Statistical End-to-end Performance Bounds for Networks under Long Memory FBM Cross Traffic

Amr Rizk and Markus Fidler
Institute of Communications Technology, Leibniz Universität Hannover

Abstract—Fractional Brownian motion (fBm) became known as a useful model for Internet traffic incorporating its self-similar and long-range dependent properties. In this paper we derive end-to-end performance bounds for a through flow in a network of tandem queues under fBm cross traffic. We build on a previously derived sample path envelope for fBm, which possesses a Weibullian decay of overflow probabilities. We employ the sample path envelope and the concept of leftover service curves to model the remaining service after scheduling fBm cross traffic at a system. Using composition results for tandem curves to model the remaining service after scheduling fBm on a previously derived sample path envelope for fBm, which networks of tandem queues under fBm cross traffic. We build similar and long-range dependent properties. In this paper we as a useful model for Internet traffic incorporating its self-

I. INTRODUCTION

Since the beginning of the Internet classical queuing theory has been used as the decisive methodology to analyze many relevant problems in computer networking [1], most prominently to prove the efficiency of packet switching over circuit switching. The significance of queuing theory stems from explicit closed-form performance measures, e.g. backlog and delay, for queuing systems under Poisson traffic1, such as the M|M|1 single server system with infinite buffer. In practice results obtained for systems with infinite buffer are frequently transferred with reasonable precision to systems with finite buffer, e.g. for buffer sizing. The outstanding features of queuing theory that are essential for analysis of networks are (i) multiplexing and demultiplexing of flows2, i.e. routing, and (ii) product form queuing networks where constituent queues can be analyzed as if in isolation.

While it has long been known that Internet traffic sources do not fulfill the memoryless property assumed by the Poisson traffic model, it has been argued that the aggregate of a large number of multiplexed flows will tend towards Poisson. This assumption is backed up by the fact that the sum of many independent Bernoulli trials converges to a Poisson random variable. However, extensive measurements [2], [3], [4], [5], [6] revealed that aggregate Internet traffic comprises long-range dependence (LRD) as well as statistical self-similarity. One possible mathematical explanation arises from observed file size distributions on storage systems [4], [5]. This can be modeled analytically through the aggregation of many heavy tailed on-off traffic sources [7].

A random process widely accepted in the literature [2], [8], [9] as a useful model for aggregate Internet traffic is fractional Brownian motion (fBm) with Hurst parameter $H \in \left( \frac{1}{2}, 1 \right)$ [10]. It captures self similarity and it is LRD for $H \in \left( \frac{1}{2}, 1 \right)$. The cases $H = \frac{1}{2}$ and $H \in \left( 0, \frac{1}{2} \right)$ correspond to standard Brownian motion and short-range dependent fBm, respectively. Using theories including effective bandwidths and large deviations significant results have been derived for the performance of single queuing systems fed with LRD fBm traffic [9], [11], [12], [13]. Regarding the analysis of networks these theories can, however, not carry the outstanding properties of queuing theory forward to fBm.

Properties similar to (i) and (ii) from queuing theory have, however, been established in the deterministic [14], [15] and stochastic [14], [16], [17], [18], [19], [20] network calculus. The network calculus uses the concept of service curves [21], [22], [23] to characterize the service provided by queuing systems. Leftover service curves are used to analyze the effects of scheduling, i.e. multiplexing and demultiplexing cross traffic (i). Service curves of tandem systems are composed by convolution enabling the analysis of networks (ii). Finally, arrival envelopes are used as a model of traffic flows [24], [25], [26], [27], [28], [16] to derive performance bounds on backlog and delay for service curve systems.

In this paper we derive end-to-end statistical performance bounds for tandem systems under LRD fBm cross traffic. We build on a rigorous sample path envelope that we derived for fBm traffic in [29], [30] and also show an approximate envelope that follows from a known asymptotic backlog bound. These envelopes are the basis for network analysis using the stochastic network calculus. Owing to the concept of leftover service curves we quantify the effects of fBm cross traffic on the performance of through flows and find that the correlation of the cross traffic has severe impact. Finally, we derive end-to-end performance bounds for a through flow that traverses $n$ tandem systems each under fBm cross traffic. We show that these bounds grow in $O(\frac{n(\log n)^{2H-1}}{n})$ . Our finding compares to [18] where end-to-end performance bounds in $O(n \log n)$ have been derived for traffic with exponentially as opposed to Weibull bounded burstiness. For $H = \frac{1}{2}$ we recover the previous result. An earlier version of this paper is available as technical report [29].

The remainder of this paper is structured as follows. In Sect. II we review different statistical envelopes for fBm traffic and
illustrate the tail decay of overflow probabilities in comparison to traffic with exponentially bounded burstiness. In Sect. III we derive a leftover service curve under fBm cross traffic. In Sect. IV we present end-to-end service curves and statistical performance bounds for tandem systems, followed by a brief review of related scaling results. Finally, we present brief conclusions in Sect. V.

II. FBM THROUGH TRAFFIC AT A SINGLE SYSTEM

In this section we review the state-of-the-art of fBm envelopes and backlog bounds at single queuing systems. Further, we provide a comparison of the tail decays of overflow probabilities for fBm traffic and traffic with exponentially bounded burstiness. We conclude this section with a corollary on affine fBm envelopes that we derive from backlog bounds.

