A Guide to the Stochastic Network Calculus

Markus Fidler, *Senior Member, IEEE*, and Amr Rizk, *Member, IEEE*

***Abstract*—The aim of the stochastic network calculus is to comprehend statistical multiplexing and scheduling of non-trivial traffic sources in a framework for end-to-end analysis of multi-node networks. To date, several models, some of them with subtle yet important differences, have been explored to achieve these objectives. Capitalizing on previous works, this paper contributes an intuitive approach to the stochastic network calculus, where we seek to obtain its fundamental results in the possibly easiest way. In detail, the method that is assembled in this work uses moment generating functions, known from the theory of effective bandwidths, to characterize traffic arrivals and network service. Thereof, affine envelope functions with an exponentially decaying overflow profile are derived to compute statistical end-to-end backlog and delay bounds for networks.**

***Index Terms*—Stochastic network calculus, end-to-end performance evaluation, moment generating functions, scheduling.**

I. Introduction

THE network calculus emerged during the 90s as a deterministic theory for quality of service analysis of packet data networks. Traffic arrivals at a networked system are modelled by upper envelope functions [2]. Minimum service guarantees that are provided by systems, such as a router, a scheduler, or a link, are characterized by so-called service curves [3]. Based on these concepts, the network calculus offers convolution forms [4], [5] that enable the derivation of worst-case performance bounds including backlog and delay. A key advantage of the convolution-based framework is that it extends immediately to networks. Any number of systems in series can be transformed into a single equivalent system by convolution of the individual systems' service curves.

A shortcoming of the deterministic model is that it generally considers the worst-case and hence, it cannot take advantage of the statistical nature of traffic flows [6]. Statistical multiplexing of traffic flows is dealt with efficiently by the theory of effective bandwidths [4], [7] that uses moment generating functions (MGFs) as a model of data traffic. The inclusion of statistical traffic models into a convolution-based framework for end-to-end analysis of networks has motivated considerable research already in the early stages of the network calculus. Since then, two basic traffic models became widely accepted, that are envelopes of MGFs [4], [8] and statistical envelopes [6], [9]–[11], respectively. Statistical envelopes relax the deterministic envelope model by allowing a violation of the envelope with a defined, small probability. Statistical envelopes follow from MGFs by use of Chernoff's bound [6], [9].

Despite the early interest in a stochastic version of the network calculus, end-to-end convolution forms for networks of systems with random service remained an open challenge for some years. The difficulty is due to the fact that the convolution evaluates entire sample paths of the traffic arrival process. Thus, it requires a statistical guarantee for sample paths that is difficult to achieve. End-to-end convolution forms for networks of systems with random service have been obtained in [12] using the statistical envelope model. Performance bounds derived thereof grow as $\Theta(n \log n)$ for n systems in series [13]. Convolution forms that are based on MGFs are established in [4], [14]. Compared to the use of statistical envelopes, MGFs utilize the additional assumption of statistical independence. Respective end-to-end performance bounds scale in $\mathcal{O}(n)$ [14].

Compared to classical queueing theory, the stochastic network calculus comprises a much larger variety of stochastic processes, including long range dependent, self-similar [15], [16], and heavy-tailed traffic [16]. This generality comes at the expense of exact solutions. Instead, the stochastic network calculus computes non-asymptotic statistical performance bounds of the type $\mathsf{P}[\text{backlog} > x] \leq \varepsilon$ or $\mathsf{P}[\text{delay} > x] \leq \varepsilon$.

With this work, we aim at an intuitive introduction to the stochastic network calculus. We seek to define a minimal framework that enables deriving the essential results of the stochastic network calculus, in particular considering networks of tandem systems. We contribute a self-contained exposition of basic methods and closed form results derived thereof. We provide frequent references for further reading, that are intended to be optional for understanding of this tutorial. To simplify the approach, we restrict the presentation to affine envelope functions of MGFs [8] and corresponding linear statistical envelope functions with an exponentially bounded burstiness (EBB) [9]. Also, we will occasionally forgo generality or precision in favor of simplicity. For more general envelope models as well as models that provide stronger guarantees, we refer the reader to the related literature, e.g., [6], [17]–[20].

Concerning the two established textbooks on the network calculus [4], [5] from 2000 and 2001, respectively, [5] focuses on the deterministic network calculus and [4] phrases stochastic tandem systems, that are essential to this work, as a problem. Stochastic end-to-end convolution results for tandem systems have been reported first in [12], [14] and have shortly afterwards been included in the textbook on the stochastic network calculus [17] from 2008. Since then, the authors of this tutorial have taught an annual course on the network calculus from which

Manuscript received October 25, 2013; revised April 17, 2014; accepted June 20, 2014. Date of publication July 31, 2014; date of current version March 13, 2015. This work was supported in part by an Emmy Noether grant from the German Research Foundation (DFG) and in part by a Starting Grant of the European Research Council (ERC). This paper was presented in part [1] at the MMBnet Workshop'13 of the German Informatics Society, Hamburg, Germany, September 2013.

The authors are with the Institute of Communications Technology, Leibniz Universität Hannover, 30167 Hannover, Germany (e-mail: markus.fidler@ikt.uni-hannover.de; amr.rizk@ikt.uni-hannover.de).

Digital Object Identifier 10.1109/COMST.2014.2337060

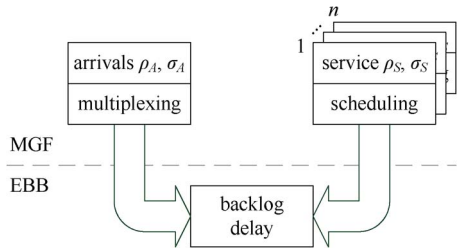

Fig. 1. Composition of arrival and service models.

the consolidated and comparably simpler approach taken in this paper emerged.

An outline of the method, that is assembled in this work, is shown in Fig. 1. We use the affine MGF envelope model from [8] to characterize arrival processes and service processes, respectively. The model has two parameters, ρ that is an envelope rate and σ that is a burstiness measure. Subscript A refers to the arrivals and S to the service, respectively. Formulas for statistical multiplexing, scheduling, and convolution of n tandem systems will be provided for this model. In the next step a transition from MGFs to the EBB model is performed using Chernoff's bound. Finally the EBB characterizations of the arrivals and of the service, respectively, are composed to compute performance bounds for backlog and delay.

We note, that a duality of MGFs and statistical envelopes exists [6]. A transition from MGFs to statistical bounding functions, such as EBB envelopes, can be made basically after any of the steps depicted in Fig. 1. Certain results can, however, be derived more easily using one or the other model. A representative example is statistical multiplexing that takes advantage of statistically independent traffic flows. Many application scenarios, such as the aggregation of flows on Internet backbone links or multiple user scheduling in wireless networks, provide reasonable grounds for the assumptions of statistically independent traffic and/or service. While the consideration of statistical multiplexing is straight-forward in case of MGFs, the EBB model is favorable in the absence of statistical independence [9], [21].

The remainder of this paper is structured as follows. In Section II, we introduce the basic queueing model of the network calculus and describe how EBB envelopes can be derived from MGFs of arrival and service processes, respectively. Backlog and delay bounds follow immediately by composition of the EBB envelopes of arrivals and service. Further, we derive a corresponding EBB result for tandem systems that is novel. In Section III, we provide a catalogue of MGF envelopes for relevant arrival and server models. We also include results for statistical multiplexing and scheduling. Section IV concludes the paper with a set of guidelines for application and an outlook. To support the applicability, we highlight a self-contained set of final results throughout the paper, using framed equations.

II. Fundamentals

This section provides an introduction to the stochastic network calculus. In Section II-A, we formulate the basic queueing model, where traffic that arrives at a system between times τ and t denoted $A(\tau,t)$, the service offered by the system $S(\tau,t)$, and the departures from the system $D(\tau,t)$ are bivariate random processes. Bivariate functions are used to express the time-varying nature of traffic and service. In Section II-B and C, statistical envelope functions of arrivals and service, respectively, are derived. The envelopes are bounds that may be violated with a defined probability. For ease of exposition, we restrict the envelopes to affine functions with an exponentially decaying violation probability. By assumption of stationarity, the envelopes become time-invariant and hence are expressed by univariate functions. In Section II-D, we show how to extend the method to multi-node networks. Section II-E provides simple expressions for backlog and delay bounds that follow from the envelope model.

A. Queueing Model

Throughout this work, we assume time is discrete, i.e., $t \in \mathbb{N}_0$. Continuous time requires an additional discretization step, see [12]. We denote $A(t)$ the cumulative number of bits arriving at a system in the time interval $[0,t]$. Clearly, $A(t)$ is a non-negative, non-decreasing function, and by convention $A(0) = 0$. We use shorthand notation $A(\tau,t) = A(t) - A(\tau)$ where $t \geq \tau \geq 0$ to denote the arrivals in $[\tau+1, t]$. Trivially, $A(t,t) = 0$ for all $t \geq 0$. Similarly, $D(t)$ denotes the cumulative departures from the system.

We characterize systems using the concept of a dynamic server [4], [22] that relates the departures of a system to its arrivals as

$$D(t) \geq \min_{\tau \in [0,t]} \{A(\tau) + S(\tau,t)\} =: A \otimes S(t), \qquad (1)$$

where $S(\tau,t)$ for $t \geq \tau \geq 0$ is a random process that defines the service offered by the system. By convention, $S(\tau,t)$ is non-negative and $S(t,t) = 0$ for all $t \geq 0$.

An example of a system that satisfies the definition of dynamic server is a lossless, work-conserving server with a time-varying capacity $S(\tau,t)$, where $S(\tau,t)$ denotes the service that is available in $[\tau+1, t]$ [4]. To see this, assume t and $\tau+1$ fall into the same busy period, such that the system is continuously backlogged during $[\tau+1, t]$. Combined with the assumption of a work-conserving system, it follows that the entire service that is available in $[\tau+1, t]$ is consumed, such that

$$D(t) = D(\tau) + S(\tau,t).$$

Now, fix $\tau = \tau^*$ to be the beginning of the last busy period before t, i.e., at τ^* the system was empty for the last time before t. Consequently, $D(\tau^*) = A(\tau^*)$ and

$$D(t) = A(\tau^*) + S(\tau^*,t). \qquad (2)$$

Finally, since τ^* is not generally known, we use the estimate

$$D(t) \geq \min_{\tau \in [0,t]} \{A(\tau) + S(\tau,t)\}$$

which proves that the system is a dynamic server (1). Moreover, generally $D(t) \leq D(\tau) + S(\tau,t)$ for $t \geq \tau \geq 0$ as the departures in $[\tau+1, t]$ cannot exceed the service. From causality we have $D(\tau) \leq A(\tau)$, such that

$$D(t) \leq A(\tau) + S(\tau,t)$$

for all $\tau \in [0,t]$ and consequently

$$D(t) \leq \min_{\tau\in[0,t]} \{A(\tau) + S(\tau,t)\}.$$

Combined with (1), it follows that

$$D(t) = \min_{\tau\in[0,t]} \{A(\tau) + S(\tau,t)\}, \tag{3}$$

i.e., the system is an exact dynamic server as (3) satisfies (1) with equality. The example of the work-conserving server with a time-varying capacity proves that the lower bound (1) is actually attained. We note that (3) implies linearity of the system, see [5] for details. Non-linear systems, such as a first-in first-out scheduler [23], satisfy only the more general definition of dynamic server (1).

