A Guide to the Stochastic Network Calculus

Markus Fidler  Amr Rizk
Institute of Communications Technology
Leibniz Universität Hannover

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. For this purpose, we will now and then trade generality or precision for simplicity. 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 exponentially decaying overflow profile are derived to compute statistical end-to-end backlog and delay bounds for networks.

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 [13]. Minimum service guarantees that are provided by systems, such as a router, a scheduler, or a link, are characterized by so-called service curves [14]. Based on these concepts, the network calculus offers convolution forms [8], [24] 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 [26]. Statistical multiplexing of traffic flows is dealt with efficiently by the theory of effective bandwidths [8], [23] 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 [7], [8] and statistical envelopes [15], [26], [37], [38], 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 [26], [37].

Despite the early interest in a stochastic version of the network calculus, end-to-end convolution forms for networks of systems 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 have been obtained in [11] using the statistical envelope model. Performance bounds derived thereof grow as \( \Theta(n \log n) \) for \( n \) systems in series [6]. Convolution forms that are based on MGFs are established in [8], [17]. Compared to the use of statistical envelopes, MGFs utilize the additional assumption of statistical independence. Respective end-to-end performance bounds scale in \( O(n) \) [17].

Compared to classical queueing theory, the stochastic network calculus comprises a much larger variety of stochastic processes, including long range dependent, self-similar [27], [33], and heavy-tailed traffic [27]. This generality comes at the expense of exact solutions. Instead, the stochastic network calculus computes statistical performance bounds of the type \( \mathbb{P}(\text{backlog} > x) \leq \varepsilon \) or \( \mathbb{P}(\text{delay} > x) \leq \varepsilon \).

With this work, we aim at an intuitive introduction to the stochastic network calculus. We seek to define a minimal framework that enables deriving known major results. We contribute a self-contained exposition of basic methods and closed form results derived thereof. To this end, we restrict the presentation to affine envelope functions of MGFs [7] and corresponding linear statistical envelope functions with an exponentially bounded burstiness (EBB) [37]. For more general envelope models as well as models that provide stronger guarantees, we refer the reader to the related literature, e.g., [19], [22], [26]. Also, we will occasionally forgo generality or precision in favor of simplicity.

An outline of the method, that is assembled in this work, is shown in Fig. 1. We use the affine MGF envelope model from [7] to characterize arrival processes and service processes, respectively. The model has two parameters, \( \rho \) that is an envelope rate and \( \sigma \) that is a burstiness measure. Formulas for statistical multiplexing, scheduling, and convolution of 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 characterization of the arrivals and of the service, respectively, is composed to compute performance bounds for backlog and delay.

We note, that a duality of MGFs and statistical envelopes exists [26]. 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 is straight-forward in case of MGFs, whereas in the absence of statistical independence the EBB model is favorable [34], [37]. Other, sometimes subtle, differences of the two models exist.

The remainder of this paper is structured as follows. In Sect. 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. We also derive corresponding results for tandem systems. Building blocks of the outlined methods are highlighted as framed equations. In Sect. III, we provide a catalogue of MGF envelopes for relevant arrival and server models. We also include results for statistical multiplexing and scheduling. Sect. IV concludes the paper with a set of guidelines for application and an outlook.

II. FUNDAMENTALS

This section provides an introduction to the stochastic network calculus. In Sect. II-A, we formulate the basic queueing model, where traffic that arrives at a system between times \(\tau\) 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 Sect. II-B and Sect. II-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 Sect. II-D, we show how to extend the method to multi-node networks. Sect. 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 [11]. We denote \(A(t)\) the 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 \(A(\tau,t) = A(t) - A(\tau)\) where \(t \geq \tau \geq 0\) to denote the arrivals in \([\tau + 1, t]\). Similarly, \(D(t)\) denotes the departures from the system.

