Network Calculus Delay Bounds in Queueing Networks with Exact Solutions

Florin Ciucu

Department of Computer Science, University of Virginia, U.S.A.
florin@cs.virginia.edu

Abstract. The purpose of this paper is to shed light on the accuracy of probabilistic delay bounds obtained with network calculus. In particular, by comparing calculus bounds with exact results in a series of M/M/1 queues with cross traffic, we show that reasonably accurate bounds are achieved when the percentage of cross traffic is low. We use recent results in network calculus and, in addition, propose novel bounds based on Doob's maximal inequality for supermartingales. In the case of single M/M/1 and M/D/1 queues, our results improve existing bounds by $\Omega\left(\frac{\log(1-\rho)^{-1}}{1-\rho}\right)$ when the utilization factor ρ converges to one. [1]

1 Introduction

Stochastic network calculus is an extension of deterministic network calculus [1–3] for analyzing network performance in terms of probabilistic backlog and delay bounds. Compared to its deterministic counterpart, the advantage of stochastic network calculus is that it can account for statistical multiplexing [4–7]. In addition, the calculus can be applied to a wide class of traffic models including deterministically regulated, Markov modulated or fractional Brownian motion [8, 9, 7]. The 'pay-bursts-only-once' property [3] observed in deterministic network calculus holds in a probabilistic setting as well [10–12].

One of the main concerns in analyzing networks with performance bounds is whether the bounds are accurate enough to be applied to practical problems. As far as network calculus bounds are concerned, there are several approaches to estimate the bounds' accuracy. For example, the authors of [4, 8, 13] use simulation results as benchmarks for calculus bounds. The admissible region of connections requiring some performance guarantees is compared with two corresponding regions: the region obtained from simulations [5, 6], and the maximal possible region based on an average rate admission control [6, 7]. Asymptotic properties of end-to-end bounds are established in both networks where arrivals and service at the nodes are either statistically independent [11], or subject to correlations [10].

[1] Adopting Landau notation for two sequences f_n and g_n, we say that $f_n \in \mathcal{O}(g_n)$ and $f_n \in \Omega(g_n)$ if the fractions f_n/g_n and g_n/f_n, respectively, are bounded. Also, $f_n \in \Theta(g_n)$ if both $f_n \in \mathcal{O}(g_n)$ and $f_n \in \Omega(g_n)$.

In this paper we take a different approach to estimate the accuracy of network calculus bounds. We apply the calculus to the derivation of end-to-end delay bounds in a network of M/M/1 queues in series, with cross traffic at each queue. We then compare the obtained bounds with exact results that are readily available in M/M/1 queueing networks [14]. This comparative study leads to accurate estimations of the bounds' behavior, yet it dispenses with computationally expensive simulations. Moreover, the presented network calculus methodology to analyze M/M/1 queueing networks can be extended to more general queueing networks where exact results are usually not available.

Applying the calculus in scenarios specific to queueing network theory contributes to an understanding of some of the complementary features between the two analytical tools. For instance, queueing networks analysis applies to a small class of scheduling algorithms (of which we only consider FIFO), whereas network calculus applies to a broader scheduling class. We derive calculus bounds for static priority (SP) scheduling assuming higher priority for cross traffic. When the percentage of cross traffic is low, we show that the obtained bounds are reasonably accurate; however, when the cross traffic dominates the traffic across the network, then the bounds may degrade significantly.

Another complementary aspect between queueing networks and network calculus is that the former requires statistical independent arrivals, whereas the latter considers both independent and correlated arrivals. By accounting for the independence of arrivals in network calculus, we show that much smaller bounds can be achieved than those holding for correlated arrivals. This indicates that the independence of arrivals may play a significant role in network calculus for practical purposes. We mention that queueing networks and network calculus have been related before in [15] where the effects of traffic shaping on queueing networks analysis are considered.

In our derivations we use recent results in network calculus, and also propose novel bounds for the special class of Lévy processes. For the first time in the context of network calculus, where service is expressed with service curves, we invoke Doob's maximal inequality for supermartingales to estimate sample path bounds. Estimating sample path bounds is a difficult problem in network calculus [7], and existing solutions generally rely on approximations using extreme value theory [13, 6], or the derivation of bounding sums with Boole's inequality [8, 2]. By using Doob's inequality we can recover exact results in the M/M/1 queue; the bounds obtained in the M/D/1 queue numerically match the corresponding exact results. Moreover, our bounds improve those obtained with Boole's inequality by $\Omega\left(\frac{\log(1-\rho)^{-1}}{1-\rho}\right)$ when the utilization factor ρ converges to one.

We structure the rest of the paper as follows. In Section 2 we derive performance bounds in a network calculus with effective bandwidth and a formulation of a statistical service curve that generalizes several existing definitions. In Section 3 we improve these bounds by exploiting the special properties of Lévy processes. In Section 4 we apply the network calculus bounds from Sections 2 and 3 to queueing networks with exact solutions, and show numerical comparisons. Finally, in Section 5, we present brief conclusions.