The operator $\otimes$ that is defined by (1) is known as the convolution under a min-plus algebra. The min-plus algebra can be regarded similar to the traditional algebra, where the minimum takes the place of the addition and the addition takes the place of the multiplication. This analogy is highly useful, as the network calculus can be viewed as a min-plus systems theory that inherits many useful properties of the classical convolution from linear systems theory [4], [5]. Among others, the min-plus convolution is associative, which enables an elegant composition of tandem systems. Consider two dynamic servers $S_1(\tau,t)$ and $S_2(\tau,t)$ in series. We use the same indices to denote the arrivals and departures of the respective systems, where $A_2(t) = D_1(t)$. By recursive insertion of (1) and by use of the associativity it holds that

$$D_2(t) \geq (A_1 \otimes S_1) \otimes S_2(t) = A_1 \otimes (S_1 \otimes S_2)(t). \tag{4}$$

As a main result, it follows that

$$S(\tau,t) = S_1 \otimes S_2(\tau,t) := \min_{\upsilon\in[\tau,t]} \{S_1(\tau,\upsilon) + S_2(\upsilon,t)\}$$

satisfies the definition of dynamic server (1), i.e., the tandem of the two systems can be substituted by a single equivalent system $S(\tau,t)$ that is composed by min-plus convolution of the individual service processes. By repeated iteration it follows that a network of n dynamic servers has an equivalent single server representation with service process

$$S_{\text{net}}(\tau,t) = S_1 \otimes S_2 \otimes \cdots \otimes S_n(\tau,t). \tag{5}$$

Since by convention $S_i(\tau,t) \geq 0$ for $t \geq \tau \geq 0$, it holds that $S_{\text{net}}(\tau,t) \geq 0$, too. As a consequence of the associativity of min-plus convolution, (5) enables to apply any result obtained for a single dynamic server to networks of dynamic servers. Finally, we note that while the min-plus convolution of univariate functions is commutative, it is not commutative in case of bivariate functions, i.e., the order of the individual dynamic servers is relevant.

We conclude this section with basic backlog and delay bounds. The backlog at time $t \geq 0$ is defined as

$$B(t) = A(t) - D(t).$$

The definition of backlog comprises bits that are stored in buffers as well as bits that are in transmission. By insertion of (1) a backlog bound that is achieved by a dynamic server follows immediately as

$$B(t) \leq \max_{\tau\in[0,t]} \{A(\tau,t) - S(\tau,t)\}. \tag{6}$$

Assuming first-come first-served (FCFS) order, the delay at time $t > 0$ is defined as

$$W(t) = \min\{\omega \geq 0 : A(t) - D(t+\omega) \leq 0\}.$$

Note that the definition of delay is not conditioned on an actual data departure but only on the time instance t, e.g., if the system is empty at t, i.e., $B(t) = 0$, then the delay is also $W(t) = 0$. By insertion of (1) a delay bound is

$$W(t) \leq \min\left\{\omega \geq 0 : \max_{\tau\in[0,t]} \{A(\tau,t) - S(\tau,t+\omega)\} \leq 0\right\}.$$

Backlog and delay both have an intuitive graphical representation, where the backlog is the vertical deviation and the delay the horizontal deviation of arrivals and departures, respectively.

B. Arrival Envelopes

To compute actual backlog and delay bounds, the deterministic network calculus uses univariate envelope functions that are defined as deterministic upper bounds of the arrivals $A(\tau,t)$ for all time intervals $[\tau+1,t]$ with $t \geq \tau \geq 0$. A widely applied model are affine envelope functions defined as $\rho(t-\tau)+b$, that are enforced by a leaky bucket shaper with rate $\rho > 0$ and burst parameter $b \geq 0$ [2]. The arrivals have a deterministic affine envelope if for all $t \geq \tau \geq 0$ it holds that

$$A(\tau,t) \leq \rho(t-\tau) + b. \tag{7}$$

Performance bounds are derived by substitution of the envelope as an upper bound for $A(\tau,t)$. As an example, a backlog bound for a work-conserving constant rate server with capacity $c > \rho$ follows by insertion of $S(\tau,t) = c(t-\tau)$ and (7) into (6) as $B(t) \leq b$ for all $t \geq 0$.

While its application is intuitive, a drawback of the deterministic envelope model is that it generally considers the worst-case. As a consequence, it cannot take advantage of the statistical nature of traffic.

Stochastic traffic models, such as the theory of effective bandwidths [4], [7], make extensive use of MGFs. MGFs uniquely determine the distribution of a random process and have the convenient property that the MGF of the sum of two or more random processes is the product of their respective MGFs. The MGF of an arrival process $A(\tau,t)$ is defined as $\mathsf{E}[e^{\theta A(\tau,t)}]$ with free parameter $\theta \geq 0$. Under the assumption of stationarity, i.e., $\mathsf{P}[A(\tau,t) \leq x] = \mathsf{P}[A(t-\tau) \leq x]$ for all $t \geq \tau \geq 0$, the MGF becomes a univariate function that depends only on the time difference $t-\tau$. We will frequently use shorthand notation $\mathsf{M}_A(\theta, t-\tau) = \mathsf{E}[e^{\theta A(\tau,t)}]$. The normalized log MGF $\ln \mathsf{M}_A(\theta,t)/(\theta t)$ is known as the effective bandwidth. For increasing $\theta > 0$ it grows from the mean rate to the peak rate

of the arrivals. Corresponding to the affine envelope model, [4], [8] define an MGF envelope for $t \geq \tau \geq 0$ as

$$\mathsf{E}\left[e^{\theta A(\tau,t)}\right] \leq e^{\theta(\rho(t-\tau)+\sigma)} \tag{8}$$

where the parameters $\rho > 0$ and $\sigma \geq 0$ are functions of $\theta \geq 0$.

A related statistical envelope, referred to as exponentially bounded burstiness (EBB), is defined in [9] to provide a guarantee of the form

$$\mathsf{P}\left[A(\tau,t) > \rho(t-\tau)+b\right] \leq \varepsilon(b) \tag{9}$$

for $t \geq \tau \geq 0$. The model relaxes the deterministic envelope (7) with parameters $\rho > 0$ and $b \geq 0$ by defining an overflow profile $\varepsilon(b) \geq 0$ that decays exponentially as

$$\varepsilon(b) = \alpha e^{-\theta b}, \tag{10}$$

where $\alpha \geq 0$.

Generalizations of the EBB model, that include different shapes of envelope functions and overflow profiles, have been provided, e.g., in [10]–[12], [20], [24], see also the survey on envelopes [18]. In general, the linear rate term $\rho \cdot (t-\tau)$ in (9) can be replaced by a non-negative, non-decreasing envelope function $E(t-\tau)$, respectively, the overflow profile $\varepsilon(b)$ can in general be a non-negative, non-increasing function with finite sum. The more general definition includes a larger set of traffic models. Also, it may improve the tightness of bounds in certain cases. For ease of exposition, we limit ourselves to linear rate EBB envelopes. We note that the basic steps of the following sample path derivations are essentially unaffected by the choice of the envelope model. We show an example of a non EBB envelope in Section III-A3.

The EBB model is directly connected to the MGF envelope by Chernoff's bound

$$\mathsf{P}[X \geq x] \leq e^{-\theta x}\mathsf{E}[e^{\theta X}] \tag{11}$$

for $\theta \geq 0$. By application of (11) to (9) and insertion of (8) it follows that $\mathsf{P}[A(\tau,t) > \rho(t-\tau)+b] \leq e^{\theta\sigma}e^{-\theta b}$. We equate the right hand side with $\varepsilon(b)$ to obtain

$$\varepsilon(b) = e^{\theta\sigma}e^{-\theta b} \tag{12}$$

that is EBB with parameter $\alpha = e^{\theta\sigma}$.

While the EBB model (9) is a natural statistical extension of (7), an important difference arises with respect to the computation of performance bounds, such as the backlog bound (6): The deterministic envelope (7) can be immediately substituted for $A(\tau,t)$ in (6), however, the EBB envelope (9) cannot. The reason is that (6) evaluates all $\tau \in [0,t]$, where the $\tau = \tau^*$ that attains the maximum is a random variable [6]. In contrast, (9) only provides a guarantee for an arbitrary, yet, fixed $\tau \in [0,t]$. To overcome this problem, a sample path argument similar to [6], [10]–[12] is required that has the form

$$\mathsf{P}\left[\exists \tau \in [0,t] : A(\tau,t) > \rho'(t-\tau)+b\right] \leq \varepsilon'(b) \tag{13}$$

for all $t \geq 0$. Throughout this work we use superscript ε' to denote sample path overflow probabilities of the type (13). Note that in the deterministic case (7) no such distinction exists.

To estimate $\varepsilon'(b)$ from (13), one can rewrite $\mathsf{P}[\exists i : X_i \geq x] = \mathsf{P}[\max_i\{X_i\} \geq x]$ and approximate the expression by its largest term as

$$\mathsf{P}\left[\max_i\{X_i\} \geq x\right] \geq \max_i\left\{\mathsf{P}[X_i \geq x]\right\}. \tag{14}$$

Note, however, that the expression only provides a lower bound of an upper bound [6]. As a consequence, one can only approximate $\varepsilon'(b) \approx \varepsilon(b)$ for $\rho' = \rho$. A true upper bound, on the other hand, follows by use of the union bound as

$$\mathsf{P}[\exists i : X_i \geq x] \leq \sum_i \mathsf{P}[X_i \geq x]. \tag{15}$$

Regarding (13), it follows as, e.g., in [11], [12] that

$$\begin{aligned}
&\mathsf{P}\left[\exists \tau \in [0,t] : A(\tau,t) > \rho'(t-\tau)+b\right] \\
&\leq \sum_{\tau=0}^{t} e^{\theta\sigma}e^{-\theta(b+\delta(t-\tau))} \\
&\leq e^{\theta\sigma}e^{-\theta b}\sum_{\tau=0}^{\infty} e^{-\theta\delta\tau} \\
&= \frac{e^{\theta\sigma}e^{-\theta b}}{1-e^{-\theta\delta}}.
\end{aligned}$$

In the second line, we used the union bound[1] (15) and substituted (12) for the expression (9) where we let $\rho' = \rho + \delta$. Parameter $\delta > 0$ can be viewed as a slack rate that is used to achieve geometrically decaying summands. Increasing parameter δ increases the envelope rate and decreases the overflow profile. In the third line, we let $t \to \infty$ to compute a steady-state bound. In the fourth line, we used that $\theta\delta > 0$ and solved the geometric sum.

Concluding, given arrivals that have MGF envelope (8) with parameters ρ and σ, the sample path envelope (13) is EBB with envelope rate $\rho' = \rho + \delta$ and overflow profile

$$\varepsilon'(b) = \frac{e^{\theta\sigma}}{1-e^{-\theta\delta}}e^{-\theta b}, \tag{16}$$

where $\theta > 0$ and $\delta > 0$ are free parameters that can be optimized.

