Non-asymptotic Performance Evaluation and Sampling-based Parameter Estimation for Communication Networks with Long Memory Traffic

Von der Fakultät für Elektrotechnik und Informatik
der Gottfried Wilhelm Leibniz Universität Hannover
zur Erlangung des akademischen Grades

Doktor-Ingenieur

genehmigte

Dissertation

von

Dipl.-Wirtsch.-Ing. Amr Rizk
geboren am 24. August 1983 in Kairo

2013

1. Referent	:	Prof. Dr.-Ing. Markus Fidler
2. Referent	:	Prof. Dr.-Ing. Jens B. Schmitt
Tag der Promotion	:	27.08.2013

Dipl.-Wirtsch.-Ing. Amr Rizk : *Non-asymptotic Performance Evaluation and Sampling-based Parameter Estimation for Communication Networks with Long Memory Traffic,* Dissertation, © 2013

ABSTRACT

Modern communication networks have become a constituent part of daily life applications. Since the design of ARPANET, which is the predecessor of the Internet, a key research focus lies on modelling and performance evaluation of communication networks. Established performance measures include quality of service (QoS) metrics known from queueing theory, e.g., the mean queue length or mean latency.

The discovery of the statistical properties of self-similarity and long memory in aggregate Internet traffic called into question the applicability of the performance analysis models and assumptions used so far. This discovery was followed by the derivation of approximations and asymptotes that indicate a significant impact of self-similarity and long memory on the queueing performance. Self-similarity and long memory are captured by the so-called Hurst parameter $H \in (0.5, 1)$.

Stochastic network calculus emerged as a promising framework for flow level performance evaluation that builds on a mathematical abstraction of data traffic flows and network elements. It comprises instrumental concepts such as the separation of the traffic characterization from the description of the service that is provided by network elements, e.g., routers. The network calculus framework permits modeling different types of statistical traffic sources as well as diverse scheduling disciplines, which contributes to its wide applicability. Moreover, composition results enable the evaluation of end-to-end performance for entire network paths. Network Calculus yields mainly non-asymptotic bounds for queueing performance metrics such as buffer occupancy and latency.

This thesis provides a performance analysis for communication networks carrying traffic that is characterized by the statistical properties of self-similarity and long memory. Based on stochastic network calculus we provide rigorous upper bounds on queueing performance metrics for single nodes with long memory input traffic. Our results strengthen conclusions from related approximations on the impact of spare capacity and buffering on the performance. Additionally, we quantify the influence of long memory traffic on the service provided to concurrent flows in resource sharing scenarios.

Based on our single node results we extend the analysis using the stochastic network calculus to multi-node scenarios. We show that end-to-end delays in network paths exhibiting long memory traffic grow super-linearly as $\mathcal{O}\left(n(\log n)^{\frac{1}{(2-2H)}}\right)$ in the number of traversed nodes n. This scaling compares to related results of $\Theta(n)$ for queueing networks under independence assumptions and $\Theta(n \log n)$ for networks carrying light tailed traffic. The derived scaling characterizes the impact of long memory on end-to-end performance over entire network paths. Our scaling result has direct implications for fundamental questions on network dimensioning and operations.

The significance of long memory traffic statistics is founded by its considerable impact on the queueing performance. However, we find that the traditional methods of acquiring such traffic properties, that are through collecting traffic traces, are challenging due to confidentiality issues and technical limitations on logging and storage speeds. This thesis provides a lightweight method to estimate properties of Internet traffic, specifically its correlations, through random sampling. We analytically demonstrate the applicability of our approach for different stochastic sampling processes. We show the impact of finite sample sizes as well as diverse sampling parameters on the estimation accuracy and provide asymptotically unbiased estimators. Complementary to trace-based approaches that reflect statistics at a single network node, we provide a method for inferring the dominant characteristics on end-to-end network paths. We substantiate our approach in a controlled testbed environment using a lightweight implementation of the presented algorithms before conducting an extensive Internet measurement campaign.

Keywords: Self-similarity, Long Range Dependence, Stochastic Network Calculus, Performance Evaluation, Quality of Service, Sampling, Network Probing.

ZUSAMMENFASSUNG

Moderne Kommunikationsnetze bilden einen elementaren Bestandteil vieler Anwendungen im alltäglichen Leben. Seit der Konzipierung des ARPANET, dem Vorgänger des Internets, ist die Modellierung und Leistungsbewertung von Datenkommunikationsnetzen Gegenstand der Forschung. Als Leistungsmetriken werden Dienstgütekriterien, die aus der Warteschlangentheorie bekannt sind, wie z.B. die mittlere Puffergröße bzw. Latenz, herangezogen.

Die Entdeckung der statistischen Eigenschaften der Selbstähnlichkeit und Langzeitkorrelation im Internetverkehr hat die Annahmen und somit die Anwendbarkeit der traditionellen Modelle bzw. Leistungsbewertungsmethoden in Kommunikationsnetzen wie dem Internet in Frage gestellt. Approximationen und asymptotische Resultate aus der Literatur weisen auf einen starken Einfluss der Selbstähnlichkeit und Langzeitkorrelation, welche mittels des sogenannten Hurst-Parameters $H \in (0.5, 1)$ beschrieben werden, auf die Dienstgüte hin.

In den letzten Jahren hat sich das stochastische Netzwerkkalkül als mathematisches Rahmenwerk für die Leistungsbewertung von Kommunikationsnetzen entwickelt. Dieses basiert auf der separaten Beschreibung von jeweils dem Datenverkehr und dem von Netzwerkkomponenten bereitgestellten Dienst. Diese Betrachtung führt zur hohen Anwendbarkeit des Netzwerkkalküls, da somit unterschiedliche Verkehrs- und Dienstmodelle kombiniert werden können. Weiterhin ermöglichen Kompositionsresultate die Abstraktion von Netzwerkpfaden als ein zusammengefasstes Äquivalentsystem und somit die Ausweitung der Analyse auf Ende-zu-Ende Pfade. Das Netzwerkkalkül erlaubt die Herleitung von oberen Schranken für Leistungsmetriken, wie z.B. die Verteilung der Puffergröße bzw. der Latenz.

In dieser Arbeit werden basierend auf dem stochastischen Netzwerkkalkül obere Schranken für Leistungsmetriken in Netzwerken mit langzeitkorreliertem Datenverkehr berechnet. Die hergeleiteten Ergebnisse werden mit Approximationen aus der Literatur verglichen. Dabei wird der Einfluss von Pufferung bzw. Überkapazitäten auf die Dienstgüte gezeigt. Weiterhin wird die Auswirkung von Langzeitkorrelationen auf die Dienstgüte konkurrierender Datenströme quantifiziert.

Auf der Basis des stochastischen Netzwerkkalküls und der vorgestellten Ergebnisse werden Leistungsschranken für Ende-zu-Ende Latenzen in Netzwerken mit langzeitkorreliertem Datenverkehr hergeleitet. Ein grundlegendes Resultat dieser Arbeit beziffert den Zuwachs von Ende-zu-Ende Latenzen für Netzwerkpfade mit n Systemen in Folge auf $\mathcal{O}\left(n(\log n)^{\frac{1}{(2-2H)}}\right)$. Vergleichbare Resultate aus der Literatur sind z.B. $\Theta(n)$ unter statistischer Unabhängigkeit und $\Theta(n \log n)$ für Datenverkehr mit exponentiell abklingender Burstartigkeit. Die Ergebnisse dieser Arbeit quantifizieren den Einfluss der Langzeitkorrelationen auf die Netzwerkperformanz. Die in geschlossener Form hergeleiteten Resultate haben eine grundlegende Auswirkung auf die Dimensionierung von Netzwerken und deren Betrieb.