2 Performance Bounds

We consider a discrete time domain with discretization step $\tau_0 = 1$. The cumulative arrivals and departures at a node are modelled with nondecreasing processes $A(t)$ and $D(t)$, where $A(0) = D(0) = 0$. We denote for convenience $A(s,t) = A(t) - A(s)$. The corresponding delay process at the node is denoted by $W(t) = \inf\{d : A(t-d) \leq D(t)\}$.

We assume that the moment generating function of the arrivals is bounded for all $s \leq t$ and some $\theta > 0$ by

$$E\left[e^{\theta A(s,t)}\right] \leq e^{\theta \rho_a(\theta)(t-s)} . \tag{1}$$

The quantity $\rho_a(\theta)$ is called *effective bandwidth* [16] and varies between the average and peak rate of the arrivals. Effective bandwidths can be obtained for a wide class of arrivals [16], or traffic descriptions with *effective envelopes* [7].

In network calculus, the service at a node is usually expressed with *service curves* that are functions specifying lower bounds on the amount of service received. We now introduce a service curve formulation that generalizes several existing formulations. This service curve is particularly useful in network scenarios where services at the nodes are either statistically independent or correlated; moreover, the service curve can account for the benefits of (partial) statistical independence. Let us first define the *convolution* operator for two doubly indexed processes f and g as $f * g(u,t) = \inf_{u \leq s \leq t} \{f(u,s) + g(s,t)\}$; also, we denote $f * g(t) := f * g(0,t)$.

We say that a nonnegative, doubly indexed random process $\mathcal{S}(s,t)$ is a *statistical service curve* if for all $t, \sigma \geq 0$

$$Pr\left(D(t) < A * [\mathcal{S} - \sigma]_+ (t)\right) \leq \varepsilon(\sigma) , \tag{2}$$

where we denoted $[x]_+ = \sup\{x, 0\}$. The nonnegative, nonincreasing function $\varepsilon(\sigma)$ is referred to as the *error* function. When $\varepsilon = 0$ then Eq. (2) recovers a service curve from [2, 11][2]; if further $\mathcal{S}(s,t)$ is non-random and stationary (depending on $t - s$, and invariant of s or t alone), then $\mathcal{S}$ is a deterministic service curve [2]. If $\mathcal{S}(s,t)$ is non-random and stationary then Eq. (2) recovers a definition from [10]. In the most general form, $\mathcal{S}(s,t)$ is random and $\varepsilon \geq 0$.

Similar to the condition on the arrivals from Eq. (1), we assume that the Laplace transform of service curves is bounded [11] for some $\theta > 0$ by

$$E\left[e^{-\theta \mathcal{S}(s,t)}\right] \leq M(\theta) e^{-\theta \rho_s(\theta)(t-s)} . \tag{3}$$

The next theorem provides delay bounds at a node where the service is given with statistical service curves. The presented bounds generalize the bounds obtained in [11] for the special case when $\varepsilon = 0$.

---
[2] We say that $\varepsilon = 0$ whenever $\varepsilon(\sigma) = 0$ for all σ.

Theorem 1. (Delay Bounds) *Consider a network node offering a statistical service curve $\mathcal{S}(s,t)$ with error function $\varepsilon(\sigma)$ to an arrival process $A(t)$. Assume that $A(t)$ and $\mathcal{S}(s,t)$ are statistically independent, and are bounded according to Eqs. (1), (3) with parameters $\rho_a(\theta)$, $M(\theta)$, $\rho_s(\theta)$, for some $\theta > 0$. If $\rho(\theta) = \rho_s(\theta) - \rho_a(\theta) > 0$, then a delay bound is given for all discrete $t, d \geq 0$ by*

$$Pr\Big(W(t) > d\Big) \leq \inf_{\sigma} \left\{ M(\theta) \frac{e^{-\theta\rho_s(\theta)d}}{\theta\rho(\theta)} e^{\theta\sigma} + \varepsilon(\sigma) \right\} \ . \tag{4}$$

Proof. In the first part of the proof we separate the estimation of the delay bound into a service curve bound and a sample path bound. The latter is estimated in the second part of the proof.

Fix σ and some discrete times t, d. Assume that for a particular sample path the following inequality holds

$$D(t) \geq A * [\mathcal{S} - \sigma]_+ (t) \ , \tag{5}$$

such that we can write

$$W(t) > d \ \Rightarrow \ A(t-d) > D(t) \ \Rightarrow \ A(t-d) > A * [\mathcal{S} - \sigma]_+ (t) \ .$$

It follows that

$$\begin{aligned} Pr\left(W(t) > d\right) &\leq Pr\Big(A(t-d) > A * [\mathcal{S} - \sigma]_+ (t)\Big) + Pr\Big(\text{Eq. (5) fails}\Big) \\ &\leq Pr\left(\sup_{0 \leq s < t-d} \{A(s, t-d) - \mathcal{S}(s,t)\} > -\sigma \right) + \varepsilon(\sigma) \ . \end{aligned} \tag{6}$$