We characterize systems using the concept of a dynamic server [8], [9] that relates the departures of a system to its arrivals as

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

(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 \([\tau + 1, t]\) [8]. To see this, assume \(t\) and \(\tau\) 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 \(\tau^*\) to be the beginning of the last busy period before \(t\), i.e., at \(\tau^*\) the system was empty for the last time before \(t\). Consequently,

\[
D(\tau^*) = A(\tau^*) + S(\tau^*,t).
\]

(2)

Finally, since \(\tau^*\) is not generally known, we 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 \(D(t) \leq D(\tau) + S(\tau,t)\) 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

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

(3)

Combined with (1), it follows that

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

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 [24] for details. Non-linear systems, such as a first-in first-out scheduler [28], 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, 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 [8], [24]. Among others, the min-
plus convolution is associative, which enables an elegant
composition of tandem systems. Consider two dynamic servers
$S_1(t, t)$ and $S_2(t, t)$ 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

$$D_2(t) = (A_1 \otimes S_1) \otimes S_2(t) = A_1 \otimes (S_1 \otimes S_2(t)).$$

(4)

As a main result, it follows that

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

satisfies the definition of dynamic server (1), i.e., the tandem
of the two systems can be substituted by a single equivalent
system of the two systems can be substituted by a single equivalent
system of the two systems can be substituted by a single equivalent
system of the two systems can be substituted by a single equivalent
system of the two systems can be substituted by a single equivalent
system of the two systems can be substituted by a single equivalent
system of the two systems can be substituted by a single equivalent
system of the two systems can be substituted by a single equivalent
system of the two systems can be substituted by a single equivalent
system of the two systems can be substituted by a single equivalent
system of the two systems can be substituted by a single equivalent
system of the two systems can be substituted by a single equivalent
system of the two systems can be substituted by a single equivalent
system of the two systems can be substituted by a single equivalent
system of the two systems can be substituted by a single equivalent
system of the two systems can be substituted by a single equivalent
system of the two systems can be substituted by a single equivalent
system of the two systems can be substituted by a single equivalent
system of the two systems can be substituted by a single equivalent
system of the two systems can be substituted by a single equivalent
system of the two systems can be substituted by a single equivalent
system of the two systems can be substituted by a single equivalent
system of the two systems can be substituted by a single equivalent
system of the two systems can be substituted by a single equivalent
system of the two systems can be substituted by a single equivalent
system of the two systems can be substituted by a single equivalent
system of the two systems can be substituted by a single equivalent
system of the two systems can be substituted by a single equivalent
system of the two systems can be substituted by a single equivalent
system of the two systems can be substituted by a single equivalent
system of the two systems can be substituted by a single equivalent
system of the two systems can be substituted by a single equivalent
system of the two systems can be substituted by a single equivalent
system of the two systems can be substituted by a single equivalent
system of the two systems can be substituted by a single equivalent
system of the two systems can be substituted by a single equivalent
system of the two systems can be substituted by a single equivalent
system of the two systems can be substituted by a single equivalent
system of the two systems can be substituted by a single equivalent
system of the two systems can be substituted by a single equivalent
system of the two systems can be substituted by a single equivalent
system of the two systems can be substituted by a single equivalent
system of the two systems can be substituted by a single equivalent
system of the two systems can be substituted by a single equivalent
system of the two systems can be substituted by a single equivalent
system of the two systems can be substituted by a single equivalent
system of the two systems can be substituted by a single equivalent
system of the two systems can be substituted by a single equivalent
system of the two systems can be substituted by a single equivalent
system of the two systems can be substituted by a single equivalent
system of the two systems can be substituted by a single equivalent
system of the two systems can be substituted by a single equivalent
system of the two systems can be substituted by a single equivalent

### B. Arrival Envelopes

To derive actual backlog and delay bounds, the deterministic
network calculus uses univariate envelope functions that are
defined as deterministic upper bounds of the arrivals $A(t, t)$
for all time intervals $[\tau, t]$ with $t \geq \tau \geq 0$. A widely applied model are affine envelope functions of the type $\rho(t - \tau) - b$
that map to the parameters $\rho > 0$ and $b \geq 0$ of a leaky bucket
shaper [13]. The arrivals have a deterministic envelope if for all $t \geq \tau \geq 0$ it holds that

$$A(t, t) \leq \rho(t - \tau) + b.$$  

(7)

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 [8], [23], 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 random processes is the product of their respective MGFs. The
MGF of an arrival process $A(t, t)$ is defined as $E[e^{\theta A(t, t)}]$ with
free parameter $\theta \geq 0$. Under the assumption of stationarity,
$P[A(t, t) \leq x] = P[A(t - \tau) \leq x]$ 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 $M_A(\theta, t - \tau) = E[e^{\theta A(t, t)}]$. The normalized log MGF
$\ln M_A(\theta, t)/(\theta t)$ is known as the effective bandwidth. For
increasing $\theta > 0$ it grows from the mean to the peak rate of
the arrivals. Corresponding to the affine envelope model, [7],
[8] defines an MGF envelope for $t \geq \tau \geq 0$ as

$$E[e^{\theta A(t, t)}] \leq e^{\theta(s(t - \tau) + \sigma)}$$  

(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 [37] to provide a
guarantee of the form

$$P[A(t, t) > \rho(t - \tau) + b] \leq \varepsilon(b)$$  

(9)

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

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

(10)

where $\alpha \geq 0$.

