End-to-End Delay Analysis for Networks with Partial Assumptions of Statistical Independence

Florin Ciucu
T-Labs / Technische Universität Berlin
florin@net.t-labs.tu-berlin.de

ABSTRACT

By accounting for statistical properties of arrivals and service, stochastic formulations of the network calculus yield significantly tighter backlog and delay bounds than those obtained in a purely deterministic framework. This paper proposes a stochastic network calculus formulation which can account for partial assumptions on statistical independence of arrivals and service across multiple network nodes. Scenarios where this can be useful are packet tandem networks with cross traffic and independent arrivals, where identical packet sizes create correlations across the nodes. As an application, the paper investigates the role of partial statistical independence on end-to-end delay bounds in four main scenarios arising by combining assumptions on the statistical independence of arrivals and packet sizes at different network nodes.

1. INTRODUCTION

The network calculus is a relatively recent theory for queueing analysis which was mostly developed and has played a significant role in the area of communication networks [4]. Its main idea is to use a bounding instead of an exact representation of the arrivals and service at queues. Besides communication networks, progress and applications of the calculus were also reported in diverse areas such as manufacturing of blocking systems [2], or real-time [23] and avionics [21] embedded systems.

Initially, the network calculus was formulated in a purely deterministic framework with strict bounds imposed on arrivals and service, and also strict bounds obtained on backlogs and delays [12]. This bounding approach unfolded in a very powerful analytical tool for deterministic queueing systems as historically difficult issues such as scheduling and multi-node analysis became more amenable to analysis. However, the statistical multiplexing inherent in packet networks cannot be captured in a deterministic framework and, consequently, applications of the calculus in networks with many flows generally results in overly pessimistic performance bounds, or, implicitly, in low utilizations of network resources.

In order to capture statistical multiplexing, and yet preserve its analytical power, the deterministic network calculus has been extended in a probabilistic framework. Some extensions of the calculus preserved the deterministic bounding of arrivals and service [22], whereas others adopted a statistical bounding of arrivals [24], or service [6], or both [13, 15] (see also [19]). These extensions have in common the concept of a probabilistic space which permits capturing statistical multiplexing using results from probability theory such as large deviations [7], or the Central Limit Theorem (CLT) [17]. In this way, significant statistical multiplexing gain was reported using the emerging statistical formulations of network calculus (see for instance [1]).

The key assumption exploited by statistical formulations of the network calculus is the statistical independence among arrival or service processes. One way to exploit the independence of many arrival processes at a node is to construct bounding functions, called statistical envelopes, for the aggregate arrival process using the CLT as in [1]. Then, by increasing the number of arrival processes, the envelope functions approach the average rate functions of the aggregate arrival process such that the subsequent analysis can yield very tight performance bounds. The statistical independence of arrival processes can also be exploited in a multi-node scenario, for instance using basic properties of moment generating functions [14]. Using similar properties, network calculus can also account for the statistical independence of arrival and service processes when modelling queueing models such as M/M/1 yielding reasonably accurate bounds [9].

A statistical network calculus can also analyze network scenarios where the statistical independence of arrival or service processes may not always hold. A bounding representation for the aggregate of non-necessarily independent arrival flows is given in [24]. Single-node performance bounds for a flow whose arrival and services processes are non-necessarily independent are derived in [18]. End-to-end delay bounds for a single flow in a tandem network are derived in [5] for a packetized service model, where correlations among the service processes at the nodes exist due to the fact that each packet maintains the same size at each traversed node. Some formulations of network calculus, e.g., [11] for a fluid service model, account for the statistical independence of the arrival processes at the same node, but do not make independence assumptions among arrival processes at different network nodes.

VALUETOOLS 2009, October 20-22, 2009 - Pisa, Italy.

In this paper we develop a continuous-time network calculus formulation, as a generalization of [11, 14], in order to exploit partial statistical independence in the service processes of a network flow. The key concept is that of a service curve defined using both a random process and a deterministic error function; such a service curve is also introduced in a discrete-time setting in [9]. The random process specifies probabilistic lower bounds on service, whereas the error function specifies the probabilities of violating the bounds by deterministic values. Our calculus formulation can be particularly useful in a tandem network with cross traffic and a packetized service model, where the traversing flow's service depends both on the cross traffic and its packets sizes. If available, the independence of cross traffic is captured in the service curve processes at the nodes, whereas the correlations induced by maintaining the size of each packet constant are captured in the error functions.

As an application of the proposed network calculus we investigate the role of partial statistical independence in a packetized tandem network with cross traffic. Concretely, we derive end-to-end delay bounds and provide numerical illustrations in four scenarios arising by combining independence assumptions on (1) the cross traffic at the nodes, and (2) the sizes of each packet at the traversed nodes. We also consider a fluid service model and investigate its accuracy relative to the packetized service model for two scenarios depending on the independence of traffic.

From a scaling perspective, the derived bounds grow as $\Theta(H)$ in the number of nodes H under complete independence assumptions [14]. Otherwise, the bounds grow as $\mathcal{O}(H \log H)$ [11]. For a packetized service model where each packet maintaining its size the $\mathcal{O}(H \log H)$ is asymptotically tight, i.e., $\Theta(H \log H)$ [5]; for the fluid service model, however, the tightness of $\mathcal{O}(H \log H)$ is still open.

The rest of the paper is organized as follows. In Section 2 we develop the main elements of a network calculus formulation, i.e., the representation of traffic and service, the derivation of single-node performance bounds, and the extension to multi-node analysis. In Section 3 we derive end-to-end delay bounds in a tandem network with cross traffic for both packetized and fluid service models in various scenarios depending on the statistical independence of cross traffic and packet sizes. Numerical illustrations of these bounds are provided in Section 4. Some brief conclusions are provided in Section 5.

2. A STOCHASTIC NETWORK CALCULUS FORMULATION

We use a continuous-time model. Network nodes have a constant service rate and infinite-sized buffers. The arrivals and departures at a node are modelled with non-decreasing, left-continuous processes. For an arrival process $A(t)$ and the corresponding departure process $D(t)$, we assume the initial condition $A(0) = 0$ and the causal condition $D(t) \leq A(t)$. For convenience we introduce the bivariate process $A(s,t) = A(t) - A(s)$. The corresponding backlog and delay processes are denoted by $B(t) = A(t) - D(t)$, and $W(t) = \inf\{d : A(t-d) \leq D(t)\}$, respectively.

2.1 Traffic Representation

Unlike the classical queueing network theory which generally uses exact traffic representations (e.g. exact distributions of packet sizes and their inter-arrival times), the network calculus uses weaker traffic representations in terms of bounds. Here we adopt the traffic representation with scaled exponential bounds on the moment generating functions of the arrival processes [8, 14].

Definition 1. *(*MGF Envelope for Arrivals*) An arrival process $A(t)$ is bounded by an MGF envelope, with rate r and scaling factor M, for some choices of a parameter $\theta > 0$, if for all $0 \leq s \leq t$*

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

Therefore, traffic is by definition essentially unknown but subject to regularity constraints [12]. Both the rate r and the scaling factor M depend on the parameter θ whose optimal value can be numerically determined. The upper limit of the range of θ is generally inversely proportional to the data unit scale such that numerical optimizations can be done over a relatively small space.

We restrict the arrivals to the case when r and M are invariant to time parameters. As such, the arrival model includes for instance many Markov-modulated and multiplexed-regulated processes, but excludes self-similar processes (e.g. fractional Brownian motion). The model also excludes heavy-tailed processes which have infinite MGFs.

As an example, consider that $A(t)$ is a compound Poisson process, i.e.,

$$A(t) = \sum_{i=1}^{N(t)} X_i \ ,$$

where $N(t)$ is a Poisson process with rate $\lambda > 0$ and X_i are i.i.d. random variables with mean $\frac{1}{\mu}$. In this case, $A(t)$ is bounded by an MGF envelope with rate and scaling factor given by

$$r = \frac{\lambda}{\mu - \theta} \ , \quad M = 1 \ ,$$

where the range of θ is $(0, \mu)$.

2.2 Service Representation

As for traffic representation, the network calculus also uses bounds for service representation. The key idea is the concept of a *service curve* which relates the arrival and departure processes of a traffic flow through a lower bound. Concretely, a service curve specifies a lower bound on the amount of service received by a flow either at a network node or across an entire network path.

Here we extend the statistical service curve model from [9] to a continuous-time setting. First, we need to define the $(min, +)$ convolution of two processes $X(s,t)$ and $Y(s,t)$ as $X * Y(s,t) = \inf_{s \leq u \leq t} \{X(s,u) + Y(u,t)\}$. Also, for a number x, we denote $[x]_+ = max\{x, 0\}$.