The utility of the EBB sample path envelope (13) is due to the fact that it can be used to substitute $\rho'(t-\tau)+b$ for the arrival process $A(\tau,t)$ in performance bounds such as the backlog bound (6). To give a first example, we consider a work-conserving constant rate server with capacity c. By insertion of $S(\tau,t) = c(t-\tau)$ into (6) and using the EBB sample path envelope (13) with envelope rate $\rho' = c$ and overflow profile (16) the statistical backlog bound

$$\mathsf{P}[B(t) > b] \leq \frac{e^{\theta\sigma}}{1-e^{-\theta\delta}}e^{-\theta b}$$

follows for all $t \geq 0$. Parameter $\delta > 0$ is determined as $\delta = c - \rho$ under the stability condition $\rho < c$. The free parameter $\theta > 0$ can be optimized to minimize the right-hand side. The remaining parameters ρ and σ are characteristics of the arrival

[1] We note that the summand at $\tau = t$ can be omitted to improve the precision as $A(t,t) = 0$ by definition.

process. Solutions for relevant traffic sources will be provided in Section III. In the following Section II-C, we will derive a similar substitution for random service processes $S(\tau,t)$.

C. Service Envelopes

We start from the definition of dynamic server (1), that defines service as a random process $S(\tau,t)$, and use the basic methods from Section II-B to derive lower envelopes thereof. We note, that a significant part of the network calculus literature is based on a notion of statistical service curves that characterize the service by non-random functions, e.g., [6], [10], [12], [17]. Statistical service curves are connected to envelopes of random service processes in [25].

A deterministic definition of service envelope for all $t \geq \tau \geq 0$ is

$$S(\tau,t) \geq \rho(t-\tau) - b$$

that defines a lower bound of the service process with parameters $\rho > 0$ and $b \geq 0$. Since by convention $S(\tau,t) \geq 0$, we can also write $S(\tau,t) \geq \rho[t-\tau-b/\rho]_+$, where the notation $[x]_+$ denotes $\max\{0,x\}$. The quotient b/ρ has the interpretation of a worst-case latency up to which the service may be zero.

A service characterization using MGFs is known in analogy to the effective bandwidth as effective capacity [26]. The model uses the negative MGF, i.e., with parameter $-\theta$ for $\theta \geq 0$ that is also known as the Laplace transform. An affine envelope of the MGF can be defined for $\theta \geq 0$ as

$$\mathsf{E}\left[e^{-\theta S(\tau,t)}\right] \leq e^{-\theta(\rho(t-\tau)-\sigma)}. \tag{17}$$

Note that although (17) is phrased as an upper bound, it defines a lower bound of the service due to the use of $-\theta$ where $\theta \geq 0$. Also, the parameters ρ and σ are functions of $-\theta$. Assuming stationarity of the service process $S(\tau,t)$, we will frequently use shorthand notation $\mathsf{M}_S(-\theta,t-\tau) = \mathsf{E}[e^{-\theta S(\tau,t)}]$. The normalized log MGF $\ln \mathsf{M}_S(-\theta,t)/(-\theta t)$ is known as the effective capacity. It decreases for increasing $\theta > 0$ from the mean rate to the minimum rate of the service.

Statistical service envelopes that mirror the concept of EBB are defined in [27] as the so-called exponentially bounded fluctuation (EBF) model with parameters $\rho > 0$, $b \geq 0$ and

$$\mathsf{P}\left[S(\tau,t) < \rho(t-\tau) - b\right] \leq \varepsilon(b), \tag{18}$$

where the deficit profile $\varepsilon(b)$ decays exponentially as $\varepsilon(b) = \alpha e^{-\theta b}$ and $\alpha \geq 0$. With Chernoff's lower bound

$$\mathsf{P}[X \leq x] \leq e^{\theta x}\mathsf{E}[e^{-\theta X}] \tag{19}$$

for $\theta \geq 0$ it follows from (17) that $\varepsilon(b) = e^{\theta\sigma}e^{-\theta b}$. Finally, the sample path envelope

$$\mathsf{P}\left[\exists \tau \in [0,t] : S(\tau,t) < \rho'(t-\tau) - b\right] \leq \varepsilon'(b) \tag{20}$$

with $\rho' = \rho - \delta$ and free parameters $\delta > 0$ and $\theta > 0$ is EBF with deficit profile

$$\varepsilon'(b) = \frac{e^{\theta\sigma}}{1-e^{-\theta\delta}}e^{-\theta b}. \tag{21}$$

The derivation uses the union bound[2] (15) and the same basic steps as in Section II-B. The free parameters $\theta > 0$ and $\delta > 0$ can be optimized.

D. Convolution-Form Networks

Due to the associativity of min-plus convolution, the network calculus can abstract a multi-node network by a single equivalent system. The corresponding service process is obtained by min-plus convolution of the individual service processes (5), giving rise to the name convolution-form networks [28]. Regarding statistical envelope functions, the recursive insertion of the departures of the first server as the arrivals of the second server (4) causes, however, additional difficulties. The reason is that the min-plus convolution evaluates sample paths of the arrivals of a server and hence requires sample path guarantees for the departures of the preceding server, see [6]. First end-to-end solutions that make use of the convolution-form appeared in the stochastic network calculus in [12], [14].

In the sequel, we derive the EBF deficit profile first for two and then by recursive insertion for n dynamic servers in tandem. The EBF result for tandem dynamic servers is novel compared to the literature [12], [14]. For the MGF of the min-plus convolution of two statistically independent and stationary service processes it is known that [4], [14]

$$\begin{aligned}\mathsf{E}\left[e^{-\theta(S_1\otimes S_2)(\tau,t)}\right] &= \mathsf{E}\left[e^{-\theta\min_{\upsilon\in[\tau,t]}\{S_1(\tau,\upsilon)+S_2(\upsilon,t)\}}\right]\\ &\leq \sum_{\upsilon=\tau}^{t}\mathsf{E}\left[e^{-\theta S_1(\tau,\upsilon)}\right]\mathsf{E}\left[e^{-\theta S_2(\upsilon,t)}\right]\\ &= \sum_{\upsilon=0}^{t-\tau}\mathsf{M}_{S_1}(-\theta,\upsilon)\mathsf{M}_{S_2}(-\theta,t-\tau-\upsilon)\\ &=: \mathsf{M}_{S_1} * \mathsf{M}_{S_2}(-\theta,t-\tau).\end{aligned}$$

In the second line, the expectation of a maximum is estimated by the sum of the individual terms. The step corresponds to the use of the union bound (15). Then, under the assumption of independence, the MGF of the sum of two random processes is the product of the individual MGFs. For stationary random processes we finally obtain the univariate convolution in classical algebra. The MGF of the service process of an n node network follows by recursive insertion as

$$\begin{aligned}&\mathsf{M}_{S_{\text{net}}}(-\theta,t) \leq \mathsf{M}_{S_1} * \mathsf{M}_{S_2} * \cdots * \mathsf{M}_{S_n}(-\theta,t)\\ &= \sum_{\tau_i\geq 0:\sum_{i=1}^{n}\tau_i=t} \mathsf{M}_{S_1}(-\theta,\tau_1)\mathsf{M}_{S_2}(-\theta,\tau_2)\ldots\mathsf{M}_{S_n}(-\theta,\tau_n).\end{aligned} \tag{22}$$

Next, we assume homogeneous MGF envelopes (17). We also provide a solution for the heterogeneous case, that basically requires additional notation. Since the convolution is order preserving, we can substitute $\mathsf{M}_{S_i}(-\theta,t) \leq e^{\theta\sigma}e^{-\theta\rho t}$ for

[2]While we apply the union bound for $\tau = 0,1,\ldots,t$, we note that the precision can be improved if the sum is computed only for $\tau = 0,1,\ldots,t-\lfloor b/\rho' \rfloor$ since $S(\tau,t)$ is non-negative.

$i = 1, 2, \ldots, n$. The sum in (22) has $\binom{t+n-1}{n-1}$ summands as there are $\binom{t+n-1}{n-1}$ different non-negative vectors $(\tau_1, \tau_2, \ldots \tau_n)$ that satisfy $\sum_{i=1}^{n} \tau_i = t$ [29], [30]. It follows that

$$\mathsf{M}_{S_{\text{net}}}(-\theta, t) \leq e^{n\theta\sigma} \binom{t+n-1}{n-1} e^{-\theta\rho t}.$$

The deficit profile of an envelope of the type of (18) for $S_{\text{net}}(\tau, t)$ follows from Chernoff's bound (19) for $\theta \geq 0$ as

$$\begin{aligned}\mathsf{P}\left[S_{\text{net}}(\tau, t) < \rho(t-\tau) - b\right] &\leq e^{\theta(\rho(t-\tau)-b)} \mathsf{M}_{S_{\text{net}}}(-\theta, t-\tau) \\ &\leq e^{n\theta\sigma} \binom{t-\tau+n-1}{n-1} e^{-\theta b}.\end{aligned}$$

Using the same basic approach as in Section II-B and C, a sample path envelope (20) can be derived as

$$\begin{aligned}&\mathsf{P}\left[\exists \tau \in [0, t] : S_{\text{net}}(\tau, t) < \rho'(t-\tau) - b\right] \\ &\leq \sum_{\tau=0}^{t} e^{n\theta\sigma} \binom{t-\tau+n-1}{n-1} e^{-\theta(b+\delta(t-\tau))} \\ &\leq e^{n\theta\sigma} e^{-\theta b} \sum_{\tau=0}^{\infty} \binom{\tau+n-1}{n-1} e^{-\theta\delta\tau} \\ &= \frac{e^{n\theta\sigma} e^{-\theta b}}{(1-e^{-\theta\delta})^n} \sum_{\tau=0}^{\infty} \binom{\tau+n-1}{n-1} \left(e^{-\theta\delta}\right)^{\tau} \left(1-e^{-\theta\delta}\right)^n \\ &= \frac{e^{n\theta\sigma} e^{-\theta b}}{(1-e^{-\theta\delta})^n}.\end{aligned}$$

In the second line, we used the union bound[3] (15) and substituted $\rho' = \rho - \delta$ where $\delta > 0$ and $\theta > 0$ are free parameters. In the third line, we let $t \to \infty$ to compute a steady-state bound. In the fourth line, we arrange terms such that the summands become the negative binomial probability mass function since $\theta\delta > 0$.

Finally, we conclude that the network service process $S_{\text{net}}(\tau, t)$ conforms to the sample path envelope (20) with envelope rate $\rho' = \rho - \delta$ and EBF deficit profile

$$\varepsilon'(b) = \left(\frac{e^{\theta\sigma}}{1-e^{-\theta\delta}}\right)^n e^{-\theta b}. \tag{23}$$

As before, $\delta > 0$ and $\theta > 0$ are free parameters. For $n = 1$, (23) recovers the single node result (21).