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

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

(11)

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

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

(12)

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

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) can not. The reason is that (6) evaluates all \( \tau \in [0, t] \), where the \( \tau^* \) that attains the maximum is a random variable [26]. 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 [11], [15], [26], [38] is required that has the form
\[
P[\exists \tau \in [0, t] : A(\tau, t) > \rho'(t - \tau) + b] \leq \varepsilon'(b) \tag{13}
\]
for all \( t \geq 0 \). Throughout this work we use superscript \( \varepsilon' \) 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 \( P[\exists i : X_i \geq x] = P[\max_i \{X_i\} \geq x] \) and approximate the expression by its largest term as
\[
P[\max_i \{X_i\} \geq x] \geq \max_i \{P[X_i \geq x]\}. \tag{14}
\]
Note, however, that the expression only provides a lower bound of an upper bound [26]. 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
\[
P[\exists i : X_i \geq x] \leq \sum_i P[X_i \geq x]. \tag{15}
\]
Regarding (13), it follows as, e.g., in [11], [38] that
\[
P[\exists \tau \in [0, t] : A(\tau, t) > \rho'(t - \tau) + b]
\leq \sum_{\tau=0}^{t} e^{\theta \sigma} e^{-\theta(b + \delta(t - \tau))}
\leq e^{\theta \sigma} e^{-\theta b} \sum_{\tau=0}^{\infty} e^{-\theta \delta \tau}
= \frac{e^{\theta \sigma} e^{-\theta b}}{1 - e^{-\theta \delta}}.
\]
In the second line, we used the union bound (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 \( \delta \) increases the envelope rate and decreases the overflow profile. We note that the summation at \( \tau = t \) can be omitted to improve the precision as \( A(t, t) = 0 \) by definition. 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 \( \rho \) and \( \sigma \), 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.

C. Service Envelopes

Unlike most of the network calculus literature, that defines (statistical) service curves as non-random functions, we take the definition of dynamic server (1), that defines service as a random process \( S(\tau, t) \), as a starting point. We will use the basic methods from Sect. II-B with some minor adaptations to derive lower service envelopes. 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 \( b/\rho \) 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 bandwidths as effective capacity [36]. 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
\[
E[e^{-\theta S(\tau, t)}] \leq e^{-\theta(\rho(t - \tau) - \sigma)} \tag{17}
\]
Note that although (17) is phrased as an upper bound, it defines a lower bound of the service due to the use of \( -\theta \) and \( \theta \geq 0 \). Also, the parameters \( \rho \) and \( \sigma \) are functions of \( -\theta \). Assuming stationarity of the service process \( S(\tau, t) \), we will frequently use shorthand notation \( M_S(-\theta, t - \tau) = E[e^{-\theta S(\tau, t)}] \). The normalized log MGF \( \ln M_S(-\theta, t)/(\theta t) \) is known as the effective capacity. It decreases for increasing \( \theta > 0 \) from the mean rate of the service to zero.

Statistical service envelopes that mirror the concept of EBB are defined in [25] as the so-called exponentially bounded fluctuation (EBF) model with parameters \( \rho > 0 \), \( b \geq 0 \) and
\[
P[S(\tau, t) < \rho(t - \tau) - b] \leq \varepsilon(b), \tag{18}
\]
where the deficit profile \( \varepsilon(b) \) decays exponentially as \( \varepsilon(b) = \alpha e^{-\theta b} \) and \( \alpha \geq 0 \). From Chernoff’s lower bound
\[
P[X \leq x] \leq e^{\theta \sigma} E[e^{-\theta X}] \tag{19}
\]
for \( \theta \geq 0 \) it follows that \( \varepsilon(b) = e^{\theta \sigma} e^{-\theta b} \). Finally, the sample path envelope
\[
P[\exists \tau \in [0, t] : S(\tau, t) < \rho'(t - \tau) - b] \leq \varepsilon'(b) \tag{20}
\]
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 (15) and the same basic steps as in Sect. II-B. 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 \( \tau = 0, 1, \ldots, t - [b/\rho'] \) since \( S(\tau, t) \) is non-negative. 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 [12]. 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 [26]. First end-to-end solutions that make use of the convolution-form appeared in the stochastic network calculus in [11], [17].

