terms, one can recognize the delay v.p. for the network made only of server 2 and flow 1.

In the second sum sign, that we denote P_2, let us apply Theorem 5 with $u_0 = u_1 = t_3 - T$, $v_1 = t_3 - T + 1$ and $v_0 = t_2$. We then obtain for all $\theta_2 \in [0, \theta_1^*] \cap [0, \theta_2^*)$,

$$
\begin{aligned}
P_2 &\leq \sum_{t_3 - T < t_2 \leq t_3} \xi_{A_1,S_1}(\theta_2) e^{\theta_2 \sigma_{S_2}(\theta_2)} e^{-\theta_2(\rho_{S_1}(\theta_2)(t_2 - t_3 + T) - \rho_{A_1}(\theta_2) + \rho_{S_2}(\theta_2)(t_3 - t_2))} \\
&\leq \xi_{A_1,S_1}(\theta_2) e^{\theta_2(\sigma_{S_2}(\theta_2) + \rho_{A_1}(\theta_2) - \rho_{S_1}(\theta_2))} \sum_{u=0}^{T-1} e^{-\theta_2(\rho_{S_1}(\theta_2)(T-1-u) + \rho_{S_2}(\theta_2)(u))} \\
&\leq [z^{T-1}] \frac{\xi_{A_1,S_1}(\theta_2) e^{\theta(\sigma_{S_2}(\theta_2) + \rho_{A_1}(\theta_2) - \rho_{S_1}(\theta_2))}}{\left(1 - e^{-\theta_2 \rho_{S_1}(\theta_2)} z\right)\left(1 - e^{-\theta_2 \rho_{S_2}(\theta_2)} z\right)}.
\end{aligned}
$$

We recognize the $T-1$-th term of the Cauchy product in Eq. (4), and with slightly modified constant term, this is the $T-1$-th term of the service bgf of the end-to-end server for flow 1. Finally, the delay v.p. can be bounded by

$$
\begin{aligned}
\mathbf{P}(d(t_3 - T) \geq T) \leq \frac{\xi_{A_1,S_1}(\theta_1) e^{\theta_1(\sigma_{S_2}(\theta_1) + \rho_{A_1}(\theta_1) - \rho_{S_2}(\theta_1)T)}}{1 - e^{-\theta_1(\rho_{S_2}(\theta_1) - \rho_{A_1}(\theta_1))}} \\
+ [z^{T-1}] \frac{\xi_{A_1,S_1}(\theta_2) e^{\theta_2(\sigma_{S_2}(\theta_2) + \rho_{A_1}(\theta_2) - \rho_{S_1}(\theta_2))}}{\left(1 - e^{-\theta_2 \rho_{S_1}(\theta_2)} z\right)\left(1 - e^{-\theta_2 \rho_{S_2}(\theta_2)} z\right)},
\end{aligned}
$$

and to minimize this bound, θ_1 and θ_2 can be optimized independently.

This computation suffers from the same limitation as for the backlog: it has not been possible yet to apply Theorem 5 to the second server directly, unless assuming that the first server is a constant-rate server.

3.3 Main Theorem for Tandem Networks

In this section, we generalize the computations presented for the two-server tandem. Following this approach, we apply the martingale analysis (and the Doob's inequality) locally at one server, and the union and Chernoff bounds for the other servers. From the discussion when computing the backlog in the two-server tandem, the martingale analysis cannot be applied to any server h, and some assumptions have to be fulfilled by server h.

Assumption 3 (Conditions for martingale analysis at server h)**.** *Server h must satisfy*

(H_6) *for all $j < h$, server j is a constant-rate server,* i.e., *there exists C_j such that $S_j(u,t) = C_j(t-u)$;*

(H_7) *for all flow $i \in \mathbb{N}_m$, $f_i \leq h \implies \ell_i \geq h$. In other words, no flow arriving before server h departs before server h.*

Assumption (H_6) is a direct consequence of the discussion for the two-server backlog: the martingale part of the analysis can be applied to the second server only if the first server is constant-rate. Here, (H_6) assumes that the all upsteam servers of server h are constant-rate.

Assumption (H_7) also relates to this. Assume for example that flow 2 crosses server 1 only (as in Figure 2 with only flows 1 and 2). For the viewpoint of flow 1, server 1 offers the service $S_1 - A_2$, which is not constant-rate anymore (unless A_2 is deterministic). Thus, the martingale part of the analysis cannot be applied to server 2. Now, if flow 2 also crosses server 2, flow 2 can be incorporated in the partial analysis.

Transforming a network by removing one server

In Section 3.2, the v.p. bound looks like the v.p. bound of a network reduced to the second server only. This holds more generally: we can separate the part where the usual MGF-SNF analysis with union bound is used (tandem minus one server) and where the Doob's inequality is applied (one server). The tandem network described next represents the part of the network we analyze with the usual MGF-SNC method.

Consider a tandem network described by the notations in (H_1) and (H_2). The tandem network obtained by removing server h is constructed as follows:

- it is made of servers $1, \dots, h-1, h+1 \dots, n$;
- for all flow i, its path of flow i is unchanged unless it crosses server h, in case it goes directly from server $h-1$ to server $h+1$;
- flows originally crossing server h only are removed;
- arrival processes and service processes of the remaining flows and servers are unchanged.

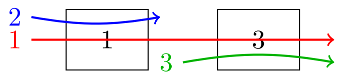


Fig. 2 Network obtained from the network in Figure 1(b) after removal of server 2.

For example, the network obtained from Figure 1(b) by removing server 2 is depicted in Figure 2. The notation $S_{e2e}^{(-h)}$ refers to the end-to-end server where server h has been removed.

We can now state our main result, proved in Section 6.