A solution for the heterogeneous case, where the service of each system $i = 1, 2, \ldots, n$ has MGF envelope (17) with parameters ρ_i and σ_i, can be derived analogously. It follows that $S_{\text{net}}(\tau, t)$ has envelope (20) with rate $\rho' = \min_{i\in[1,n]}\{\rho_i\} - \delta$ and deficit profile

$$\varepsilon'(b) = \frac{e^{\theta \sum_{i=1}^{n} \sigma_i}}{(1-e^{-\theta\delta})^n} e^{-\theta b},$$

i.e., the network path is characterized by the minimum of the envelope rates $\min_{i\in[1,n]}\{\rho_i\}$ and the sum of the burstiness measures $\sum_{i=1}^{n} \sigma_i$ of the individual systems $i = 1, 2, \ldots, n$.

[3]As before, to improve the precision, the sum can be evaluated only for $\tau = 0, 1, \ldots, t - \lfloor b/\rho' \rfloor$ as $S(\tau, t)$ is non-negative.

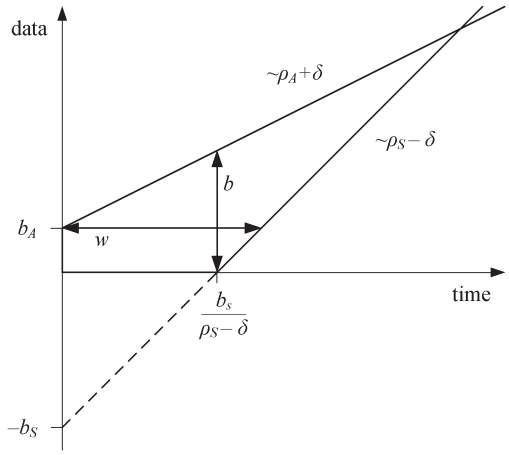

Fig. 2. Backlog and delay bound.

E. Backlog and Delay Bounds

So far, we considered envelope models of arrivals and service independently. For computation of non-asymptotic performance bounds, the network calculus offers convenient methods to compose the partial results derived so far. To distinguish parameters of the arrivals and of the service, we will use subscript A and S, respectively.

Consider arrivals with envelope (13). Fix $t \geq 0$ and assume a sample path where

$$A(\tau, t) \leq \rho'_A(t-\tau) + b_A \tag{24}$$

for all $\tau \in [0, t]$. Also, consider service with envelope (20) and assume a sample path where

$$S(\tau, t) \geq \rho'_S(t-\tau) - b_S \tag{25}$$

and generally $S(\tau, t) \geq 0$ for all $\tau \in [0, t]$. By insertion into (6), a backlog bound $B(t) \leq b$ follows as

$$b = \max_{\tau\in[0,t]} \{\rho'_A(t-\tau) + b_A - [\rho'_S(t-\tau) - b_S]_+\}, \tag{26}$$

where b is finite under the stability condition $\rho'_A \leq \rho'_S$. Since (24) and (25) may fail with probability $\varepsilon'_A(b_A)$ (16) and $\varepsilon'_S(b_S)$ (23), respectively, it follows by application of the union bound for any $t \geq 0$ that

$$\mathsf{P}\left[B(t) > b\right] \leq \varepsilon'_A(b_A) + \varepsilon'_S(b_S) = \varepsilon'. \tag{27}$$

Next, we compute (26). We substitute $\rho'_A = \rho_A + \delta$ and $\rho'_S = \rho_S - \delta$ where $\delta > 0$ is a free parameter, see Section II-B and D. For stability $\delta \leq (\rho_S - \rho_A)/2$ and

$$\boxed{\rho_A < \rho_S} \tag{28}$$

is required. The backlog bound follows as

$$\boxed{b = b_A + b_S \tfrac{\rho_A+\delta}{\rho_S-\delta}.} \tag{29}$$

Fig. 2 illustrates the backlog bound graphically as the maximal vertical deviation of the arrival envelope (13) and the service envelope (20), where we used that $S(\tau, t)$ is non-negative. The backlog comprises two terms: b_A as a measure

of the burstiness of the arrivals; and $b_S(\rho_A+\delta)/(\rho_S-\delta)$ that is the amount of data that is accumulated at the rate of the arrival envelope $\rho_A+\delta$ during the latency $b_S/(\rho_S-\delta)$ that is caused by the variability of the service. In addition, Fig. 2 depicts a delay bound as the maximal horizontal deviation of the two envelopes. The delay bound follows under the same stability condition (28) as

$$\mathsf{P}\left[W(t) > w\right] \leq \varepsilon'_A(b_A) + \varepsilon'_S(b_S) = \varepsilon',$$

where

$$w = \tfrac{b_A+b_S}{\rho_S-\delta}. \tag{30}$$

Finally, we fix $\varepsilon'_A = \varepsilon'_S = \varepsilon'/2$ and derive the quantity b_A by inversion of (16) as

$$b_A = \sigma_A - \tfrac{1}{\theta}\left(\ln\left(\tfrac{\varepsilon'}{2}\right) + \ln\left(1-e^{-\theta\delta}\right)\right), \tag{31}$$

and b_S from (23) as

$$b_S = n\sigma_S - \tfrac{1}{\theta}\left(\ln\left(\tfrac{\varepsilon'}{2}\right) + n\ln\left(1-e^{-\theta\delta}\right)\right). \tag{32}$$

The three summands of (31) and (32) are due to the burstiness measure σ, the violation probability ε', and the sample path derivation using slack rate δ. Regarding n-node networks, (32) exhibits a linear dependence on n. As an important consequence, backlog and delay bounds derived thereof grow as

$$b,\ w \in \mathcal{O}(n).$$

Concluding, the framed equations specify how to compute backlog and delay bounds with a defined violation probability ε' from the rate and burstiness parameters of the traffic arrivals ρ_A, σ_A and the service ρ_S, σ_S, respectively. In a final step, the free parameters $\theta > 0$ and $0 < \delta \leq (\rho_S-\rho_A)/2$ can be optimized to minimize b (29) and w (30). The two parameter characterization of different types of arrivals and service will be provided in the following section.

III. Traffic and Server Models

In this section, we provide the traffic and service parameters that are input to the performance bounds established by (28)–(32) for relevant cases. We investigate elementary traffic models in Section III-A, rules for multiplexing in Section III-B, server models in Section III-C, and a basic model for scheduling in Section III-D. For all figures, we optimized the free parameters θ and δ numerically.

As a basis, we first consider traffic arrivals at a work-conserving constant rate server, such as a constant rate link with capacity c, i.e., $S(\tau,t) = c(t-\tau)$ for all $t \geq \tau \geq 0$. Formally, expressed as an MGF envelope (17), the service has envelope rate $\rho_S = c$ and $\sigma_S = 0$. It is EBF (20) with $\rho'_S = c$ and deficit profile $\varepsilon'_S(b_S) = 0$ for all $b_S \geq 0$. Hence, we set $b_S = 0$ and obtain the backlog bound from (29) and the delay bound from (30)

$$b = b_A, \quad \text{and} \quad w = \frac{b_A}{c}, \tag{33}$$

with violation probability $\varepsilon' = \varepsilon'_A(b_A)$. The stability condition is $\rho_A < c$ and by choice of parameter $\delta = c - \rho_A$ we obtain

$$b_A = \sigma_A - \frac{1}{\theta}\left(\ln\varepsilon' + \ln\left(1-e^{-\theta(c-\rho_A)}\right)\right). \tag{34}$$

We will use the bounds obtained for the constant rate server to evaluate different traffic models in the following sections.

A. Traffic Models

The stochastic network calculus comprises a large variety of traffic models, including the extensive body of effective bandwidth results [4], [7]. In this tutorial, we include three fundamental traffic models: Poisson traffic, that enables a comparison with exact results from classical queueing theory; Markovian traffic, that is frequently used to model the On-Off characteristics of certain sources such as voice; and fractional Brownian motion, that captures the self-similarity and long range dependence observed for aggregated Internet data traffic.

1) Poisson: We denote $N(t)$ the number of packets arriving at a queueing system in the interval $[0,t]$. The counting process $N(t)$ is a Poisson process, if the inter-arrival times are memoryless, i.e., exponential. The Poisson process has distribution $\mathsf{P}[N(t)=k] = e^{-\lambda t}(\lambda t)^k/k!$ for $t > 0$ and $N(0) = 0$ where λ is the mean arrival rate. The MGF of the Poisson process is known as [29]

$$\mathsf{M}_N(\theta,t) = e^{\lambda t(e^\theta-1)}. \tag{35}$$

Given arrivals of constant size $1/\nu$, the cumulative number of bits that arrive in $[0,t]$ becomes $A(t) = N(t)/\nu$. For the MGF $\mathsf{M}_A(\theta,t) = \mathsf{E}[e^{\theta A(t)}]$ it follows that $\mathsf{M}_A(\theta,t) = \mathsf{M}_N(\theta/\nu,t)$ and by insertion of (35) $A(t)$ has an envelope (8) with parameters $\sigma = 0$ and rate

$$\rho = \frac{\lambda(e^{\theta/\nu}-1)}{\theta}$$

for $\theta > 0$. The combination of the Poisson arrival process with a constant rate server with capacity c corresponds to the $M|D|1$ model, where the service time is $1/(\nu c)$.

The well-known $M|M|1$ model results if the arrivals are independent and identically distributed (iid) exponential random variables X_k with mean $1/\nu$ and MGF $\mathsf{M}_X(\theta) = \nu/(\nu-\theta)$ for $\theta < \nu$. In this case, the arrival process is the doubly stochastic process

$$A(t) = \sum_{k=1}^{N(t)} X_k.$$

It has conditional MGF $\mathsf{E}[e^{\theta A(t)}|N(t)=k] = (\mathsf{M}_X(\theta))^k$ [29], such that by unconditioning

$$\mathsf{E}[e^{\theta A(t)}] = \mathsf{E}\left[(\mathsf{M}_X(\theta))^{N(t)}\right] = \mathsf{E}\left[e^{\ln(\mathsf{M}_X(\theta))N(t)}\right].$$

By substitution of $\vartheta = \ln(\mathsf{M}_X(\theta))$, it follows that [31]

$$\mathsf{M}_A(\theta,t) = \mathsf{M}_N(\vartheta,t) = \mathsf{M}_N(\ln\mathsf{M}_X(\theta),t).$$

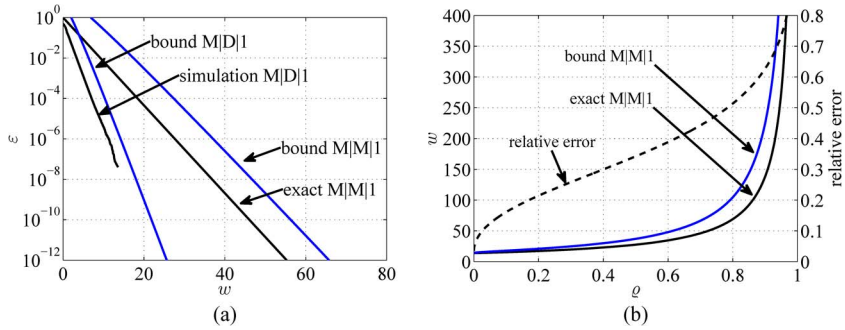

Fig. 3. Delay bounds for an $M|D|1$ and an $M|M|1$ queue compared to simulation results, respectively, exact results from queueing theory. (a) Tail decay. (b) Impact of the utilization.