In the sequel, we derive the EBF deficit profile first for two and then by recursive insertion for n dynamic servers in tandem. For the MGF of the min-plus convolution of two statistically independent and stationary service processes we estimate [8], [17]
\[
E[e^{-\theta (S_1 \oplus S_2)(\tau,t)}] = E\left[ e^{-\theta \min_{v \in [\tau,t]} [S_1(\tau,v)+S_2(\tau,v)]} \right] \\
\leq \sum_{v=\tau}^{t} E\left[ e^{-\theta S_1(\tau,v)} \right] E\left[ e^{-\theta S_2(\tau,v)} \right] \\
= \sum_{v=0}^{t-\tau} M_{S_1}(\theta,v) M_{S_2}(\theta,t-\tau-v) \\
= M_{S_1} \ast M_{S_2}(\theta,t-\tau).
\]

In the second line, we estimated the expectation of a maximum 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
\[
M_{S_{\text{net}}}(\theta,t) \leq M_{S_1} \ast M_{S_2} \ast \cdots \ast M_{S_n}(\theta,t) \\
= \sum_{\tau_i \geq 0, \sum_{i=1}^{n} \tau_i = t} M_{S_1}(\theta,\tau_1) M_{S_2}(\theta,\tau_2) \cdots M_{S_n}(\theta,\tau_n). \quad (22)
\]

Next, we assume homogeneous MGF envelopes (17). The non-homogeneous case requires additional notation. Since the convolution is order preserving, we can substitute \(M_{S_i}(\theta,t) \leq e^{\theta t} e^{-\theta t} \) for \(i = 1, 2, \ldots, n\). 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 \((\tau_1, \tau_2, \cdots \tau_n)\) that satisfy \(\sum_{i=1}^{n} \tau_i = t\) [35]. It follows that
\[
M_{S_{\text{net}}}(\theta,t) \leq e^{n\theta \tau} \binom{t+n-1}{n-1} e^{-\theta t}.
\]

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
\[
P[S_{\text{net}}(\tau,t) < \rho(t-\tau) - b] \leq e^{X(\rho(t-\tau) - b) M_{S_{\text{net}}}(\theta,t) - \theta b}.
\]

Using the same basic approach as in Sect. II-B and Sect. II-C, a sample path envelope (20) can be derived as
\[
P[\exists \tau \in [0,t] : S_{\text{net}}(\tau,t) < \rho(t-\tau) - b] \\
\leq \sum_{\tau=0}^{t} e^{\theta \tau} \binom{t+n-1}{n-1} e^{-\theta \tau} \\
\leq e^{n\theta \tau} \sum_{\tau=0}^{\infty} \binom{t+n-1}{n-1} e^{-\theta \tau} \\
= e^{n\theta \tau} \sum_{\tau=0}^{\infty} \binom{t+n-1}{n-1} e^{-\theta \tau} \\
= e^{n\theta \tau} \binom{t+n-1}{n-1} e^{-\theta \tau}.
\]

In the second line, we substituted \(\rho' = \rho - \delta\) where \(\delta > 0\) and \(\theta > 0\) are free parameters and used the union bound (15). In fact, the sum has to be evaluated only for \(\tau = 0, 1, \ldots, t - \lfloor b/\rho' \rfloor\) as \(S(\tau,t)\) is non-negative. 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
\[
e(\theta) = \left( \frac{e^{\theta \rho}}{1-e^{-\theta \delta}} \right)^n e^{-\theta b}.
\]

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

E. Backlog and Delay Bounds

So far, we considered envelope models of arrivals and service independently. For computation of 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) and assume a sample path where
\[
A(\tau,t) \leq \rho_A'(t-\tau) + b_A \quad (24)
\]
for all \(\tau \in [0,t]\) and \(t \geq 0\). Also, consider service with envelope (20) and assume a sample path where
\[
S(\tau,t) \geq \rho_S'(t-\tau) - b_S \quad (25)
\]
and generally \(S(\tau,t) \geq 0\) for all \(\tau \in [0,t]\) and \(t \geq 0\). By insertion into (6), a backlog bound \(B(t) \leq b\) for any \(t \geq 0\) follows as
\[
b = \max_{\tau \in [0,t]} \{\rho_A'(t-\tau) + b_A - [\rho_S'(t-\tau) - b_S]_+\}. \quad (26)
\]
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 that

\[
P [ B(t) > b ] = \varepsilon'_A(b_A) + \varepsilon'_S(b_S) = \varepsilon'.
\]