A. State-of-the-art FBM Backlog Bounds and Envelopes

fBm is a stochastic process that comprises self-similarity and can be short or long-range dependent based on its Hurst parameter $H \in (0, \frac{1}{2})$ or $H \in (\frac{1}{2}, 1)$, respectively. The second case is of great interest and is widely accepted for the modeling of aggregate Internet traffic [2], [8], [9]. An fBm process $Z(t)$ is a zero-mean process with stationary Gaussian increments, and in case of LRD, a superlinearly increasing variance $E[Z(t)^2] = \sigma^2 t^{2H}$ for all $t \geq 0$, where $\sigma > 0$ is its standard deviation at $t = 1$. The long-range dependence of fBm for $H \in (\frac{1}{2}, 1)$ is established by the diverging sum $\sum_{\tau} v(t) = \infty$ of the auto-covariance of the increments $v(t) \approx \sigma^2 H(2H-1)t^{2H-2}$ as $t \to \infty$. For a comprehensive introduction see e.g. the textbook [31].

We denote cumulative data arrivals of a traffic source in an interval $[\tau, t]$ by $A(\tau, t)$, and we use $A(t)$ to mean $A(0, t)$ for convenience. A statistical bound on the steady state backlog $B$ at a lossless constant rate work-conserving server with capacity $C$ can be derived for $t \to \infty$ as

$$P[B > b] = P \left[ \sup_{\tau \in [0, t]} \{ A(\tau, t) - C(t - \tau) \} > b \right]. \quad (1)$$

For the case of fBm traffic the arrivals can be modeled as a superposition of a mean rate $\lambda$ and an fBm process $Z(t)$ such that $A(t) = \lambda t + Z(t)$. Using approaches such as large deviations theory [32], [11], [14] or the largest term approximation [8], [9], significant results have been derived for the stochastic backlog bound (1), such as the asymptote $P[B > b] = \varepsilon_a$ for $b \to \infty$ where

$$\varepsilon_a = \exp \left( -\frac{1}{2\sigma^2} \left( \frac{C - \lambda}{H} \right)^{2H} \left( \frac{b}{1 - H} \right)^{2-2H} \right) \quad (2)$$

with mean rate $\lambda$, standard deviation $\sigma$ at $t = 1$ and Hurst parameter $H$. The result was proven to be logarithmically asymptotical generally and exact for $H = \frac{1}{2}$. The asymptote for infinite backlogs was further refined in [12], [33]. Considering finite backlogs (2) provides only an approximate result since its derivation [8], [9] builds on interchanging the probability and the supremum in (1), which makes it strictly a lower bound of an upper bound. This approximation is referred to as the max-approximation in [34] as it involves taking the supremum of the overflow probabilities at different times considering only the most probable time scale for overflow.

To provide non-asymptotic performance bounds we resort to the concept of effective envelopes in combination with the stochastic network calculus [27], [28], [16], [17], [18], [20]. Effective envelopes are stochastic upper bounds of the arrivals such that for any $t \geq \tau \geq 0$

$$P[A(\tau, t) - E(t - \tau) > 0] \leq \varepsilon_p. \quad (3)$$

The envelope in (3) is point-wise such that it can be violated at each point in time with probability $\varepsilon_p$. Such a point-wise envelope for fBm traffic was developed in [35], [36], [16] as

$$E(t) = \lambda t + \sqrt{-2\log \varepsilon_p} \sigma t^H. \quad (4)$$

The use of stochastic network calculus for the derivation of bounds implies, however, the use of sample path envelopes that are valid for all $\tau \in [0, t]$ such as

$$P \left[ \sup_{\tau \in [0, t]} \{ A(\tau, t) - E(t - \tau) \} > 0 \right] \leq \varepsilon_s \quad (4)$$

see e.g. [16] for details. fBm sample path envelopes as well as backlog and delay bounds for constant rate servers with fBm through traffic were provided in [29] first and more recently in [37]. In [29], [30] Boole’s inequality is used in a discrete time model to derive a sample path envelope of the form

$$E(t) = \lambda t + \sqrt{-2\log \eta \sigma t^{H+\beta}}$$

with sample path overflow probability according to (4)

$$\varepsilon_s = \frac{\Gamma(\frac{1}{2H})}{2\beta(-\log \eta)^{\frac{1}{2H}}} \quad (5)$$

and free parameters $\beta \in (0, 1-H)$ and $\eta \in (0, 1)$. Further, on, for a lossless work-conserving constant rate server with capacity $C$ and fBm through traffic, an upper bound $b$ on the steady state backlog $B$ with overflow probability $P[B > b] \leq \varepsilon_s$ is given by (5) where $\beta \in (0, 1-H)$ is a free parameter and

$$\eta = \exp \left( -\frac{1}{2\sigma^2} \left( \frac{C - \lambda}{H + \beta} \right)^{2(H+\beta)} \left( \frac{b}{1 - (H + \beta)} \right)^{2-2(H+\beta)} \right).$$

The use of Boole’s inequality is referred to as the sum approximation in [34] since (1) is bounded by the sum of the probabilities for all $\tau \in [0, t]$. Note that the sum approximation provides a rigorous upper bound for (1) as opposed to the max-approximation.

B. Weibullian Decay of FBM Overflow Probabilities

Sample path envelopes of the type of (4) are essential for application of the stochastic network calculus. They are required to quantify the service left over by cross traffic as well as to derive end-to-end performance bounds for through traffic. For traffic of the class of exponentially bounded burstiness (EBB) performance bounds for single systems were derived in [25]. A framework for stochastic network calculus with
sample path envelopes was developed in [18], [38], where end-to-end performance bounds for EBB traffic are presented. The aim of our work is to provide end-to-end performance bounds for the class of long-range dependent fBm traffic, which is challenging due to the fundamentally different and much slower tail decay of overflow probabilities compared to EBB traffic. We comment on the tail decay in this subsection.