Insertion of the exponential MGF into the Poisson MGF (35) gives $\mathsf{M}_A(\theta,t)=e^{(\theta\lambda t)/(\nu-\theta)}$ that satisfies (8) with $\sigma=0$ and envelope rate

$$\rho=\frac{\lambda}{\nu-\theta}$$

for $0\leq\theta<\nu$. For $\theta=0$ we obtain the mean rate λ/ν.

In Fig. 3, we depict the delay bound w from (33) and (34) for Poisson arrivals with rate λ at a constant rate server with capacity c. We use arrivals of constant size $(M|D|1)$ and of exponentially distributed size $(M|M|1)$, respectively, both with mean $1/\nu$. The mean service time follows as $1/(\nu c)$, where c is the capacity. For comparison, we show simulation results for the response time of the $M|D|1$ queue and the exact result for the $M|M|1$ queue. For the response time of the $M|M|1$ queue it is known from queueing theory [32] that $\varepsilon=e^{-\nu c(1-\varrho)w}$, where $\varrho=\lambda/(\nu c)$ is the utilization. For numerical evaluation we use $\nu=1$ and $c=1$. We depict the tail decay of ε for $\lambda=0.5$, that corresponds to a utilization of $\varrho=0.5$, in Fig. 3(a) and show the impact of λ, respectively, ϱ on delays for $\varepsilon=10^{-6}$ in Fig. 3(b).

The curves in Fig. 3(a) show an exponential decay that is characteristic for EBB arrival processes such as Poisson traffic. Clearly, delays are smaller in case of the $M|D|1$ model. Compared to the simulation results for the $M|D|1$ queue and the exact result for the $M|M|1$ queue, the bounds from the stochastic network calculus are conservative. This is due to a relaxation of assumptions compared to queueing theory, which enables including a much larger set of traffic models than Poisson. For a restricted set of traffic models, that permit the construction of exponential supermartingales, the tightness of bounds is improved in [33]. Notably, the bounds in Fig. 3(a) show the same exponential decay as the exact results. For a numerical example, consider the $M|M|1$ queue and $\varepsilon=10^{-6}$ where the delay bound is approximately 37 compared to the exact result of 28 resulting in a relative error of 0.3. The relative error decreases for smaller ε. Also, Fig. 3(b) shows that the relative error is smaller for moderate to low utilizations and stays below 0.5 up to a utilization of about 0.8.

Further, it can be observed in Fig. 3(a) that the simulation results for the $M|D|1$ queue bend down for $\varepsilon<10^{-6}$. This is a general problem when obtaining tail probabilities from simulations due to the inherently restricted sample size. Further comparisons of bounds derived from the stochastic network calculus with simulation results are also provided in [13], [34].

For interpretation of the units, note that in the discrete time model delays are measured in units of timeslots. As an example, given packets of 10 kbit size and a link with 10 Mbit/s capacity, the timeslot can be fixed as the transmission time of one packet, e.g., 1 ms.

2) Markov On-Off: Next, we consider a Markov modulated arrival process with a two state Markov chain. Compared to the memoryless Poisson arrival process, Markov processes have first-order memory, where the current state depends (only) on the previous state. In state 1 (Off) the source generates no arrivals, and in state 2 (On) it generates arrivals with rate r. The steady state probability of the On state is $p_{\rm on}=p_{12}/(p_{12}+p_{21})$, where p_{ij} for $i,j=1,2$ are the transition probabilities from state i to state j. The mean arrival rate follows as $p_{\rm on}r$. In addition, the arrivals can be characterized by a burstiness parameter $T=1/p_{12}+1/p_{21}$ that is the mean time to change state twice. The MGF of the Markov On-Off process satisfies (8) with $\sigma=0$ and envelope rate [4], [12]

$$\rho=\frac{1}{\theta}\ln\left(\frac{p_{11}+p_{22}e^{\theta r}+\sqrt{(p_{11}+p_{22}e^{\theta r})^2-4(p_{11}+p_{22}-1)e^{\theta r}}}{2}\right)$$

for $\theta>0$. For the special case of a memoryless On-Off process it holds that $p_{11}=p_{21}$ and $p_{12}=p_{22}$, so that $p_{\rm on}=p_{22}$ and

$$\rho=\frac{\ln(p_{\rm on}e^{\theta r}+1-p_{\rm on})}{\theta} \tag{36}$$

for $\theta>0$. As Markov traffic falls into the EBB class, it shows the same characteristic exponential decay as observed for Poisson traffic in Fig. 3, where the burstiness parameter T of the Markov On-Off source determines the slope. We will show results on the impact of the burstiness in Fig. 7.

3) Fractional Brownian Motion: While many relevant arrival processes fall into the EBB class, defined by (9) and (10), we cover fractional Brownian motion (fBm) as an example of a process that is not EBB, to draw some important conclusions. FBm is frequently used as a model of aggregated Internet data traffic to analyze the impact of long range dependence on networks. Fbm is a self-similar arrival process with correlated Gaussian increments. It has MGF [7]

$$\mathsf{M}_A(\theta,t)=e^{\theta\left(\lambda t+\frac{\theta\varsigma^2}{2}t^{2h}\right)}, \tag{37}$$

where λ is the mean rate, and ς^2 the variance of the increments. Parameter h is the Hurst parameter where $h\in(0.5,1)$ denotes

long range dependence (LRD). If $h = 0.5$, fBm becomes standard Brownian motion that has envelope rate (8)

$$\rho = \lambda + \frac{\theta\varsigma^2}{2}.$$

In case of LRD, i.e., $h \in (0.5, 1)$, (37) grows superlinearly with t, such that no affine MGF envelope as defined by (8) exists. Consequently, fBm does not fall into the EBB class.

To derive performance bounds for fBm traffic in the stochastic network calculus, a generalized definition of statistical envelope functions $E(t)$ can be used [6], [35]–[37]. By Chernoff's bound (11) it holds for $\theta \geq 0$ that

$$\mathsf{P}\left[A(\tau,t) > E(t-\tau)\right] \leq e^{-\theta E(t-\tau)} \mathsf{M}_A(\theta, t-\tau) = \varepsilon.$$

After solving for $E(t)$, a minimal envelope function follows by optimization over $\theta > 0$ as [6]

$$E(t) = \inf_{\theta>0} \left\{ \frac{1}{\theta} \left(\ln \mathsf{M}_A(\theta,t) - \ln \varepsilon \right) \right\}$$

for $t \geq 0$. By insertion of $\mathsf{M}_A(\theta,t)$ from (37), the minimum can be obtained at $\theta = \sqrt{-2\ln\varepsilon}/(\varsigma t^h)$, such that [6], [35], [36]

$$E(t) = \lambda t + \sqrt{-2\ln\varepsilon}\varsigma t^h. \tag{38}$$

Following the steps of Section II-B, we have to construct a sample path envelope of the form $\mathsf{P}[\exists\tau \in [0,t] : A(\tau,t) > E(t-\tau)] \leq \varepsilon'$ to be able to derive performance bounds. A respective solution is provided in [15]. Instead, in this work, we use the much simpler approximation by the largest term (14) to estimate $\varepsilon' \approx \varepsilon$. In this case, a backlog bound $\mathsf{P}[B(t) > b] \approx \varepsilon$ at a constant rate server with capacity c follows from (6) by substitution of $E(t-\tau)$ from (38) for $A(\tau,t)$ and $S(\tau,t) = c(t-\tau)$. Letting $t \to \infty$, a backlog bound is

$$b = \max_{\tau \geq 0} \left\{ \lambda\tau + \sqrt{-2\ln\varepsilon}\varsigma\tau^h - c\tau \right\}.$$

The maximum is attained at $\tau = \tau^*$, where [36]

$$\tau^* = \left(\frac{\sqrt{-2\ln\varepsilon}\varsigma h}{c-\lambda} \right)^{\frac{1}{1-h}}.$$

By insertion of τ^* and after solving for ε the main result

$$\varepsilon = \exp\left(-\frac{1}{2\varsigma^2} \left(\frac{c-\lambda}{h} \right)^{2h} \left(\frac{b}{1-h} \right)^{2-2h} \right), \tag{39}$$

that was first reported in [38], [39], is recovered in the stochastic network calculus.

In Fig. 4, we depict the violation probability ε of a backlog bound b from (39) for $h = 0.5$, 0.6, and 0.7. The remaining parameters are $c = 1$, $\lambda = 0.5$, and $\varsigma = 0.5$. For $h = 0.5$, i.e., standard Brownian motion that falls into the EBB class, the curve shows an exponential decay. A fundamentally different behavior can, however, be observed under LRD, i.e., for $h > 0.5$, where the decay is much slower and exhibits a Weibull tail. The same log-asymptotic decay of ε' with b is also obtained from the sample path analysis [15]. The Weibull tail significantly impacts resource dimensioning as it demonstrates the inefficiency of buffering LRD traffic [36], [39]. Regarding (39), the spare capacity $c - \lambda$ and the buffer size b are equally important if $h = 0.5$, whereas spare capacity becomes more important and buffering less efficient for increasing h, as applicable for the Internet.

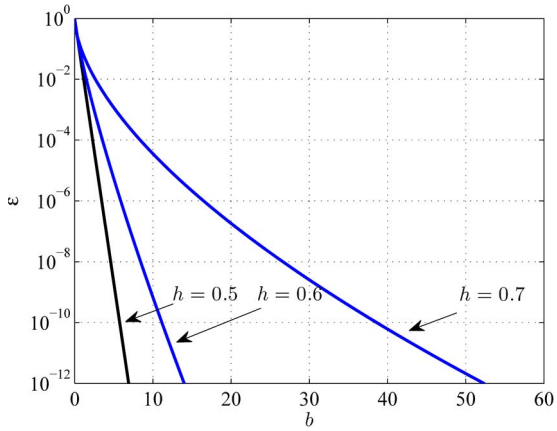

Fig. 4. Backlog bounds for fBm traffic with Hurst parameter h at a constant rate server.