Theorem 6. *Consider a tandem network satisfying notations and assumptions given in* (H_1)*–*(H_7)*, and* $\Theta = [0, \theta_h^*] \cap [0, \inf_{j \neq h} \theta_j^*)$*. For all* $\theta, \theta_1 \in \Theta$ *and* $\theta_2 \in [0, \theta_h^*]$*,*

1. *the violation probability of the backlog for flow 1 satisfies*

$$\mathbf{P}(q(t) \geq b) \leq \frac{\xi_{(A_i)_{i\in\mathrm{Fl}(h)},S_h}(\theta)e^{-\theta\sigma_{A_1}(\theta)}}{e^{\theta(\sum_{i\in\mathrm{Fl}(h)-H}\sigma_{A_i}(\theta))}} F_{S_{e2e}^{(-h)}}(\theta, e^{\theta\rho_{A_1}(\theta)})e^{-\theta b},$$

2. *a delay bounding generating series for the delay of flow 1 is*

$$\frac{\xi_{(A_i)_{i\in\mathrm{Fl}(h)},S_h}(\theta_1)}{e^{\theta_1\sum_{i\in\mathrm{Fl}(h)\setminus H}\sigma_{A_i}(\theta_1)}} F_{d,S_e2e^{(-h)}}(\theta_1, z) + \frac{\xi_{(A_i)_{i\in\mathrm{Fl}(h)},S}(\theta_2)e^{-\theta_2(\rho_{S_h}(\theta_2)-\sum_{i\in\mathrm{Fl}(h)}\rho_{A_i}(\theta_2))}}{e^{\theta_2(\sigma_{S_h}(\theta_2)+\sum_{i\in\mathrm{Fl}(h)\setminus\{1\}}\sigma_{A_i}(\theta_2))}} zF_{S_{e2e}}(\theta_2, z),$$

where H is the set of flows crossing only server h, and the bounding generating functions are computed as in Theorems 1 and 4.

The v.p. of the backlog is similar to the one with MGF-based SNC obtained by combining Theorems 1 and 4 when server h is removed. This can be explained as follows: the end-to-end service bgf $F_{S_{e2e}}(\theta, z)$ in Theorem 4 is the product of n geometric functions, each representing one server. The product of generating function represents the Cauchy product of series, and in our case, corresponds to the union bound at each server. When martingale analysis is used locally at one server, then there is no union bound for that server, but the union bound is still used for the other servers and $F_{S_{e2e}}^{(-h)}(\theta, z)$ naturally appears. The prefactor is some reorganization of the terms, and compensating the $e^{\theta\sigma_{A_i}(\theta)}$ that are already taken into account in a different way in $\xi_{(A_i)_{i\in\mathrm{Fl}(h)},S_h}(\theta)$. The pre-factor $\frac{\xi_{(A_i)_{i\in\mathrm{Fl}(h)},S_h}(\theta)e^{-\theta\sigma_{A_1}(\theta)}}{e^{\theta(\sum_{i\in\mathrm{Fl}(h)-H}\sigma_{A_i}(\theta))}} \leq e^{\theta(\sigma_{S_h}(\theta)+\sum_{i\in H}\sigma_{A_i}(\theta))}$, which are terms that would appear in the MGF-SNC analysis of the complete end-to-end tandem, but not when server h is removed.

The v.p. of the delay has two parts. The interpretation of the first term is similar to the one of the v.p. of the backlog. The second term comes from the case distinction that was done. In the second case, the sum has a finite number of terms, and the factor

$e^{-\theta\rho_{A_1}(\theta)}$ does not appear with a power of some time variables. This means that this term will be similar to the end-to-end service.

4 An Alternative Analysis with Martingales

Another way of using martingales was attempted in the literature, first in Angrishi and Killat (2011), and recently used in Yu et al (2022). In this section we briefly explain why this approach is not sound. More details can be found in Bouillard (2024).

The authors study a tandem network in the absence of cross traffic ($m = 1$ and flow 1 crosses all the servers), and where each server j is a work-conserving server with i.i.d. increments.

To simplify the notations, we adapt Angrishi and Killat (2011) and Yu et al (2022) to the computation of the backlog v.p. instead of the delay. The first step of the analysis is:

$$\begin{aligned}\mathbf{P}(q(t_{n+1}) \geq b) \leq \mathbf{P}\big([\sup_{t_1 \leq t_{n+1}} & A(t_1, t_{n+1}) - \rho_A(\theta^*)(t_{n+1} - t_1)] \\ & + \sum_{j=1}^{n-1}[\sup_{t_j \leq t_{j+1} \leq t_{n+1}} \rho_{S_j}(\theta^*)(t_{j+1} - t_j) - S_j(t_j, t_{j+1})] \\ & + [\sup_{t_n \leq t_{n+1}} \rho_{S_n}(\theta^*)(t_{n+1} - t_n) - S_j(t_j, t_{j+1})] \geq b\big). \end{aligned} \tag{12}$$

The second step is bounding tightly, for all $j < n$,

$$P_j = \mathbf{P}\big(\sup_{t_j \leq t_{j+1} \leq t_{n+1}} \rho_{S_j}(\theta^*)(t_{j+1} - t_j) - S_j(t_j, t_{j+1}) \geq x\big).$$

With $\widetilde{M}_j(t_{j+1}) = \sup_{0 \leq t_j \leq t_{j+1}} e^{-\theta^*(\rho_{S_j}(\theta^*)(t_{j+1}-t_j) - S_j(t_j, t_{j+1}))}$, the authors then claim that $P_j \leq ee^{-\theta^* x}$, and use the inequality $\mathbf{E}[\widetilde{M}_j(t_{n+1})] \leq e\mathbf{E}[M_j(0, t_{n+1})]$, refering to Theorem 3.7 by Rao Rao (2007) stating that if $(X_t)_{t\in\mathbb{N}}$ is a demi-submartingale, then for all $r > 0$, $\mathbf{E}[e^{r \max_{0\leq u\leq t} X_u}] \leq e\mathbf{E}[e^{rX_t}]$.

However, to obtain $\mathbf{E}[\sup_{0\leq t_j\leq t_{n+1}} M_j(\theta^*, t_j, t_{n+1})] \leq e\mathbf{E}[M_j(\theta^*, 0, t_{n+1})]$, one would require that $(\sup_{0\leq t_j\leq t_{n+1}} \theta^*(\rho_{S_j}(\theta^*)(t_{n+1} - t_j) - S_j(t_j, t_{n+1})))_{t_{n+1}\geq 0}$ is a sub-martingale.