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 Sect. II-B and Sect. II-D. For stability \( \delta \leq (\rho_S - \rho_A)/2 \) and

\[
\rho_A < \rho_S
\]

is required. The backlog bound follows as

\[
b = b_A + b_S \frac{\rho_A + \delta}{\rho_S - \delta}.
\]

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

\[
P [ W(t) > w ] \leq \varepsilon'_A(b_A) + \varepsilon'_S(b_S) = \varepsilon',
\]

where

\[
w = b_A + b_S \frac{\rho_A + \delta}{\rho_S - \delta}.
\]

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

\[
b_A = \sigma_A - \frac{1}{\theta} \left( \ln \left( \frac{\varepsilon'}{2} \right) + \ln \left( 1 - e^{-\delta_\theta} \right) \right).
\]

and \( b_S \) from (23) as

\[
b_S = n \sigma_S - \frac{1}{\theta} \left( \ln \left( \frac{\varepsilon'}{2} \right) + n \ln \left( 1 - e^{-\theta_\delta} \right) \right).
\]

The three summands of (31) and (32) are due to the burstiness measure \( \sigma \), the violation probability \( \varepsilon' \), and the sample path derivation using slack rate \( \delta \). Regarding \( n \)-node networks, (32) exhibits a linear dependence on \( n \). As an important consequence, backlog and delay bounds derived thereof grow in \( \mathcal{O}(n) \). Finally, we optimize \( \theta > 0 \) and \( 0 < \delta \leq (\rho_S - \rho_A)/2 \) to minimize \( b \) (29) and \( w \) (30).

III. TRAFFIC AND SERVER MODELS

In this section, we consider basic server models (Sect. III-A), traffic models (Sect. III-B), rules for multiplexing (Sect. III-C), and a basic model for scheduling (Sect. III-D).

A. Server Models

1) Constant Rate Server: First, we consider a deterministic constant rate server with capacity \( c \), i.e., \( S(\tau, t) = c(t - \tau) \) for all \( t \geq \tau \geq 0 \). Formally expressed as 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 \( b_S = 0 \) to find the backlog bound from (29) and the delay bound from (30) as

\[
b = b_A, \quad \text{and} \quad w = \frac{b_A}{c},
\]

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

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

We will use these bounds to evaluate different traffic models in the following sections.

2) Memoryless On-Off Server: A memoryless server has service process for \( t \geq \tau \geq 0 \)

\[
S(\tau, t) = \sum_{v=\tau+1}^{t} X_v,
\]

where the increments \( X_v \) are independent and identically distributed (iid) random variables. From the independence, it follows that

\[
M_S(-\theta, t) = (M_X(-\theta))^t
\]

for \( t \geq 0 \). For the special case of an On-Off server the \( X_i \) are iid Bernoulli trials with probability mass function \( p_X(r) = p_{on} \) and \( p_X(0) = 1 - p_{on} \). The MGF, respectively, Laplace transform follows as

\[
M_X(-\theta) = \sum_{r=0}^{\infty} e^{-\theta r} p_X(r) = p_{on} e^{-\theta r} + 1 - p_{on}.
\]

It follows that \( M_S(-\theta, t) = (p_{on} e^{-\theta r} + 1 - p_{on})^t \) for \( t \geq 0 \) is binomial and has an envelope (17) with parameters \( \sigma = 0 \) and rate

\[
\rho = \ln(p_{on} e^{-\theta r} + 1 - p_{on})
\]

for \( \theta > 0 \).
B. Traffic Models

1) Poisson: Let $N(t)$ be a Poisson process with mean rate $\lambda$, i.e., $P[N(t) = k] = e^{-\lambda t}(\lambda t)^k/k!$ for $t > 0$ and $N(0) = 0$. The process $N(t)$ counts the number of arrivals in $[0, t]$. The MGF of this process is known as \[35\]
$$M_N(\theta, t) = e^{\lambda(e^\theta - 1)}.$$  

Given arrivals of constant size $1/\nu$ it follows that $A(t) = N(t)/\nu$, so that $M_A(\theta, t)$ has an envelope (8) with parameters $\sigma = 0$ and rate $\rho = \lambda(\theta/\nu - 1)/\theta$ 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 $M|D|1$ model results if the arrivals are iid exponential random variables $X_k$ with mean $1/\nu$ and MGF $M_X(\theta) = \nu/(\nu - \theta)$ for $\theta < \nu$. In this case the arrival process is $A(t) = \sum_{k=1}^{N(t)} X_k$.