Der erhebliche Einfluss der statistischen Eigenschaften von langzeitkorreliertem Datenverkehr auf die Dienstgüte in Warteschlangensystemen begründet die Erforschung von Methoden zur Schätzung dieser Eigenschaften. Traditionelle Methoden aus der Literatur basieren auf einer vollständigen Aufnahme des Datenverkehrs an Netzwerkkomponenten, wie z.B. Routern. Im Rahmen dieser Arbeit wird eine weniger datenintensive Schätzmethode für die Langzeitkorrelationen des Datenverkehrs vorgestellt, die auf zufälligem Abtasten des Verkehrs basiert. Die Rückgewinnung der statistischen Eigenschaften des Datenverkehrs aus den zufälligen Beobachtungen wird analytisch bewiesen. Der Einfluss einer begrenzten Anzahl an Beobachtungen auf die Schätzgenauigkeit wird quantifiziert. Weiterhin werden asymptotisch erwartungstreue Schätzer für die Langzeitkorrelationen des Datenverkehrs vorgestellt. Diese Arbeit ergänzt traditionelle Schätzmethoden aus der Literatur durch ein Verfahren, welches Testpakete in das Netzwerk injiziert, auf zufälligem Abtasten aufbaut und keine administrative Berechtigung an Routern erfordert. Im Gegensatz zu traditionellen Methoden, die die statistischen Eigenschaften an einem einzelnen Router wiedergeben, ermöglicht das vorgestellte Verfahren die dominanten Eigenschaften entlang eines untersuchten Pfades zu schätzen. Das präsentierte Verfahren wird in einer kontrollierten Umgebung verifiziert und es werden Ergebnisse einer Langzeit-Internetstudie vorgestellt.

Schlagwörter: Selbstähnlichkeit, Langzeitkorrelationen, stochastisches Netzwerkkalkül, Dienstgüte, Leistungsbewertung, zufälliges Abtasten, Netzwerktestverfahren.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

GLOSSARY OF ACRONYMS

ARMA	Autoregressive Moving Average
ARPANET	Advanced Research Projects Agency Network
Bm	Brownian Motion
CBR	Constant Bit Rate
CLT	Central Limit Theorem
DAG	Data Acquisition and Generation
EBB	Exponentially Bounded Burstiness
EDF	Earliest Deadline First
FBm	Fractional Brownian Motion
FIFO	First-in First-out
FILab	Future Internet Lab
Gbps	Gigabit per second
GPS	Generalized Processor Sharing
ICMP	Internet Control Message Protocol
IETF	Internet Engineering Task Force
iid	Independent and Identically Distributed
kbps	Kilobit per second
LRD	Long Range Dependence
Mbps	Megabit per second
MGF	Moment Generating Function
NIMASTA	Non-intrusive Mixing Arrivals See Time Averages
PASTA	Poisson Arrivals See Time Averages
PDF	Probability Density Function
PMF	Probability Mass Function
QoS	Quality of Service
RTT	Round Trip Time
SBB	Stochastically Bounded Burstiness
SP	Static Priority
SRD	Short Range Dependence
TCP	Transmission Control Protocol
UDP	User Datagram Protocol

$\Gamma(x)$	Gamma function
$\alpha(\theta, t)$	Effective bandwidth of a traffic flow
β	Free parameter $\in (0, 1-H)$
$\delta(t)$	Kronecker delta
ε_a	Approximate and Asymptotic overflow probability
ε_p	Point-wise overflow probability
ε_s	Sample path overflow probability
η	Free parameter $\in (0, 1)$
ϵ_Y^{rel}	Relative estimation error of the traffic autocovariance
λ	Mean arrival rate
ω	Period of the periodic sampling process
σ^2	Variance of the LRD arrival process $V(t)$ at $t = 1$
τ	Lag index of the autocovariance function
θ	Space parameter of the moment generating function
υ, ι	Parameters of the Gamma sampling distribution
ϱ	Positive slack rate
$A(t)$	Arrival process
$B(t)$	Backlog process
C	Capacity
$\tilde{c}_{(\cdot)}(\tau)$	Sample autocovariance of process $(\cdot)$ at lag τ
$c_{(\cdot)}(\tau)$	Autocovariance of process $(\cdot)$ at lag τ
$D(t)$	Departure process
$\mathsf{E}[(\cdot)]$, $\mu_{(\cdot)}$	Expected value of $(\cdot)$
$E(t)$	Envelope function
e	Euler's constant
$f^{(*k)}$	The k-fold self-convolution of the function f
G	Peak rate of an on-off Markov source
g	Mean rate of an on-off Markov source
H	Hurst parameter
i, j	general indexes
K, k_0, k_1, k_2	Positive constants
$\mathsf{M}_N(\theta)$	Moment generating function of random variable N
$\log(x)$	Natural logarithm of x

m	Number of flows
$\tilde{\mu}_{(\cdot)}$	Sample mean of $(\cdot)$
$\mathcal{N}\left(\mu, \sigma^2\right)$	Normal distribution with mean μ and variance σ^2
n	Number of nodes along a network path
p	Parameter of the geometric sampling distribution
r	Affine envelope rate
$S(t)$	Service curve function
s, t, u	Time indexes
T	Sample size
U	Burstiness parameter of an on-off Markov source
$V(t)$	Zero mean fBm process
v	Parameter of the uniform sampling distribution
$\mathcal{W}$	Lambert's $\mathcal{W}$ function
$W(t)$	Delay process
$X(t)$	Sampling process
$x(t), y(t), z(t)$	Sample paths of $X(t), Y(t), Z(t)$
$Y(t)$	LRD traffic increment process
$Z(t)$	Observed process
$\approx$	Approximately equal
$:=$	Equal by definition
$\stackrel{d}{=}$	Equal in distribution
$[x]_+$	Positive part of x, i.e., $\max\{0, x\}$
$l(x) \simeq h(x)$	Asymptotic behavior as $\lim_{x\to\infty} l(x)/h(x) = 1$
$*$	Classical convolution operator
$\otimes$	Min-plus convolution operator
$\oslash$	Min-plus deconvolution operator
$(\cdot)_i$	Subscript referring to the node number $i \in [1, n]$
$(\cdot)^{cr}$	Superscript referring to the cross traffic
$(\cdot)^{lo}$	Superscript referring to the leftover service
$(\cdot)^{net}$	Superscript referring to a network path
$(\cdot)^{\varrho}$	Superscript referring to the slack rate ϱ
$(\cdot)^{th}$	Superscript referring to the through traffic
$l(x) = \mathcal{O}(h(x))$	Landau Big-Oh notation
$l(x) = \Theta(h(x))$	Landau Big-Theta notation
$(\cdot)_W$	Subscript referring to the observed process $W(t)$
$(\cdot)_X$	Subscript referring to the sampling process $X(t)$
$(\cdot)_Y$	Subscript referring to the traffic process $Y(t)$
■	Halmos symbol to end a proof

Part I

DISSERTATION

1 INTRODUCTION

The statistical property known as long memory, or long range dependence (LRD), has been observed for a variety of natural phenomena in the past century. Long memory data series exhibit strong correlations that decay very slowly. Statisticians working with empirical records in different fields such as economics, hydrology and computer engineering reported for diverse metrics slower than exponential decay of correlations between distant observations. This phenomenon was reported in a prominent study of the Nile River water levels coining the so-called "Hurst" effect. The effect named after H. Hurst is a measure for the slow decay of the correlations that belong to long memory data series [56]. The seminal works by B. B. Mandelbrot [85, 86, 87] introduced long memory to stochastic modeling paving the way for numerous research works on the application of long memory processes to diverse scientific fields. In this thesis we provide a non-asymptotic performance evaluation of communication networks that exhibit long memory data traffic. Complementary to the performance evaluation we contribute a framework for estimating long memory traffic properties from sampled data subsets.