We remark that the points $s = t-d, \ldots, t$ do not contribute to the supremum in Eq. (6) (due to the positivity constraint). Next, to estimate the sample path bound in Eq. (6), we apply Boole's inequality and the Chernoff bound

$$\begin{aligned} &Pr\left(\sup_{0 \leq s < t-d} \{A(s, t-d) - \mathcal{S}(s,t)\} > -\sigma \right) \leq \sum_{s=0}^{t-d-1} E\left[e^{\theta(A(s,t-d) - \mathcal{S}(s,t))} \right] e^{\theta\sigma} \\ &\leq M(\theta) \sum_{s=0}^{t-d-1} e^{\theta\rho_a(\theta)(t-d-s)} e^{-\theta\rho_s(\theta)(t-s)} e^{\theta\sigma} \leq M(\theta) \frac{e^{-\theta\rho_s(\theta)d}}{\theta\rho(\theta)} e^{\theta\sigma} \ . \end{aligned} \tag{7}$$

In Eq. (7) we first used the independence of A and $\mathcal{S}$. Then we substituted the bounds from Eqs. (1) and (3), and finally applied the inequality $\sum_{s \geq 1} e^{-as} \leq 1/a$, for $a > 0$. The proof is completed by minimizing over σ. □

Consider now a flow along a network path with H nodes. Assume that the service given to the flow at each node is expressed by a statistical service curve $\mathcal{S}^h(s,t)$ with error function $\varepsilon^h(\sigma)$. Then, the service given to the flow by the network as a whole can be expressed using a *statistical network service curve*, such that end-to-end performance bounds can be derived using single node bounds. If $\varepsilon^h = 0$, then the network service curve is given by $\mathcal{S}^{net} = \mathcal{S}^1 * \ldots * \mathcal{S}^H$ [11]. A similar expression can be constructed in the case when $\varepsilon^h \geq 0$ [10].

For the rest of the section we show how to construct *leftover* service curves for a tagged flow at a node, in terms of the capacity left unused by the remaining flows. Consider a workconserving network node operating at a constant rate R. We denote by $A(t)$ a tagged flow (or aggregate of flows) at the node, and by $A_c(t)$ the aggregate of the remaining flows; $A_c(t)$ is referred to as cross traffic. We assume SP scheduling with $A(t)$ getting lower priority.

If $A(t)$ and $A_c(t)$ are statistically independent, then a leftover service curve is given by

$$\mathcal{S}(s,t) = R(t-s) - A_c(s,t) \ , \tag{8}$$

with error function $\varepsilon = 0$ [11]. Assume now that $A(t)$ and $A_c(t)$ are not necessarily independent, and that $A_c(t)$ is bounded according to Eq. (1) with parameter $\rho_c(\theta) < R$, for some $\theta > 0$. Then for any choice of $\delta > 0$, a leftover service curve is given by

$$\mathcal{S}(t) = (R - \rho_c(\theta) - \delta)\, t \ \text{ with } \ \varepsilon(\sigma) = \frac{1}{\theta\delta} e^{-\theta\sigma} \ . \tag{9}$$

The proof of Eq. (9) proceeds similarly as the proof of Theorem 3 in [10] and is omitted here (the main difference is that here we use effective bandwidth to describe arrivals, whereas [10] uses *statistical envelopes*).

Although leftover service curves give a worst case view on the per-flow service, they have the advantage of leading to simple, closed-form expressions for the performance bounds of interest. Tighter per-flow service curves can be derived for GPS or EDF schedulers, but their notation increases and the differences with SP service curves at a single-node are small [7].

3 Performance Bounds for Lévy Processes

In this section we assume that the arrivals and service curves are Lévy processes. Using the special properties of Lévy processes, i.e., independent and stationary increments, we show that we can improve the performance bounds obtained in the previous section. We discretize Lévy processes (that are defined in a continuous time domain) with discretization step $\tau_0 = 1$.

Theorem 2. (Delay Bounds for Lévy processes) *Consider the hypothesis from Theorem 1. In addition, assume that $A(t)$ and $\mathcal{S}(s,t)$ are Lévy processes and that the following condition holds*

$$M(\theta)e^{-\theta\rho(\theta)} \leq 1 \ . \tag{10}$$

Then, a statistical delay bound is given for all discrete $t, d \geq 0$ by

$$Pr\Big(W(t) > d\Big) \leq \inf_{\sigma} \Big\{ M(\theta) e^{-\theta\rho(\theta)} e^{-\theta\rho_s(\theta) d} e^{\theta\sigma} + \varepsilon(\sigma) \Big\} \ . \tag{11}$$

The delay bounds obtained in Theorem 2 are smaller than those obtained in Theorem 1; the reason is that $e^{-\theta\rho(\theta)} < (\theta\rho(\theta))^{-1}$, for all $\theta > 0$. Note that the difference becomes large when $\theta\rho(\theta) \to 0$ (i.e. at very high utilizations).