Let us define $N(t) = \sup_{0\leq u\leq t} \rho_{S_j}(\theta^*)(t-u) - S_j(u, t)$. First, with $s_t = S(t, t+1)$, we have $\mathbf{E}[N(t+1)|N(t)] = \mathbf{E}[N(t) + \rho_{S_j}(\theta) - s_t \vee 0 \mid N(t)]$. But from Jensen's inequality, since $f : x \mapsto e^{-\theta^* x}$ is convex, $\mathbf{E}[e^{-\theta^* s_t}] = e^{-\theta^* \rho_{s_j}(\theta^*)} \geq e^{-\theta^* \mathbf{E}[s_t]}$ and $\mathbf{E}[s_t] \geq \rho_{s_j}(\theta^*)$. So we cannot conclude that $N(t)$ is a sub-martingale (without the maximum with 0, one could easily conclude that $(N(t))_{t\geq 0}$ is on the contrary a super-martingale).

Figure 3 illustrates the incorrect bound, with i.i.d. Poisson distribution with parameter 1 and $\theta^* = 0.5$. The bound $ee^{-\theta^* x}$ is depicted in dashed line. In solid lines, the v.p. P_j in function of x strongly depends on the values of t_{n+1} is obtained by simulation. The bounds is less respected as t_{n+1} increases.

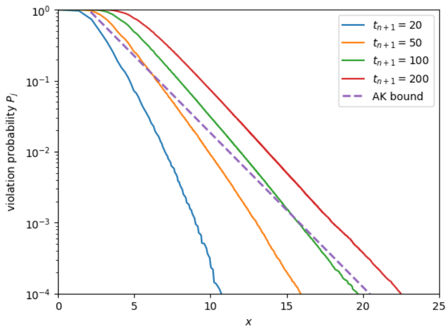


Fig. 3 Illustration of the erroneous bound for P_j in Angrishi and Killat (2011).

5 Numerical Evaluation

In this section, we compare our bounds against simulation and the state of the art. It has been demonstrated in Bouillard et al (2022) that the PMOO analysis outperforms by far the other SNC methods from the state-of-the-art. We then only compare against this method, with two types of experiment: 1) compare the v.p. for different values of target delays. For this type of experiments, we run 10 independent simulations with 10^8 time steps; 2) use a free parameter (transmission probability or capacity of a server) and observe the quality of the bound for a v.p. of 10^{-4}. For this type of experiment, we run one simulation with 10^7 time steps.

We also chose to only compare the end-to-end delays for the networks. The results for the backlog bounds show similar comparisons.

5.1 Two-server case

Let us first consider the network of Figure 1(a). Arrivals follow a MMOO process, With $P_{\texttt{Off},\texttt{On}} = 0.7$, $P_{\texttt{On},\texttt{Off}} = 0.1$, and a Poisson distribution with intensity 2 in the `On` state. Services are i.i.d.: server 1 serves 5 packets with probability p, and server 2, 6 packets with probability q (and no service otherwise).

Figure 4(left) compares the v.p. in function of a target delay obtained by simulation, PMOO and with our martingale bound, when $p = q = 0.5$. Our new methods improves the PMOO analysis. For example, with a v.p. of 10^{-4}, the simulation delay is 27, with PMOO 54 and with our method 37. The gap is then reduced by 63%.

In Figure. 4(right), we fix $q = 0.5$, the v.p. 10^{-4} and vary p. When p is large, server 2 is the bottleneck, and the improvement of the martingale analysis is limited On the contrary, when server 1 is the bottleneck (small values of p), the improvement becomes large.

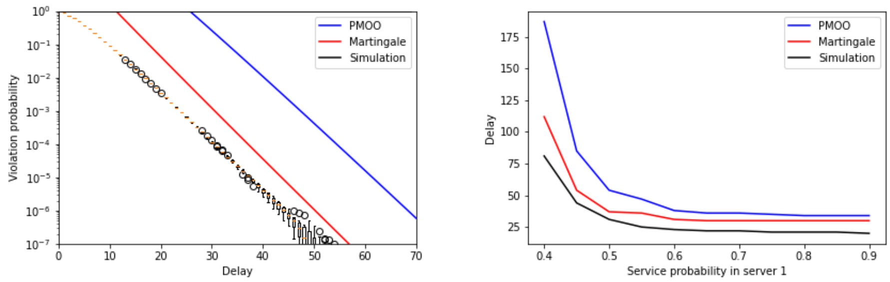


Fig. 4 Two-server tandem network with Bernoulli service process. (left) v.p. in function of the target delay; (right) delay bound in in function of the service probability of server 1.

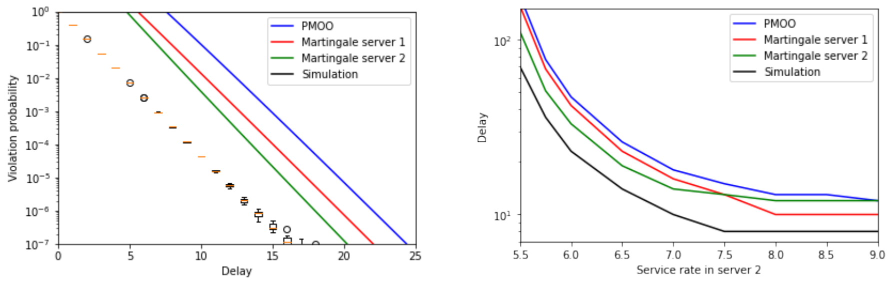


Fig. 5 Interleaved tandem network with constant-rate service. (left) v.p. in function of the target delay; (right) delay bound in function of the capacity of server 2.

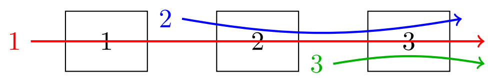
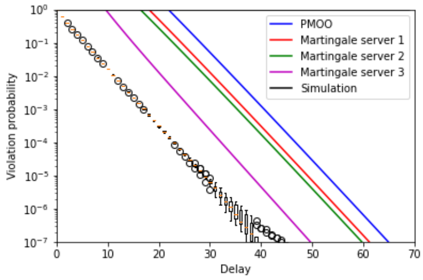


Fig. 6 Example of sink-tree tandem network.

5.2 Interleaved tandem network

Let us now focus on the network of Figure 1(b) and assume servers are constant-rate. Theorem 6 can be applied for $h \in \{1, 2\}$. In Figure 5(left), we set the rates of the server as $C_1 = 5$, $C_2 = 7$ and $C_3 = 6$. The bottleneck is the server 2. The figure compares simulation, PMOO and the bound obtained by Theorem 6 applied respectively to server 1 and 2. As expected, the gain is more important when applying the martingale analysis to server 2.