1.1 FEATURES AND MODELING OF INTERNET TRAFFIC

In recent years, communication networks have become an integral part of human life, shaping basic actions such as social interaction, business and entertainment. The dimensioning and optimization of communication networks, as well as the deployment of heterogeneous applications by, e.g., service providers establish the need for quantifying the performance of such networks. The performance analysis of modern communication networks has its roots in the mathematical investigation of telephone networks by A. K. Erlang at the beginning of the twentieth century [42, 43]. Erlang's studies led to the introduction and development of queueing theory as a framework of applied mathematics for performance evaluation that was later deployed in diverse fields such as telecommunications and operations research. In

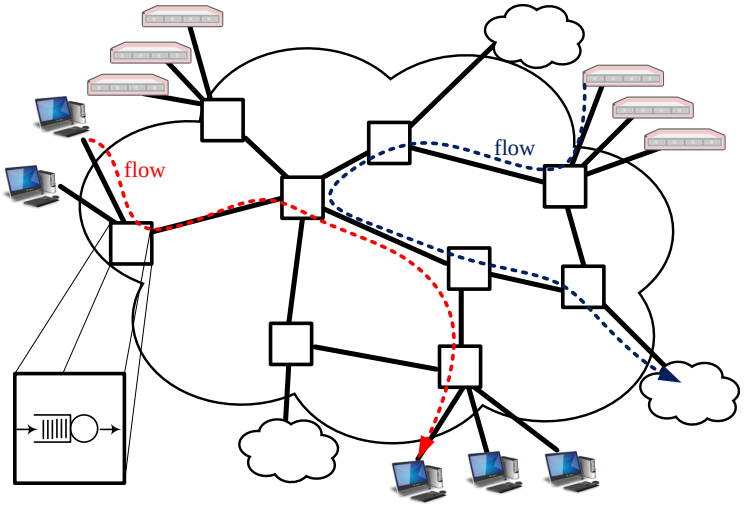


Figure 1.1: Network abstraction with traffic flows, internal nodes, e.g., routers, and end-hosts at the network edge. Note that for every node each output port is modeled by the queue-server abstraction in the bottom left.

Figure 1.1 we depict an example of a communication network with end-hosts at the network edge and internal elements denoted as nodes in the sequel that are for example routers or switches. We also depict packet data traffic flowing through the network and sharing resources at different nodes.

Queueing theory abstracts network nodes as servers with queues as depicted in the lower left corner of Figure 1.1. It provides closed form results on performance metrics such as the distribution of the packet waiting time for single queues, respectively, networks of queues using independence assumptions on arrival and service times [60, 67]. Prominent queueing theory results use the Poisson traffic model, which is based on data packets with exponentially distributed inter-arrival times. Hence, the traffic is modelled as a memoryless process. Using the Poisson traffic model the output queue of a network node, e.g., a router, can be abstracted as an $M|M|1$ queue [67, 68]. See [10] for an overview on queueing theory results. In the sequel we will rely on the node abstraction depicted in Figure 1.1.

For a long time the analysis of communication networks was based on the Poisson traffic model, which was justified through the multiplexing of a huge number of statistically independent traffic flows. The convergence of aggregate traffic to the Poisson model, that is known for independent Bernoulli sources, is argued in the limit for other types of sources in [20].

Comprehensive measurements in the 1990s revealed that aggregate Internet traffic possesses LRD and statistical self-similarity [32, 45, 72, 100, 128]. LRD and self-similarity are captured by the so-called Hurst parameter H, that can be estimated using numerous methods from given time series, respectively traffic traces [8, 123, 126].

A stochastic process $V(t)$ is called self-similar if it retains its finite dimensional distribution on different time scales apart from a stretching factor that depends on $H \in (0,1)$ [[8] p. 48]. Thus, for a self-similar stochastic process $V(t)$ it holds that $a^H V(t) \stackrel{d}{=} V(at)$ for any $a > 0$, with $\stackrel{d}{=}$ denoting equality in distribution. Self-similar LRD processes exhibit heavy temporal correlations, i.e., the related increment process $Y(t) := V(t+1) - V(t)$ with variance σ_Y^2 possesses an autocovariance[1] function that decays as

$$c_Y(\tau) \simeq H(2H-1)\sigma_Y^2 \tau^{2H-2} \quad \text{for } \tau \to \infty, \tag{1.1}$$

with lag τ and Hurst parameter $H \in (0.5, 1)$. LRD manifests itself in the slow decay of the correlations such that they are not summable $\sum_\tau c_Y(\tau) = \infty$ [[8] p. 52]. This stands in contrast to the independence assumption of Poisson traffic and the exponential decay of the correlations of Markov and autoregressive moving average (ARMA) models, respectively [8, 97].

The emergence of LRD and statistical self-similarity in network traffic is mathematically explained in [124] by the aggregation of many on-off sources that exhibit heavy tailed on and off periods, which corresponds to file size distributions observed on storage systems [32, 128]. A recent large-scale experimental study on this relation recovered the properties of self-similarity and LRD in a controlled environment using open- and closed-loop traffic [82]. Despite the continuous change in the traffic mix carried through the Internet due to perpetual emerging applications and protocols [7, 9, 51], recent measurements in [54] show that self-similarity and LRD still hold.

A widely adopted model for aggregate Internet traffic exhibiting self-similarity and LRD [44, 72, 95, 96] is fractional Brownian motion (fBm). FBm was introduced to

1 We use the definition from [97] for the autocovariance of the process $Y(t)$ at lag τ: $c_Y(\tau) := \mathsf{E}[Y(t)Y(t+\tau)] - \mathsf{E}[Y(t)]\mathsf{E}[Y(t+\tau)]$.

stochastic modeling by B. B. Mandelbrot in [86]. Depending on the Hurst parameter an fBm stochastic process denoted $V(t)$ may be long- or short-range dependent (SRD), i.e., for $H \in (\frac{1}{2}, 1)$ or $H \in (0, \frac{1}{2})$. SRD stochastic processes have a summable autocovariance $\sum_\tau c_Y(\tau) < \infty$. For $H = 0.5$ the process degenerates to standard Brownian motion (Bm) with independent increments. FBm possesses a stationary increment process $Y(t)$ denoted fractional Gaussian noise (fGn) [38]. Although the Gaussian increment distribution may in some cases be physically difficult to interpret, it simplifies calculations and makes the traffic model parsimonious [66, 72]. The established model for aggregate Internet traffic which is known from literature [44, 72, 95, 96] describes cumulative data traffic using the sum of a linear mean rate process and a zero-mean fBm process $V(t)$. The fBm process $V(t)$ features the following basic properties: $V(0) = 0$, $\mathsf{E}[V(t)] = 0$, and $\mathsf{E}[V(t)^2] = \sigma^2 t^{2H}$ for all $t \geq 0$ where $\sigma^2 > 0$ is its variance at $t = 1$ and denoted in the sequel variance parameter.

There exist numerous statistical tools to test time series for self-similarity and LRD and to estimate the related Hurst parameter. The estimators are based on the distributional, time and frequency domain characteristics of self-similar LRD series. The metrics and methods used for the estimation include: rescaled range, aggregate variance, periodogram, maximum likelihood and wavelets. See [8, 123, 126] and references therein for a detailed performance and robustness review of the aforementioned estimators. These statistical estimators have been deployed to provide evidence for self-similarity and LRD in Internet traffic traces and to estimate the related Hurst parameters in [32, 54, 72, 126, 128].

Next, we will review prominent performance evaluation techniques for single queueing systems with self-similar LRD traffic input as known from literature.

1.2 APPROACHES TO PERFORMANCE ANALYSIS WITH LRD TRAFFIC

Classical queueing theory was deployed as conclusive methodology to analyze packet data networks since its emergence in the 1960s [67]. It features fundamental results on scheduling and multiplexing as well as explicit closed-form performance

measures such as the exact distributions of backlog and delay for single queues as well as networks of queues given Poisson traffic. Queueing theory provides results on product form queueing networks that enable the analysis of tandem queues as if in isolation [10, 60].

The discovery of self-similarity and LRD in aggregate Internet traffic inspired many researchers to investigate the impact of these statistical properties on network performance [24, 40, 44, 57, 66, 74, 88, 90, 94, 96, 106]. Experimental investigations in [44] using self-similar LRD traffic traces show a strong degradation in queueing performance, e.g., in the distribution of the buffer occupancy. Studies as [44] showed practical evidence that networks carrying LRD traffic show fundamentally different performance in comparison to networks with memoryless or Markovian traffic.