The proof's main idea is to estimate sample path bounds using Doob's maximal inequality for supermartingales. This technique is applied in a classic note by Kingman [17] to the derivation of exponential backlog bounds in GI/GI/1 queues. Since Kingman's note, several works use related supermartingales techniques to derive exponential bounds (e.g. in queueing systems with Markovian arrivals [18, 19], or in stochastic linear systems under the (max,+) algebra [20]). Here we integrate the technique with supermartingales in network calculus, where service is expressed with service curves. Using the properties of service curves, supermartingales can then be directly applied to analyze many scheduling algorithms and multi-node networks.

Proof. We adopt the first part of the proof of Theorem 1. The rest of the proof estimates the sample path bound from Eq. (6) by first constructing a supermartingale, and then invoking Doob's maximal inequality.

Fix t, d, σ. For positive s with $s \leq t - d$ we construct the process

$$T(s) = e^{\theta(A(t-d-s,t-d)-\mathcal{S}(t-d-s,t))},$$

with the associated $\sigma-$algebras $\mathcal{F}_s$ generated by $A(t-d-s, t-d)$ and $\mathcal{S}(t-d-s, t)$. We can write

$$\begin{aligned} E\left[T(s+1) \parallel \mathcal{F}_s\right] &= E\left[T(s)e^{\theta(A(t-d-s-1,t-d-s)-\mathcal{S}(t-d-s-1,t-d-s))} \parallel \mathcal{F}_s\right] \\ &= T(s)E\left[e^{\theta(A(1)-\mathcal{S}(1))}\right] \leq T(s)M(\theta)e^{-\theta\rho(\theta)} \leq T(s) \ . \end{aligned} \tag{12}$$

In Eq. (12) we first used the fact that A and $\mathcal{S}$ are independent Lévy processes, then we substituted the bounds from Eqs. (1) and (3), and finally we used the condition from Eq. (10).

From Eq. (12) we obtain that $T(1),\ T(2), \ldots,\ T(t-d)$ form a supermartingale. We can now estimate the sample path bound from Eq. (6) as follows

$$\begin{aligned} &Pr\left(\sup_{0\leq s<t-d}\{A(s,t-d)-\mathcal{S}(s,t)\} > -\sigma\right) \leq Pr\left(\sup_s T(s) > e^{-\theta\sigma}\right) \\ &\leq E\left[T(1)\right]e^{\theta\sigma} \leq M(\theta)e^{-\theta\rho(\theta)}e^{-\theta\rho_s(\theta)d}e^{\theta\sigma} \end{aligned} \tag{13}$$

In Eq. (13) we first invoked Doob's inequality (see [17]) for the supermartingale $T(s)$, and the rest follows as in Eq. (12). The proof is completed by minimizing over σ. □

Finally we show how to exploit the properties of Lévy processes to the construction of leftover service curves. Consider the scenario from the end of Section 2, with a node serving a tagged flow $A(t)$ and some cross traffic $A_c(t)$. If, in addition, the cross traffic $A_c(t)$ is a Lévy process, then a leftover service curve for the tagged flow $A(t)$ is now given by

$$\mathcal{S}(t) = (R - \rho_c(\theta))\,t \ \text{ with } \ \varepsilon(\sigma) = e^{-\theta\sigma} \ . \tag{14}$$

The proof for Eq. (14) can be constructed by invoking Doob's maximal inequality, similarly as in the proof of Theorem 2. Note that the service curve in Eq. (14) is tighter than the one given in Eq. (9); the difference becomes significant when the rate of $A_c(t)$ approaches the rate R.

4 Applications to Queueing Networks with Exact Solutions

In this section we apply network calculus to the derivation of delay bounds in queueing networks. For single M/M/1 and M/D/1 queues we show that by using the special properties of Lévy processes, the derived bounds match the exact results. In the multi-node case we investigate the bounds' behavior depending on factors such as the traffic mix in the network and the statistical independence of arrivals.

We assume that exogenous flows at a node (queue) consist of packets arriving according to a Poisson process $N(t)$ with rate λ. Since a Poisson process is given in a continuous time domain, we discretize time as in Sections 2 and 3 with step $\tau_0 = 1$. Each node serves packets at rate μ and each flow is locally FIFO. For stability, we assume that the utilization factor $\rho = \lambda/\mu$ is less than one. To fit a queueing model with network calculus, we construct the arrival process $A(t) = \sum_{i=1}^{N(t)} X_i$, where X_i represents the service time of the i'th packet [16].