Definition 2. *(*Statistical Service Curve*) A doubly-indexed random process $S(s,t)$ is a statistical service curve with error function $\varepsilon(\sigma)$ for an arrival process $A(t)$ if the corresponding departure process $D(t)$ satisfies for all $t \geq 0$ and σ*

$$Pr\Big(D(t) < A * [S - \sigma]_+ (t + \tau_0)\Big) \leq \varepsilon(\sigma) \ ,$$

where $\tau_0 \geq 0$ is a discretization parameter.

For each sample path the random process $S(s,t)$ is decreasing in s, increasing in t, and satisfies $S(s,t) = S(s,u) + S(u,t)$ for all $0 \le s \le u \le t$. The error function $\varepsilon(\sigma)$ is nonnegative and non-increasing, and satisfies

$$\varepsilon(\sigma) \ge 1 \quad \text{for all} \quad \sigma < S(\tau_0) \ . \tag{1}$$

The service curve model imposes a positivity constraint in order to simplify the analysis of scenarios with negative service curves. It also lets the convolution span the time interval $[0, t+\tau_0]$, rather than the interval $[0,t]$ used especially in discrete-time service models, in order to simplify formulas arising from the discretization of continuous sample paths with the parameter τ_0. Consequently, the process $S(s,t)$ or the function $\varepsilon(\sigma)$ depend on τ_0 which is generally subject to optimizations. The next lemma shows how to make the transition from service curves satisfying Definition 2 with $\tau_0 = 0$ to more less common service curves defined with $\tau_0 \ge 0$.

LEMMA 1. *Consider an arrival process $A(t)$, the corresponding departure process $D(t)$, and an error function $\varepsilon(\sigma)$. If a function $\hat{S}(s,t)$ satisfies*

$$Pr\Big(D(t) < A * \Big[\hat{S} - \sigma\Big]_+ (t)\Big) \le \varepsilon(\sigma) \ ,$$

and $\varepsilon(\sigma) \ge 1$ for all $\sigma < \hat{S}(0)$, then for any $\tau_0 > 0$ the function

$$S(s,t) = \hat{S}(s, t-\tau_0)$$

for $t - s \ge \tau_0$, and $S(s,t) = 0$ for $t - s < \tau_0$, is a statistical service curve in the sense of Definition 2.

PROOF. Using the positivity of $A(t)$, the proof immediately follows using

$$A * [S - \sigma]_+ (t + \tau_0) \le A * \Big[\hat{S} - \sigma\Big]_+ (t) \ ,$$

for all $t, \tau_0 \ge 0$, and all σ. The condition from Eq. (1) on the error function $\varepsilon(\sigma)$ is also satisfied because $S(\tau_0) = \hat{S}(0)$. □

Along with its discrete-time counterpart [9], the service model from Definition 2 generalizes existing service models by letting $S(s,t)$ be a random process and $\varepsilon(\sigma) \ge 0$. Particularizing with $S(s,t)$ non-random and $\varepsilon(\sigma) \ge 0$ it reduces to the service model from [11]. Also, by letting $S(s,t)$ be random, $\varepsilon(\sigma) = 0$, and $\tau_0 = 0$, it reduces to a service model from [8].

Our generalized service model is motivated by the need to deal with partial statistical independence assumptions on the service received by an arrival flow across different network nodes. For instance, in a two-node scenario, if the services are not necessarily independent then at least one of the service curves is non-random and the corresponding error functions are positive in order to carry out the convolution of the service curves (see Subsection 2.4). Otherwise, if the services are independent, then both service curves are random and the error functions are zero. As we will see in Subsections 3.2 and 3.3, there also exist scenarios where the service curves are random and the error functions are positive in order to exploit partial assumptions of statistical independence.

In order to deal with the convolution of multiple service curves, or with the derivation of performance bounds, MGF envelope models are used to bound service curves defined with random processes [8, 14].

DEFINITION 3. *(MGF BOUND FOR SERVICE CURVES) A statistical service curve $S(s,t)$ has an MGF bound, with rate r and scaling factor M, for some choices of a parameter $\theta > 0$, if for all $0 \le s \le t$*

$$E\Big[e^{-\theta S(s,t)}\Big] \le M e^{-\theta r (t-s)} \ .$$

Alike in the MGF envelope model for arrivals from Definition 1, both the rate r and the scaling factor M depend on θ. On the other hand, unlike bounding the arrivals from above, the MGF envelope model from Definition 3 bounds service curves from below. The reason is that the arrival and service processes of a flow have opposite signs in the derivation of performance bounds. This can be more clearly seen in the next lemma which will be used to the derivation of single and multi-node performance bounds.

LEMMA 2. *(SAMPLE-PATH BOUNDS) Suppose that an arrival process $A(t)$ is bounded for some choice of $\theta > 0$ by an MGF envelope with rate r_a and scaling factor M_a. For some parameter τ_0, let a service curve $S(s,t)$ independent of $A(t)$. For the same θ, $S(s,t)$ has an MGF bound with rate r_s and scaling factor $M_s = M_s' \binom{\lfloor \frac{t-s}{\tau_0} \rfloor + H - 1}{H-1}$ for some integer $H > 0$, where M_s' does not depend on $t - s$. Denote $M = M_a M_s'$, $r = r_s - r_a$, and assume for stability that $r > 0$. Then for all $t \ge 0$ and σ*

$$Pr\left(\sup_{0 \le s \le t} \{A(s,t) - S(s, t+\tau_0)\} > \sigma\right) \le \varepsilon(\sigma) \ ,$$

where $\varepsilon(\sigma) = M \left(\frac{1}{\theta r \tau_0}\right)^H e^{-\theta\sigma}$.

The error function of the service curve is not needed for the lemma's purpose. In applications, H corresponds to the number of nodes. The case when M_s does not depend on $t - s$ corresponds to $H = 1$. The complementary case corresponds to $H > 1$; the dependency is caused by a binomial factor arising in the evaluation of multi-node convolutions (for further technical details see Theorem 4).

PROOF. Fix $t \ge 0$ and σ. For $0 \le s \le t$ we let $j = \lfloor \frac{t-s}{\tau_0} \rfloor$ be the integer part of $\frac{t-s}{\tau_0}$, so that $[t - (j+1)\tau_0]_+ < s \le t - j\tau_0$. We can write

$$\begin{aligned}
&Pr\left(\sup_{0 \le s \le t} \left\{A(s,t) - [S(s,t+\tau_0) - \sigma]_+\right\} > 0\right) \\
&\quad \le Pr\Big(\sup_{j \ge 0} \Big\{A\left([t - (j+1)\tau_0]_+, t\right) \\
&\qquad\qquad - S\left(t - j\tau_0, t + \tau_0\right)\Big\} > \sigma\Big) \\
&\quad \le M \sum_{j \ge 1} \binom{j + H - 1}{H - 1} e^{-\theta r j \tau_0} e^{-\theta\sigma} \\
&\quad \le M \left(1 - \left(\frac{1}{1 - e^{-\theta r \tau_0}}\right)^H\right) e^{-\theta\sigma} \\
&\quad \le M \left(\frac{1}{\theta r \tau_0}\right)^H e^{-\theta\sigma} \ .
\end{aligned}$$

In the fourth line we applied Boole's inequality. In the fifth line we used $\sum_{j \ge 0} \binom{j+H-1}{H-1} a^j = \left(\frac{1}{1-a}\right)^H$ for all $0 < a < 1$ (see [14]). Last we used that $\left(\frac{1}{1-e^{-x}}\right)^H - 1 \le \left(\frac{1}{x}\right)^H$ for all $x > 0$. The proof is thus complete. □

2.2.1 Leftover Fluid Service Curves

Here we construct service curves for the lowest-priority flow, or an aggregate of flows, at a static-priority (SP) scheduler serving with constant rate. These service curves are suggestively referred to as leftover service curves, since they express the capacity left unused by the higher priority flows. They provide thus a worst-case description of service and have the property that they are guaranteed by any workconserving scheduling mechanism. The next theorem provides such constructions for a fluid service model. This model dispenses with the size of a packet, i.e., the service unit is infinitesimal; in other words, a fraction of a packet becomes available for service as soon as processed upstream.