In [29], [30] we showed that the sample path overflow probability \( \varepsilon_s \) (5) has the same log-asymptotic tail behavior\(^3\) as \( \varepsilon_a \) (2), i.e. \( \log \varepsilon_s \simeq -c\theta^{2-2H} \) in \( b \), with \( c \) being a positive constant. The result agrees with the earlier finding in [34] that uses the sum approximation with sampling at exponential time scales. Another sample path envelope for fBm traffic was derived recently in [37] under geometric time sampling, which also agrees in the log-asymptotic decay in \( b \) described above.

To illustrate the tail behavior we compare the overflow probability of backlog bounds for fBm traffic and EBB traffic. EBB traffic is characterized by the exponential decay of the point-wise overflow probabilities at a constant rate envelope. For generation of EBB traffic we use a discrete time Markov model with two states: on and off. In the off state (state 1) no traffic is generated and in the on state (state 2) traffic is generated with peak rate \( P \). The steady state probability of the on state is \( p_{on} = p_{12}/(p_{12} + p_{21}) \) where \( p_{ij} \) are the state transition probabilities from state \( i \) to state \( j \). The mean rate becomes \( \lambda = p_{on}P \). As [18] we characterize the burstiness of an on-off source by the average time for two state changes to occur \( T = 1/p_{12} + 1/p_{21} \). The aggregate of \( m \) on-off sources has a point-wise envelope (3) given as \( E(t) = b + m\rho(t) \) with overflow profile \( \varepsilon_p(b) = e^{-\theta b} \), see [18]. The envelope rate \( \rho(\theta) \) of a single discrete time on-off source is given in [14] as

\[
\frac{1}{\theta} \log \left( \frac{p_{11} + p_{22}e^{\theta P}}{2} + \sqrt{(p_{11} + p_{22}e^{\theta P})^2 - 4(p_{11} + p_{22} - 1)e^{\theta P}} \right)
\]

for \( \theta > 0 \) where \( \rho(\theta) \) ranges between the mean and the peak rate of the source. Employing the envelope a backlog bound follows from (1) by application of Boole’s inequality and integration of the overflow profile, see [18], [38] for details. Given aggregate traffic generated by \( m \) on-off sources fed into a server with capacity \( C \) the overflow probability \( P[B > b] \leq \varepsilon_s \) of the backlog bound \( b \) becomes

\[
\varepsilon_s = \int_0^\infty e^{-\theta (b + (C - m\rho(\theta)))} dT = \frac{e^{-\theta b}}{\theta (C - m\rho(\theta))}
\]

for any \( \theta > 0 \) that satisfies \( m\rho(\theta) < C \).

Throughout this paper we set the time slot of the discrete time model to 10 \( \mu s \) that is the transmission time of a 10 kb sized packet at a server with 1 Gbps speed. For Fig. 1 we use through traffic with parameters given in Tab. I at a server with capacity \( C = 1 \) Gbps. We use a traffic aggregate consisting of 100 EBB flows. We optimized the parameter \( \theta \) in (6) numerically. In case of fBm we use only a single flow. Recall that fBm is typically used as a model for aggregate traffic, e.g. it has been related to the superposition of on-off sources with heavy-tailed on and off periods [7], [4], [5]. In contrast the on-off sources used for EBB traffic have geometrically distributed on and off periods.

Fig. 1 clearly shows the Weibullian decay of the fBm traffic where \( \log \varepsilon \simeq -c\theta^{2-2H} \) compared to the exponential decay for EBB traffic.

C. Affine FBM Envelopes

We conclude this section with a corollary on affine fBm envelopes that follow from the backlog bounds. The relation between arrival envelopes and backlog bounds has been elaborated in [28] where a traffic model referred to as generalized Stochastically Bounded Burstiness (gSBB) is defined. The gSBB traffic characterization uses an affine sample path envelope with defined overflow profile. A general definition of sample-path envelope with overflow profile \( \varepsilon(b) \) [23], [18] extends (4) to

\[
P \left[ \sup_{\tau \in [0,t]} \left\{ A(\tau,t) - E(t-\tau) \right\} > b \right] \leq \varepsilon_s(b).
\]
The gSBB envelope is the special case where \( E(t) = rt \). The finding in [28] is that given a server with capacity \( C \) and statistical backlog bound \( b \) with overflow probability \( P[B > b] \leq \varepsilon(b) \) the arrivals of the system are gSBB and satisfy the definition of sample path envelope (7) with \( E(t) = Ct \) and identical overflow profile \( \varepsilon(b) \). It is argued in [28] that a gSBB envelope for fBM traffic follows immediately from (2).

**Corollary 1 (Affine FBM Envelopes).** Given fBM traffic \( E(t) = rt \) is a sample path envelope (7) with overflow profile

\[
\varepsilon_s(b) = \frac{\Gamma(\frac{r}{2})}{2b^\frac{r}{2}} b^{1-(H+\beta)}
\]

where \( \beta \in (0,1-H) \) is a free parameter and

\[
\vartheta = \frac{1}{2\sigma^2} \left( \frac{r - \lambda}{H + \beta} \right)^{2(H+\beta)} \left( \frac{1}{1 - (H + \beta)} \right)^{2 - 2(H + \beta)}.
\]

An approximate overflow profile is \( \varepsilon_a(b) = e^{-\vartheta b^{2-2H}} \) where

\[
v = \frac{1}{2\sigma^2} \left( \frac{r - \lambda}{H} \right)^{2H} \left( \frac{1}{1 - H} \right)^{2 - 2H}.
\]

The first statement of the corollary uses the backlog bound (5) that is derived based on the sample path envelope in [29], [30]. The second statement uses (2) as proposed in [28]. Note, however, that (2) is based on the approximation by the largest term, i.e. it does not provide a sample path envelope.