Theories such as large deviations and effective bandwidths [40, 66, 90, 96] indicate that the tail decay of the buffer occupancy in the presence of LRD is slower than exponential. In comparison, memoryless or Markovian traffic are known for exponential queueing behavior. Next, we review a prominent result on the probability that the buffer occupancy B exceeds a given threshold b for systems fed with LRD traffic. The buffer overflow probability $\mathsf{P}[B > b]$ at a constant rate server with capacity C with incoming fBm traffic with mean rate λ, variance parameter σ^2 and Hurst parameter $H \in (0.5, 1)$ is given as

$$\mathsf{P}[B > b] \approx \exp\left(-\frac{1}{2\sigma^2}\left(\frac{C-\lambda}{H}\right)^{2H}\left(\frac{b}{1-H}\right)^{2-2H}\right) := \varepsilon_a. \tag{1.2}$$

The result (1.2) is derived as approximation in [95, 96] and as asymptotic bound for $b \to \infty$ in [40]. Thus, we use the subscript in ε_a to denote the approximation, respectively, asymptote in (1.2). Observe that the buffer overflow probability has a Weibull tail for $H \in (0.5, 1)$ and reduces to an exponential decay for the Bm case of $H = 0.5$. The result (1.2) is derived using the Gaussian distribution of the fBm increments and the largest term approximation in [95, 96]. The largest term approximation, that considers only the most probable time scale for buffer overflow, provides, strictly speaking, a lower bound on the buffer overflow probability. We will investigate the largest term approximation in more detail in Section 2.2.

The result (1.2) is deduced in [40] using large deviations theory in conjunction with the largest term approximation as $\mathsf{P}[B > b] \approx \varepsilon_a$ for $b \to \infty$. Large deviations theory focuses on finding tail asymptotics for distributions of steady-state performance measures such as queue length or delay [23, 40, 52, 99]. The asymptotic behavior given above is refined in [24, 49, 57, 88, 90, 94, 106] without altering the Weibull tail.

Large deviation theory is directly related to the theory of effective bandwidths [40, 127]. The notion of effective bandwidths arose in the context of resource allocation problems such as admission control of traffic flows [22, 41, 50, 65, 92]. Statistical multiplexing phenomena can be elegantly expressed using effective bandwidths. The effective bandwidth of the sum of independent flows is given by the sum of the effective bandwidths of the individual flows, see [23, 66]. Consider a traffic flow A that is described by the cumulative data arrivals between two time points, e.g., between times 0 and t by $A(0,t)$ or in shorthand $A(t)$. The effective bandwidth for some traffic flow $A(t)$ with stationary increments is defined in [66] as $\alpha(\theta,t) = \frac{1}{\theta t}\log \mathsf{E}\left[e^{\theta A(\tau,\tau+t)}\right]$ and varies between the mean and peak rate depending on the free parameter $\theta > 0$. It provides a measure for the resource requirements of traffic flows at various time scales. The effective bandwidth is related to the moment generating function (MGF) that is defined as $\mathsf{M}_N(\theta) := \mathsf{E}\left[e^{\theta N}\right]$ with parameter $\theta \in \mathbb{R}$ for a random variable N [[53] p.181]. It is derived in [66] for fBm that is characterized through the tuple $\{\lambda, \sigma^2, H\}$ as

$$\mathsf{M}_A(\theta,t) = e^{\lambda\theta t + \frac{\theta^2\sigma^2}{2}t^{2H}}, \tag{1.3}$$

such that the effective bandwidth for fBm is given by $\alpha(\theta,t) = \lambda + \frac{\theta\sigma^2}{2}t^{2H-1}$ [66]. Note the continuous growth of $\alpha(\theta,t)$ in t for LRD.

Large deviations theory and effective bandwidths provide the approximate, respectively, asymptotic result (1.2). We find that using the theory of network calculus it is possible to introduce the outstanding concepts of queueing theory on scheduling and multi-node performance measures to fBm. Network calculus with its deterministic [12, 23, 33] and stochastic [21, 23, 47, 64, 71, 89] branches uses the notion of envelopes to characterize traffic [11, 33, 69, 74, 92, 121, 130, 131] and the notion of service curves to characterize queueing systems while taking scheduling into

account [11, 18, 34, 47, 98, 118]. A set of powerful results [12, 18, 19, 23, 28, 34, 35, 70, 75, 119, 120, 131] enables the derivation of non-asymptotic bounds on queueing performance measures in single as well as multi-node scenarios. In this thesis we provide a non-asymptotic performance evaluation for networks with LRD traffic.

In the next section we will provide a detailed description of the contributions of this thesis.

1.3 THESIS CONTRIBUTIONS

In this thesis we provide a non-asymptotic performance evaluation of communication networks carrying long memory traffic, e.g., the Internet [72, 95, 96], as well as a lightweight traffic parameter estimation method without administrative network support. The contributions of this thesis have theoretical as well as practical implications.

First, we demonstrate a lightweight and practical method to extract long memory traffic parameters from *random* traffic samples. We specify the influence of the chosen statistical sampling process on the observations and determine the impact of sampling parameters on the accuracy of the estimates. We find that the relative error in the traffic autocovariance estimates increases with lag τ as τ^{2-2H} with Hurst parameter $H \in (0.5, 1)$. We provide a method for acquiring the relevant samples without administrative support through actively injecting packet probes into the network. Complementary to trace driven analysis, which is based on a *single* vantage point, we deploy our approach to characterize entire end-to-end paths. Finally, we provide testbed and Internet measurement campaign results for synthetic as well as real world network traffic.

Second, we propose a novel traffic envelope formulation in the context of stochastic network calculus that enables deriving non-asymptotic performance bounds for long memory traffic. Our result shows that queueing systems with long memory traffic exhibit a fundamentally different behavior than the well understood exponentially bounded burstiness (EBB) traffic analyzed in [28, 130]. Using the concept of leftover service curves [28, 74, 103] we specify the impact of long memory on the

service provided to concurrent traffic under scheduling constraints. We analyze the asymptotic behavior of our rigorous performance bounds proving that it retains a Weibull tail.

Third, we capitalize on the derived single node performance bounds to infer end-to-end performance bounds for networks carrying long memory traffic. We consider a general class of communication networks that are established in [28] and coined as convolution-form networks in [29] to derive a corresponding network service curve to describe entire paths under long memory traffic. Further, we complement related work of queueing theory and network calculus on the scaling of performance measures in communication networks. Queueing theory provides the exact distribution of the steady-state end-to-end delays based on statistical independence assumptions of arrival and service times [68]. This exact result can be transformed into a growth of end-to-end delays in the number of traversed nodes n of $\Theta(n)$ [26]. For EBB traffic, i.e., with queueing performance decaying exponentially, a fundamental result [19, 28] proves a scaling of $\Theta(n \log n)$ without assumptions on statistical independence. We show that for networks under long memory traffic with $H \in (0.5, 1)$ end-to-end performance bounds grow as $\mathcal{O}(n(\log n)^{\frac{1}{2-2H}})$ in the number of traversed nodes n. This reveals the impact of LRD on the performance of entire network paths. Our scaling result recovers the $\mathcal{O}(n \log n)$ result for the EBB case of $H = 0.5$. We complement the scaling of $\mathcal{O}(n(\log n)^{\frac{1}{2-2H}})$ by a large buffer asymptotic scaling result for multi-node scenarios. An elementary finding of our work is that over-provisioning is essential for reliable network performance in the presence of LRD.

1.4 THESIS OUTLINE

In Chapter 2 we review related work on the performance evaluation of communication networks using the network calculus framework. We present the network scenarios analyzed throughout this work and give a mathematical description of the arrival traffic and the service provided by the network nodes. We review existing definitions and results known from deterministic network calculus before

motivating its stochastic extension. Further, we recapitulate statistical models for arrivals and service curves and derive single node performance bounds before showing state-of-the-art results on the construction of network service curves and the derivation of end-to-end performance bounds.

In Chapter 3 we formulate the research problems addressed in this thesis. After introducing the individual research questions we show the significance of the corresponding contributions.

In Chapter 4 we motivate the problem of inferring traffic characteristics in communication networks from sampling observations with or without administrative support. We derive a framework for sampling LRD traffic and show the reversibility of the observations for different statistical sampling processes. Further, we assess the impact of the sampling parameters on the estimates and show trade-offs that impact the design of measurement campaigns. Finally, we show real world results that shed light on the properties of Internet traffic on end-to-end paths.

In Chapter 5 we derive a sample path envelope for LRD fBm traffic and provide non-asymptotic upper bounds on backlog and delay for servers fed with LRD traffic. We compare the overflow probabilities for single node performance bounds with LRD traffic to EBB class results known from literature to show that LRD traffic has a fundamentally different queueing behavior. Further, we derive a leftover service curve under fBm cross traffic, which is elementary to the subsequent end-to-end performance evaluation results.