THEOREM 1. *(LEFTOVER SERVICE CURVE) Consider a node with capacity C serving two arrival processes $A(t)$ and $A_c(t)$, whose corresponding departure processes are $D(t)$ and $D_c(t)$, respectively. Assume that $A_c(t)$ is bounded by an MGF envelope with rate $r_c < C$ and scaling factor 1 for some choice of $\theta > 0$. Then we have the following two constructions for some $\tau_0 > 0$.*

1. *The random process*

$$S(s,t) = \left[C\left(t-s-\tau_0\right) - A_c\left(s, t-\tau_0\right)\right]_+ \qquad (2)$$

is a statistical service curve for $A(t)$ with error function $\varepsilon(\sigma) = 0$. It has an MGF bound with rate $C - r_c$ and scaling factor $e^{\theta(C-r_c)\tau_0}$.

2. *For any choice of $\delta > 0$ the non-random function*

$$S(s,t) = \left[C - r_c - \delta\right]_+ (t-s) \qquad (3)$$

is a statistical service curve for $A(t)$ with error function $\varepsilon(\sigma) = \frac{e^{\theta C \tau_0}}{\theta\delta\tau_0} e^{-\theta\sigma}$.

The first construction, as a continuous-time extension of a construction from [14], is useful when $A(t)$ and $A_c(t)$ are statistically independent. The second construction extends a similar construction from [11] for arrivals described with MGF envelopes and is useful when $A(t)$ and $A_c(t)$ are not necessarily independent; note that in this case $S(s,t)$ is non-random. The proof of Eq. (2) follows from [14] and Lemma 1. The proof of Eq. (3) follows by extending a proof from [11] to bivariate service curves.

2.2.2 Packetization Service Curves

Here we construct service curves for a packetized service model which takes into account packet sizes. These constructions complement the constructions from the previous subsection for the fluid service model.

Consider a network node with capacity C serving a through arrival process $A(t)$, and possibly some cross arrival process $A_c(t)$. To account for the packetized service received by $A(t)$ we represent the node as a concatenation between a fluid server with rate C and a statistical packetizer denoted here by P^μ (see Figure 1). The fluid server serves packets according to the fluid service model, whereas the packetizer has the role of a delay element by ensuring that packets become available for service downstream after they were fully processed upstream. Unlike packetizers used in the literature [20, 3] for bounded packet sizes, P^μ herein deals with packet sizes described by probability distributions.

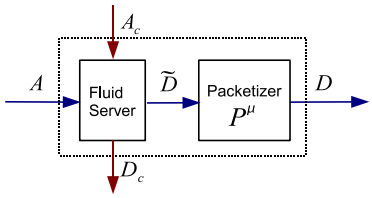


Figure 1: A statistical packetizer P^μ at a node with cross traffic.

We define $A(t)$ as the compound process

$$A(t) = \sum_{i=1}^{N(t)} X_i \ , \qquad (4)$$

where $N(t)$ is a counting process and X_i are i.i.d. random variables (the packets sizes) with mean $1/\mu$. Then, the output process $\widetilde{D}(t)$ satisfies for all $t \geq 0$

$$\widetilde{D}(t) = \sum_{i=1}^{M(t)} X_i + X_f(t) \ ,$$

where $M(t)$ denotes the number of packets fully processed by time t, and $X_f(t)$ denotes the processed fraction of the packet (if any) currently in service at time t; if the server is idle at time t, then $X_f(t) = 0$. The process $\widetilde{D}(t)$ is thus a virtual output process which represents the fluid output of the through flow at the fluid server.

Furthermore, the packetizer P^μ takes the fluid output $\widetilde{D}(t)$ as input and produces the packetized output

$$D(t) = \sum_{i=1}^{M(t)} X_i \ .$$

This accounts for the fact that a downstream node can start processing a packet no sooner than the packet was completely processed by the fluid server at the next upstream node. It then follows inductively that packetizers account for packetization in the entire network. The possible cross traffic is not required to pass through packetizers [3] and leftover fluid service curves for $A(t)$ at the fluid server can be constructed with Theorem 1.

The next lemma gives two statistical service curve representations for the packetizer P^μ, which will be useful depending on the statistical independence assumptions on packet sizes across a network.

LEMMA 3. *Consider a network node modelled as in Figure 1. Then the function*

$$S^\mu(s,t) = \left[C(t-s) - X_f(t)\right]_+ \qquad (5)$$

is a statistical service curve for the packetizer P^μ with error function $\varepsilon^\mu(\sigma) = 0$, in the sense of Definition 2 with $\tau_0 = 0$. If the packets sizes are exponentially distributed then the function

$$S^\mu(t) = Ct \qquad (6)$$

is a service curve for P^μ with error function $\varepsilon^\mu(\sigma) = e^{\mu C \tau_0} e^{-\mu\sigma}$.

The service curve from Eq. (5) is useful when the sizes of each packet are statistically independent at the traversed nodes (Kleinrock's independence assumption [16]). In turn,

the service curve from Eq. (6), also obtained in [5], is useful when the sizes of each packet are identical across the nodes; the reason is that the non-randomness of $S^{\mu}(t)$ circumvents correlations in the service times across the nodes.

PROOF. Fix $t \geq 0$ and denote by s the beginning of the last busy period before t at the fluid server. For proving the claim for $S^{\mu}(t)$ from Eq. (5) we observe that

$$u = t - \frac{X_f(t)}{C}$$

is the starting processing time of the packet currently serviced at time t. It then follows that

$$\begin{aligned} D(t) &= D(s) + C(u-s) \\ &= A(s) + C\left(t - \frac{X_f(t)}{C} - s\right) \\ &= A(s) + S^{\mu}(s,t) \\ &\geq A * S^{\mu}(t) \ , \end{aligned}$$

which completes the first part of the proof. For the second part we assume that for some σ and $\tau_0 \geq 0$

$$X_f(t) \leq \sigma - C\tau_0 \ .$$

Using the first part it follows that

$$D(t) \geq \inf_{0 \leq s \leq t+\tau_0} \left\{ A(s) + [C(t+\tau_0-s) - \sigma]_+ \right\} \ .$$

The proof is complete by taking probabilities. □

2.3 Single-Node Performance Bounds

The next theorem provides single-node performance bounds for a traffic flow with arrivals described by MGF envelopes and service described by service curves and MGF bounds.

THEOREM 2. *(PROBABILISTIC PERFORMANCE BOUNDS) Consider a flow with arrivals and departures processes $A(t)$ and $D(t)$, respectively. For some discretization parameter τ_0, the flow has the statistical service curve $S(s,t)$, independent of $A(t)$, with error function $\varepsilon^s(\sigma)$. Assume that $A(t)$ has an MGF envelope with rate r_a and scaling factor M_a for some $\theta > 0$. Also, $S(s,t)$ has an MGF bound with rate r_s and scaling factor $M_s = M_s'\left(\lfloor \frac{t-s}{\tau_0} \rfloor + H - 1 \atop H-1\right)$ for the same θ, some integer $H > 0$, and where M_s' does not depend on time parameters. Denote $M = M_a M_s'$, $r = r_s - r_a$, and assume for stability that $r > 0$. Define the error function*

$$\varepsilon(\sigma) = \inf_{\sigma^a + \sigma^s = \sigma} \left\{ M \left(\frac{1}{\theta r \tau_0}\right)^H e^{-\theta\sigma^a} + \varepsilon^s(\sigma^s) \right\} \ .$$

Then we have the following probabilistic bounds.

1. OUTPUT MGF ENVELOPE: *If $H = 1$ and $\varepsilon^s(\sigma) = 0$ for all σ, then the output process $D(t)$ is bounded by an MGF envelope with rate r_a and scaling factor $M\left(\frac{1}{\theta r \tau_0}\right)^H$.*

2. BACKLOG BOUND: *A bound on the backlog process $B(t)$ is given for all $t, \sigma \geq 0$ by*

$$Pr\Big(B(t) > \sigma\Big) \leq \varepsilon(\sigma) \ . \tag{7}$$

3. DELAY BOUND: *A bound on the delay process $W(t)$ is given for all $t, \sigma \geq 0$ by*

$$Pr\Big(W(t) > \frac{\sigma}{r_s}\Big) \leq \varepsilon(\sigma) \ . \tag{8}$$

The theorem generalizes existing results from the literature. If the service curve is the non-random function $S(s,t) = r_s(t-s)$ similar bounds were obtained in [11]. If $\varepsilon^s(\sigma) = 0$ similar bounds were derived in [14] in a discrete-time setting.

In this paper we only use the delay bounds; the other two bounds are provided here for completeness. We also point out that the results from the theorem depend on the discretization parameter τ_0 which will be subject to convex optimization.