In Figure 5(right), C_2 is varying from 5.5 to 9, so depending on its value, server 1 or server 2 is the bottleneck. When server 2 is the bottleneck ($C_2 \leq 7.5$), it is better to to apply the martingale analysis at server 2 and at server 1 when it is the bottleneck ($C_2 \geq 7.5$). The reduction of the pessimism gap compared to PMOO ranges from 33% to 75%. The reduction is the smallest for $C = 7.5$, when servers 1 and 2 are bottlenecks.



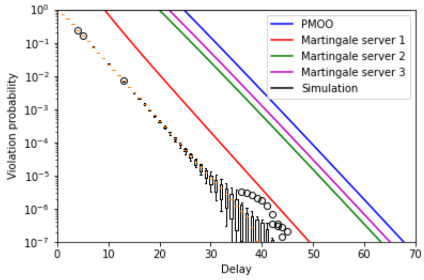


Fig. 7 Sink-tree tandem network. Violation probability in function of the target delay when server 3 (left) or server 1 (right) is the bottleneck.

5.3 Sink-trees tandem network

A sink-tree tandem network is a tandem network for which all flows end at the last server, as in Figure 6 use here. Assumption (H_6) and (H_7) are then always satisfied for constant-rate servers. The arrival processes are the same as previously. In Figure 7(left), the service rates are $C_i = 3+i$, $i \in \{1,2,3\}$. In that case, $\theta_3^* < \theta_2^* < \theta_1^*$ and server 3 is the bottleneck, and it is best to to apply the martingale analysis to server 3 (pessimism gap reduced by more than 50%). Applying it on server 1 or 2 only marginally improves the bounds.

In Figure 7(right), the service rates are $C_i = 3i-1$, $i \in \{1,2,3\}$, and $\theta_1^* < \theta_2^* < \theta_3^*$. Server 1 is then the bottleneck, and the results are reversed: applying the martingale at server 1 reduces the pessimism gap the most.

6 Proof of Theorem 6

From Theorem 3, an end-to-end dynamic server for flow 1 is

$$S_{e2e}(t_1, t_{n+1}) = \Big[\inf_{\forall j,\ t_j \le t_{j+1}} \sum_{j=1}^{n} S_j(t_j, t_{j+1}) - \sum_{i=2}^{n} A_i(t_{f_i}, t_{\ell_i+1})\Big]_+.$$

From Theorem 1, one can express the violation of the backlog bound b and the delay bound T as

$$
\begin{aligned}
q(t_{n+1}) \geq b \Rightarrow & \sup_{t_1 \leq t_{n+1}} A_1(t_1, t_{n+1}) - S_{2e2}(t_1, t_{n+1}) \geq b \\
& \Rightarrow \sup_{\substack{1 \leq j \leq n \\ t_j \leq t_{j+1}}} \sum_{i=1}^{m} A_i(t_{f_i}, t_{\ell_i+1}) - \sum_{j=1}^{n} S_j(t_j, t_{j+1}) \geq b, \text{ and}
\end{aligned}
$$

$$
\begin{aligned}
d(t_{n+1} - T + 1) \geq T \Rightarrow & \exists t_1 \leq t_{t+1} - T,\ A_1(t_1, t_{n+1} - T + 1) > S_{e2e}(t_1, t_{n+1}) \\
& \Rightarrow \exists t_1 \leq t_{n+1} - T, \exists t_1 \leq t_2 \leq \cdots \leq t_{n+1}, \qquad (13) \\
& A_1(t_1, t_{n+1} - T + 1) > \sum_{j=1}^{n} S_j(t_j, t_{j+1}) - \sum_{i=2}^{m} A_i(t_{f_i}, t_{\ell_i+1}).
\end{aligned}
$$

Let us assume that (H_6) and (H_7) hold for server h. For fixed values of $t_1 \leq \cdots \leq t_{n+1}$, let us define the new notations $k_1, \ldots, k_{h-1}, \tau$ as:

- $k_j = t_{j+1} - t_j$ for all $j < h$;
- $\tau = t_{h+1} - t_h$.

Using this transformation between variables, we have the equivalence between the two sets $\{(t_1, \ldots, t_{n+1}) \mid t_1 \leq \cdots \leq t_{n+1}\}$ and $\{(k_1, \ldots, k_{h-1}, \tau, t_{h+1}, \ldots, t_{n+1}) \mid k_1, \ldots, k_{h-1}, \tau \in \mathbb{N}, t_{h+1} \leq \cdots \leq t_{n+1}\}$ (see Figure 8).

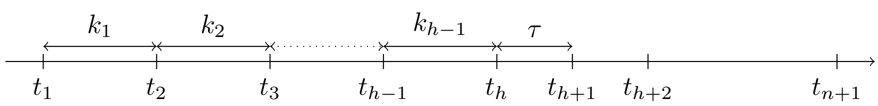


Fig. 8 Variable change.

We use the notation $k_j^{j'} = k_j + \cdots + k_{j'}$, with the convention $k_j^{j'} = 0$ if $j' < j$. For example, for a flow i crossing server h, we have

$$t_{f_i} = t_{h+1} - \sum_{j=f_i}^{h} (t_{j+1} - t_j) = t_{h+1} - \sum_{j=f_i}^{h-1} k_j - \tau = t_{h+1} - k_{f_i}^{h-1} - \tau.$$

Let us also define the sets $\mathcal{K} = \{(k_1, \ldots, k_{h-1}, t_{h+1}, t_{n+1}) \in \mathbb{N}^n \mid t_{h+1} \leq \cdots \leq t_{n+1}\}$ and $\Theta = [0, \theta_h^*] \cap [0, \inf_{j \neq h} \theta_j^*)$.

Each computation in the next paragraphs has the main three steps: 1) application of Theorem 5 for each element of $\mathcal{K}$; 2) summing the terms (union bound); 3) rewriting and simplifying the formula.

To simplify the formulas for the end of the proof, we introduce the following notations to group together the quantities related to servers:

- $A_{[h]} = (A_i)_{i \in \mathrm{Fl}(h)}$, $\sigma_{A[-h]} = \sum_{i \notin \mathrm{Fl}(h)} \sigma_{A_i}$, $\sigma_{S_{>h}} = \sum_{j>h} \sigma_{S_j}$;
- for each server j, $\rho_j = \rho_{S_j} - \sum_{i \in \mathrm{Fl}(j)} \sigma_{A_i}$, $\rho'_j = \rho_{S_j} - \sum_{i \in \mathrm{Fl}(j) \setminus \{1\}} \rho_{A_i}$.