In Sect. III we will use the sample path envelope to derive a leftover service curve to analyze systems where fBM cross traffic is multiplexed, scheduled, and de-multiplexed afterwards. In Sect. IV we will compose these service curves by convolution to explore tandem systems, each under fBM cross traffic. We will use the affine envelopes established by Cor. 1. The envelope from Th. 1 in [29], [30] can be used in the same way and may yield tighter bounds at the cost of additional complexity.

### III. LEFTOVER SERVICE UNDER FBM CROSS TRAFFIC

The network calculus uses the concept of service curves \( S(t) \) to model the service provided by a system. The service curve relates a system’s departures \( D(t) \) to its arrivals \( A(t) \) [22]. A system is said to offer a deterministic lower service curve \( S(t) \) if it holds for all \( t \geq 0 \) that \( D(t) \geq A \otimes S(t) \) where the operation \( f \otimes g(t) := \inf_{\tau \in [0,t]} \{ f(\tau) + g(t - \tau) \} \) is referred to as the convolution under the min-plus algebra. A fundamental stochastic service curve is defined in [18] and likewise in [23]

\[
P\left[D(t) < \inf_{\tau \in [0,t]} \{ A(\tau) + |S(t - \tau - b)| \} \right] \leq \varepsilon(b) \quad (8)
\]

where \( |x|_+ = \max\{0,x\} \). Compared to the deterministic case the guarantee provided by the stochastic service curve may be violated at most with probability \( \varepsilon \). The definition of stochastic service curve is subject to a deficit profile \( \varepsilon(b) \) that is decreasing in \( b \). For the special case of deterministic systems (8) has overflow profile \( \varepsilon(b) = 0 \) for all \( b \geq 0 \) and reduces to the deterministic definition \( D(t) \geq A \otimes S(t) \).

So-called leftover service curves can effectively characterize the service that remains for a through flow at a system after scheduling cross traffic [27], [18]. A basic leftover service curve that does not make any assumptions about the order of scheduling can be deduced by subtracting the sample path envelope of the cross traffic from the service provided by the system. Given a server with capacity \( C \) and cross traffic with sample path envelope \( E(t) \) and overflow profile \( \varepsilon_s(b) \) as in (7).

A leftover service curve that satisfies (8) with deficit profile \( \varepsilon_s(b) \) is \( S(t) = Ct - E(t) \). The following corollary uses the affine fBM sample path envelope from Cor. 1 to characterize the leftover service under fBM cross traffic.

**Corollary 2 (FBM Leftover Service Curve).** Consider a server with capacity \( C \) under fBM cross traffic. A service curve (8) for through traffic is \( S(t) = (C-r)t \) with deficit profile \( \varepsilon(b) \). The deficit profile equals the overflow profile \( \varepsilon_a(b) \) of the fBM envelope in Cor. 1. It is approximated by \( \varepsilon_a(b) \), Cor. 1.

The definitions of service curve and sample path arrival envelope facilitate the derivation of performance bounds. Given arrivals with sample path envelope \( E(t) \) (7) at a system with service curve \( S(t) \) (8), the steady state virtual backlog \( B \) is stochastically bounded by [23], [18]

\[
P\left[B > \sup_{\tau \geq 0} \{ E(\tau) - S(\tau) \} + b \right] \leq \varepsilon(b)
\]

where \( \varepsilon(b) = \varepsilon^{th} \otimes \varepsilon^{cr} \). We use superscripted \( \varepsilon^{th} \) and \( \varepsilon^{cr} \) to distinguish overflow profiles of through and cross traffic. The latter matches the deficit profile of the service curve. The first-come first-serve waiting time \( W \) is bounded by \( P[W > w] \leq \varepsilon(b) \) where

\[
w = \inf \{ \tau \geq 0 : S(t + \tau) \geq E(t) + b \ \forall t \geq 0 \} \quad (9)
\]

Backlog and delay can be visualized as the vertical, respectively, horizontal deviation of the arrival envelope and the service curve subject to the overflow and deficit profiles.

We derive performance bounds for a through flow under fBM cross traffic using the leftover service curve from Cor. 2. We consider three fundamentally different types of through traffic: (a) constant bit rate (CBR), (b) traffic with exponentially bounded burstiness (EBB) [25], and (c) fBM with LRD, i.e. slower than exponential burstiness decay.

All three types of through traffic are configured to have the same mean rate \( \lambda \). We generally use affine sample path envelopes (7). Trivially, the envelope of the CBR traffic is \( E(t) = rt \) with \( r = m\lambda \) and overflow profile \( \varepsilon_s(b) = 0 \) for all \( b \geq 0 \). For generation of EBB traffic we use the discrete time Markov model introduced in Sect. II. A sample path envelope (7) for the aggregate traffic generated by \( m \) on-off sources follows from the backlog bound (6) as \( E(t) = rt \) with overflow profile \( \varepsilon_s(b) = e^{-\vartheta b^{2-2H}}(\theta(r - m\lambda)) \) for any \( \theta > 0 \) that satisfies \( m\lambda \theta < r \). For the fBM through traffic we use the envelope from Cor. 1.