In the single node-case is sufficient to model the service with a deterministic service curve $\mathcal{S}(t) = t$ that induces a fluid view of the service (infinitesimal service unit). However, in the multi-node case the output from a node h may be the input at the next node $h+1$. Consequently, we introduce packetizers [21] to enforce that, for each packet, the starting processing time at node $h+1$ can be no sooner than the completion time at node h. Packetizers can be ignored at the last node [21], hence no packetizer is needed in the single-node case. A packetizer can be described with the service curve $\mathcal{S}(s,t) = \left[t - s - 1 - X^f(t)\right]_+$, where $X^f(t)$ denotes the time already spent in service by the packet currently in service at time t. The substraction of 1 in the expression of $\mathcal{S}(s,t)$ is a consequence of discretizing continuous time processes.

Along with the derivation of bounds, we provide numerical comparisons with exact results for the following setting. Each node has a service rate $R = 100\ Mbps$ and the average size of packets is 400 *Bytes*. We optimize the delay bounds over the parameter τ_0. We show the delays on a milliseconds time scale, and with violation probability $\varepsilon = 10^{-6}$. The numerical comparisons reflect the sensitivity of the bounds to factors such as different network loads or number of nodes.

4.1 Single M/M/1 and M/D/1 Queues

Here we apply network calculus to the analysis of two of the most common queueing models, namely the M/M/1 and M/D/1 queues.

In the M/M/1 queue the service times X_i are exponentially distributed ($X_i \sim exp(\mu)$). The distribution of the steady state delay $W = \lim_{t\to\infty} W(t)$ is given by [14]

$$P\left(W > d\right) = e^{-\mu(1-\rho)d} \ . \tag{15}$$

Next we derive two network calculus delay bounds for the M/M/1 queue. First, the conditions from Eqs. (1) and (3) yield $\rho_a(\theta) = \frac{\lambda}{\mu-\theta}$, $M(\theta) = 1$, and $\rho_s(\theta) = 1$ for all $0 < \theta < \mu$. One delay bound is obtained by plugging in these values into Eq. (4) from Theorem 1 (recall that $\varepsilon(\sigma) = 0$). Moreover, since $A(t)$ is a Lévy process, a second delay bound can be obtained with Eq. (11) from Theorem 2. Remarkably, by choosing $\theta = \mu - \lambda$, the latter delay bound recovers the exact result from Eq. (15). We note that the same bound is obtained by Kingman in [17], but for the waiting time in the *queue*.

For some fixed violation probability ε, let us solve for the ε-quantiles in Eqs. (4) and (11) yielding d_1 and d_2, respectively. Then we have $d_1 - d_2 \geq \frac{1}{\theta} \log \frac{1}{\theta(1-\rho_a(\theta))}$, implying that $d_1 - d_2 \in \Omega\left(\frac{\log(1-\rho)^{-1}}{1-\rho}\right)$ as $\rho \to 1$.

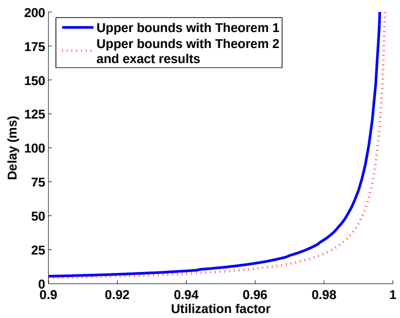
In the M/D/1 queue the service times X_i are constant. The distribution of the steady state delay W is given by [22]

$$P(W > d) = 1 - (1-\rho)e^{\lambda d} \sum_{k=0}^{T} \frac{(k\rho - \lambda d)^k}{k!} e^{-(k-1)\rho} \ , \qquad (16)$$

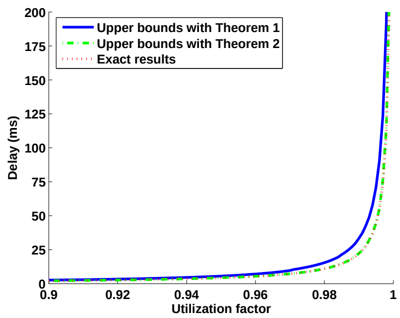
where $T = \lfloor d\mu \rfloor$ denotes the largest integer less than or equal to $d\mu$. This formula poses numerical complications when ρ is close to unity, due to the appearance of large alternating, very nearly cancelling terms (note that the factor $k\rho - \lambda d$ is negative). We evaluate Eq. (16) using a numerical algorithm from [22].

Next we derive delay bounds for the M/D/1 queue with network calculus. The conditions from Eqs. (1) and (3) give $\rho_a(\theta) = \frac{\lambda}{\theta}\left(e^{\frac{\theta}{\mu}} - 1\right)$, $M(\theta) = 1$, and $\rho_s(\theta) = 1$ for all $\theta > 0$ satisfying $\rho_s(\theta) - \rho_a(\theta) > 0$. One delay bound is obtained by plugging in these values into Eq. (4) from Theorem 1. Since $A(t)$ is a Lévy process, a second delay bound is obtained with Eq. (11) from Theorem 2. As shown above, the latter delay bound improves the former by $\Omega\left(\frac{\log(1-\rho)^{-1}}{1-\rho}\right)$.