In Chapter 6 we present a network service curve based on the single node results from Chapter 5. Subsequently, we provide non-asymptotic statistical end-to-end performance bounds for networks under fBm cross traffic. We show that for theses networks, performance bounds scale in $\mathcal{O}(n(\log n)^{\frac{1}{2-2H}})$ in the number of traversed nodes n. The application of the derived results within network dimensioning shows the significance of spare capacity and over-provisioning in communication networks.

Chapter 7 contains the conclusions and an outlook on future work.

2

FLOW LEVEL PERFORMANCE EVALUATION IN COMMUNICATION NETWORKS

In this chapter we review the state-of-the-art flow level performance analysis in communication networks using the network calculus theory. The framework of network calculus builds on a system theoretic description of network elements. Figure 1.1 depicts an abstraction of a communication network, where the individual network elements, e.g., routers are modeled as queueing systems that consist of a queue and a server. Data traffic traversing the network goes through tandem queueing systems, where it may share resources with other traffic or it may be shaped or policed [12].

In the following, we discuss established metrics for flow level performance analysis and review relevant analytical results. The exposition is based on data traffic volumes that are measured in bits and timing information given in seconds.

The network calculus framework relates the cumulative traffic arrivals of a considered data traffic flow at a given node, or multiple tandem nodes, to the cumulative traffic departures as exemplified in Figures 2.1 and 2.2. We denote the incoming cumulative arrivals in the time interval $[s,t)$ as $A(s,t)$, i.e., $A(0,t) - A(0,s)$. Similarly, $D(s,t)$ denotes the cumulative departures of the system in the time interval $[s,t)$, i.e., $D(0,t) - D(0,s)$. We denote $A(0,t)$ as $A(t)$ and $D(0,t)$ as $D(t)$ for convenience. We assume lossless systems and unless stated otherwise use a continuous time fluid traffic model implying infinitely divisible data. Functions of time are assumed non-negative, non-decreasing, left continuous and pass through the origin, e.g., $0 \leq A(s) \leq A(t), \forall s \leq t$ and $A(0) = 0$. From causality we have $D(t) \leq A(t)$ for all t. An arrival traffic flow $A(t)$ may be an aggregate of multiple flows as addressed later on in this chapter.

Unless stated otherwise we assume nodes are work conserving constant rate servers with capacity C, i.e., not idling when work is present. We assume stability at each node such that the mean arrival rate is strictly smaller than the node capacity. A considered flow may share resources with cross traffic at a single node as depicted

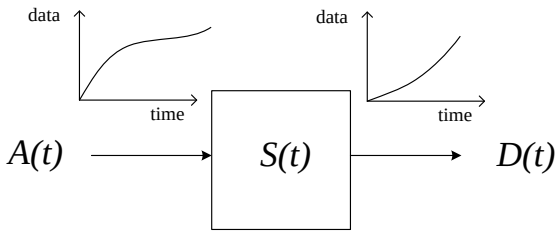


Figure 2.1: A queueing system with cumulative arrivals $A(t)$ and departures $D(t)$. The departures are related to the arrivals through the service curve $S(t)$. The graph shows sample path examples for cumulative arrivals and departures.

in Figure 2.3, or at the different nodes along a given path as shown in Figure 2.4. We restrict the composition to feed forward topologies, i.e., without cyclic dependencies. Note that arbitrary feed-forward topologies can be transformed into line topologies such as in Figure 2.4, see [47] for further details and a discussion.

2.1 BACKGROUND ON DETERMINISTIC NETWORK CALCULUS

The seminal work by Cruz [33] laid the foundation for network performance evaluation in a simple and mathematically elegant framework. Next, we briefly recapitulate relevant mathematical definitions and techniques that provide the basis for this thesis. A thorough treatment of the following concepts can be found in [12, 23, 33].

Deterministic Service Curves and Traffic Envelopes

Deterministic network calculus provides a general representation for service provided by a network node as in Figure 2.1. We build on the characterization for arrivals and departures specified at the beginning of this chapter. The concept of service curves was introduced in [98] and formalized in [34, 36, 118] as follows:

Definition 2.1 (Deterministic Service Curve) A system offers a lower service curve $S(t)$ to an arrival process $A(t)$ if it holds for the corresponding departures $D(t)$ that for all $t \geq 0$

$$D(t) \geq \inf_{s \in [0,t]} \{A(s) + S(t-s)\} =: (A \otimes S)(t). \tag{2.1}$$

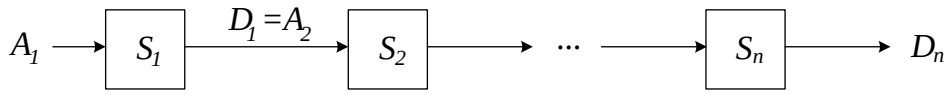


Figure 2.2: A flow traversing a network of tandem nodes with service curves $S_i(t)$ for $i \in [1, n]$. Traffic arrivals to the network are denoted $A_1(t)$, while network departures are given by $D_n(t)$.

The operation $\otimes$ is known as min-plus convolution [[12] p. 111]. For convenience, we will leave out the brackets around the operation $\otimes$ when possible. The lower service curve is a lower bound on the amount of service the arrivals $A(t)$ receive. If for a given system the lower bound on the departures $D(t)$ is also an upper bound, (2.1) holds with equality and the service curve is called *exact*, i.e.,

$$D(t) = \inf_{s \in [0,t]} \{A(s) + S(t-s)\}. \tag{2.2}$$

Constant rate servers with capacity C have an exact service curve $S(t) = Ct$ [23, 70].

Similar to the input-output relation of queueing systems given in (2.2), classical system theory relates the output $l(t)$ of a linear time-invariant system to the input $k(t)$ through classical convolution with the system impulse response $h(t)$ [[97] p. 398]. Accordingly, the system output is described as

$$l(t) = \int_{-\infty}^{\infty} k(s)h(t-s)ds := (k * h)(t). \tag{2.3}$$

Comparing (2.3) to (2.2) reveals a noticeable similarity in the sense that summation and multiplication in the classical convolution are replaced in (2.2) by the infimum operation and summation, respectively. A thorough treatment of the mathematical framework, the so-called $(min, +)$ algebra, that is deployed in network calculus is given in [4, 12, 23].

Traffic envelopes can be regarded as counterpart to the notion of service curves. Traffic envelopes and service curves are the main ingredients for the derivation of performance bounds for queueing systems as will be shown later on in this chapter. A deterministic envelope function $E(t)$ is an upper bound on the traffic arrivals $A(t)$ in the sense that for all $0 \leq s \leq t$ it holds that

$$A(t) - A(s) \leq E(t-s). \tag{2.4}$$

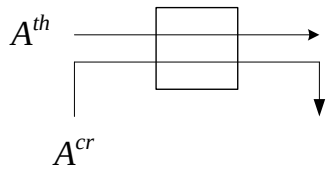


Figure 2.3: A through flow $A^{th}(t)$ sharing resources with a cross flow $A^{cr}(t)$.
A through flow $A^{th}(t)$ sharing resources with a cross flow $A^{cr}(t)$. Both flows are scheduled on the same output port.

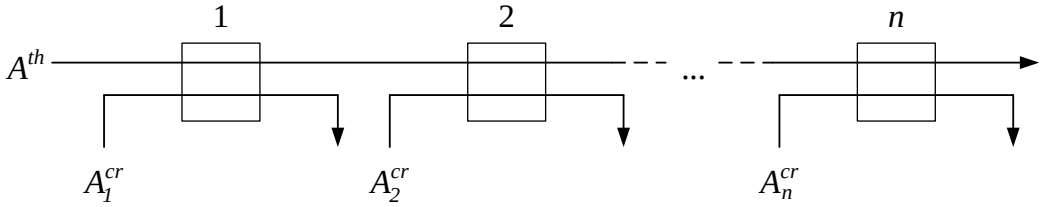


Figure 2.4: A through flow $A^{th}(t)$ traversing a network of tandem nodes with single node persistent cross traffic $A_i^{cr}(t)$ for $i \in [1, n]$.