The delay bound follows from (9) for \( \varepsilon^{th} + \varepsilon^{cr} \leq C \) as

\[
P\left[W > \frac{b^{th} + b^{cr}}{C - r^{cr}} \right] \leq \varepsilon^{th}(b^{th}) + \varepsilon^{cr}(b^{cr}).
\]
Here, \( b^{th, rth} \) are the parameters of the through traffic envelope and \( b^{cr}, r^{cr} \) are the parameters of the cross traffic envelope that determine the leftover service curve. Note that the parameters of the envelopes are free parameters where different sets of parameters can yield different violation probabilities for the same delay bound. To obtain the best possible result we optimize the parameters of the envelopes \( b^{th, rth} \) and \( b^{cr}, r^{cr} \) as well as the parameters \( \beta \) for fBm and \( \theta \) for EBB numerically.

In Fig. 2 we display the violation probability of a delay bound of 1 ms for CBR, EBB, and fBm through traffic at a server with capacity \( C = 1 \) Gb/s under fBm cross traffic. Through and cross traffic are configured to have the same mean rate. The traffic parameters are summarized in Tab. II. Fig. 2 shows the huge impact of the Hurst parameter \( H^{cr} \) of the cross traffic on the performance of through flows. While type and burstiness of the through traffic largely determine the violation probability of the delay bound, the influence becomes much less pronounced for cross traffic with large \( H^{cr} \), i.e. the LRD of the cross traffic becomes the dominating effect.

### IV. END-TO-END PERFORMANCE BOUNDS

The particular strength of the network calculus is its ability to characterize the end-to-end service of tandem systems. The service curves of individual systems can be composed by min-plus convolution into a network service curve. This network service curve models a whole network as if it were a single system. Hence, it facilitates the derivation of end-to-end performance bounds using single system results, e.g. (9).

#### A. Sample Path FBM Leftover Service Curve

The easy composition of tandem systems in the network calculus is due to the associativity of min-plus convolution. Given a network of two systems with deterministic service curves \( S_1(t) \) and \( S_2(t) \) in series the departures of the first system \( D_1(t) \) and \( D_2(t) \) are the arrivals to the second system \( D_2(t) = D_1(t) \) and by recursive insertion it follows that \( D_2(t) \geq A^2 \otimes S^{net}(t) \), where \( S^{net}(t) = S^1 \otimes S^2(t) \) is the network service curve.

The composition of stochastic service curves is, however, much more involved. The difficulty is due to the fact that a recursive insertion of the stochastic service curve (8) is not possible since (8) uses sample paths of the arrivals \( A(t) \) but makes only a point-wise statement for the departures \( D(t) \). To derive stochastic network service curves an extended definition of stochastic service curve that makes sample path guarantees for the departures is required. The problem does not occur in the deterministic case, see [18], [38] for details.

A fundamental stochastic network service curve is derived in [18], [38]. We make a marginal adaptation of the service curve to discrete time that is used in this work. For \( n \) systems in series each with service curve \( S_i(t) \) and deficit profile \( \varepsilon^i(t) \) according to (8) a network service curve is

\[
S^{net}(t) = S^1 \otimes S^2_\delta \otimes \cdots \otimes S^n_{(n-1)\delta}(t) \tag{10}
\]

where \( S_{\delta}(t) = S(t) - \delta t \) and \( \delta \geq 0 \) is a free parameter. The network service curve satisfies (8) with deficit profile

\[
\varepsilon^{net}(t) = \varepsilon^1_\delta \otimes \varepsilon^2_\delta \otimes \cdots \varepsilon_{n-1}^n_\delta \otimes \varepsilon^n(t). \tag{11}
\]

The deficit profiles \( \varepsilon^j_\delta(b) \) stem from an extended definition of stochastic service curve that makes guarantees for entire sample paths of the departures \( D(t) \) as opposed to (8) that only makes a point-wise statement. To derive such sample path guarantees [18], [38] contributes an essential sample path service curve that is relaxed by parameter \( \delta > 0 \). The definition of sample path service curve states that

\[
P\left[ \sup_{t \in [0,u]} \left\{ \inf_{\tau \in [0,t]} \left[ A(\tau) + [S(t-\tau) - \delta(u-t-b)]_+ \right] - D(t) \right\} > 0 \right] \]

is smaller equal \( \varepsilon_\delta(b) \). The relaxation by \( \delta \) permits deriving the sample path deficit profile from the point-wise deficit profile of the departures using Boole’s inequality as [18], [38]

\[
\varepsilon_\delta(b) = \frac{1}{\delta} \int_b^\infty \varepsilon(x) dx. \tag{12}
\]

We use the concept of network service curve to analyze tandem systems where fBm cross traffic is multiplexed and de-multiplexed. To this end, we derive the sample path deficit profile (12) for a leftover service curve under fBm cross traffic.
Corollary 3 (Sample Path FBM Leftover Service Curve). Consider a server with capacity $C$ under fBm cross traffic as given in Cor. 1. The leftover service curve $S(t) = (C - r)t$ relaxed by rate $\delta > 0$ has the sample path deficit profile (12)

$$\varepsilon_{\delta}(b) = \frac{\Gamma(\frac{1}{2\beta})}{2\delta \vartheta^{\pi}} b^{-(H+2\beta)}$$

where $\beta \in (0, \frac{1-H}{2})$ is a free parameter and $\vartheta$ is defined in Cor. 1.

Cor. 3 follows by insertion of the deficit profile from Cor. 2, respectively, Cor. 1 into (12) as

$$\varepsilon_{\delta}(b) = \frac{\Gamma(\frac{1}{2\beta})}{2\delta \vartheta^{\pi}} \int_{b}^{\infty} x^{-(H+2\beta)} dx$$

that has a finite solution if $\beta < \frac{1-H}{2}$. Alternatively, the approximate overflow profile $\varepsilon_{a}(b)$ from Cor. 1 can be used to derive a similar result where $b^{2-2H}$ becomes, however, the second argument of an incomplete Gamma function.