PROOF. We only provide the proof for the delay bound; the other two proofs are similar. Fix $\tau_0 > 0$ and choose $t, \sigma \geq 0$ and σ^a, σ^s such that $\sigma^a + \sigma^s = \sigma$. Denote $d = \frac{\sigma}{r_s}$, and assume that for a particular sample-path

$$A(s, t-d) \leq [S(s, t-d+\tau_0) + S(t-d+\tau_0, t+\tau_0) - \sigma^s]_+ \tag{9}$$

holds for all $0 \leq s \leq t-d$. Also, assume that

$$D(t) \geq A * [S - \sigma^s]_+ (t+\tau_0) \ . \tag{10}$$

holds.

From Eq. (9) we successively obtain

$$\begin{aligned} &\sup_{0 \leq s \leq t-d} \left\{ A(s,t-d) - [S(s,t+\tau_0) - \sigma^s]_+ \right\} \leq 0 \\ &\quad \Rightarrow \sup_{0 \leq s \leq t+\tau_0} \left\{ A(s,t-d) - [S(s,t+\tau_0) - \sigma^s]_+ \right\} \leq 0 \\ &\quad \Rightarrow A(t-d) \leq D(t) \Rightarrow W(t) \leq d \ . \end{aligned}$$

In the second line we extended the range of the supremum using the positivity constraints. In the third line we applied Eq. (10), and then we used the definition of delay.

Since we started with Eqs. (9) and (10) we arrive at

$$\begin{aligned} Pr\left(W(t) > d\right) &\leq P\left(\text{Eqs. (9) or (10) fail}\right) \\ &\leq M\left(\frac{1}{\theta r \tau_0}\right)^H e^{-\theta r_s d} e^{\theta \sigma^s} + \varepsilon^s(\sigma^s) \\ &\leq M\left(\frac{1}{\theta r \tau_0}\right)^H e^{-\theta\sigma^a} + \varepsilon^s(\sigma^s) \ . \end{aligned}$$

We first applied Lemma 2 with $\sigma = S(t-d+\tau_0, t+\tau_0) - \sigma^s$. The proof is complete after minimizing over $\sigma^a + \sigma^s = \sigma$. □

2.4 Statistical Network Service Curve

Here we analyze the multi-node case. The next theorem gives the construction of a statistical network service curve for a flow traversing a network, i.e., a service curve describing the service given to the flow as if it traversed a single-node only. Having such a network service curve, end-to-end performance bounds can be obtained with Theorem 2.

Let us first introduce two useful notations. For a process $X(t)$ and a real number δ we define the process

$$X_{\delta}(t) = X(t) + \delta t \ .$$

Also, for an integrable error function $\varepsilon(\sigma)$ and a positive number a we define the function

$$\tilde{\varepsilon}_a(\sigma) = \frac{1}{a} \int_{\sigma}^{\infty} \varepsilon(u) du,$$

as an upper bound for the discrete sum $\sum_{j=1}^{\infty} \varepsilon\left(\sigma + ja\right)$.

THEOREM 3. (STATISTICAL NETWORK SERVICE CURVE). *Consider a traffic flow traversing a network with H nodes in series. For some discretization parameter $\tau_0 > 0$ assume that $S^h(s,t)$ are statistical service curves for the flow with error functions $\varepsilon^h(\sigma)$ at the nodes $h = 1, \ldots, H$. If $\varepsilon^h(\sigma) = 0$ for all h and σ, then the process*

$$S^{net}(s,t) = S^1 * S^2 * \ldots * S^H(s,t) \ . \tag{11}$$

is a (statistical) network service curve for the flow. Otherwise, if $\varepsilon^h(\sigma) \geq 0$ and are integrable then the corresponding statistical network service curve is given for any choice of $\delta > 0$ by

$$S^{net}(s,t) = S^1 * S^2_{-\delta} * \ldots * S^H_{-(H-1)\delta}(s,t) \ , \tag{12}$$

with the error function

$$\varepsilon^{net} = \tilde{\varepsilon}^1_{\delta\tau_0} * \ldots * \tilde{\varepsilon}^{H-1}_{\delta\tau_0} * \varepsilon^H \tag{13}$$

The first construction from Eq. (11) extends the construction from [8] to a continuous-time setting. The second construction from Eq. (12) generalizes a construction from [11] to service curves defined as random processes. For a discussion on the motivation of introducing the additional parameter δ in Eqs. (12) and (13) see [11]. The proofs for the two constructions are similar as in [8, 11] and are omitted here.

As we have seen in Theorem 2, the derivation of performance bounds requires the existence of MGF bounds on the service curves. The next theorem provides MGF bounds for the two statistical network service curves from Theorem 3. To keep the notation simple we only consider the case when all the service curves have the same distributions and error functions.

THEOREM 4. (MGF BOUND FOR STATISTICAL NETWORK SERVICE CURVE). *Consider the scenario from Theorem 3. For some choice of $\theta > 0$, assume that the service curves $S^h(s,t)$ are independent, and each has an MGF bound with rate r_s and scaling factor M_s that does not depend on time parameters. Then the flow's statistical network service curve has the MGF bound*

$$E\left[e^{-\theta S^{net}(s,t)}\right] \leq M^{net} e^{-\theta r_s (t-s)} \ , \tag{14}$$

where the scaling factor M^{net} depends on the construction of the network service curve.

1. *If the statistical network service curve is given by Eq. (11) then*

$$M^{net} = M_s^H \binom{\lfloor \frac{t-s}{\tau_0} \rfloor + H - 1}{H-1} e^{(H-1)\theta r_s \tau_0} \ . \tag{15}$$

2. *If the statistical network service curve is given by Eq. (12) then*

$$M^{net} = M_s^H \binom{\lfloor \frac{t-s}{\tau_0} \rfloor + H - 1}{H-1} e^{(H-1)\theta\left(r_s+\delta+\delta\lfloor \frac{t-s}{\tau_0} \rfloor\right)\tau_0} \ . \tag{16}$$

The constructions from Eqs. (14) and (15) for the network service curve from Eq. (11) extend a corresponding result from [14] to a continuous-time setting.

PROOF. Fix $\delta, \tau_0 > 0$ and $0 \leq s \leq t$. In the first case we can expand the MGF of $S^{net}(s,t)$ by applying Boole's inequality and the discretization technique used in the proof of Lemma 2.

$$\begin{aligned}
&E\left[e^{-\theta S^{net}(s,t)}\right] \\
&\leq E\left[\sup_{s\leq x_1 \leq \cdots \leq x_{H-1} \leq t} e^{-\theta\left(S^1(s,x_1)+\cdots+S^H\left(x_{H-1},t\right)\right)}\right] \\
&\leq \sum_{0\leq j_1 \leq \cdots \leq j_{h-1} \leq \lfloor \frac{t-s}{\tau_0} \rfloor} E\Bigg[e^{-\theta\left(S^1\left(s,[t-(j_1+1)]_+\tau_0\right)+\cdots\right.} \\
&\qquad\qquad \left. e^{\cdots+S^H\left((t-j_{H-1})\tau_0,t\right)}\right)\Bigg] \\
&\leq M^H e^{(H-1)\theta r_s \tau_0} e^{-\theta r_s (t-s)} \sum_{0\leq j_1 \leq \cdots \leq j_{h-1} \leq \lfloor \frac{t-s}{\tau_0} \rfloor} 1 \\
&\leq M^H \binom{\lfloor \frac{t-s}{\tau_0} \rfloor + H - 1}{H-1} e^{(H-1)\theta r_s \tau_0} e^{-\theta r_s (t-s)} \ .
\end{aligned}$$

In the fifth line we expanded the MGF of $S^h(s,t)$ by using statistical independence and then collected terms. In the sixth line the binomial coefficient is the number of combinations with repetitions.

For the second case we proceed similarly as before

$$\begin{aligned}
&E\left[e^{-\theta S^{net}(s,t)}\right] \\
&\leq E\left[\sup_{s\leq x_1 \leq \cdots \leq x_{H-1} \leq t} e^{-\theta\left(S^1(s,x_1)+\cdots+S^H_{-(H-1)\delta}\left(x_{H-1},t\right)\right)}\right] \\
&\leq E\Bigg[\sup_{s\leq x_1 \leq \cdots \leq x_{H-1} \leq t} e^{-\theta\left(S^1(s,x_1)+\cdots+S^H\left(x_{H-1},t\right)\right)} \\
&\qquad\qquad e^{\theta\delta\left((H-1)t-(x_1+\cdots+x_{H-1})\right)}\Bigg] \\
&\leq \sum_{0\leq j_1 \leq \cdots \leq j_{H-1} \leq \lfloor \frac{t-s}{\tau_0} \rfloor} E\Bigg[e^{-\theta\left(S^1(s,(t-(j_1+1))\tau_0)+\cdots\right.} \\
&\qquad\qquad \left. e^{\cdots+S^H\left((t-j_{H-1})\tau_0,t\right)}\right) e^{\theta\delta\sum_{h=1}^{H-1}(j_h+1)\tau_0}\Bigg] \\
&\leq M^H \binom{\lfloor \frac{t-s}{\tau_0} \rfloor + H - 1}{H-1} e^{(H-1)\theta\left(r_s+\delta+\delta\lfloor \frac{t-s}{\tau_0} \rfloor\right)\tau_0} e^{-\theta r_s (t-s)} \ .
\end{aligned}$$

In the last line we bounded each j_h by $\lfloor \frac{t-s}{\tau_0} \rfloor$. The proof is thus complete. □

3. APPLICATIONS: DERIVATION OF END-TO-END DELAY BOUNDS

In this section we apply the network calculus formulation from Section 2 to the derivation of end-to-end delay bounds. The main goal is to illustrate the derivation of the bounds in four different scenarios depending on four different types of statistical independence assumptions.