This condition translates into that the envelope $E(t)$ is a deterministic upper bound for the arrivals on all time intervals. This can be rewritten as $A(t) \leq A \otimes E(t)$. Traffic envelopes may be enforced by traffic shapers, e.g., a leaky bucket shaper that takes unknown arrivals and delays traffic that violates the envelope only as much as required for the output of the shaper to conform with $E(t)$. The leaky bucket shaper has the envelope $E(t) = q + \rho t$, where q is the maximum instantaneous burst of arrivals permitted and ρ is the envelope rate, i.e., an upper bound on the mean rate of arrivals.

A remarkable feature of the network calculus framework is its ability to capture multiplexing and scheduling of traffic flows using traffic envelopes and service curves, respectively. Multiplexing m flows $A_i(t)$ each with envelope $E_i(t)$ for $i \in [1, m]$ results into an aggregate $A(t) = \sum_i A_i(t)$ that possesses an envelope $E(t) = \sum_i E_i(t)$. The concept of service curves allows the characterization of different scheduling algorithms including static priority scheduling (SP) [12, 23], first-in first-out (FIFO) [12, 35], earliest deadline first (EDF) [75] and generalized processor sharing (GPS) [23, 70]. Consider multiple traffic flows as in Figure 2.3 sharing the provided resources on the output link of some node. We denote the arrivals of a considered flow by $A^{th}(t)$, i.e., *through* traffic flow and denote the remaining traffic arrivals by $A^{cr}(t)$, i.e., *cross* traffic. The characterizations of the service curves for $A^{th}(t)$ and $A^{cr}(t)$ depend on the scheduling discipline.

In the sequel, we describe the service provided to the through traffic with no assumptions on the scheduling algorithm. This model is denoted by blind multiplexing in [[12] p. 176] and is used to model SP scheduling. Consider the node in Figure 2.3 and assume that $A^{th}(t)$ receives the lowest priority. The service provided to the through traffic is captured by the so-called *leftover service curve*. The node in Figure 2.3 is assumed to maintain the order of bits within each traffic flow, which is referred to as locally-FIFO. Further, we regard a particular type of service curves the so-called strict service curve [2, 12], which is a function $S(t)$ for which it holds

$$D(t) - D(s) \geq S(t-s), \tag{2.5}$$

for all $0 \leq s \leq t$ falling into the same backlog period, i.e., the server does not idle in-between. The strict service curve $S(t)$ guarantees a minimum service over the backlogged period. It is obvious that (2.1) follows from (2.5) but not vice versa [[12] p. 177].

Based on the strict service curve characterization (2.5) and assuming an empty system at $s = 0$, we can find the following service curve formulation for the through traffic [[12] p. 176]

$$S^{lo}(t) = [S(t) - E^{cr}(t)]_+, \tag{2.6}$$

with envelope $E^{cr}(t)$ for the cross traffic $A^{cr}(t)$ satisfying (2.4). The service curve formulation (2.6) reflects the service *leftover* by the higher priority cross traffic, hence, the superscript lo. The constraint $[x]_+$ denotes the positive part of x, i.e., $\max\{0, x\}$. Now, we are able to replace the system in Figure 2.3 with an equivalent system as the one depicted in Figure 2.1 providing $S^{lo}(t)$ from (2.6) as a service curve to the arrivals $A^{th}(t)$.

If the deployed scheduling algorithm is known, the service curve representation can be improved with respect to the blind multiplexing model (2.6), for example, for FIFO and GPS scheduling the particular service curves can be found in [12, 35] and in [23, 70], respectively.

Deterministic Performance Bounds and Concatenation Results

Deterministic network calculus provides simple formulations for worst case bounds on performance metrics of queueing systems. Next, we review results on two performance metrics, namely backlog and delay. The backlog $B(t)$ describes the number of bits "in flight" at t, which is the amount of data that entered the system and still has not left the system yet. Thus, it is defined in [[12] p. 5] as

$$B(t) = A(t) - D(t), \tag{2.7}$$

which can be visualized as vertical deviation of $A(t)$ and $D(t)$. The definition (2.7) allows the backlog to increase infinitely large or for a non-busy system to become zero. For constant rate servers of capacity C we can insert (2.2) with $S(t) = Ct$ into (2.7) to relate the backlog $B(t)$ to the arrivals $A(t)$ and the server capacity C as

$$B(t) = \sup_{0 \le s \le t} \{A(s,t) - C(t-s)\}. \tag{2.8}$$

The expression (2.8) is obtained using Reich's equation in [[70] p. 127] and using Lindley's recursion in [[23] p. 7].

Similarly, the (virtual) delay is defined for locally-FIFO systems in [[12] p. 5] as

$$W(t) = \inf\{s \ge 0 : A(t) \le D(t+s)\}. \tag{2.9}$$

The delay in (2.9) can be visualized as horizontal deviation of $A(t)$ and $D(t)$. Note that the virtual delay in (2.9) may become zero, whereas a data packet always sees a positive delay. However, it was shown in [[26] p. 110] that for numerical purposes the difference between these two measures is negligible.

The following definition from [[12] p. 122] helps phrasing the desired performance bounds on the backlog and delay metrics (2.8) and (2.9).

Definition 2.2 (Min-Plus Deconvolution) Given two functions $k(t)$ and $l(t)$, that both are assumed non-decreasing and pass through the origin. The min-plus deconvolution of $k(t)$ and $l(t)$ is defined as

$$\sup_{0\leq s}\{k(t+s)-l(s)\} := (k \oslash l)(t). \tag{2.10}$$

Next, we review backlog and delay bounds from [[12] p. 22] derived for a single node scenario as depicted in Figure 2.1. Consider arrivals that possess an envelope $E(t)$ according to (2.4) at a system that offers a service curve $S(t)$ according to (2.1). The backlog $B(t)$ is then upper bounded for all $t \geq 0$ as

$$\begin{aligned} B(t) &= A(t) - D(t) \\ &\leq A(t) - \inf_{0\leq s\leq t}\{A(s)+S(t-s)\} \\ &= \sup_{0\leq s\leq t}\{A(t)-A(s)-S(t-s)\} \\ &\leq \sup_{0\leq s\leq t}\{E(t-s)-S(t-s)\} \\ &\leq \sup_{0\leq u}\{E(u)-S(u)\} := E \oslash S(0). \end{aligned} \tag{2.11}$$

The derivation is based on inserting first the service curve definition (2.1), then the envelope definition (2.4) and lastly the definition of the min-plus deconvolution (2.10).

Similarly, a bound on the delay $W(t)$ can be found as the maximum horizontal deviation between the envelope $E(t)$ and the service curve $S(t)$. Thus, the upper bound on the delay $W(t)$ for all $t \geq 0$ can be written for $s \geq 0$ as

$$\begin{aligned} W(t) &\leq \inf\{s : \sup_{0\leq u}\{E(u)-S(u+s)\} \leq 0\} \\ &= \inf\{s : E \oslash S(-s) \leq 0\}. \end{aligned} \tag{2.12}$$

A further strong feature of the network calculus framework is the derivation of end-to-end performance bounds for entire network paths as depicted in Figure 2.2 through composition. Similar to classical systems theory, concatenated systems as in Figure 2.2 can be consolidated into a *single* equivalent system enabling the use of

single node performance bounds as in (2.11) and (2.12). We note that the alternative of summing up per node bounds delivers much looser bounds as shown in [12, 26].

Regard the network depicted in Figure 2.2. The departures of the last node can be expressed using (2.1) as $D_n(t) \geq A_n \otimes S_n(t)$, where the subscript denotes the node index. Using the fact that the arrivals at each of the last $n-1$ nodes equal the departures of the respectively preceding node, we can utilize the following recursive insertion $A_{n-i+1}(t) = D_{n-i}(t) \geq A_{n-i} \otimes S_{n-i}(t)$, for $i \in [1, n-1]$. By recursive insertion in conjunction with the associativity property of the min-plus convolution from [[12] p. 111], we can write

$$\begin{aligned} D_n(t) &\geq (A_1 \otimes S_1) \otimes S_2) \otimes \dots) \otimes S_n \\ &= A_1 \otimes (\underbrace{S_1 \otimes S_2 \otimes \dots \otimes S_n}_{:=S^{net}})(t). \end{aligned} \tag{2.13}$$

The service provided by the path is characterized by the network service curve $S^{net}(t)$ given by the min-plus convolution of the single node service curves $S_i(t)$ for $i \in [1, n]$. As depicted in Figure 2.2 the arrivals to the network are $A_1(t)$ whereas the departures of the network are given by $D_n(t)$. Collapsing the nodes of the given network path into a *single* equivalent system described by $S^{net}(t)$ enables finding end-to-end performance bounds by substituting $S^{net}(t)$ for $S(t)$ in (2.11) and (2.12). For through traffic on network paths with cross traffic as depicted in Figure 2.4 the scenario is transformed first into the equivalent configuration depicted in Figure 2.2 using the leftover service curve formulation $S_i^{lo}(t)$ from (2.6) for $i \in [1, n]$. Next, $S^{net}(t)$ is found in (2.13) by convolution, that is through substituting for $S_i(t)$ by the respective leftover service curves $S_i^{lo}(t)$ from (2.6). Finally, applying (2.11) and (2.12) yields single node performance bounds. It was shown in [26] that delay bounds derived through the network service curve scale in the number of traversed nodes n as $\mathcal{O}(n)$ in comparison to adding per node delay bounds which scales in $\mathcal{O}(n^2)$.