For the sake of concision, we will also omit the dependence in θ of the parameters ρ, σ and ξ.

6.1 Backlog

Step 1: Application of Theorem 5.

For all $K = (k_1, \ldots, k_{h-1}, t_{h+1}, \ldots t_{n+1}) \in \mathcal{K}$, let us define

$$\begin{aligned} W_\tau^K &= \sum_{i \in \mathrm{Fl}(h)} A_i(t_{f_i}, t_{\ell_i+1}) - S_h(t_h, t_{h+1}) \\ &= \sum_{i \in \mathrm{Fl}(h)} A_i(t_{h+1} - k_{f_i}^{h-1} - \tau, t_{\ell_i+1}) - S_h(t_{h+1} - \tau, t_{h+1}), \end{aligned}$$

and the random variable

$$
\begin{aligned}
Y^K &= b + \sum_{j \neq h} S_j(t_j, t_{j+1}) - \sum_{i \notin \mathrm{Fl}(h)} A_i(t_{f_i}, t_{\ell_i+1}) \\
&= b + \sum_{j<h} k_j C_j + \sum_{j>h} S_j(t_j, t_{j+1}) - \sum_{i \notin \mathrm{Fl}(h)} A_i(t_{f_i}, t_{\ell_i+1}).
\end{aligned}
$$

As (H_7) holds, for all flow $i \notin \mathrm{Fl}(h)$, $f_i > h$. Moreover processes $(A_i)_i$ and $(S_j)_j$ are mutually independent. Therefore, Y^K is independent of W_τ^K and Theorem 5 can be applied: for all θ such that $\rho_h(\theta) = \rho_{S_h}(\theta) - \sum_{i \in \mathrm{Fl}(h)} \rho_{A_i}(\theta) \geq 0$, there exists $\xi_{A[h],S_h}(\theta)$ such that

$$
\begin{aligned}
\mathbf{P}(\sup_{\tau \geq 0} W_\tau^K \geq Y^K) &\leq \xi_{A[h],S_h} \mathbf{E}[e^{-\theta Y^K}] e^{\theta \sum_{i \in \mathrm{Fl}(h)} \rho_{A_i}(t_{\ell_i+1} - t_{h+1} + k_{f_i}^{h-1})} \\
&\leq \xi_{A[h],S_h} e^{-\theta b} e^{\theta \sum_{i \in \mathrm{Fl}(h)} \rho_{A_i}(t_{\ell_i+1} - t_{h+1} + k_{f_i}^{h-1})} \prod_{j=1}^{h-1} e^{-\theta C_j k_j} \\
&\quad \cdot \prod_{j=h+1}^{n} e^{\theta(\sigma_{S_j} - \rho_{S_j}(t_{j+1} - t_j))} \prod_{i \notin \mathrm{Fl}(h)} e^{\theta(\sigma_{A_i} + \rho_{A_i}(t_{\ell_i+1} - t_{f_i}))} \\
&= \xi_{A[h],S_h} e^{\theta(\sigma_{S_{>h}} + \sigma_{A_{[-h]}})} e^{-\theta b} \prod_{j=1}^{h-1} e^{-\theta C_j k_j} \\
&\quad \cdot \prod_{i \in \mathrm{Fl}(h)} e^{\theta \rho_{A_i} k_{f_i}^{h-1}} \prod_{j=h+1}^{n} e^{-\theta(\rho_{S_j}(t_{j+1} - t_j))} \prod_{i=1}^{m} e^{\theta \rho_{A_i}(t_{\ell_i+1} - t_{f_i \vee (h+1)})} \\
&= \xi_{A[h],S_h} e^{\theta(\sigma_{S_{>h}} + \sigma_{A_{[-h]}})} e^{-\theta b} \prod_{j=1}^{h-1} e^{-\theta \rho_j k_j} \prod_{j=h+1}^{n} e^{-\theta \rho_j (t_{j+1} - t_j))}.
\end{aligned}
$$

To obtain the last equality, we combine all the contributions that include $t_{j+1} - t_j$ or k_j, for each server j. They consists of factors related to the service server j itself and the arrivals of the flows crossing that server.

Step 2: Union bound.

Using the right equality of Eq. (4), we obtain the v.p., for all $\theta \in \Theta$,

$$\mathbf{P}(q(t) \geq b) \leq \mathbf{P}(\exists K \in \mathcal{K},\ \sup_{\tau \geq 0} W_\tau^K \geq Y^K) \overset{\textbf{(UB)}}{\leq} \sum_{K \in \mathcal{K}} \mathbf{P}(\sup_{\tau \geq 0} W_\tau^K \geq Y^K)$$
$$\overset{\textbf{Eq. }(4)}{\leq} \xi_{A[h],S_h} e^{\theta(\sigma_{S_{>h}} + \sigma_{A_{[-h]}})} e^{-\theta b} \cdot \prod_{j \neq h} \frac{1}{1 - e^{-\theta \rho_j}}.$$

Step 3: Rewriting terms.

The end-to-end service bgf of flow 1 (cf. Theorem 4) for the tandem network where server h has been removed is:

$$F_{S_{e2e(-h)}}(\theta, z) = e^{\theta(\sum_{i \notin H \cup \{1\}} \sigma_{A_i} + \sum_{j \neq h} \sigma_{S_j})} \prod_{j \neq h} \frac{1}{1 - e^{-\theta \rho'_j} z},$$

where H is the set of flows crossing server h only. As a consequence,

$$F_{S_{e2e}^{(-h)}}(\theta, e^{\theta \rho_{A_1}}) = e^{\theta(\sum_{i \notin H \cup \{1\}} \sigma_{A_i} + \sum_{j \neq h} \sigma_{S_j})} \prod_{j \neq h} \frac{1}{1 - e^{-\theta \rho_j}}.$$

Finally, including $F_{S_{e2e}^{(-h)}}(\theta, e^{\theta \rho_{A_1}})$ in the formulation leads to the desired result:

$$\mathbf{P}(q(t) \geq b) \leq \frac{\xi_{A_{[h]},S_h} e^{-\theta \sigma_{A_1}}}{e^{\theta(\sum_{i \in \mathrm{Fl}(h) \setminus H} \sigma_{A_i})}} F_{S_{e2e}^{(-h)}}(\theta, e^{\theta \rho_{A_1}}) e^{-\theta b}.$$