Concretely, we consider the tandem network with cross traffic from Figure 2. A through flow traverses H nodes and each node is also transited by a cross flow; the notation for the flows is as in the figure. Each node has capacity C and serves the packets in a SP manner giving the cross flow's packets higher priorities. The through flow and each

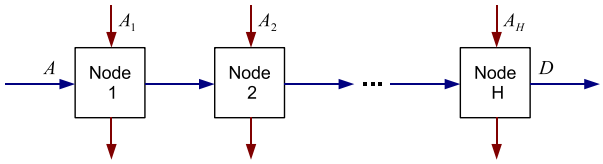


Figure 2: A tandem network with cross traffic

of the cross flows consist of packets arriving according to Poisson processes with rates λ and λ_c, respectively. The size of each packet is exponentially distributed with mean $1/\mu$. The network is stable, i.e., the utilization factor $\rho = (\lambda + \lambda_c)/(\mu C)$ is less than one.

We represent the arrivals by compound Poisson processes as in Subsection 2.1. As such, the through flow $A(t)$ is bounded by an MGF envelope with rate and scaling factor given by

$$r_a = \frac{\lambda}{\mu - \theta} \ , \quad M_a = 1 \ . \tag{17}$$

Similarly, each cross flow $A_h(t)$ is bounded by an MGF envelope with rate $r_c = \frac{\lambda_c}{\mu-\theta}$ and scaling factor $M_c = 1$ for the same choice of θ with $0 < \theta < \mu$.

In the next four subsections (3.1-3.4) we analyze four scenarios by combining independence assumptions on (1) the arrival processes, and (2) the sizes of the through flow's packets at the nodes. As mentioned earlier, if each packet of the through flow has identical sizes at the nodes then the corresponding services are not statistically independent.

In addition to treating a packetized service model, the first two subsections also treat the case of a fluid service model. By deriving delay bounds in both packetized and service models, the goal is to offer insight into the justification of using the (approximative) fluid models which are generally easier to be carried out analytically.

We also present some technical considerations in Subsection 3.5 on how the network calculus deals with (lack of) assumptions of statistical independence.

3.1 Independent arrivals / Independent service times

Here we assume statistical independence everywhere: the through and the cross flows are independent processes, whereas the sizes of the through packets are independently regenerated at each node. We first analyze the packetized service model and then the fluid service model.

Let us consider the representation of each of the network nodes as in Figure 1 (i.e. as the concatenation between a fluid server and a packetizer P^{μ}). By enforcing the condition that $\theta < \mu - \lambda_c/C$, we can invoke Theorem 1 (by dispensing with the discretization parameter τ_0) and obtain that the function

$$T^h(s,t) = [C(t-s) - A_h(s,t)]_+ \tag{18}$$

is a statistical leftover service curve at the h^{th} fluid server. Then, by using the service curve representation of each packetizer from Eq. (5) in Lemma 3, we further obtain with Eq. (11) that each node in the virtual network from Figure 2 can be described with the statistical network service curve

$$\begin{aligned} S^h(s,t) &= T^h * S^{\mu,h}(s,t) \\ &= \inf_{s \le u \le t} \left\{ [C(u-s) - A_h(s,u)]_+ + \Big[C(t-u) - X_f^h(t)\Big]_+ \right\} \\ &\ge \Big[C(t-s) - A_h(s,t) - X_f^h(t)\Big]_+ \ , \end{aligned}$$

where $X_f^h(t)$ denotes the fraction already processed of the packet currently in service (if any) at node h at time t. Moreover, each service curve $S^h(s,t)$ has an MGF bound with rate and scaling factor given by

$$r_s = C - r_c \ , \quad M_s = \frac{\mu}{\mu - \theta} \ ,$$

where we used that $E\left[e^{\theta X_f^h(t)}\right] \le \frac{\mu}{\mu-\theta}$.

Next we can construct the statistical network service curve for the through flow along the H nodes. At this point we make the transition to a service curve representation with the discretization parameter τ_0. Using Eq. (11) and Lemma 1 we obtain the statistical network service curve

$$S^{net}(s,t) = S^1 * \ldots * S^H(s, t-\tau_0) \ ,$$

that has (according to Eq. (15) from Theorem 4) an MGF bound with rate and scaling factor given by

$$r^{net} = r_s \ , \quad M^{net} = \left(\frac{\mu}{\mu-\theta} e^{2\theta r_s \tau_0}\right)^H \binom{\lfloor \frac{t-s}{\tau_0} \rfloor + H - 1}{H-1} \ .$$

We remark that the contribution of using Lemma 1 to the scaling factor M^{net} from Eq. (15) in Theorem 4 is $e^{H\theta r_s \tau_0}$.

Finally, having the through flow's MGF envelope description from Eq. (17) and the network service curve S^{net}, we can invoke Theorem 2 and derive the delay bounds. Denote

$$r = r_s - r_a$$

and enforce the stability condition

$$r > 0 \quad \Leftrightarrow \quad \theta < \mu(1-\rho) \ .$$

Then Eq. (8) gives the delay bound for all $\sigma \ge 0$

$$P\left(W^{net}(t) > \frac{\sigma}{r^{net}}\right) \le \left(\frac{e^{2\theta r_s \tau_0}}{\theta r \tau_0} \frac{\mu}{\mu-\theta}\right)^H e^{-\theta\sigma} \ .$$

Optimizing the discretization parameter $\tau_0 = \frac{1}{2\theta r_s}$, replacing σ with $d \cdot r^{net}$, and letting $t \to \infty$ we obtain the steady-state delay bound for all $d \ge 0$

$$P\left(W^{net} > d\right) \le \left(e \frac{2r_s}{r} \frac{\mu}{\mu-\theta}\right)^H e^{-\theta\left(C - \frac{\lambda_c}{\mu-\theta}\right)d} \ . \tag{19}$$

Next we consider a fluid service model. Then we can view each node in the network from Figure 2 as a fluid server, and consequently derive the leftover service curves

$$S^h(s,t) = [C(t-s) - A_h(s,t)]_+ \ ,$$

i.e., the expression for $T^h(s,t)$ from Eq. (18).

To derive end-to-end delay bounds we can proceed as before, with the difference that the scaling factor of the MGF bound of $S^h(s,t)$ is now $M_s = 1$ instead of $M_s = \frac{\mu}{\mu-\theta}$.

The steady-state delay bound assuming the fluid service model thus becomes

$$P\left(W^{net} > d\right) \le \left(e \frac{2r_s}{r}\right)^H e^{-\theta\left(C - \frac{\lambda_c}{\mu-\theta}\right)d} \ . \tag{20}$$

3.2 Correlated arrivals / Independent service times

Here we dispense with the statistical assumptions on arrivals but still assume the independent regeneration of the packet sizes of the through flow. Note that the assumption of correlated arrivals makes inapplicable the product property of MGFs (i.e. $E[XY] = E[X]E[Y]$ for independent r.v.'s X and Y) used in the previous section to derive the MGF bound of $S^{net}(s,t)$; however, the product property of MGFs can be still used for packet sizes.