B. Statistical Multiplexing

Statistical multiplexing is the reason for the resource efficiency of packet data networks. In brief, the aggregate of independent traffic flows becomes smoother as the number of flows increases. As a consequence, each additional flow requires less resources, where the resource requirement of a flow approaches its mean rate. MGFs provide a convenient and efficient model to take advantage of this effect. The aggregate arrival process of the superposition of m arrival processes is

$$A_{\text{agg}}(\tau,t) = \sum_{i=1}^{m} A_i(\tau,t).$$

Under the assumption of statistical independence it holds for the aggregate arrivals that

$$\mathsf{E}\left[e^{\theta A_{\text{agg}}(\tau,t)} \right] = \prod_{i=1}^{m} \mathsf{E}\left[e^{\theta A_i(\tau,t)} \right].$$

If the arrival processes A_i each have MGF envelope (8) with parameters ρ_A and σ_A it follows that

$$\mathsf{E}\left[e^{\theta A_{\text{agg}}(\tau,t)} \right] \leq e^{\theta(m\rho_A(t-\tau) + m\sigma_A)},$$

where we considered the homogeneous case for notational simplicity. The aggregate arrivals have MGF envelope (8) with parameters

$$\boxed{\begin{aligned} \rho_{A_{\text{agg}}} &= m\rho_A, \\ \sigma_{A_{\text{agg}}} &= m\sigma_A. \end{aligned}} \tag{40}$$

In general, the parameters of the MGF envelope model are additive, i.e., for the heterogeneous case where the arrival processes

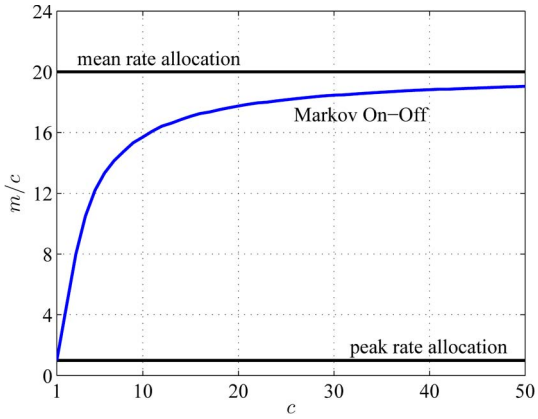

Fig. 5. Number of admissible Markov On-Off flows per unit of capacity compared to the mean rate and the peak rate allocation, respectively.

A_i have individual MGF envelopes (8) with parameters ρ_{A_i} and σ_{A_i} it follows that

$$\rho_{A_{\text{agg}}} = \sum_{i=1}^{m} \rho_{A_i},$$

$$\sigma_{A_{\text{agg}}} = \sum_{i=1}^{m} \sigma_{A_i}.$$

For a numerical example, we consider the number of admissible Markov On-Off sources m at a constant rate server with capacity c. Flows are admitted as long as a target delay bound of $w = 100$ is violated at most with probability $\varepsilon' = 10^{-3}$. The sources are statistically independent and have peak rate $r = 1$, mean rate $p_{\text{on}} r = 0.05$, and burstiness parameter $T = 300$. The delay bound is computed from (33) and (34) where we use the aggregate traffic parameters from (40).

Fig. 5 depicts the number of flows per unit capacity m/c for increasing c. For comparison, an allocation that considers only the peak rate has $m/c = 1$ and the mean rate $m/c = 20$, respectively. The number of flows that are actually admissible grows from the peak rate to the mean rate allocation. The effect is due to statistical multiplexing, that makes the aggregate traffic smoother as the number of flows increases. In the mathematical model, the statistical multiplexing gain is realized by optimizing the free parameter $\theta > 0$.

C. Server Models

In Section III-A and B we considered different types of traffic at a constant rate server. Next, we investigate variable rate servers where the service is a random process, e.g., due to the characteristics of a wireless channel or due to scheduling of cross traffic. One of the key contributions of the network calculus, compared to the theory of effective bandwidths, is that it comprehends a variety of server models and provides results for their composition. We first show an elementary model of a memoryless On-Off server. More elaborate models, such as Markov On-Off or general Markovian servers, are dual to the respective traffic models shown before and are therefore omitted. Next, we illustrate, using the example of the On-Off server, how to derive a basic characterization of a Rayleigh fading channel. More elaborate models of wireless channels that follow the same fundamental approach can be found, e.g., in [17], [30], [34], [40], [41].

1) Memoryless On-Off Server: First, we investigate the elementary model of a lossless work-conserving server with a time-varying capacity, where $X(t)$ denotes the service available in timeslot $t \geq 0$. The cumulative service in $[\tau + 1, t]$ for $t \geq \tau \geq 0$ follows as

$$S(\tau, t) = \sum_{\upsilon=\tau+1}^{t} X(\upsilon).$$

If the increments $X(t)$ are iid random variables, the server is memoryless and it follows that

$$\mathsf{M}_S(-\theta, t) = (\mathsf{M}_X(-\theta))^t$$

for $t \geq 0$. For the special case of an On-Off server, the increments $X(t)$ are iid Bernoulli trials with probability mass function $p_X(r) = p_{\text{on}}$ and $p_X(0) = 1 - p_{\text{on}}$. The MGF, respectively, Laplace transform is $\mathsf{M}_X(-\theta) = \sum_x e^{-\theta x} p_X(x) = p_{\text{on}} e^{-\theta r} + 1 - p_{\text{on}}$. It follows that $\mathsf{M}_S(-\theta, t) = (p_{\text{on}} e^{-\theta r} + 1 - p_{\text{on}})^t$ for $t \geq 0$ is the binomial MGF that has an envelope (17) with parameters $\sigma = 0$ and rate

$$\rho = \frac{\ln\left(p_{\text{on}} e^{-\theta r} + 1 - p_{\text{on}}\right)}{-\theta}$$

for $\theta > 0$. Note how the envelope rate of the service process parallels the corresponding rate of the On-Off arrival process (36). The On-Off server can be parameterized by choice of r and p_{on} to characterize specific systems, such as a Rayleigh fading channel in the following Section III-C2.

2) Rayleigh Fading: We consider a system that transmits data at a fixed rate r over a wireless channel. Communications is possible if r does not exceed the channel capacity C. The channel is characterized by a Rayleigh block fading process that causes fluctuations of the instantaneous channel capacity $C(t)$ where $t \in \mathbb{N}_0$, i.e., $C(t)$ is a random process. If $C(t) \geq r$, the data transmitted in timeslot t can be successfully decoded by the receiver. Otherwise, the data cannot be decoded and are retransmitted in timeslot $t + 1$. Consequently, the system behaves like an On-Off server with parameters r and p_{on} where the transmission rate r determines $p_{\text{on}} = \mathsf{P}[C(t) \geq r]$. Given the distribution of $C(t)$, the system can optimize the free parameter r, e.g., to maximize the average rate of successful transmission $p_{\text{on}} r$.

Following the approach in [30], the instantaneous channel capacity is estimated from the signal-to-noise ratio (SNR) by the Shannon capacity as $C(t) = \text{ld}(1 + \gamma(t))$, where the capacity is normalized and measured in bit/Hz/s. Given a Rayleigh fading channel, the SNR $\gamma(t)$ is exponentially distributed with mean value $\overline{\gamma}$.

In Fig. 6(a) we illustrate the relationship between r and p_{on}. Clearly, p_{on} decreases with increasing r, where $p_{\text{on}} r$ reaches a maximum for $r \approx 1.7$. Also in Fig. 6(b), we depict the delay bound w from (30) with $\varepsilon' = 10^{-6}$ for Poisson traffic that is transmitted via the Rayleigh fading channel. The Poisson traffic has arrival rate $\lambda = \{0.5, 0.6, 0.7\}$ and unit sized packets, i.e.,

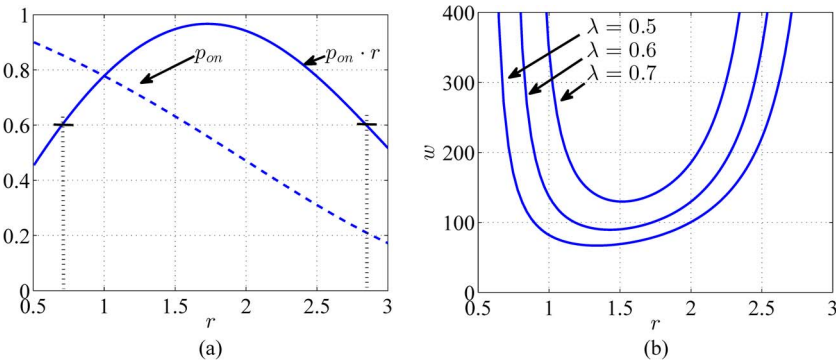

Fig. 6. Rayleigh fading channel with an average SNR of 6 dB. (a) The transmission rate r determines the probability that data can be decoded at the receiver $p_{\rm on}$. The average throughput $p_{\rm on}r$ is maximized for $r \approx 1.7$. (b) Delay bound w versus transmission rate r for different arrival rates λ. The transmission rate that minimizes the delay bound depends on the arrival rate. (a) Successful transmission rate. (b) Delay bound.

$1/\nu = 1$. The u-shape of the curves is due to the fact that delays grow unboundedly if the average rate of successful transmission $p_{\rm on}r$ approaches λ. As Fig. 6(a) shows for the example $\lambda = 0.6$, there exist two values of r, i.e., $r \approx 0.7$ and $r \approx 2.8$, such that $p_{\rm on}r = \lambda$. All r in between these two extreme points are feasible as they achieve the stability criterion $p_{\rm on}r > \lambda$. Interestingly, the choice of r that maximizes $p_{\rm on}r$ does not minimize w. Instead, smaller r become favorable with decreasing λ. The effect is caused by the increase of $p_{\rm on}$ with decreasing r that makes the transmission less variable and hence helps reduce w.

D. Scheduling

The network calculus uses a notion of leftover service to characterize schedulers that offer a certain amount of service to a flow depending on the presence of other traffic. We show a blind scheduling model from [14] that is conservative in general, as it does not make any assumptions about the order of serving traffic. Given a work-conserving system with a time-varying capacity $S(\tau,t)$. Let $A(t) = A_{\rm th}(t) + A_{\rm cr}(t)$ and $D(t) = D_{\rm th}(t) + D_{\rm cr}(t)$ be composed of through traffic, i.e., the flow of interest, and cross traffic, i.e., other traffic. From (2), it follows after some reordering that

$$D_{\rm th}(t) \geq A_{\rm th}(\tau^*) + S(\tau^*,t) - (D_{\rm cr}(t) - A_{\rm cr}(\tau^*))$$

for $t \geq \tau^* \geq 0$, where τ^* is the beginning of the last busy period before t. By substitution of $D_{\rm cr}(t) \leq A_{\rm cr}(t)$ for causality and since $D_{\rm th}(t) \geq D_{\rm th}(\tau^*) = A_{\rm th}(\tau^*)$ by choice of τ^*, it holds that

$$D_{\rm th}(t) \geq A_{\rm th}(\tau^*) + [S(\tau^*,t) - A_{\rm cr}(\tau^*,t)]_+ .$$

Finally, it follows for all $t \geq 0$ that

$$D_{\rm th}(t) \geq \min_{\tau\in[0,t]} \left\{A_{\rm th}(\tau) + [S(\tau,t) - A_{\rm cr}(\tau,t)]_+\right\},$$

such that for $t \geq \tau \geq 0$

$$S_{\rm lo}(\tau,t) = [S(\tau,t) - A_{\rm cr}(\tau,t)]_+$$

is a leftover service process that satisfies the definition of dynamic server (1) for the through traffic. Under the assumption of statistical independence of $S(\tau,t)$ and $A_{\rm cr}(\tau,t)$, it follows for the MGF of the leftover service that

$$\mathsf{E}\left[e^{-\theta S_{\rm lo}(\tau,t)}\right] \leq \mathsf{E}\left[e^{-\theta S(\tau,t)}\right] \mathsf{E}\left[e^{\theta A_{\rm cr}(\tau,t)}\right],$$

for $t \geq \tau \geq 0$. Given the service $S(\tau,t)$ has an MGF envelope (17) with parameters ρ_S, σ_S and the cross traffic arrivals $A_{\rm cr}$ have an MGF envelope (8) with parameters $\rho_{A_{\rm cr}}$,$\sigma_{A_{\rm cr}}$, it holds for $t \geq \tau \geq 0$ that

$$\mathsf{E}\left[e^{-\theta S_{\rm lo}(\tau,t)}\right] \leq e^{-\theta((\rho_S - \rho_{A_{\rm cr}})(t-\tau) - (\sigma_S + \sigma_{A_{\rm cr}}))},$$

such that the leftover service process $S_{\rm lo}$ has MGF envelope (17) with parameters

$$\boxed{\begin{array}{l} \rho_{S_{\rm lo}} = \rho_S - \rho_{A_{\rm cr}}, \\ \sigma_{S_{\rm lo}} = \sigma_S + \sigma_{A_{\rm cr}}. \end{array}} \tag{41}$$

We show delay bounds obtained for through traffic that is scheduled with cross traffic at a constant rate server. The delay bounds are computed from (30) where we use the leftover service parameters from (41). Further, we let the cross traffic parameters be the parameters of aggregated traffic from (40).