Figures 1.(a) and (b) show that the bounds obtained with Theorem 2 improve the bounds obtained with Theorem 1 at very high utilizations, as a consequence of accounting for the special properties of Lévy processes. For small to high



(a) M/M/1



(b) M/D/1

Fig. 1. Delay bounds at a node as a function of the utilization factor (node's service rate $R = 100\ Mbps$, average packet size $400\ B$, violation probability $\varepsilon = 10^{-6}$).

utilizations, the bounds closely match and are not depicted. This indicates that the use of Boole's inequality in estimating sample path bounds can lead to conservative bounds, but only at very high utilizations. Figure 1.(b) also shows that, at all utilizations, the M/D/1 delay bounds obtained with Theorem 2 exactly match the exact results from Eq. (16).

4.2 M/M/1 Queues in Series

Now we analyze a network with H nodes arranged in series. A Poisson *through* flow $A(t)$ with rate λ traverses the entire network; moreover, a Poisson *cross* flow $A_h(t)$ with rate λ_c transits each node h, and exits the network thereafter. Each packet has independent and exponentially distributed service times at each traversed node [14]; also, the flows and the service times of packets are assumed independent. The utilization factor is now $\rho = (\lambda + \lambda_c)/\mu$.

This network is an M/M/1 queueing network where exact results are available. In particular, considering FIFO scheduling, the steady-state end-to-end delay W^{net} of the through flow has a Gamma distribution $\Gamma(\mu(1-\rho), H)$ [14]:

$$P(W^{net} > d) = \left(\sum_{k=0}^{H-1} \frac{(\mu(1-\rho)d)^k}{k!}\right) e^{-\mu(1-\rho)d} \ . \tag{17}$$

Next we derive two end-to-end delay bounds for SP scheduling ($A(t)$ gets lower priority) with network calculus. The first one uses the independence of $A(t)$ and $A_h(t)$, and is constructed using techniques from [11]. The second bound is obtained using techniques from [10], that apply for both independent or correlated arrivals. Observe first that the condition from Eq. (1) gives $\rho_a(\theta) = \frac{\lambda}{\mu-\theta}$.

Using the statistical independence of arrivals: Using Eq. (8), a leftover service curve for the through flow at node h is given by $T^h(s,t) = [t-s-A_h(s,t)]_+$. Convolving $T^h(s,t)$ with the service curve corresponding to the packetizer at each node, we obtain that the service at node h is described with the service curve $\mathcal{S}^h(s,t) = \left[t-s-A_h(s,t)-1-X_h^f(t)\right]_+$. Therefore, the service given by the network to the through flow can be expressed with the network service curve $\mathcal{S}^{net}(s,t) = \mathcal{S}^1 * \mathcal{S}^2 * \ldots * \mathcal{S}^H(s,t)$ [11]. Using $E\left[e^{\theta X_h^f(t)}\right] = \frac{\mu}{\mu-\theta}$ and denoting $K = \frac{e^\theta \mu}{\mu-\theta}$, the Laplace transform of $\mathcal{S}^{net}(s,t)$ gives

$$\begin{aligned} E\left[e^{-\theta \mathcal{S}^{net}(s,t)}\right] &\leq \sum_{s=x_0\leq x_1\leq\cdots\leq x_H=t} E\left[e^{-\theta\left(t-s-\sum A_h(x_h,x_{h+1})-H-\sum X_h^f(x_{h+1})\right)}\right] \\ &\leq \binom{t-s+H-1}{H-1} K^H e^{-\theta\left(1-\frac{\lambda_c}{\mu-\theta}\right)(t-s)} \ . \end{aligned} \tag{18}$$

The binomial coefficient is the number of combinations with repetitions. Matching the last equation with Eq. (3) yields $M(\theta) = \binom{t-s+H-1}{H-1}K^H$ and $\rho_s(\theta) = 1 - \frac{\lambda_c}{\mu-\theta}$. Since $M(\theta)$ depends on $t-s$, Theorems 1 and 2 do not apply. However, we can use the proof of Theorem 1 and plug $M(\theta)$ into Eq. (7). Using

$\sum_s \binom{s+H-1}{H-1} a^s = \left(\frac{1}{1-a}\right)^H$ for $0 < a < 1$ [11], $\left(1+\frac{1}{x}\right)^x \le e$ for $x > 0$, and optimizing $\tau_0 = \frac{1}{\theta\rho(\theta)} \log(1+\rho(\theta))$ where $\rho(\theta) = \rho_s(\theta) - \rho_a(\theta) > 0$, we obtain

$$Pr\left(W^{net}(t) > d\right) \le \left(e \frac{\mu}{\mu-\theta} \frac{1+\rho(\theta)}{\rho(\theta)}\right)^H e^{-\theta\rho_s(\theta)d} . \quad (19)$$

Lastly, the parameter θ is optimized numerically.