B. Scaling of End-to-end Performance Bounds

Cor. 3 enables the derivation of end-to-end service curves for networks under fBm cross traffic. In the remainder of this section we consider the line topology shown in Fig. 3 where fBm cross traffic is multiplexed and de-multiplexed at each hop. Note that decomposition results exist that facilitate the transformation of arbitrary feed-forward topologies into line topologies with single hop persistent cross traffic as in Fig. 3, see [39] for a discussion. For ease of notation we assume $n$ homogeneous systems in series, each with capacity $C$ and fBm cross traffic with identical parameters $\lambda$, $\sigma$, and $H^{cr}$. We show the growth of end-to-end performance bounds for $n$ tandem systems under fBm cross traffic.

The leftover service curve at each system is $S^{s}(t) = (C - r^{cr})t$ where $r^{cr}$ is the envelope rate of the fBm cross traffic. From (10) we obtain $S^{net}(t) = (C - r^{cr} - \Delta) t$ where $\Delta = (n-1)\delta$ is in $(0, C - r^{cr})$ and used as a constant. The deficit profile follows from (11) by insertion of Cor. 2 and Cor. 3 as

$$\varepsilon^{net}(b) = \frac{\Gamma(\frac{1}{2\beta})}{2\delta \vartheta^{\pi}} \inf_{x} \left\{ \frac{(n-1)^{2}x^{\pi}}{2\beta} \frac{1}{\Delta(1-H+2\beta)} + x^{\beta} \right\}$$

where $x \in (0, b)$ and $\beta \in (0, \frac{1-H}{2})$.

Fig. 4 shows end-to-end delay bounds for a CBR through flow that traverses $n$ tandem systems each with capacity $C = 1$ Gb/s and fBm cross traffic. The delay bound is computed from (9) using the network service curve and CBR through traffic as $P[W > b/(C - r^{cr} - \Delta)] \leq \varepsilon^{net}(b)$ under the constraint that $r^{th} + r^{cr} + \Delta \leq C$. The parameters of through and cross traffic are given in Tab. II. As before, we optimize the free parameters of the envelopes numerically.

Theorem 1 (Scaling of Performance Bounds). Given $n$ homogeneous tandem systems, each under fBm cross traffic with Hurst Parameter $H$. For a fixed violation probability the end-to-end backlog and delay bounds scale as

$$b \in O\left(n \left(\log n \right)^{\frac{1-H}{2\beta}}\right)$$

in the number of tandem systems $n$.

Proof: We derive an analytical result on the scaling of end-to-end delay bounds with the number of systems in series using Stirling’s formula $\Gamma(x) \approx \sqrt{2\pi/x} (x/e)^{x}$ for $x \gg 1$, which is exact in the limit for $\beta \rightarrow 0$. The growth of end-to-end delay bounds with $n$ is determined by the tail decay of the fBm overflow, respectively, deficit profiles. We neglect the irregularity in (11) that is due to the last hop and estimate $\varepsilon^{n}$ by $\varepsilon^{\delta}_{b}$ to obtain the simplified deficit profile

$$\varepsilon^{net}(b) = \frac{\Gamma(\frac{1}{2\beta})}{2\delta \vartheta^{\pi}} \inf_{x} \left\{ \frac{(n-1)^{2}x^{\pi}}{2\beta} \frac{1}{\Delta(1-H+2\beta)} + x^{\beta} \right\}$$

Assuming $\beta \ll 1$ we use Stirling’s formula to obtain

$$\varepsilon^{net}(b) = \frac{\sqrt{\pi/2}}{2\beta \vartheta^{\pi}} \Delta(1-H+2\beta)^{\frac{1-H}{2\beta}}.$$

We use minor simplifications for small $\beta$ to minimize $\varepsilon^{net}(b)$ over $\beta$. We find that the minimum of (13) is approached at $\beta = W\left(n^{2-2H}/2(-\log \varepsilon_{a})\right)$ where $W(z)$ denotes Lambert’s
W function that is the inverse of $z = xe^x$ and $e_a$ is defined in Cor. 1. The term $e_a$ is used here only to shorten notation and to express $\varepsilon^{\text{net}}$ as a multiple of $e_a$ in the sequel. We emphasize that we do not use the approximation by the largest term (2). Since $\beta$ is assumed to be small a good approximation of the optimal solution is

$$\beta^* = \frac{n^{2-2H}}{2(-\log e_a)}$$

where we estimated the Lambert W function by a linear segment. We define $\psi = (1 - H)/(1 - (H + 2\beta))$ and $\chi$ as

$$\chi = \left(\frac{H}{H + \beta(1 - (H + \beta))}\right)^2$$

which is in $\left[\frac{4}{3\sqrt{3}}, 1\right]$ for $H \in (\frac{1}{2}, 1)$, $\beta \in (0, \frac{1-2H}{2})$ and converges against 1 for small $\beta$. We find by insertion of $\beta^*$ into (13) that

$$\varepsilon^{\text{net}}(n) = \frac{n - 1}{nH} \left(\frac{H}{1 - H}\right)^H \sqrt{\pi n} \sigma c_1^{1+H} \Delta(r^{cr} - \lambda)^{1+H} e_{(1+\log c_a)n^{2H-2}}.$$ (14)