In Fig. 7, we illustrate the impact of the cross traffic burstiness on the delay bound for Poisson through traffic. The constant rate server has capacity $c = 1$. The mean arrival rate of the Poisson through traffic is fixed to $\lambda = 0.25$ in Fig. 7(a) and varied in Fig. 7(b) where we fix $\varepsilon' = 10^{-6}$. The size of the arrivals is deterministic and the arrivals are unit sized, i.e., $1/\nu = 1$. The cross-traffic consists of 10 independent Markov On-Off flows, each with peak rate $r = 0.15$, mean rate 0.025, and different burstiness parameters $T = 10, 20, 40,$ and 80. We observe that the burstiness of the cross traffic directly impacts the delay bound of the through traffic, where it alters the slope of the exponential decay.

In Fig. 8, we show end-to-end delay bounds for Poisson through traffic that traverses a tandem of n homogeneous constant rate servers, each with independent Markov On-Off cross traffic. The scenario is illustrated in Fig. 9. The traffic and service parameters are the same as for Fig. 7(a). We fix the end-to-end violation probability $\varepsilon' = 10^{-6}$. Clearly, the delay bounds grow linearly with n, where the slope is determined by the burstiness of the cross traffic.

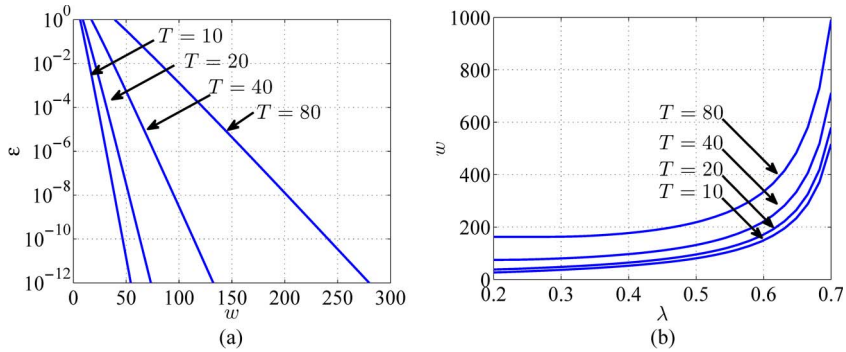

Fig. 7. Delay bounds for Poisson through traffic with arrival rate λ that is scheduled with Markov On-Off cross traffic with burstiness parameter T at a constant rate server. (a) Tail decay. (b) Impact of the arrival rate.

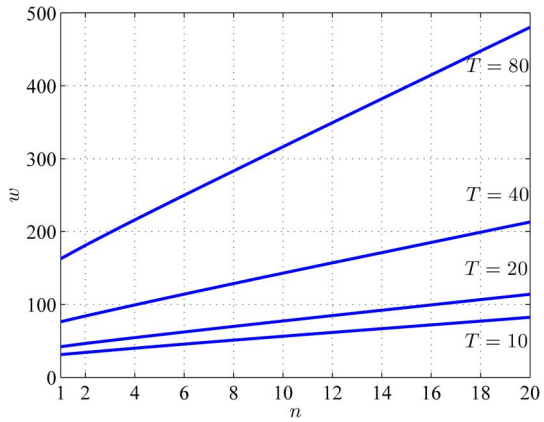

Fig. 8. Growth of end-to-end delay bounds for Poisson through traffic at a tandem of n constant rate servers, each with independent Markov On-Off cross traffic.

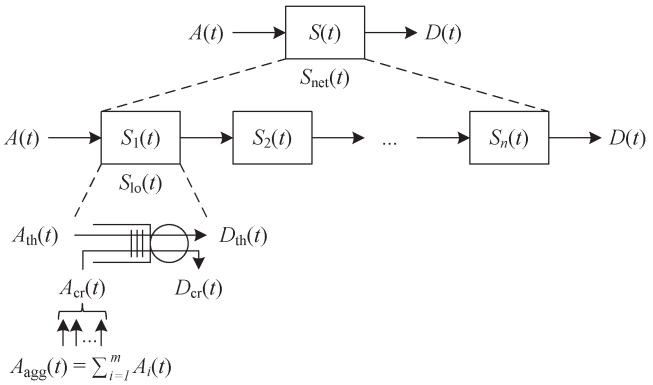

Fig. 9. Equivalent single system representation of a network.

In this section, we showed numerical results for the concatenation of tandem systems with scheduling for the special case of constant rate servers. Due to the modularity of the network calculus approach, variable rate servers can be dealt with in the same way, e.g., to analyze tandem Rayleigh fading channels with cross traffic. We summarize the basic steps of the method in Section IV-A.

IV. Conclusion and Outlook

We conclude this tutorial with a summary of the methods that we presented (Section IV-A), pointers to related works (Section IV-B), and an outlook on open challenges (Section IV-C).

TABLE I
Building Blocks for Composition

basic steps of the method			
1.a)	arrival parameters	ρ_A, σ_A	Sect. III-A
b)	service parameters	ρ_S, σ_S	Sect. III-C
2.a)	EBB arrival envelope	ρ_A, b_A	Eq. (31)
b)	EBF service envelope	ρ_S, b_S	Eq. (32)
3.a)	stability	$\rho_A < \rho_S$	Eq. (28)
b)	backlog	$\mathsf{P}[B(t) > b] \leq \varepsilon'$	Eq. (29)
c)	delay	$\mathsf{P}[W(t) > w] \leq \varepsilon'$	Eq. (30)
d)	optimize parameters	$\theta > 0, 0 < \delta < (\rho_S - \rho_A)/2$	
extensions of respective steps			
1.a)	multiplexing	$\rho_{A_{\text{agg}}}, \sigma_{A_{\text{agg}}}$	Eq. (40)
b)	scheduling	$\rho_{S_{\text{lo}}}, \sigma_{S_{\text{lo}}}$	Eq. (41)
2.b)	multi-node networks	$n > 1$	Eq. (32)

A. Toolbox

The results that we presented include stochastic arrival and server models and rules for their composition. The composition including multiplexing, scheduling, and series connection is shown in Fig. 9. The topology in Fig. 9 is also known as a line topology with single-hop persistent cross traffic. For arbitrary feed-forward topologies there exist methods that transform the network into the line topology in Fig. 9, see [19] for an overview. The general approach is to consider the leftover service of the systems on the path of the through traffic. For this purpose, envelopes of the cross traffic at each of the systems are computed in an iterative fashion.

In the following, we provide a summary of the lessons learned to guide the reader through the presented method. Table I gives an outline of the basic steps of the method and provides pointers to the respective equations for quick access. In step 1, the parameters of the MGF envelopes of the arrivals ρ_A, σ_A and the service ρ_S, σ_S are defined. Optionally, rules for multiplexing and scheduling can be applied. In step 2, we make the transition from MGF to EBB envelopes. The outcome of step 2 is an equivalent single system representation of the network as depicted in Fig. 9. The resulting system is fully characterized by the EBB parameters of the through traffic arrival envelope ρ_A, b_A and the network service envelope ρ_S, b_S. Step 3 provides backlog and delay bounds under the stability condition $\rho_A < \rho_S$. Finally, the violation probability ε' can be fixed and the free parameters $\theta > 0$ and $0 < \delta \leq (\rho_S - \rho_A)/2$ can be optimized.

TABLE II
PROS AND CONS OF MGF VERSUS EBB ENVELOPES

	MGF	EBB	this work
statistical independence	mandatory	optional	MGF
statistical multiplexing	simple	complex	MGF
schedulers	blind	many	MGF
multi-node networks	$\mathcal{O}(n)$	$\Theta(n \log n)$	MGF
backlog and delay	involved	simple	EBB

B. Duality of Envelope Models

A design decision of the method that we presented is the transition from MGF to EBB envelopes in step 2, see Table I. In fact, the transition to EBB could as well take place in any other step, resulting, however, in a method with different qualities. Table II considers this aspect and compares the pros and cons of the MGF and the EBB envelope models. Next, we highlight some major differences.

Statistical Independence and Multiplexing: While technically MGFs of sums of non-independent random processes can be computed, MGFs are in general applied under the assumption of statistical independence. Regarding EBB, overflow profiles can be added by the union bound without assumption of independence as, e.g., in (27). On the other hand, the overflow profiles are essentially complementary cumulative distribution functions (CCDFs) that can be convolved under the assumption of independence, to take advantage of statistical multiplexing [9], [17], [21]. As the MGF transforms convolution into multiplication, it provides the computationally simpler approach to make use of statistical independence.

Scheduling: We presented the blind scheduling model from [14] that is based on MGFs. The model does not make any assumptions about the order of serving cross traffic and through traffic. Hence, it is conservative in general. Solutions for specific schedulers are derived, e.g., in [6] using statistical envelopes.

Multi-Node Networks: We showed end-to-end performance bounds for n statistically independent systems in series that grow in $\mathcal{O}(n)$ [14]. Without assumption of independence, an upper bound $\mathcal{O}(n \log n)$ is derived in [12] using the EBB model. A corresponding lower bound is proven in [13].

Backlog and Delay: Finally, we mention that the transition from MGF to EBB envelopes that is used in this paper can be omitted. In [14], MGFs of backlogs and delays are derived and in a final step Chernoff's bound is used to compute performance bounds thereof. The final computation is, however, more involved and less intuitive than in case of the EBB envelope model.

C. Open Challengess

We conclude this paper with an outlook on open challenges in the stochastic network calculus.

Packet Loss: The definition of dynamic server (1) assumes a lossless system, i.e., a system that generally provides sufficient buffer space to store backlogged data. Statistical backlog bounds $\mathsf{P}[B(t) > b] \leq \varepsilon'$ (27) can be interpreted as the probability of buffer overflow, given a buffer of limited size b, but provide only an approximation [4]. Solutions for server models that include loss are still open.