It has conditional MGF $E[e^{\theta A(t)} | N(t) = k] = (M_X(\theta))^k$ [35], such that
$$E[e^{\theta A(t)}] = E[(M_X(\theta))^{N(t)}] = E[e^{\theta \ln(M_X(\theta)) N(t)}].$$

By choice of $\theta = \ln(M_X(\theta))$, it follows that [21]
$$M_A(\theta, t) = M_N(\ln(M_X(\theta)), t).$$

Insertion of the exponential MGF into the Poisson MGF (35) gives $M_A(\theta, t) = e^{\theta \ln(M_X(\theta))}$ that satisfies (8) with $\sigma = 0$ and envelope rate $\rho = \lambda/\nu - \theta$ for $0 \leq \theta < \nu$. For $\theta = 0$ we obtain the mean rate $\lambda/\nu$.

In Fig. 3, we depict the delay bound $w$ from (33) for Poisson arrivals 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) with mean $1/\nu$, respectively. The mean service time follows as $1/(\nu c)$. For comparison, we show the exact response time from queueing theory [4] $\varepsilon' = e^{-\nu c(1-\varphi)}w$, where $\varphi = \lambda/(\nu c)$ is the utilization. We fixed $\lambda = 0.8$, $\nu = 1$, and $c = 1$, such that $\varphi = 0.8$. We optimized the free parameter $\theta$ numerically. The curves in Fig. 3 show an exponential decay that is characteristic for EBB arrival processes. Clearly, delays are smaller in case of the M|D|1 model. Compared to the exact M|M|1 result, the bound from the stochastic network calculus is conservative, but shows the same exponential decay. In the discrete time model, delays are measured in units of timeslots. Given packets of say 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. 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 $p_\text{on} = p_{12}/(p_{12} + p_{21})$, where $p_{ij}$ for $i, j = 1, 2$ are the transition probabilities from state $i$ to state $j$. The mean arrival rate follows as $p_\text{on}r$. 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 On-Off Markov process satisfies (8) with $\sigma = 0$ and envelope rate [8]
$$\rho = \frac{1}{\theta} \ln \left( p_{11} + p_{22} e^{\theta r} + \sqrt{(p_{11} + p_{22} e^{\theta r})^2 - 4(p_{11} + p_{22} - 1) e^{2\theta r}} \right)$$
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 $p_\text{on} = p_{22}$ and
$$\rho = \frac{\ln(p_\text{on} e^{\theta r} + 1 - p_\text{on})}{\theta}$$
for $\theta > 0$. Note how the envelope rate of the arrival process parallels the corresponding rate of the service process (34).

3) Fractional Brownian Motion: While many relevant arrival processes fall into the EBB class (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 a self-similar arrival process with correlated Gaussian increments. It has MGF
$$M_A(\theta, t) = e^{\theta(M + \frac{\theta^2}{2} \zeta^2 t^2)},$$
(36)

where $\lambda$ is the mean rate, and $\zeta^2$ the variance of the increments. Parameter $h$ is the Hurst parameter where $h \in (0, 1)$ denotes long range dependence (LRD). If $h = 0.5$, fBm becomes standard Brownian motion that has envelope rate (8)
$$\rho = \lambda + \frac{\theta \zeta^2}{2}.$$  

In case of LRD, i.e., $h \in (0.5, 1)$, (36) grows superlinearly with $t$, such that no affine MGF envelope as defined in (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 applied [20], [26], [31]. By Chernoff’s bound (11) it holds for \( \theta \geq 0 \) that

\[
P[A(\tau, t) > E(t - \tau)] \leq e^{-\theta E(t - \tau) M_A(\theta, t - \tau)} = \varepsilon.
\]

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

\[
E(t) = \inf_{\theta > 0} \left\{ \frac{1}{\theta} (\ln M_A(\theta, t) - \ln \varepsilon) \right\}
\]

for \( t \geq 0 \). By insertion of (36), the minimum can be obtained at \( \theta = \sqrt{-2\ln \varepsilon} / (c t^h) \), such that [20], [26], [31]

\[
E(t) = \lambda t + \sqrt{2\ln \varepsilon} c t^h.
\]