Note that both $\chi$ and $\psi$ approach 1 as $\beta \to 0$. Assuming $\beta^* \ll (1 - H)/2$ we can generally find constants that bound $\chi$ and $\psi$ from below, respectively, from above. Inserting $e_a$ as defined in Cor. 1 and using positive constants $c_i$ we obtain

$$\varepsilon^{\text{net}}(n) \leq n^{1-H} c_1^{1+H} e^{-c_2 n^{2H-2} n^{2H-2}}.$$ We let $b = n(c_0 \log n)^{\frac{1}{2-2H}}$ for $n \geq 2$ and find

$$\varepsilon^{\text{net}} \leq (c_0 \log n)^{\frac{1-H}{2-2H}} c_1 n^{1-2c_2}.$$ Generally, there exists $c_0 > 2/c_2$ such that $\varepsilon^{\text{net}}$ is upper bounded by a constant for all $n$. Fixing $\varepsilon^{\text{net}}$ it follows that $b$ scales as stated in Th. 1. This scaling holds for the backlog as well as for the delay bound which we compute as $P[W > b/(C - r^{cr} - \Delta)] \leq \varepsilon(b)$.

Finally, we verify that $\beta^*$ decreases with $n$ where the decay is proportional to $1/\log(n)$. This confirms the assumption that given $\beta^*$ is small it also remains small with increasing $n$. Recall that the above approximation using Stirling’s formula is exact for $\beta \to 0$.

The result derived in Th. 1 carries forward the impact of LRD fBm cross traffic to performance bounds derived for entire network paths. It describes the growth of end-to-end performance bounds in the number of tandem systems $n$ under fBm cross traffic to retain a fixed violation probability. It can be interpreted as the delay allocation over a network path that is required to keep the violation probability constant.

For comparison we use the same network, however, with EBB instead of fBm cross traffic. As before, the EBB traffic is an aggregate of $m$ on-off sources. From the EBB sample path envelope in Sect. III it follows that $S(t) = (C - r^{cr}) t$ is a leftover service curve with deficit profile $\varepsilon(b) = e^{-\theta b}/(\theta (r^{cr} - m\rho(\theta)))$. After Relaxation of the service curve by rate $\delta > 0$ the sample path deficit profile follows from (12) as $\varepsilon_S(b) = e^{-\theta b}/(\delta^2 (r^{cr} - m\rho(\theta)))$ for any $\theta$ that satisfies $m\rho(\theta) < r^{cr}$.

The network service curve under EBB cross traffic becomes $S^{\text{net}}(t) = (C - r^{cr} - \Delta) t$ with deficit profile

$$\varepsilon^{\text{net}}(b) = \frac{1}{r^{cr} - m\rho(\theta)} \inf_{x} \left\{ \frac{(n - 1)^2 e^{\theta b} - \frac{\theta b}{r^{cr}}}{\Delta^2 \theta^2} + e^{-\theta x} \right\}$$

where $x \in (0, b)$ and $\theta$ such that $m\rho(\theta) < r^{cr}$.

Solving $\varepsilon^{\text{net}}(b)$ for $b$ yields that end-to-end performance bounds under EBB cross traffic are in $O(n \log n)$. This result is derived in [18] and also proven as a lower bound in [40]. In contrast, our result for fBm cross traffic is a scaling in $O(n \log n)^\frac{1}{1-2H}$. The scaling is largely determined by the decay rate of overflow probabilities. The difference to EBB is caused by the Weibullian decay for fBm traffic as opposed to the exponential decay for EBB traffic, see Sect. II in particular Fig. 1. We find that the poly-logarithmic scaling component that increases with $H$ is due to LRD. For the special case $H = \frac{1}{2}$ the overflow probability of fBm traffic decays exponentially fast and we recover the scaling $O(n \log n)$ for EBB cross traffic. While we conclude that LRD has significant impact on single system performance bounds, we find that the additional effect due to concatenation of tandem systems is moderate.

In Fig. 5 we show end-to-end delay bounds for CBR through traffic under either EBB or fBm cross traffic. We use the same parameters and the same network as for Fig. 4, however, we normalize the delay by the number of systems in series $n$ to compare the logarithmic and poly-logarithmic scaling under EBB and fBm cross traffic, respectively. The traffic parameters are given in Tab. II. As before, we optimize the parameters of the traffic envelopes numerically. Fig. 5 clearly confirms the logarithmic growth of the normalized delay bounds. Moreover, Fig. 5 shows that a variation of the parameter $T^*$, that determines the correlation of the EBB cross traffic, mainly reproduces staggered versions of $w/n$. In contrast, varying
parameter $H^{cr}$ of fBm cross traffic changes the exponent of the poly-logarithmic scaling and hence alters the shape of $w/n$.

Further on, we observe that under fBm cross traffic $\varepsilon_{net}$ has a log-asymptotic decay in $b$ that can be expressed in the form

$$\lim_{b \to \infty} \frac{\log \varepsilon_{net}}{\log \varepsilon_s} = n^{2H-2}$$

(15)

where we used the sample path bound $\varepsilon_s$ (14) in [29], respectively, (12) in [30] and $\varepsilon_{net}$ from (14). The result implies that the end-to-end deficit profile of the network path $\varepsilon_{net}$ has the same log-asymptotic decay in $b$ as the sample path bound $\varepsilon_s$ as well as the largest term approximation $\varepsilon_a$ (2), however, impaired by the length of the network path $n$ combined with the grade of self-similarity $H$ of the fBm cross traffic. For the special case $H = \frac{1}{2}$ (15) reduces to $\frac{1}{n}$, which agrees with results for EBB traffic that can be easily derived from [18].