As in the previous section we start with the packetized service model. Let a positive number θ_c such that $\theta_c < \mu - \lambda_c/C$, and denote

$$r_s(\theta_c) = C - \frac{\lambda_c}{\mu - \theta_c} \ .$$

It then follows that the function

$$\mathcal{T}^h(s,t) = r_s(\theta_c)(t-s)$$

is a statistical leftover service curve for the through flow at the h^{th} fluid server with error function $\varepsilon^s(\sigma) = e^{-\theta_c \sigma}$. This is a refinement of the result from Eq. (3) in Theorem 1 by accounting for the independent increments property of the compound arrival processes (see also [9]). Since the error function corresponding to the packetizer's service curve $S^{\mu,h}$ in Lemma 3 is zero, we further obtain for some $\tau_0 > 0$ that the function

$$\begin{aligned} S^h(s,t) &= \mathcal{T}^h * S^{\mu,h}(s, t-\tau_0) \\ &\geq \left[r_s(\theta_c)(t-s) - X_f^h(t) - r_s(\theta_c)\tau_0 \right]_+ \end{aligned}$$

is a statistical service curve for the through flow at the h^{th} node in the network from Figure 2 with error function $\varepsilon^s(\sigma)$ (as before, $X_f^h(t)$ denotes the fraction already processed of the packet currently in service at node h at time t).

Next we construct the statistical network service curve $S^{net}(s,t)$ as in Eq. (12) for the through flow along the H nodes. The corresponding error function is given as in Eq. (13) from Theorem 3, i.e.,

$$\varepsilon^{net}(\sigma) = \underbrace{\tilde{\varepsilon}^s_{\delta\tau_0} * \ldots * \tilde{\varepsilon}^s_{\delta\tau_0}}_{H-1 \text{ times}} * \varepsilon^s(\sigma) \ ,$$

where

$$\tilde{\varepsilon}^s_{\delta\tau_0}(\sigma) = \frac{1}{\delta\tau_0} \int_\sigma^\infty e^{-\theta_c u} du = \frac{1}{\theta_c \delta\tau_0} e^{-\theta_c \sigma} \ ,$$

for some $\delta > 0$. Using Lemma 3 from [11] we can optimize the expression of the error function as

$$\varepsilon^{net}(\sigma) = H \left(\frac{1}{\theta_c \delta\tau_0} \right)^{\frac{H-1}{H}} e^{-\frac{\theta_c}{H}\sigma} \ .$$

Next we have from Theorem 4 (more exactly Eq. (16)) that $S^{net}(s,t)$ has an MGF bound with rate and scaling factor given by

$$\begin{aligned} r^{net} &= r_s(\theta_c) - (H-1)\delta \ , \\ M^{net} &= n \left(\frac{\mu}{\mu - \theta} \right)^H e^{(2H-1)\theta r_s(\theta_c)\tau_0} e^{(H-1)\theta\delta\tau_0} \ , \end{aligned} \quad (21)$$

where $n = \binom{\lfloor \frac{t-s}{\tau_0} \rfloor + H - 1}{H-1}$.

Finally, having the through flow's rate envelope description from Eq. (17) and the network service curve just derived, we can invoke Theorem 2 and derive delay bounds. Let us first denote

$$r = r^{net} - r_a$$

and enforce the stability condition that

$$r > 0 \quad \Leftrightarrow \quad \delta < \frac{1}{H-1} \left(C - \frac{\lambda}{\mu - \theta} - \frac{\lambda_c}{\mu - \theta_c} \right) \ .$$

Then Eq. (8) from Theorem 2 gives the delay bound for all $\sigma \geq 0$

$$P\left(W^{net}(t) > \frac{\sigma}{r^{net}} \right) \leq \inf \left\{ M' \left(\frac{1}{\theta r \tau_0} \right)^H e^{-\theta\sigma^a} + \varepsilon^{net}(\sigma^s) \right\} \ ,$$

where the infimum is taken after $\sigma^a + \sigma^s = \sigma$; also, $M' = \left(\frac{\mu}{\mu - \theta} \right)^H e^{H(2\theta r_s(\theta_c) + \delta)\tau_0}$ is obtained by slightly relaxing the term after the binomial factor in Eq. (21).

We can optimize this expression using Lemma 3 from [11]. Then, replacing σ with $d \cdot r^{net}$, and letting $t \to \infty$ we obtain the steady-state delay bound for all $d \geq 0$

$$Pr\left(W^{net} > d \right) \leq K e^{-\frac{\theta\theta_c}{\alpha}\left(C - \frac{\lambda_c}{\mu - \theta_c} - (H-1)\delta \right) d} \ , \quad (22)$$

where

$$\begin{aligned} K &= \frac{\alpha}{\theta_c} \left(\frac{\mu}{\mu - \theta} \right)^{\frac{H\theta_c}{\alpha}} \left(\frac{He\theta_c \left(2r_s(\theta_c) + \delta \right)}{\beta r} \right)^{\frac{\beta}{\alpha}} \\ &\quad \cdot \left(\frac{r}{\delta} \right)^{\frac{(H-1)\theta}{\alpha}} \left(\frac{\theta_c}{\theta} \right)^{\frac{\theta}{\alpha}} \\ \alpha &= H\theta + \theta_c, \quad \beta = (H-1)\theta + H\theta_c \ . \end{aligned}$$

In the following we consider the fluid service model. As shown at the end of Subsection 3.1, the derivation of the corresponding bounds proceeds as before with the difference that the term $\frac{\mu}{\mu - \theta}$ is to be replaced by 1. Consequently, the steady-state delay bound takes the form

$$Pr\left(W^{net} > d \right) \leq K e^{-\frac{\theta\theta_c}{\alpha}\left(C - \frac{\lambda_c}{\mu - \theta_c} - (H-1)\delta \right) d} \ , \quad (23)$$

where

$$K = \frac{\alpha}{\theta_c} \left(\frac{He\theta_c \left(2r_s(\theta_c) + \delta \right)}{\beta r} \right)^{\frac{\beta}{\alpha}} \left(\frac{r}{\delta} \right)^{\frac{(H-1)\theta}{\alpha}} \left(\frac{\theta_c}{\theta} \right)^{\frac{\theta}{\alpha}} \ ,$$

whereas α and β are as above.

3.3 Independent arrivals / Identical service times

Here we assume that the arrival processes are independent, but that the sizes of each of the through packets are identical at the traversed nodes.

Following the same steps as in the previous two subsections, we first use Eq. (2) in Theorem 1 and Eq. (6) in Lemma 3, and obtain that the function

$$S^h(s,t) = \left[C(t-s) - A_h(s,t) \right]_+$$

is a statistical service curve for the through flow at the h^{th} node with error function $\varepsilon^h(\sigma) = e^{\mu C \tau_0} e^{-\mu\sigma}$, for some $\tau_0 > 0$. The service curve has an MGF bound with rate and scaling factor given by

$$r_s = C - r_c \ , \quad M_s = 1 \ ,$$

for some positive θ with $\theta < \mu - \frac{\lambda_c}{C}$.

Next we construct the statistical network service curve $S^{net}(s,t)$ as in Eq. (12) for the through flow along the H nodes. The corresponding error function is given as in Eq. (13) from Theorem 3, and can be written after optimizations with Lemma 3 from [11] as

$$\varepsilon^{net}(\sigma) = He^{\mu C\tau_0}\left(\frac{1}{\mu\delta\tau_0}\right)^{\frac{H-1}{H}} e^{-\frac{\mu}{H}\sigma} .$$

for some $\delta > 0$.

Furthermore, we have from Eq. (16) in Theorem 4 that $S^{net}(s,t)$ has an MGF bound with rate and scaling factor given by

$$r^{net} = r_s - (H-1)\delta , \quad M^{net} = e^{(H-1)\theta(r_s+\delta)\tau_0}\binom{\lfloor\frac{t-s}{\tau_0}\rfloor + H - 1}{H-1} .$$

Finally, having the through flow's rate envelope description from Eq. (17) and the network service curve just derived, we can invoke Theorem 2 and derive the delay bounds. First, let us denote

$$r = r^{net} - r_a$$

and enforce the stability condition that

$$r > 0 \quad \Leftrightarrow \quad \delta < \frac{1}{H-1}\left(C - \frac{\lambda+\lambda_c}{\mu-\theta}\right) .$$

Then Eq. (8) from Theorem 2 yields the following delay bound for all $\sigma \geq 0$

$$P\left(W^{net}(t) > \frac{\sigma}{r^{net}}\right) \leq \inf\left\{\left(\frac{e^{\theta(r_s+\delta)\tau_0}}{\theta r\tau_0}\right)^H e^{-\theta\sigma^a} + \varepsilon^{net}(\sigma^s)\right\},$$

where the infimum is taken after $\sigma^a + \sigma^s = \sigma$. We can optimize this using Lemma 3 from [11]. Then, replacing σ with $d \cdot r^{net}$, and letting $t \to \infty$ we obtain the steady-state delay bound for all $d \geq 0$

$$Pr\left(W^{net} > d\right) \leq Ke^{-\frac{\theta\mu}{\alpha}\left(C - \frac{\lambda_c}{\mu-\theta} - (H-1)\delta\right)d} , \qquad (24)$$

where

$$\begin{aligned} K &= \frac{\alpha}{\mu}\left(\frac{He\mu\,(C+r_s+\delta)}{\beta r}\right)^{\frac{\beta}{\alpha}}\left(\frac{r}{\delta}\right)^{\frac{(H-1)\theta}{\alpha}}\left(\frac{\mu}{\theta}\right)^{\frac{\theta}{\alpha}} \\ \alpha &= H\theta + \mu, \quad \beta = (H-1)\theta + H\mu . \end{aligned}$$

3.4 Correlated arrivals / Identical service times

Here we consider the most pessimistic scenario from this section, whereby we dispense with the statistical independence on arrivals and also assume that the sizes of each of the through packets are identical at each traversed node.