(37)

Following the steps of Sect. II-B, we have to construct a sample path envelope of the form \( P[\exists \tau \in [0, t] : A(\tau, t) > E(t - \tau)] \leq \varepsilon' \) to be able to derive performance bounds. A respective solution is provided in [33]. Instead, in this work, we use the much simpler approximation by the largest term (14) to approximate \( \varepsilon' \approx \varepsilon \). In this case, a backlog bound \( P[B(t) > b] \approx \varepsilon \) at a constant rate server with capacity \( c \) follows from (6) by substitution of \( E(t - \tau) \) from (37) for \( A(\tau, t) \) and \( S(\tau, t) = c(t - \tau) \). Letting \( t \rightarrow \infty \), a backlog bound is

\[
b = \max_{\tau \geq 0} \{ \lambda \tau + \sqrt{2\ln \varepsilon} c t^h - c \tau \}.
\]

The maximum is attained at [20]

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

By insertion of \( \tau^* \) and after solving for \( \varepsilon \) the result

\[
\varepsilon = \exp \left( \frac{1}{2\sigma^2} \left( \frac{c - \lambda}{h} \right)^{2h \left( \frac{b}{1 - h} \right)^{2 - 2h}} \right),
\]

(38)

that was first reported in [16], [32], is recovered in the stochastic network calculus.

In Fig. 4, we depict results from (38) for \( h = 0.5, 0.6, 0.7 \). 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 \( \varepsilon' \) with \( b \) is also obtained from the sample path analysis [33]. The Weibull tail significantly impacts resource dimensioning as it demonstrates the inefficiency of buffering LRD traffic [20], [32]. Regarding (38), 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 \( h \).

C. Statistical Multiplexing

The aggregate arrival process of the superposition of \( m \) arrival processes is

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

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

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

If the arrival processes \( A_i \) have MGF envelope (8) with parameters \( \rho \) and \( \sigma \) it follows that

\[
E[e^{\theta A_{agg}(\tau, t)}] \leq e^{\theta (m \rho (t - \tau) + m \sigma)},
\]

where we considered the homogeneous case for notational simplicity. In general, the parameters of the MGF envelope model are additive. The aggregate arrivals have MGF envelope (8) with parameters

\[
\sigma_{A_{agg}} = m \sigma_A,
\]

\[
\rho_{A_{agg}} = m \rho_A.
\]

(39)
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 $10^{-3}$. The sources are statistically independent and have peak rate $r = 1$, mean rate $0.05$, and burstiness parameter $T = 300$. The delay bound is computed from (33) where we use the aggregate traffic parameters from (39). 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. The statistical multiplexing gain is achieved by optimizing the free parameter $\theta$.

### D. Scheduling

We show a general scheduling model from [17] 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 $A(t) = A_{th}(t) + A_{cr}(t)$ and $D(t) = D_{th}(t) + D_{cr}(t)$ 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

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

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

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

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

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

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

$$S_0(\tau, t) = [S(\tau, t) - A_{cr}(\tau, t)]_+$$

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 $A_{cr}(\tau, t)$ the MGF is derived for $t \geq \tau \geq 0$ as

$$E[e^{-\theta S_0(\tau, t)}] \leq E[e^{-\theta S(\tau, t)}] E[e^{\theta A_{cr}(\tau, t)}].$$

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

$$E[e^{-\theta S_0(\tau, t)}] \leq e^{-\theta((\rho_S - \rho_{A_{cr}})(t-\tau) - (\sigma_S + \sigma_{A_{cr}}))},$$

such that the leftover service process $S_0$ has MGF envelope (17) with parameters

$$\sigma_{S_0} = \sigma_S + \sigma_{A_{cr}},$$

$$\rho_{S_0} = \rho_S - \rho_{A_{cr}}.$$  \hspace{1cm} (40)

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 (40). Further, we let the cross traffic parameters be the parameters of aggregated traffic from (39).

In Fig. 6, 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 $\lambda = 0.25$ and the size of the arrivals is $1/\nu = 1$. 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 $T = 10, 20, 40, 80$. We observe that the burstiness of the cross traffic directly impacts the delay bound of the through traffic, where it changes the slope of the exponential decay.
In Fig. 7, we show end-to-end delay bounds for Poisson traffic through cross traffic that traverses a tandem of $n$ homogeneous constant rate servers, each with independent Markov On-Off cross traffic. The traffic and service parameters are the same as for Fig. 6. 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.