6.2 Delay

From Eq. (13), the inequality $t_1 \leq t_{n+1} - T$ is equivalent to $\tau \geq t_{h+1} - t_{n+1} - k_1^{h-1} + T$ and one can write

$$d(t_{n+1} - T + 1) \geq \Rightarrow \exists k_1, \ldots, k_{h-1} \geq 0,\ t_{h+1} \leq \cdots \leq t_{n+1},$$

$$\sup_{\tau \geq [t_{h+1}-t_{n+1}+T-k_1^{h-1}]_+} \Big(A_1(t_{h+1}-k_1^{h-1}-\tau, t_{n+1}-T+1) + \sum_{i\in \mathrm{Fl}(h), i\neq 1} A_i(t_{h+1}-k_{f_i}^{h-1}-\tau, t_{\ell_i+1}) - S_h(t_{h+1}-\tau, t_{h+1})\Big) \geq \sum_{j<h-1} C_j k_j + \sum_{j>h+1} S_j(t_j, t_{j+1}) - \sum_{i\notin \mathrm{Fl}(h)} A_i(t_{f_i}, t_{\ell_i+1}).$$

Step 1: Application of Theorem 5.

With $t' = [t_{h+1} - t_{n+1} + T - k_1^{h-1}]_+$ and τ replaced by $\tau' + t'$, let us define for $K = (k_1, \dots, k_{h-1}, t_{h+1}, \dots, t_{n+1}) \in \mathcal{K}$ and $\tau' \geq 0$,

$$W_{\tau'}^K = A_1(t_{h+1} - k_1^{h-1} - t' - \tau', t_{n+1} - T + 1) + \sum_{i\in \mathrm{Fl}(h)\setminus\{1\}} A_i(t_{h+1} - k_{f_i}^{h-1} - t' - \tau', t_{\ell_i+1}) - S_h(t_{h+1} - t' - \tau', t_{h+1}),$$

and $Y^K = \sum_{j<h-1} C_j k_j + \sum_{j>h+1} S_j(t_j, t_{j+1}) - \sum_{i\notin \mathrm{Fl}(h)} A_i(t_{f_i}, t_{\ell_i+1})$.

The random variable Y^K and the process $(W_{\tau'}^K)_{\tau'\geq 0}$ are independent, so by applying Theorem 5, for all θ such that $\rho_h(\theta) = \rho_{S_h}(\theta) - \sum_{i\in\mathrm{Fl}(h)} \rho_{A_i}(\theta) \geq 0$, there exists $\xi_{A_{[h]},S_h}(\theta)$ such that

$$\mathbf{P}(\sup_{\tau'\geq 0} W_{\tau'}^K \geq Y^K) \leq \xi_{A_{[h]},S_h} \mathbf{E}[e^{-\theta Y^K}] \, e^{\theta(\rho_{A_1}(t_{n+1}-t_{h+1}+k_1^{h-1}+t'-T+1)+\sum_{i\in\mathrm{Fl}(h)\setminus\{1\}} \rho_{A_i}(t_{\ell_{i+1}}-t_{h+1}+k_{f_i}^{h-1}+t')-\rho_{S_h} t')}.$$

If $t' = 0$, this gives

$$\mathbf{P}(\sup_{\tau'\geq 0} W_{\tau'}^K \geq Y^K) \leq \xi_{A_{[h]},S_h}(\theta) \mathbf{E}[e^{-\theta Y^K}] \, e^{\theta(\rho_{A_1}(t_{n+1}-t_{h+1}+k_1^{h-1}-T+1)+\sum_{i\in\mathrm{Fl}(h)\setminus\{1\}} \rho_{A_i}(t_{\ell_{i+1}}-t_{h+1}+k_{f_i}^{h-1}))},$$

and if $t' = t_{h+1} - t_{n+1} + T - k_1^{h-1}$, then

$$\mathbf{P}(\sup_{\tau' \geq 0} W_{\tau'}^K \geq Y^K) \leq \xi_{A_{[h]},S_h}(\theta)\mathbf{E}[e^{-\theta Y^K}]e^{\theta \rho_{A_1}(\theta)}$$
$$e^{\theta(\sum_{i \in \mathrm{Fl}(h)\setminus\{1\}} \rho_{A_i}(t_{\ell_{i+1}} - t_{n+1} + T - k_1^{f_i - 1}) - \rho_{S_h}(t_{h+1} - t_{n+1} + T - k_1^{h-1}))}.$$

Now, one can express the v.p. of the delay and separate the terms of the union bound into two groups:

$$\mathbf{P}(d(t_{n+1} - T + 1) \geq T) \leq \mathbf{P}(\exists K \in \mathcal{K},\ \sup_{\tau' \geq 0} W_{\tau'}^K \geq Y^K) \leq \sum_{K \in \mathcal{K}} \mathbf{P}(\sup_{\tau' \geq 0} W_{\tau'}^K \geq Y^K)$$
$$\leq \sum_{K \in \mathcal{K}, t'=0} \mathbf{P}(\sup_{\tau' \geq 0} W_{\tau'}^K \geq Y^K) + \sum_{K \in \mathcal{K}, t'>0} \mathbf{P}(\sup_{\tau' \geq 0} W_{\tau'}^K \geq Y^K). \tag{14}$$

***Step 2a: Union bound** [$t' = 0$].*

Let us first give a bound on the left-hand sum term of Eq. (14). Having $t' = 0$ is equivalent to $t_{n+1} - t_{h+1} + k_1^{h-1} - T \geq 0$. In line (15), we sum over all possible values u; in line (16), we regroup the terms per servers and use $k_j = t_{j+1} - t_j$; and in line (17), we recognize the term of the end-to-end service bgf using Eq. (4). For all $\theta_1 \in \Theta$,