We first apply Eq. (3) in Theorem 1 to derive a statistical leftover service curve for the through flow at each fluid server h for $h = 1, \ldots, H$. The service curves are given for any $\delta > 0$ by the functions

$$T^h(t) = \left(C - \frac{\lambda_c}{\mu-\theta_c} - \delta\right)t ,$$

with error functions

$$\varepsilon^{T,h}(\sigma) = \frac{e^{\theta_c C\tau_0}}{\theta_c\delta\tau_0}e^{-\theta_c\sigma} .$$

Next, having the description of the packetizers from Eq. (6) in Lemma 3, we invoke Theorem 3 to derive a statistical network service curve for the concatenation between a fluid server and the corresponding packetizer. The service curve is given at each node h by the function

$$\begin{aligned} S^h(t) &= T^h * S^\mu_{-\delta}(t) \\ &= \inf_{0\leq s\leq t}\left\{\left(C - \frac{\lambda_c}{\mu-\theta_c} - \delta\right)s + (C-\delta)(t-s)\right\} \\ &= \left(C - \frac{\lambda_c}{\mu-\theta_c} - \delta\right)t . \end{aligned}$$

The corresponding error function is

$$\begin{aligned} \varepsilon^h(\sigma) &= \tilde{\varepsilon}^{T,h}_{\delta\tau_0} * \varepsilon^\mu(\sigma) \\ &= \inf_{\sigma_1+\sigma_2=\sigma}\left\{e^{\theta_c C\tau_0}\left(\frac{1}{\theta_c\delta\tau_0}\right)^2 e^{-\theta_c\sigma_1} + e^{\mu C\tau_0}e^{-\mu\sigma_2}\right\} \\ &= \frac{\theta_c+\mu}{\theta_c}\left(\left(\frac{e^{\theta_c C\tau_0}}{\theta_c\delta\tau_0}\right)^2\frac{\theta_c}{\mu}\right)^{\frac{\mu}{\theta_c+\mu}} e^{-\frac{\theta_c\mu}{\theta_c+\mu}\sigma} . \end{aligned}$$

In the last line we applied Lemma 3 from [11].

Next we can derive the statistical network service curve $S^{net}(t)$ for the whole network as in Eq. (12). Finally, using Theorem 2 and letting $t \to \infty$ we obtain for all $d \geq 0$

$$Pr\left(W^{net} > d\right) \leq Ke^{-\frac{\theta\theta_c\mu}{\gamma}\left(C - \frac{\lambda_c}{\mu-\theta_c} - H\delta\right)d} , \qquad (25)$$

where

$$\begin{aligned} K &= \frac{\gamma}{\theta}\left(\frac{2HCe\theta\mu}{\delta\beta}\right)^{\frac{\beta}{\gamma}}(\theta_c+\mu)^{\frac{(H-1)\theta(\theta_c+\mu)}{\gamma}} \\ &\quad \cdot\left(\frac{1}{\theta_c}\right)^{\frac{H\theta\theta_c}{\gamma}}\left(\frac{1}{\mu}\right)^{\frac{\beta-H\theta\mu}{\gamma}} , \\ \gamma &= H\theta(\theta_c+\mu) + \theta_c\mu \text{ , and} \\ \beta &= (3H-1)\theta\mu + (H-1)\theta\theta_c + \theta_c\mu . \end{aligned}$$

3.5 Considerations on dealing with (lack of) statistical independence

In Subsection 3.2 we briefly mentioned that the statistical independence of independent random variables is exploited by using the product property of MGFs. This is applied in conjunction with the Chernoff bound, i.e., for n independent r.v. X_i

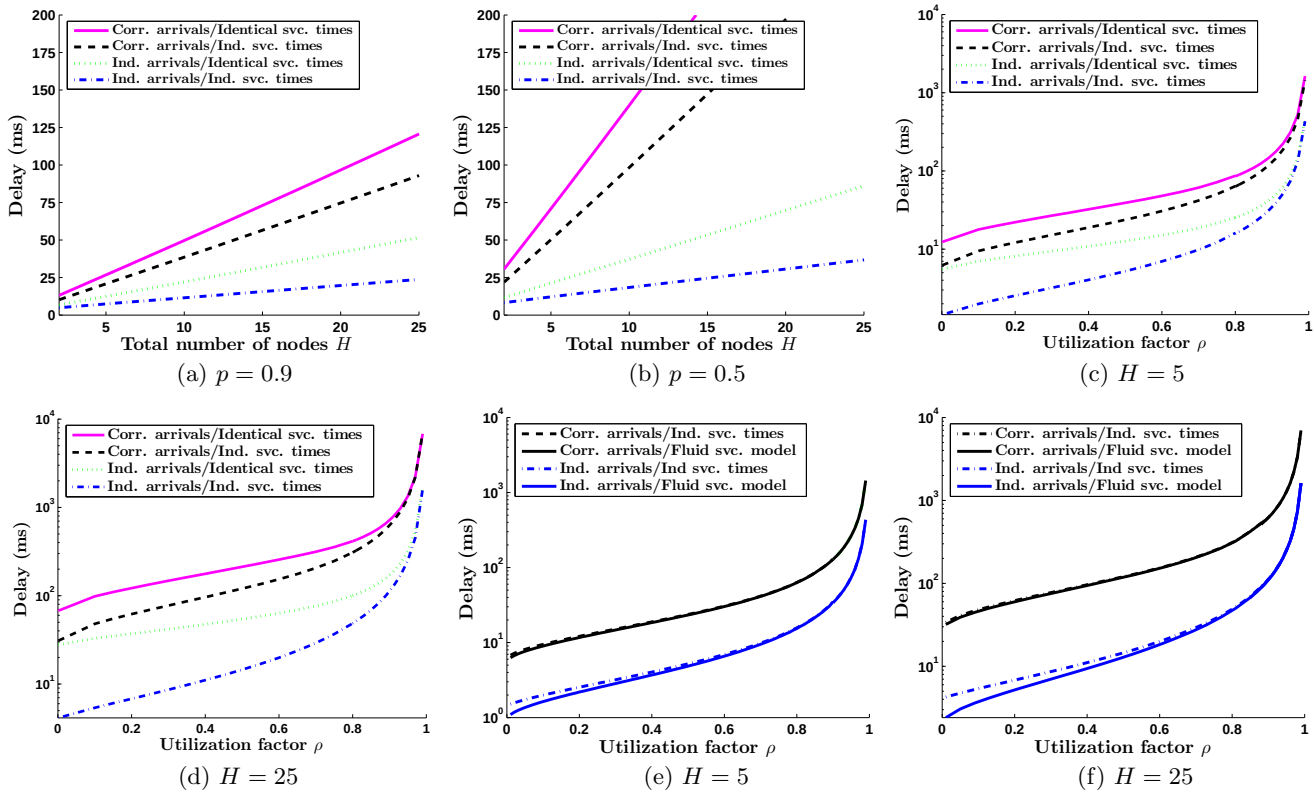
$$Pr\Big(\sum_{i=1}^n X_i > z\Big) \leq \prod_{i=1}^n E\left[e^{\theta X_i}\right]e^{-\theta z} , \qquad (26)$$

holds for all z and some $\theta > 0$. In Subsections 3.1-3.3 the X_i's can represent either arrival processes, or packet sizes, or both, depending on independence assumptions.

In the case of lack of statistical independence, then the left-hand side of Eq. (26) is bounded in turn by

$$\begin{aligned} Pr\Big(\sum_{i=1}^n X_i > z\Big) &\leq \inf_{\Sigma_{i=1}^n x_i = z}\sum_{i=1}^n Pr\left(X_i > x_i\right) \\ &\leq n\left(\prod_{i=1}^n E\left[e^{\theta X_i}\right]\right)^{\frac{1}{n}} e^{-\frac{\theta}{n}z} , \qquad (27) \end{aligned}$$

for all z and some $\theta > 0$. This is a worst case bound as it dispenses with any information on the joint distribution of X_i's. The inequality is used in Subsections 3.2-3.4. For instance, in Subsection 3.2, Eq. (27) is used with $n = H$ (i.e. the number of nodes) and X_i's representing the cross arrival processes.