Without the statistical independence of arrivals: Now we derive delay bounds that hold for both independent and correlated arrivals. Using the Lévy properties of $A_h(t)$, we first get the leftover service curve $\mathcal{T}^h(s,t) = \left(1 - \frac{\lambda_c}{\mu-\theta_c}\right)(t-s)$ with error function $\varepsilon^h(\sigma) = e^{-\theta_c\sigma}$ for some $\theta_c > 0$ (see Eq. (14)). Taking into account the packetizers, the service at node h is given by the service curve $\mathcal{S}^h(s,t) = \left[\left(1 - \frac{\lambda_c}{\mu-\theta_c}\right)(t-s) - 1 - X_h^f(t)\right]$ with error function $\varepsilon^h(\sigma)$. Then, the network service curve [10] is given by $\mathcal{S}^{net}(s,t) = \left[\rho_s(\theta_c)(t-s) - H - \sum Y_h\right]_+$ with error function $\varepsilon^{net}(\sigma) = H\left(\frac{1}{\theta_c\delta}\right)^{\frac{H-1}{H}} e^{-\frac{\theta_c}{H}\sigma}$, where $\delta > 0$, $\rho_s(\theta_c) = 1 - \frac{\lambda_c}{\mu-\theta_c} - (H-1)\delta$ and $Y_h \sim exp(\mu)$. Proceeding as before and optimizing $\tau_0 = \frac{1}{\theta\rho(\theta,\theta_c)} \log(1+\rho(\theta,\theta_c))$ where $\rho(\theta,\theta_c) = \rho_s(\theta_c) - \rho_a(\theta) > 0$, we obtain

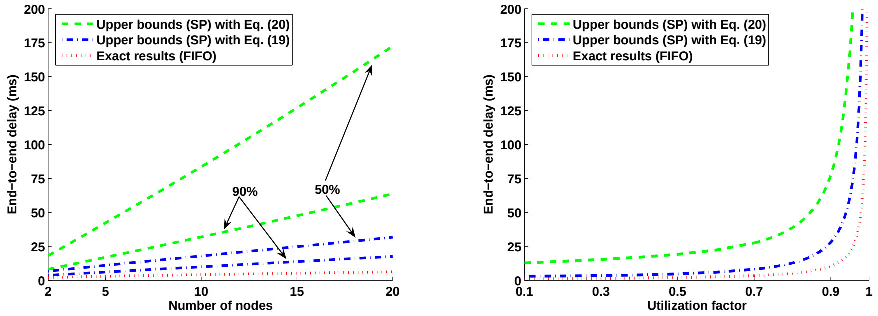
$$Pr\left(W^{net}(t) > d\right) \le \frac{\alpha}{\theta_c}\left(\frac{1}{\delta}\right)^{\frac{H\theta}{\alpha}} \left(e \frac{\mu}{\mu-\theta} \frac{1+\rho(\theta,\theta_c)}{\rho(\theta,\theta_c)}\right)^{\frac{H\theta_c}{\alpha}} e^{-\frac{\theta\theta_c}{\alpha}\rho_s(\theta_c)d} , \quad (20)$$

where $\alpha = H\theta + \theta_c$. The parameter δ can be optimized as in [10]. Lastly, the parameters θ and θ_c are optimized numerically.

Figure 2.(a) illustrates the delay bounds from Eqs. (19), (20) for fixed $\rho = 75\%$, through traffic percentages of 50% and 90%, and different number of nodes H. The bounds approach the exact results from Eq. (17) when the percentage of cross traffic is low (less than 10%), and when accounting for the independence of arrivals (Eq. (19)). Increasing the cross traffic mix leads to more conservative bounds, due to the higher priority given to cross traffic. The decay of the bounds is more visible when dispensing with the independence of arrivals (Eq. (20)). This indicates that the leftover service curves holding for adversarial arrivals give much smaller service than those holding for independent arrivals.

In Figure 2.(b) we illustrate the delay bounds for 10 nodes, 90% through traffic, and variable utilization factor ρ. This figure shows that, at all utilizations, the independence of arrivals leads to much smaller bounds than those holding for adversarial arrivals. Therefore, the independence of arrivals appears to play a significant role in network calculus for practical purposes. From an asymptotic point of view, the delay bounds from Eq. (19) grow as $\Theta(H)$, whereas the delay bounds from Eq. (20) grow as $\Theta(H \log H)$; the extra logarithmic factor stems from dispensing with the independence of arrivals [23].

Finally, we remark that the network calculus bounds derived in this section can be extended to more general queueing networks where exact results are usually not available. Such an extension reduces to the derivation of bounds on the



(a) $\rho = 75\%$ (b) H=10

Fig. 2. End-to-end delay bounds in a M/M/1 network as a function of (a) number of nodes and (b) utilization factor; the through traffic percentages are $50\%, 90\%$ in (a), and 90% in (b); nodes' service rate $R = 100\ Mbps$, average packet size $400\ B$, $\varepsilon = 10^{-6}$.

moment generating functions of $A(t) = \sum_{i=1}^{N(t)} X_i$ and X_i. Moreover, one can adapt the presented calculus to analyze queueing networks with constant service times of packets at each traversed node. A solution consists in describing packetizers with non-random service curves whose convolution can be analytically expanded without independence requirements, as needed for Eq. (19) (see [23]).