$$P_1(\theta_1) = \sum_{K \in \mathcal{K}, t_{n+1} - t_{h+1} + k_1^{h-1} - T \geq 0} \mathbf{P}(W_\tau^K \geq Y^K)$$
$$\leq \sum_{u \geq 0} \xi_{A_{[h]},S_h} e^{\theta_1(\rho_{A_1}(u+1) + \sigma_{A_{[-h]}} + \sigma_{S_{>h}})} \tag{15}$$
$$\cdot \sum_{K \in \mathcal{K}, t_{n+1} - t_{h+1} + k_1^{h-1} - T = u} e^{\theta_1(\sum_{i=2}^m \rho_{A_i}(t_{\ell_i + 1} - t_{f_i \vee h+1} - k_{f_i}^{h-1}) - \sum_{j<h-1}^n C_j k_j - \sum_{j>h} \rho_{S_j}(t_{j+1} - t_j))}$$
$$= \sum_{u \geq 0} \xi_{A_{[h]},S_h} e^{\theta_1(\rho_{A_1}(u+1) + \sigma_{A_{[-h]}} + \sigma_{S_{>h}})} \cdot \sum_{\sum_{j \neq h} k_j = u+T} \prod_{j \neq h} e^{\theta_1 \rho'_j k_j} \tag{16}$$
$$\leq \sum_{u \geq 0} \xi_{A_{[h]},S_h} e^{\theta_1(\rho_{A_1}(u+1) - \sum_{i \in \mathrm{Fl}(h)\setminus(H \cup \{1\})} \sigma_{A_i})}[z^{u+T}]F_{S_{e2e}}^{(-h)}(\theta_1, z). \tag{17}$$

***Step 3a: Rewriting terms** [$t' = 0$].*

One can now notice that

$$\frac{e^{\theta_1 \sum_{i\in \mathrm{Fl}(h)\setminus H} \sigma_{A_i}}}{\xi_{A_{[h]},S_h}} P_1(\theta_1) \leq \sum_{u\geq 0} e^{\theta_1(\sigma_{A_1}+\rho_{A_1}(u+1))}[z^{u+T}]F^{(-h)}_{S_{e2e}}(\theta_1, z),$$

and we recognize the right-hand side as the T-th term of the delay bgf for the arrival process A_1 and dynamic $S^{(-h)}_{e2e}$-server according to (Bouillard et al, 2022, Eq.(11)), so

$$\forall \theta_1 \in \Theta, \quad P_1(\theta_1) \leq \frac{\xi_{A_{[h]},S_h}(\theta_1)}{e^{\theta_1 \sum_{i\in \mathrm{Fl}(h)\setminus H} \sigma_{A_i}(\theta_1)}}[z^T]F_{d,S^{(-h)}_{e2e}}(\theta_1, z).$$

***Step 2b: Union bound** [$t' > 0$].*

In the right-hand sum term of Eq. (14), if $t' > 0$, then $t_{n+1} - t_{h+1} + k_1^{h-1} < T$, and the only possible values for t' are $\{1, \dots, T\}$. Rewriting the third line with $k_h = t'$ and $k_j = t_{j+1} - t_j$ for $j > h$, and $C_j = \rho_j(\theta)$, we recognize the expression of the product of geometric series of Eq. (4) and for all $\theta_2 \in [0, \theta_h^*]$,

$$\begin{aligned}
P_2(\theta_2) &= \sum_{K\in\mathcal{K}, t'>0} \mathbf{P}(\sup_{\tau\geq 0} W_\tau^K \geq Y^K) \\
&\leq \sum_{\substack{t_{n+1}-t_{h+1}+k_1^{h-1}+t'=T \\ (k_1,\dots,t_{n+1})\in\mathcal{K}}} \xi_{A_{[h]},S_h} e^{\theta_2(\sum_{i\in \mathrm{Fl}(h)\setminus\{1\}} \rho_{A_i}(t_{\ell_i+1}-t_{h+1}+k_{f_i}^{h-1}+t')-\rho_{S_h}t')} \\
&\quad e^{\theta_2(\rho_{A_1}+\sigma_{A_{[-h]}}+\sigma_{S_{>h}})} e^{\theta_2(\sum_{i\notin \mathrm{Fl}(h)} \rho_{A_i}(t_{\ell_i+1}-t_{f_i})-\sum_{j<h} C_j k_j - \sum_{j>h}\rho_{S_j}(t_{j+1}-t_j))} \\
&= \xi_{A_{[h]},S_h} e^{\theta_2(\rho_{A_1}+\sigma_{A_{-[h]}}+\sigma_{S_{>h}})} \sum_{\substack{k_j\geq 0, k_h>0 \\ \sum_j k_j = T}} e^{-\theta_2(\sum_{j=1}^n \rho'_j k_j)} \\
&\leq \xi_{A_{[h]},S_h} e^{\theta_2(\rho_{A_1}+\sigma_{A_{[-h]}}+\sigma_{S_{>h}})} [z^{T-1}] \prod_{j=1}^n \frac{e^{-\theta_2\rho'_h}}{1-e^{\theta_2\rho'_j}z}.
\end{aligned}$$

***Step 3b: Rewriting terms** [$t' > 0$]**.*

The bounding generating function for the end-to-end service curve for flow 1 is

$$F_{S_{e2e}}(\theta_2, z) = e^{\theta_2(\sum_{i=2}^{n} \sigma_{A_i} + \sum_{j\geq h} \sigma_{S_j})} \prod_{j=1}^{n} \frac{1}{1 - e^{\theta_2 \rho'_j} z},$$

so $P_2(\theta_2) \leq \frac{\xi_{A_{[h]},S} e^{-\theta_2(\rho_{S_h} - \sum_{i\in \text{Fl}(h)} \rho_{A_i})}}{e^{\theta_2(\sigma_{S_h} + \sum_{i\in \text{Fl}(h)\setminus\{1\}} \sigma_{A_i})}} [z^{T-1}] F_{S_{e2e}}(\theta_2, z)$. Noticing that $[z^{T-1}]f(z) = [z^T]zf(z)$ and summing the two bounds concludes the proof.

7 Conclusion

In this paper, we have presented a method to compute probabilistic end-to-end performance bounds in networks that combines two types of analysis from SNC: we locally use the martingale analysis at one server and use the more classical MGF-based SNC results. In particular, this can avoid the use of the union bound at the bottleneck of the network and simulations show that, for small networks, the gap between simulation and best results from the state of the art is drastically reduced. Of course, the improvement may vanish when considering larger networks, and further investigations are needed to achieve tight bounds. The challenge is to bound the maximum of random variables that cannot be, or have not yet been, expressed as a martingale. Another research direction is to investigate other service policies. Indeed, the MGF-based SNC has until now mainly focused on blind multiplexing, which encompasses all the possible service policies, hence the more pessimistic for the flow of interest.