(a) $p = 0.9$ (b) $p = 0.5$ (c) $H = 5$

(d) $H = 25$ (e) $H = 5$ (f) $H = 25$

Figure 3: The impact of relaxing the statistical independence assumptions of arrivals and packet sizes in the network from Figure 2 with Poisson arrivals and exponentially distributed packet sizes in (a), (b), (c), (d); the impact of using a fluid service model in (e), (f). End-to-end delays as functions of (1) the number of nodes H with utilization factor $\rho = 0.75$ and percentages of through traffic $p = 0.9$ and $p = 0.5$ in (a), (b), and (2) the utilization factor with percentage of through traffic $p = 0.5$ and number of nodes $H = 5$ and $H = 25$ in (c), (d), (e), (f) ($C = 100$ Mbps, average packet size $\mu^{-1} = 400$ Bytes, $z = 1 - 10^{-9}$)

Equating Eqs. (26) and (27) to some fixed violation probability ε and solving for z yields an $\mathcal{O}(\log n)$ increase in the latter case. Therefore, when $n = \mathcal{O}(H)$, the end-to-end delay bounds obtained with Eq. (27) (in Subsections 3.2-3.4) gain an $\mathcal{O}(\log H)$ factor when compared to those obtained only with Eq. (26) (in Subsection 3.1) and which grow as $\mathcal{O}(H)$. The $\mathcal{O}(H \log H)$ scaling behavior of end-to-end delays in scenarios with partial statistical independence assumptions was first reported in [11] for the fluid service model and later proven to be asymptotically tight in [5] for the packetized service model. It is still open whether the $\mathcal{O}(H \log H)$ growth of the delay bounds from Eq. (23) is asymptotically tight.

4. NUMERICAL RESULTS

In this section we numerically illustrate the behavior of the network calculus bounds derived in Subsections 3.1-3.4 when (1) using a packetized service model and relaxing the independence assumptions of arrivals and/or packet sizes, and (2) using the approximative fluid service model and relaxing the independence assumptions on arrivals.

First we consider the case of a packetized service model. In Figures 3.(a-d) we illustrate the bounds by relaxing the statistical independence assumptions of arrivals and/or packet sizes. In Figures 3.(a-b) we plot the end-to-end delay bounds as functions of the number of nodes H and consider two cases: (a) large amount of through traffic (percentage $p = 0.9$), and (b) medium amount of through traffic ($p = 0.5$). The plots correspond to Eqs. (19), (24), (22), and (25), respectively, in an increasing order of the bounds. The plots show that dispensing with the independence of packet sizes has a similar effect on the bounds for both independent and correlated arrivals. Dispensing with the independence assumption of arrivals has a much more noticeable effect in Figure 3.(b), due to the increase in the amount of cross traffic. The bounds obtained for correlated arrivals but independent packet sizes are more pessimistic than the bounds obtained for independent arrivals but identical packet sizes, i.e., correlations within arrivals have a more noticeable effect on the bounds than correlations within service.

Similar conclusions can be drawn from Figures 3.(c-d) which show the end-to-end delay bounds, on a logarithmic scale, as functions of the utilization factor ρ for two cases: (c) small number of nodes ($H = 5$), and (d) large number of nodes ($H = 25$). In both (c) and (d) we let the same percentage of through and cross traffic ($p = 0.5$). Remarkably, the two plots indicate that Kleinrock's independence assumptions is justified at high utilizations for both independent and correlated arrivals. Using simulations, this observation was also pointed out in the context of M/M/1 queueing networks with independent arrivals [16].

Finally, Figures 3.(e-f) illustrate the effects of dispensing with the packetized service model at the nodes. We consider a small number of nodes ($H = 5$) in (e) and a large number of nodes ($H = 25$) in (f). The two figures consider both correlated and independent arrivals, and show the end-to-end delay bounds as functions of the utilization factor. The plots correspond to Eqs. (20), (19), (23), and (22), in an increasing order of the bounds. For the case of correlated arrivals, the bounds obtained for the packetized and fluid service models closely match. A similar behavior is observed for independent arrivals, with the difference that the fluid model predicts more optimistic bounds than the packetized model but only at very low utilizations. The plots indicate that using a fluid service model is generally justified for both independent and correlated arrivals.

5. CONCLUSIONS

In this paper we have developed a stochastic network calculus by formulating the arrival and service models and analyzing the single-node and multi-node cases. This calculus generalizes existing formulations in the literature by providing a unified framework to deal with partial assumptions on the statistical independence of arrivals and service at the network nodes. This can be particularly useful in analyzing packet networks where the fact that each packet has the same size in the network creates subtle correlation among service at network nodes. We have applied our calculus to investigate the behavior of end-to-end delay bounds in a tandem network with high-priority cross traffic by relaxing the assumptions of independence of arrivals and packet sizes. Also, we have investigated the behavior of the bounds by using an approximative fluid service model for both cases of independent and correlated arrivals.

Acknowledgments

The author is grateful to Jörg Liebeherr and Almut Burchard for their contribution to [10] which contains most of the results presented in this paper.

6. REFERENCES

[1] R. Boorstyn, A. Burchard, J. Liebeherr, and C. Oottamakorn. Effective envelopes: Statistical bounds on multiplexed traffic in packet networks. In *Proceedings of IEEE Infocom 2000*, pages 1223–1232, Mar. 2000.

[2] A. Bose, X. Jiang, B. Liu, and G. Li. Analysis of manufacturing blocking systems with network calculus. *Performance Evaluation*, 63(12):1216–1234, 2006.

[3] J.-Y. Le Boudec. Some properties of variable length packet shapers. *IEEE/ACM Transactions on Networking*, 10(3):329–337, June 2002.

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

[5] A. Burchard, J. Liebeherr, and F. Ciucu. On $\Theta(H \log H)$ scaling of network delays. In *Proceedings of IEEE Infocom*, May 2007.

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

[7] C.-S. Chang. Stability, queue length, and delay of deterministic and stochastic queueing networks. *IEEE Transactions on Automatic Control*, 39(5):913–931, May 1994.

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

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

[10] F. Ciucu. *Scaling Properties in the Stochastic Network Calculus*. PhD thesis, University of Virginia, Charlottesville, VA, 2007.

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

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

[13] R. L. Cruz. Quality of service management in integrated services networks. In *1st Semi-Annual Research Review, CWC, University of California at San Diego*, June 1996.

[14] M. Fidler. An end-to-end probabilistic network calculus with moment generating functions. In *IEEE 14th International Workshop on Quality of Service (IWQoS)*, pages 261–270, June 2006.

[15] Y. Jiang. A basic stochastic network calculus. In *ACM Sigcomm*, pages 123–134, Sept. 2006.

[16] L. Kleinrock. *Communication Nets : Stochastic Message Flow and Delay*. McGraw-Hill, Inc., 1964.

[17] E. W. Knightly. Second moment resource allocation in multi-service networks. In *Proceedings of ACM Sigmetrics*, pages 181–191, June 1997.

[18] J. Liebeherr, S. Patek, and A. Burchard. Statistical per-flow service bounds in a network with aggregate provisioning. In *Proceedings of IEEE Infocom*, Mar. 2003.

[19] Y. Liu, C.-K. Tham, and Y. Jiang. A calculus for stochastic QoS analysis. *Performance Evaluation*, 64(6):547–572, 2007.

[20] A. K. Parekh and R. G. Gallager. A generalized processor sharing approach to flow control in integrated services networks: The multiple node case. *IEEE/ACM Transactions on Networking*, 2(2):137–150, April 1994.

[21] J.-L. Scharbarg, F. Ridouard, and C. Fraboul. A probabilistic analysis of end-to-end delays on an AFDX avionic network. *IEEE Transactions on Industrial Informatics*, 5(1):38–49, Feb 2009.

[22] M. Vojnovic and J.-Y. Le Boudec. Stochastic analysis of some expedited forwarding networks. In *Proceedings of IEEE Infocom*, pages 1004–1013, June 2002.

[23] E. Wandeler, A. Maxiaguine, and L. Thiele. Quantitative characterization of event streams in analysis of hard real-time applications. *Real-Time Systems*, 29(2-3):205–225, 2005.

[24] O. Yaron and M. Sidi. Performance and stability of communication networks via robust exponential bounds. *IEEE/ACM Transactions on Networking*, 1(3):372–385, June 1993.