5 Conclusions

We have explored the accuracy of stochastic network calculus bounds by comparing them with exact results available in product-form networks. The single-node analysis showed that the bounds are tight at most utilizations and, by using the independent increments property of arrivals we could recover exact M/M/1 results and numerically match M/D/1 results. The multi-node analysis showed that for some scenarios (low percentage of cross traffic and accounting for independence of arrivals), the obtained bounds are reasonably accurate. Nevertheless, there exist complementary scenarios where the calculus may yield conservative bounds, due to the worst-case representation of service with leftover service curves and dispensing with statistical independence.

Acknowledgements

The research in this paper is supported in part by the National Science Foundation under grant CNS-0435061. The author gratefully acknowledges valuable comments from Almut Burchard and Jörg Liebeherr. Also, the author thanks the anonymous reviewer for pointing out a technical error in an earlier manuscript.

References

1. R. Cruz, "A calculus for network delay, parts I and II," *IEEE Transactions on Information Theory*, vol. 37, pp. 114–141, Jan. 1991.
2. C. S. Chang, *Performance Guarantees in Communication Networks*. Springer Verlag, 2000.
3. J. Y. Le Boudec and P. Thiran, *Network Calculus*. Springer Verlag, Lecture Notes in Computer Science, LNCS 2050, 2001.
4. J. Kurose, "On computing per-session performance bounds in high-speed multi-hop computer networks," in *ACM Sigmetrics*, pp. 128–139, 1992.
5. E. W. Knightly, "Second moment resource allocation in multi-service networks," in *ACM Sigmetrics*, pp. 181–191, 1997.
6. R. Boorstyn, A. Burchard, J. Liebeherr, and C. Oottamakorn, "Statistical service assurances for traffic scheduling algorithms," *IEEE Journal on Selected Areas in Comm. Special Issue on Internet QoS*, vol. 18, pp. 2651–2664, 2000.
7. C. Li, A. Burchard, and J. Liebeherr, "A network calculus with effective bandwidth," Tech. Rep. CS-2003-20, University of Virginia, Nov. 2003.
8. O. Yaron and M. Sidi, "Performance and stability of communication networks via robust exponential bounds," *IEEE/ACM Trans. on Net.*, vol. 1, pp. 372–385, 1993.
9. D. Starobinski and M. Sidi, "Stochastically bounded burstiness for communication networks," *IEEE Transactions of Information Theory*, vol. 46, pp. 206–212, 2000.
10. F. Ciucu, A. Burchard, and J. Liebeherr, "Scaling properties of statistical end-to-end bounds in the network calculus," *IEEE Transactions on Information Theory*, vol. 52, pp. 2300–2312, June 2006.
11. M. Fidler, "An end-to-end probabilistic network calculus with moment generating functions," in *IEEE 14th Int. Workshop on Quality of Service*, pp. 261–270, 2006.
12. Y. Jiang, "A basic stochastic network calculus," in *ACM Sigcomm*, pp. 123–134, 2006.
13. E. Knightly, "Enforceable quality of service guarantees for bursty traffic streams," in *IEEE Infocom*, pp. 635–642, 1998.
14. L. Kleinrock, *Queueing Systems*, vol. 1. John Wiley and Sons, 1975.
15. K. Pandit, J. Schmitt, and R. Steinmetz, "Network calculus meets queueing theory - a simulation based approach to bounded queues," in *IEEE 12th International Workshop on Quality of Service*, pp. 114–118, 2004.
16. F. 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*, pp. 141–168, Oxford University Press, 1996.
17. J. F. C. Kingman, "A martingale inequality in the theory of queues," in *Cambridge Philos. Soc.*, vol. 59, pp. 359–361, 1964.
18. E. Buffet and N. G. Duffield, "Exponential upper bounds via martingales for multiplexers with Markovian arrivals," *J. of App. Prob.*, vol. 31, pp. 1049–1060, 1994.
19. Z. Liu, P. Nain, and D. Towsley, "Exponential bounds with applications to call admission," *J. of the ACM*, vol. 44, pp. 366–394, 1997.
20. C. S. Chang, "On the exponentiality of stochastic linear systems under the max-plus algebra," *IEEE Trans. on Automatic Control*, vol. 41, pp. 1182–1188, 1996.
21. J.-Y. L. Boudec, "Some properties of variable length packet shapers," *IEEE/ACM Transactions on Networking*, vol. 10, pp. 329–337, 2002.
22. V. Iversen and L. Staalhagen, "Waiting time distribution in M/D/1 queueing systems," *Electronics Letters*, vol. 35, no. 25, pp. 2184–2185, 1999.
23. A. Burchard, J. Liebeherr, and F. Ciucu, "On $\Theta(H \log H)$ scaling of network delays," in *Proceedings of IEEE Infocom*, 2007.