Feedback Control: The deterministic network calculus features an elegant formulation of feedback control such as window flow control. The feedback controlled arrivals that are input to the network are $A_{\mathrm{fc}}(t) = \min[A_{\mathrm{uc}}(t), D(t) + x]$, where $A_{\mathrm{uc}}(t)$ are the uncontrolled, external arrivals and x is the window size [4], [5], [42]. In the stochastic network calculus, the difficulty of this model is due to the fact that sample paths of the departures determine the arrivals to the network.

Wireless Channels: Non-equilibrium models of wireless channels receive growing interest, see, e.g., recent works in the area of effective capacity [26], [43] and in the stochastic network calculus [17], [30], [34], [40], [41]. Common channel models that have been explored so far, are Markov or memoryless processes, such as in Section III-C1, that are calibrated using, e.g., a fading process. The results provided help to understand the impact of essential aspects of wireless systems, such as the fading speed, multiple antennas, or hybrid ARQ, on packet delays.

MAC and ARQ Protocols: The service provided by random access protocols as well as automatic repeat request protocols is inherently random. It lends itself to an analysis using the stochastic network calculus, see for example the works on ALOHA [44] and on CSMA/CA [45]. A possible approach to model the overhead due to retransmissions of lost packets is [28].

REFERENCES

[1] M. Fidler and A. Rizk, "A guide to the stochastic network calculus," presented at the Proc. GI MMBnet, Hamburg, Germany, Sep. 2013, Invited Paper.

[2] R. L. Cruz, "A calculus for network delay part I and II: Network elements in isolation and network analysis," *IEEE Trans. Inf. Theory*, vol. 37, no. 1, pp. 114–141, Jan. 1991.

[3] R. L. Cruz, "Quality of service guarantees in virtual circuit switched networks," *IEEE J. Sel. Areas Commun.*, vol. 13, no. 6, pp. 1048–1056, Aug. 1995.

[4] C.-S. Chang, *Performance Guarantees in Communication Networks*. London, U.K.: Springer-Verlag, 2000.

[5] J.-Y. Le Boudec and P. Thiran, *Network Calculus A Theory of Deterministic Queuing Systems for the Internet*, vol. 2050. Berlin, Germany: Springer-Verlag, 2001, ser. LNCS.

[6] C. Li, A. Burchard, and J. Liebeherr, "A network calculus with effective bandwidth," *IEEE/ACM Trans. Netw.*, vol. 15, no. 6, pp. 1442–1453, Dec. 2007.

[7] F. P. Kelly, *Notes on Effective Bandwidths*. Oxford, U.K.: Clarendon, 1996, ser. Number 4 in Royal Statistical Society Lecture Notes, pp. 141–168.

[8] C.-S. Chang, "Stability, queue length and delay of deterministic and stochastic queueing networks," *IEEE Trans. Autom. Control*, vol. 39, no. 5, pp. 913–931, May 1994.

[9] O. Yaron and M. Sidi, "Performance and stability of communication networks via robust exponential bounds," *IEEE/ACM Trans. Netw.*, vol. 1, no. 3, pp. 372–385, Jun. 1993.

[10] R. L. Cruz, "Quality of service management in integrated services networks," presented at the Proc. Semi-Annual Research Review, Center Wireless Communication, UCSD, San Diego, CA, USA, Jun. 1996.

[11] Q. Yin, Y. Jiang, S. Jiang, and P. Y. Kong, "Analysis of generalized stochastically bounded bursty traffic for communication networks," in *Proc. IEEE LCN*, Nov. 2002, pp. 141–149.

[12] F. Ciucu, A. Burchard, and J. Liebeherr, "Scaling properties of statistical end-to-end bounds in the network calculus," *IEEE/ACM Trans. Netw.*, vol. 52, no. 6, pp. 2300–2312, Jun. 2006.

[13] A. Burchard, J. Liebeherr, and F. Ciucu, "On superlinear scaling of network delays," *IEEE/ACM Trans. Netw.*, vol. 19, no. 4, pp. 1043–1056, Aug. 2011.

[14] M. Fidler, "An end-to-end probabilistic network calculus with moment generating functions," in *Proc. IWQoS*, Jun. 2006, pp. 261–270.
[15] A. Rizk and M. Fidler, "Non-asymptotic end-to-end performance bounds for networks with long range dependent FBM cross traffic," *Comput. Netw.*, vol. 56, no. 1, pp. 127–141, Jan. 2012.
[16] J. Liebeherr, A. Burchard, and F. Ciucu, "Delay bounds in communication networks with heavy-tailed and self-similar traffic," *IEEE Trans. Inf. Theory*, vol. 58, no. 2, pp. 1010–1024, Feb. 2012.
[17] Y. Jiang and Y. Liu, *Stochastic Network Calculus*. London, U.K.: Springer-Verlag, Sep. 2008.
[18] S. Mao and S. S. Panwar, "A survey of envelope processes and their applications in quality of service provisioning," *IEEE Commun. Surveys Tuts.*, vol. 8, no. 3, pp. 2–20, Jul. 2006.
[19] M. Fidler, "A survey of deterministic and stochastic service curve models in the network calculus," *IEEE Commun. Surveys Tuts.*, vol. 12, no. 1, pp. 59–86, 2010.
[20] F. Ciucu and J. Schmitt, "Perspectives on network calculus—No free lunch but still good value," in *Proc. ACM SIGCOMM*, Aug. 2012, pp. 311–322.
[21] A. Rizk and M. Fidler, "On multiplexing models for independent traffic flows in single- and multi-node networks," *IEEE Trans. Netw. Serv. Manage.*, vol. 10, no. 1, pp. 15–28, Mar. 2013.
[22] C.-S. Chang, R. L. Cruz, J.-Y. Le Boudec, and P. Thiran, "A min, + system theory for constrained traffic regulation and dynamic service guarantees," *IEEE/ACM Trans. Netw.*, vol. 10, no. 6, pp. 805–817, Dec. 2002.
[23] J. Liebeherr, M. Fidler, and S. Valaee, "A system theoretic approach to bandwidth estimation," *IEEE/ACM Trans. Netw.*, vol. 18, no. 4, pp. 1040–1053, Aug. 2010.
[24] D. Starobinski and M. Sidi, "Stochastically bounded burstiness for communication networks," *IEEE Trans. Inf. Theory*, vol. 46, no. 1, pp. 206–212, Jan. 2000.
[25] R. Lübben, M. Fidler, and J. Liebeherr, "A foundation for stochastic bandwidth estimation of networks with random service," in *Proc. IEEE INFOCOM*, Apr. 2011, pp. 1817–1825.
[26] D. O. Wu and R. Negi, "Effective capacity: A wireless link model for support of quality of service," *IEEE Trans. Wireless Commun.*, vol. 2, no. 4, pp. 630–643, Jul. 2003.
[27] K. Lee, "Performance bounds in communication networks with variable-rate links," in *Proc. ACM SIGCOMM*, Aug. 1995, pp. 126–136.
[28] F. Ciucu, J. Schmitt, and H. Wang, "On expressing networks with flow transformations in convolution-form," in *Proc. IEEE INFOCOM*, 2011, pp. 1979–1987.
[29] S. Ross, *A First Course in Probability*, 6th ed. Upper Saddle River, NJ, USA: Prentice-Hall, 2002.
[30] H. Al-Zubaidy, J. Liebeherr, and A. Burchard, "A (min,x)-network calculus for multi-hop fading channels," in *Proc. IEEE INFOCOM*, Apr. 2013, pp. 1833–1841.
[31] G. Grimmett and D. Stirzaker, *Probability and Random Processes*, 3rd ed. London, U.K.: Oxford Univ. Press, 2001.
[32] G. Bolch, S. Greiner, H. de Meer, and K. S. Trivedi, *Queueing Networks and Markov Chains*, 2nd ed. Hoboken, NJ, USA: Wiley, 2006.
[33] F. Ciucu, "Exponential supermartingales for evaluating end-to-end backlog bounds," in *Proc. MAMA Workshop ACM SIGMETRICS*, Jun. 2007, pp. 21–23.
[34] M. Fidler, "A network calculus approach to probabilistic quality of service analysis of fading channels," in *Proc. IEEE GLOBECOM*, Nov. 2006, pp. 1–6.
[35] G. Mayor and J. Silvester, "Time scale analysis of an ATM queueing system with long-range dependent traffic," in *Proc. IEEE INFOCOM*, Apr. 1997, pp. 205–212.
[36] N. Fonseca, G. Mayor, and C. Neto, "On the equivalent bandwidth of self-similar sources," *ACM Trans. Model. Comput. Simul.*, vol. 10, no. 2, pp. 104–124, Apr. 2000.
[37] C. Melo and N. Fonseca, "Envelope process and computation of the equivalent bandwidth of multifractal flows," *Comput. Netw.*, vol. 48, no. 3, pp. 351–375, Jun. 2005.
[38] N. G. Duffield and N. O'Connell, "Large deviations and overflow probabilities for the general single-server queue, with applications," *Math. Proc. Camb. Phil. Soc.*, vol. 118, no. 2, pp. 363–375, Sep. 1995.
[39] I. Norros, "On the use of fractional Brownian motion in the theory of connectionless networks," *IEEE J. Sel. Areas Commun.*, vol. 13, no. 6, pp. 953–962, Aug. 1995.
[40] K. Mahmood, A. Rizk, and Y. Jiang, "On the flow-level delay of a spatial multiplexing MIMO wireless channel," in *Proc. IEEE ICC*, Jun. 2011, pp. 1–6.
[41] R. Lübben and M. Fidler, "On the delay performance of block codes for discrete memoryless channels with feedback," in *Proc. IEEE Sarnoff Symp.*, May 2012, pp. 1–6.
[42] R. Agrawal, R. L. Cruz, C. M. Okino, and R. Rajan, "Performance bounds for flow control protocols," *IEEE/ACM Trans. Netw.*, vol. 7, no. 3, pp. 310–323, Jun. 1999.
[43] S. Akin and M. C. Gursoy, "Effective capacity analysis of cognitive radio channels for quality of service provisioning," *IEEE Trans. Wireless Commun.*, vol. 9, no. 11, pp. 3354–3364, Nov. 2010.
[44] F. Ciucu, "On the scaling of non-asymptotic capacity in multi-access networks with bursty traffic," in *Proc. IEEE ISIT*, Aug. 2011, pp. 2547–2551.
[45] M. Bredel and M. Fidler, "Understanding fairness and its impact on quality of service in IEEE 802.11," in *Proc. IEEE INFOCOM*, Apr. 2009, pp. 1098–1106.

Markus Fidler (M'04–SM'08) received the Doctoral degree in computer engineering from RWTH Aachen University, Germany, in 2004. He was a Post-Doctoral Fellow of NTNU Trondheim, Norway, in 2005 and the University of Toronto, Toronto, ON, Canada, in 2006. During 2007 and 2008, he was an Emmy Noether Research Group Leader at Technische Universität Darmstadt, Germany. Since 2009, he has been a Professor of communications networks at Leibniz Universität Hannover, Germany.

Amr Rizk (M'13) graduated in electrical engineering and business administration in 2008 from TU Darmstadt. He received the Doctoral degree Dr.-Ing. from the Leibniz Universität Hannover, Germany. His research interests are in the areas of network performance evaluation, stochastic modeling and teletraffic theory.