8 Conflict of Interest

The authors have no conflict of interest to declare that are relevant to this article.

References

Angrishi K, Killat U (2011) An approach using Demisubmartingales for the stochastic analysis of networks. CoRR abs/1111.3063. 1111.3063

Beck M (2016) Advances in Theory and Applicability of Stochastic Network Calculus. PhD Thesis of the University of Kaiserslautern

Bouillard A (2024) Stochastic network calculus with localized application of martingales. CoRR abs/2211.05657. 2211.05657

Bouillard A, Nowak T (2015) Fast symbolic computation of the worst-case delay in tandem networks and applications. Perform Eval 91:270–285. https://doi.org/10.1016/j.peva.2015.06.016

Bouillard A, Gaujal B, Lagrange S, et al (2008) Optimal routing for end-to-end guarantees using Network Calculus. Performance Evaluation 65(11-12):883–906. https://doi.org/10.1016/j.peva.2008.04.008

Bouillard A, Boyer M, Le Corronc E (2018) Deterministic Network Calculus: From Theory to Practical Implementation. ISTE

Bouillard A, Nikolaus P, Schmitt JB (2022) Unleashing the power of paying multiplexing only once in stochastic network calculus. Proc ACM Meas Anal Comput Syst 6(2):31:1–31:27. https://doi.org/10.1145/3530897

Chang CS (1994) Stability, queue length, and delay of deterministic and stochastic queueing networks. IEEE Transactions on Automatic Control 39(5):913–931. https://doi.org/10.1109/9.284868

Chang CS (2000) Performance Guarantees in Communication Networks. TNCS, Springer-Verlag

Ciucu F, Poloczek F (2019) Two Extensions of Kingman's GI/G/1 Bound. In: Nahum EM, Bonald T, Duffield N (eds) Abstracts of the 2019 SIGMETRICS/Performance Joint International Conference on Measurement and Modeling of Computer Systems. ACM, pp 63–64, https://doi.org/10.1145/3309697.3331494

Ciucu F, Burchard A, Liebeherr J (2006) Scaling properties of statistical end-to-end bounds in the network calculus. IEEE/ACM Transactions on Networking (ToN) 14(6):2300–2312

Ciucu F, Poloczek F, Chen LY, et al (2021) Practical analysis of replication-based systems. In: 40th IEEE Conference on Computer Communications, INFOCOM 2021, pp 1–10, https://doi.org/10.1109/INFOCOM42981.2021.9488829

Cruz RL (1995) Quality of Service Guarantees in Virtual Circuit Switched Networks. IEEE Journal on selected areas in communication 13:1048–1056. https://doi.org/10.1109/49.400660

Duffield NG (1994) Exponential bounds for queues with Markovian arrivals. Queueing Syst Theory Appl 17(3-4):413–430. https://doi.org/10.1007/BF01158702

Fidler M (2006) An End-to-End Probabilistic Network Calculus with Moment Generating Functions. In: 14th IEEE International Workshop on Quality of Service, pp 261–270, https://doi.org/10.1109/IWQOS.2006.250477

Fidler M, Rizk A (2015) A guide to the stochastic network calculus. IEEE Communications Surveys Tutorials 17(1):92–105. https://doi.org/10.1109/COMST.2014.2337060

Kingman JFC (1964) A martingale inequality in the theory of queues. Mathematical Proceedings of the Cambridge Philosophical Society 60(2):359?361. https://doi.org/10.1017/S0305004100037841

Le Boudec JY, Thiran P (2001) Network Calculus: A Theory of Deterministic Queuing Systems for the Internet, vol LNCS 2050. Springer-Verlag, https://doi.org/10.1007/3-540-45318-0, revised version 4, May 10, 2004

Liu Y, Jiang Y (2008) Stochastic Network Calculus. Springer

Nikolaus P, Schmitt J (2017) On Per-Flow Delay Bounds in Tandem Queues under (In)Dependent Arrivals. In: Proc. 16th IFIP Networking Conference (NETWORKING'17), Stockholm, Sweden, pp 1–9

Nikolaus P, Schmitt J (2020) Improving delay bounds in the stochastic network calculus by using less stochastic inequalities. In: Proc. 13th EAI International Conference on Performance Evaluation Methodologies and Tools (VALUETOOLS 2020). Association for Computing Machinery, New York, NY, USA, VALUETOOLS '20, p 96?103, https://doi.org/10.1145/3388831.3388848

Poloczek F, Ciucu F (2014) Scheduling analysis with martingales. Performance Evaluation 79:56 – 72. https://doi.org/https://doi.org/10.1016/j.peva.2014.07.004, special Issue: Performance 2014

Poloczek F, Ciucu F (2015) Service-martingales: Theory and applications to the delay analysis of random access protocols. In: 2015 IEEE Conference on Computer Communications, INFOCOM 2015, pp 945–953, https://doi.org/10.1109/INFOCOM.2015.7218466

Rao BLSP (2007) On some maximal inequalities for demisubmartingales and N-demisuper martingales. Journal of Inequalities in Pure & Applied Mathematics 8

Rizk A, Fidler M (2013) On multiplexing models for independent traffic flows in single- and multi-node networks. IEEE Trans Netw Serv Manag 10(1):15–28. https://doi.

org/10.1109/TNSM.2012.091012.120234

Schmitt JB, Zdarsky FA, Martinovic I (2008) Improving Performance Bounds in Feed-Forward Networks by Paying Multiplexing Only Once. In: Bause F, Buchholz P (eds) Proceedings 14th GI/ITG Conference on Measurement, Modelling and Evaluation of Computer and Communication Systems (MMB 2008), March 31 - April 2, 2008, Dortmund, Germany. VDE Verlag, pp 13–28

Sedgewick R, Flajolet P (1996) An Introduction to the Analysis of Algorithms. Addison-Wesley Longman Publishing Co., Inc., USA

Yaron O, Sidi M (1993) Performance and stability of communication networks via robust exponential bounds. IEEE/ACM Trans Netw 1(3):372–385. https://doi.org/10.1109/90.234858

Yu B, Chi X, Liu X (2022) Martingale-based bandwidth abstraction and slice instantiation under the end-to-end latency-bounded reliability constraint. IEEE Communications Letters 26(1):217–221. https://doi.org/10.1109/LCOMM.2021.3122305