IV. CONCLUSIONS AND OUTLOOK

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

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. 8. Tab. 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 $\rho_A, \sigma_A$ and the service $\rho_S, \sigma_S$ 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. 8. The resulting system is fully characterized by the EBB parameters of the arrival envelope $\rho_A, b_A$ and the network service envelope $\rho_S, b_S$. Step 3 provides backlog and delay bounds under the stability condition $\rho_A < \rho_S$. Finally, the violation probability $\varepsilon'$ can be fixed and the free parameters $\theta > 0$ and $0 < \delta \leq (\rho_S - \rho_A)/2$ can be optimized.

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 Tab. I. In fact, the transition to EBB could as well take place at any other step, resulting, however, in a method with different qualities. Tab. 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 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 [22], [34], [37]. As the MGF transforms convolution into multiplication, it provides the computationally simpler approach.

**Scheduling:** We presented the general scheduling model from [17] 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 in [26] using statistical envelopes.

**Multi-node networks:** We showed end-to-end performance bounds for $n$ statistically independent systems in series that grow in $\Theta(n)$ [17]. Without assumption of independence, an upper bound $\Theta(n \log n)$ is derived in [11] using the EBB model. A corresponding lower bound is proven in [6].

**Backlog and delay:** Finally, we mention that the transition from MGF to EBB envelopes that is used in this paper can be omitted. In [17], MGFs of backlogs and delays are derived and in a final step Chernoff’s bound is used to compute performance bounds. The final computation is, however, more involved and less intuitive than in case of EBB.

---

**TABLE I**

<table>
<thead>
<tr>
<th>Building Blocks for Composition.</th>
</tr>
</thead>
<tbody>
<tr>
<td>basic steps of the method</td>
</tr>
<tr>
<td>1.a) arrival parameters $\rho_A, \sigma_A$</td>
</tr>
<tr>
<td>1.b) service parameters $\rho_S, \sigma_S$</td>
</tr>
<tr>
<td>2.a) EBB arrival envelope $\rho_A, b_A$</td>
</tr>
<tr>
<td>2.b) EBB service envelope $\rho_S, b_S$</td>
</tr>
<tr>
<td>3.a) stability $\rho_A &lt; \rho_S$</td>
</tr>
<tr>
<td>3.b) backlog $P[B(t) &gt; b] \leq \varepsilon'$</td>
</tr>
<tr>
<td>3.c) delay $P[V(t) &gt; w] \leq \varepsilon'$</td>
</tr>
<tr>
<td>3.d) optimize parameters $\theta &gt; 0, 0 &lt; \delta &lt; (\rho_S - \rho_A)/2$</td>
</tr>
</tbody>
</table>

**TABLE II**

<table>
<thead>
<tr>
<th>Pros and Cons of MGF versus EBB Envelopes.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
<tr>
<td>------------------------------------------</td>
</tr>
<tr>
<td>statistical independence</td>
</tr>
<tr>
<td>statistical multiplexing</td>
</tr>
<tr>
<td>schedulers</td>
</tr>
<tr>
<td>multi-node networks</td>
</tr>
<tr>
<td>backlog and delay</td>
</tr>
</tbody>
</table>

---
C. Open Challenges

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 \( P[B(t) > b] \leq \varepsilon \) (27) can be interpreted as the probability of buffer overflow, given a buffer of limited size \( b \), but provide only an approximation \[8\]. 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 \( A_{fe}(t) = \min[A_{un}(t), D(t)+x] \), where \( A_{un}(t) \) are the uncontrolled arrivals and \( x \) is the window size \[1\], \[8\], \[24\]. In the stochastic network calculus, the difficulty of this model is due to the fact that sample path guarantees for the departures are required to estimate the arrivals of the network.

Wireless channels: Non-equilibrium models of wireless channels receive growing interest, see, e.g., recent works in the area of effective capacity \[2\], \[36\] and in the stochastic network calculus \[3\], \[18\], \[22\], \[29\], \[30\]. Common channel models that have been explored so far, are Markov or memoryless processes, such as in Sect. III-A2, that are calibrated using, e.g., a fading process.

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 \[10\] and on CSMA/CA \[5\]. A possible approach to model the overhead due to retransmissions of lost packets is \[12\].

REFERENCES