2.2 RELATED WORK ON STOCHASTIC NETWORK CALCULUS

The deterministic network calculus framework presented above is useful for applications with stringent performance guarantees as it builds on worst case characterizations of traffic and service in communication networks as given in (2.4) and (2.1). The deterministic framework provides elegant formulations for performance bounds such as (2.11) and (2.12) and an outstanding composition result (2.13) enabling end-to-end network analysis. However, worst case performance analysis may be, in general, too pessimistic for most of the real world applications, as it fails to capture the natural statistical properties of arrivals and service. We highlight this property using the following example.

Consider a simple application generating traffic with independent random packet lengths at discrete time slots. At each time slot one packet arrives with a length drawn from a discrete uniform distribution between 1 and a maximum packet length $L > 1$, $L \in \mathbb{N}$. The smallest envelope for the arrivals that satisfies (2.4) is $E(t) = Lt$. However, this envelope is too pessimistic as it considers the worst case arrivals, i.e., packets of size L at *every* time slot. Note that the probability of M consecutive packet arrivals each of maximum length L decreases exponentially in M, i.e., $\frac{1}{L^M}$. Surely, the independent packet size assumption behind this simplistic example is unrealistic as real world protocols and applications exhibit memory, see, for example, the coding dependency in the MPEG-1 video element denoted "group of pictures" [59]. However, this simple example shows the high potential for resource saving and performance metric improvement if the worst case is excluded.

Furthermore, the formulation (2.4) does not permit taking advantage of statistical multiplexing. In the previous section we calculated the deterministic envelope for the sum of m arrival flows as the sum of the individual envelopes. However, this envelope is highly inefficient as it does not exploit temporal fluctuations of independent multiplexed flows. In the context of the previous example consider the aggregate of m independent sources with independent uniformly distributed packet sizes between 1 and L. A deterministic envelope for the aggregate is given by $E(t) = mLt$. However, for one time slot the probability that all sources transmit

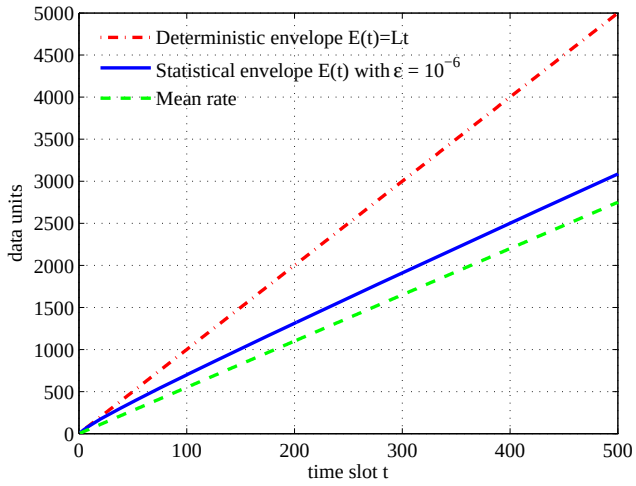


Figure 2.5: Comparison of deterministic and stochastic traffic envelopes with violation probability ε. Traffic comprises independent increments drawn from a uniform distribution between 1 and $L = 10$.

a packet of maximum size L is $\frac{1}{L^m}$. Intuitively, statistical multiplexing profits from the temporal fluctuations of the individual sources. Hence, the aggregation of independent sources gives rise to central limit theorem (CLT) results.

The aforementioned example displays the limitations of the deterministic network calculus framework and motivates its stochastic extension, which has its roots in [21, 71]. Stochastic network calculus is a performance analysis framework that delivers results excluding the worst case and incorporating many practical traffic, scheduling and service models, see [11, 12, 18, 19, 23, 26, 27, 28, 29, 33, 34, 47, 64, 69, 74, 89, 120, 121, 130, 131] and references therein. Two fundamental concepts contribute to the applicability of the stochastic network calculus, namely (i) multiplexing and scheduling results, and (ii) convolution form networks, i.e., results on end-to-end performance measures where tandem queues can be collapsed into a single equivalent system. This compares to results from queueing theory on product form networks [10, 60].

In stochastic network calculus the objective of the analysis is finding, or bounding, the probability that a certain performance measure exceeds a given threshold. A bound on the steady-state[1] virtual backlog B defined in (2.7) that is formulated as

$$P[B > b] \leq \varepsilon, \tag{2.14}$$

approximates the buffer overflow probability [26]. Similarly, a bound on the steady-state virtual delay W defined in (2.9) approximates the outage probability for the steady-state packet delay. Typically, ε is chosen to be small, e.g., 10^{-6}.

We apply this concept to the introducing example at the beginning of this section. For traffic with uniformly distributed independent increments we find that this small violation probability allows remarkable bound improvements. We derive an envelope function $E(t) < Lt$ that is violated at most by a small probability ε. We visualize the significant improvement in Figure 2.5 by plotting the single source deterministic envelope function Lt together with an envelope that is at most violated by $\varepsilon = 10^{-6}$. Note that the long term rate of the statistical envelope equals $\limsup_{t\to\infty} E(t)/t < L$.

The construction of such statistical envelopes will be discussed in the following sections. The deterministic restriction can be recovered by setting $\varepsilon = 0$. Contrary to its deterministic counterpart, the stochastic framework enables taking advantage of statistical multiplexing effects that improve performance bounds. Next, we review basic results on the construction of stochastic traffic envelopes and service curves.

Stochastic Traffic Envelopes and Service Curves

Stochastic envelopes are statistical bounds on arriving traffic volumes over time periods [11, 22, 69, 121, 130]. We assume discrete time stationary arrivals such that $\mathsf{P}[A(0,t) > x] = \mathsf{P}[A(s,s+t) > x]$ holds for all time indexes s, t, for $x \geq 0$,

1 We seek a time independent bound for $B(t)$, thus, it also holds for some steady-state to which the backlog distribution converges as $\lim_{t\to\infty} B(t)$.

and $s, t \in \mathbb{N}_0$. A widely adopted formulation for stochastic traffic envelopes [26, 64] is given as

$$\mathsf{P}\left[A(s,t) > E(t-s) + b\right] \leq \varepsilon_p(b), \tag{2.15}$$

with the so-called overflow profile $\varepsilon_p(b)$. The subscript p denotes that the envelope in (2.15) can be violated with point-wise probability $\varepsilon_p(b)$ by any arrival realization at one time *point*. Setting $b = 0$ in (2.15) yields an expression which is an immediate stochastic extension of (2.4), however, the parameter b is important to the derivation of performance bounds as will be shown later on in this chapter.

Two important traffic envelope models exist in the literature: exponentially bounded burstiness (EBB) [130] and stochastically bounded burstiness (SBB) [121] models. In the context of (2.15), the EBB model follows for $E(t) = \rho t$ and $\varepsilon_p(b) = \phi e^{-\theta b}$ with linear rate ρ and positive ϕ and θ. Here, the overflow profile $\varepsilon_p(b)$ decays exponentially in b. Closely related is the $(\sigma(\theta), \rho(\theta))$ model from [[23] p. 241] that builds on traffic characterized by an MGF satisfying $\mathsf{M}_A(\theta, t) \leq e^{\theta(\sigma(\theta)+\rho(\theta)t)}$ for $\theta, t \geq 0$. The $(\sigma(\theta), \rho(\theta))$ model is linked to the EBB model through the application of Chernoff's theorem.[2] The EBB class comprises Poisson traffic in addition to traffic from Markov modulated sources. The more general SBB traffic model implies $E(t) = \rho t$ and that $\varepsilon_p(b)$ is n-times integrable. The SBB model includes fBm traffic with Hurst parameter $H \in (0.5, 1)$ [121]. A thorough overview of traffic envelopes can be found in [89] and references therein.