Next, we illustrate the scaling results from Th. 1 and (15). Fig. 6 depicts the end-to-end deficit profiles $\varepsilon_{net}$ for topologies as in Fig. 3 with $C = 1$ Gb/s. We use fBm cross traffic with parameters from Tab. II and a Hurst parameter $H^{cr} = 0.75$. In each step we quadruple the length of the network path $n$ to show the different scalings with the number of tandem systems. These are best illustrated on logarithmic scale both for $\varepsilon_{net}$ and $b$. For large $b$ the vertical arrow indicates the ratio of the log-asymptotic decay in (15). For $H^{cr} = 0.75$ it simplifies to $1/\sqrt{n}$ such that quadrupling $n$ halves log $\varepsilon_{net}$. As exemplified in Fig 6 for $b \approx 200 \text{kbit} \log \varepsilon_{net}$ approximately halves from $-64 \rightarrow -30$. The ratio becomes exact for $b \rightarrow \infty$, respectively, $\varepsilon_{net} \rightarrow 0$.

Further, given a fixed violation probability $\varepsilon_{net}$, the scaling of $b$ with the number of tandem systems $n$ from Th. 1 is indicated by the horizontal arrow in Fig. 6. For $H^{cr} = 0.75$ the scaling becomes $n(\log n)^{2}$ as $n \rightarrow \infty$. Taking the logarithmic scale of $b$ into account, Fig. 6 reconfirms the super-linear growth with $n$.

Finally, we show how end-to-end delay bounds vary with the amount of spare capacity provided on average. To this end, we carry forward a result for buffering at a single system from [9]. Consider $\varepsilon_{net}$ that is phrased as a multiple of $\varepsilon_a$ in (14). In order to achieve constant $\varepsilon_a$ if $r - \lambda$ is reduced $b$ has to grow by an amount that is determined by $H$. Halving $r - \lambda$ requires doubling $b$ for $H = 0.5$, whereas it requires increasing $b$ e.g. eightfold for $H = 0.75$. Similarly, Fig. 7 shows the impact of mean spare capacity, i.e. $C$ minus the mean rate $\lambda$ of both through and cross traffic, on end-to-end delay bounds subject to $\varepsilon = 10^{-12}$. The network consists of $n = 10$ tandem systems according to Fig. 3 with CBR through traffic and fBm cross traffic as in Tab. II. As exemplified by the markers halving the spare capacity doubles the delay, if the cross traffic is uncorrelated, i.e. $H^{cr} = 0.5$, whereas the delay increases tenfold for $H^{cr} = 0.75$. Compared to the previous argument for $\varepsilon_a$, the more pronounced growth is due to the additional pre-factor in (14). Clearly, under LRD providing spare capacity cannot be considered over-provisioning per se as it is essential for good network performance.

C. Related Scaling Results and Open Research Questions

The growth of performance bounds with the length of the network path $n$ provides significant insights into networking, e.g. super-linear scaling results question the common perception that the cost metric of a network path is simply its hop count. The first scaling result obtained from the stochastic network is $O(n \log n)$ for EBB traffic [18]. For fBm traffic our result $O(n \log n \pi - \pi)$ makes the significant impact of LRD explicit and recovers the $O(n \log n)$ result for $H = \frac{1}{2}$. We reported this scaling first in the technical report [29]. Since then, it has been recovered also in the report [37]. Further on, [37] provides the scaling $O(n \pi - \pi \log n)$ for networks with heavy tailed Pareto traffic with tail index $\alpha$. 

![Fig. 6. Scaling of the end-to-end deficit profile. For a fixed large $b$ the vertical arrow shows the ratio of logarithmic decay $n^{2H-2}$ of the deficit profile $\varepsilon_{net}$ according to (15), while for fixed $\varepsilon_{net}$ and large $n$, the scaling from Th. 1 is represented by the horizontal arrow.](image)

![Fig. 7. Impact of mean spare capacity on end-to-end delay bounds for $n = 10$ tandem systems. Halving the spare capacity doubles the delay for $H^{cr} = 0.5$, whereas the delay increases e.g. tenfold for $H^{cr} = 0.75$, see the markers. Under LRD spare capacity is essential for good network performance.](image)
While the scaling result for EBB is proven both as an upper and a lower bound [40], a lower bound that complements the scaling for fBm still remains to be derived. Note that the scaling results mentioned so far are derived without making assumptions about statistical independence of the service left over at individual systems. Under the additional assumption of statistical independence end-to-end performance bounds that scale in $O(n)$ are derived in [19] for $(\sigma(\theta), \rho(\theta))$ constrained cross traffic [14] that is closely related to the EBB traffic model. The effect of statistical independence on the scaling under fBm cross traffic is also an unresolved research question. For an elaboration on known scaling results see [38].

V. CONCLUSIONS

The contribution of this paper are end-to-end statistical performance bounds for a through flow in a network under fBm cross traffic with LRD. To this end, we used the framework of the stochastic network calculus. We developed an affine sample path envelope for fBm traffic that complements a previous sample path envelope and a known approximation using the largest term. From the derived envelopes we obtained the service curve left over by fBm cross traffic at a system. By convolution of these leftover service curves we derived a network service curve. We showed a numerical evaluation of the impact of fBm cross traffic on the end-to-end performance of through flows. We proved that end-to-end performance bounds for $n$ systems in series grow in $O(n(\log n)^{1/(1+\eta)})$.

ACKNOWLEDGEMENTS

This work was supported by an Emmy Noether grant of the German Research Foundation (DFG).

REFERENCES