We review the derivation of statistical envelopes for stationary traffic with known MGF [74]. Fix b in (2.15) to zero and apply Chernoff's theorem as $\mathsf{P}[A(s, s+t) \geq E(t)] \leq e^{-\theta E(t)} \mathsf{M}_A(\theta, t)$, where θ is a free parameter. Fix the right hand side to ε_p and solve for $E(t)$ to find

$$E(t) = \inf_{\theta \geq 0} \left\{ \frac{1}{\theta} \log \left(\frac{\mathsf{M}_A(\theta, t)}{\varepsilon_p} \right) \right\}. \tag{2.16}$$

This result is named as effective envelope in [74]. The literature comprises analytical expressions of the MGF $\mathsf{M}_A(\theta, t)$ for various traffic sources including Poisson,

2 Chernoff's theorem states that for a stationary random process $A(s, s+t)$ it holds $\mathsf{P}[A(s, s+t) \geq x] \leq e^{-\theta x} \mathsf{M}_A(\theta, t)$ [114].

Markov, regulated and fBm traffic sources [66]. Note that it is not always possible to analytically optimize the free parameter θ in (2.16).

We recall that the objective of the performance analysis in stochastic network calculus is bounding the probability that a regarded performance measure such as the backlog exceeds a given threshold. We consider arrivals that can be described by a statistical envelope as reviewed above at the ingress of a work-conserving constant rate server with capacity C. A probabilistic extension of (2.8) is given as the left hand side of (2.14) and can be rewritten as

$$\mathsf{P}[B > b] = \mathsf{P}\left[\sup_{s\in[0,t]} \{A(s,t) - C(t-s)\} > b\right]. \tag{2.17}$$

The difficulty in the derivation of a time independent bound in (2.17) lies in finding the argument $\widehat{s}$ that attains the supremum, since $\widehat{s}$ is a random variable. For an elaboration on "what makes performance evaluation using statistical network calculus hard" see [74]. There exist approximations and asymptotes for the queue length tail behavior that use the principle of the largest term, i.e.,

$$\mathsf{P}[B > b] \approx \sup_{s\in[0,t]} \left\{\mathsf{P}\left[A(s,t) - C(t-s) > b\right]\right\}. \tag{2.18}$$

It is obvious that the right hand side of (2.18) is in fact a lower bound to the buffer overflow probability. In Chapter 1.2 we reviewed a prominent result (1.2) that is derived through the approximation by the largest term for LRD fBm traffic.

An approach to upper bound (2.17) is based on an envelope $E(t) \leq b + Ct$ for all t such as we can write

$$\mathsf{P}\left[\sup_{s\in[0,t]} \{A(s,t) - C(t-s)\} > b\right] \leq \mathsf{P}\left[\sup_{s\in[0,t]} \{A(s,t) - E(t-s)\} > 0\right]. \tag{2.19}$$

Clearly, the envelope in (2.19) is directly related to the derivation of performance bounds. Finding an upper bound on the right hand side of (2.19) yields directly a stochastic backlog bound. Note that the right hand side of (2.19) demands a bound for $A(s,t)$ for all $s \in [0,t]$, a so-called sample path bound. This so-called sample

path effective envelope given above relates to the effective envelope given in (2.16) and leads to a desired sample path bound in the form of

$$\mathsf{P}\left[\sup_{s\in[0,t]}\{A(s,t)-E(t-s)\}>0\right]\leq\varepsilon_s. \tag{2.20}$$

Note the subscript $(\cdot)_s$ denoting the sample path violation probability, that is at *any* point in time along the sample path. It is important that the sample path envelope in (2.20) holds for $t\to\infty$, i.e., dispensing with assumptions on critical time scales and time scale bounds [74].

A generalization of (2.20) that provides a sample path envelope with overflow profile $\varepsilon_s(b)$ is given in [28] and likewise in [34] as

$$\mathsf{P}\left[\sup_{s\in[0,t]}\{A(s,t)-E(t-s)\}>b\right]\leq\varepsilon_s(b). \tag{2.21}$$

The so-called generalized Stochastically Bounded Burstiness (gSBB) model provides a sample path bound satisfying (2.21) with affine sample path envelopes, i.e., substituting $E(t)=rt$ in (2.21). Hence, a valid statistical backlog bound implies a gSBB traffic characterization [63, 131]. Given a backlog bound $\mathsf{P}[B>b]\leq\varepsilon(b)$ the arrivals are gSBB satisfying (2.21) with $E(t)=Ct$ and overflow profile $\varepsilon(b)$ [63].

The authors of [28, 34, 131] present solutions to the research problem of sample path envelope construction. First, we sketch the basic steps behind the envelope construction before providing an example calculation for EBB traffic. In essence, consider (2.15) with integrable overflow profile $\varepsilon_p(b)$, i.e., $\int_0^\infty\varepsilon_p(b)db<\infty$. A slack rate $\varrho>0$ is substituted into b in (2.15) as $b=b'+\varrho u$ where we use u to substitute for the time span $t-s$ in (2.15). In this case the envelope is relaxed by the slack rate ϱ and the corresponding point-wise violation probability ε_p decays with increasing u. An upper bound on the overflow probability for the sample path envelope formulation in (2.21) is found by invoking Boole's inequality[3] over all $u>0$. For each summand in the sum returned by Boole's inequality we can invoke (2.15) with violation probability $\varepsilon_p(b'+\varrho u)$. The decreasing nature of $\varepsilon_p(b)$ in b enables upper bounding the sum returned by Boole's inequality $\sum_{u=1}^\infty\varepsilon_p(b'+\varrho u)$

3 Boole's inequality states that for any countable set of events x_i for $i\in\mathbb{N}$ it holds that $\mathsf{P}\left[\bigcup_{i=1}^\infty x_i\right]\leq\sum_{i=1}^\infty\mathsf{P}\left[x_i\right]$ [53].

by the integral $\frac{1}{\varrho}\int_{b'}^{\infty}\varepsilon_p(k)dk$ [28]. The result from evaluating the integral is a valid upper bound $\varepsilon_s(b)$ on the sample path violation probability in (2.21).

Next, we review an example for stochastic performance bounds derived for EBB class traffic. In the following chapters we will often refer to the EBB traffic model for comparison. EBB traffic is defined in [130] through a linear rate envelope and exponentially decaying overflow profile as $\varepsilon_p(b)$ in (2.15). Consider an EBB traffic source that is described by a discrete time Markov model. The Markov model possesses two states, namely the *off* state (state 1), where no traffic is generated, and the *on* state (state 2), where traffic is generated with peak rate G. The state transition probabilities are denoted p_{ij} resembling the transition from state i to state j for $i,j \in \{1,2\}$. Thus, the steady-state probability of the *on* state is $p_{\text{on}} = p_{12}/(p_{12}+p_{21})$. The source produces traffic with a mean rate $g = p_{\text{on}}G$. The source burstiness can be characterized by the average time to change states twice [28], i.e., $U = 1/p_{12} + 1/p_{21}$. A Markov source possesses an envelope $E(t) = \rho(\theta)t$ that satisfies (2.15) with overflow profile $\varepsilon_p(b) = e^{-\theta b}$ [28]. The linear envelope rate $\rho(\theta)$ is given in [[23] p. 246] by

$$\rho(\theta) = \frac{1}{\theta}\log\left(\frac{1}{2}\left(p_{11}+p_{22}e^{\theta G}+\sqrt{(p_{11}+p_{22}e^{\theta G})^2-4(p_{11}+p_{22}-1)e^{\theta G}}\right)\right) \quad (2.22)$$

for $\theta > 0$. Note that $g \leq \rho(\theta) \leq G$.

For m multiplexed sources the arrivals are expressed as $A(t) = \sum_{i=1}^{m} A_i(t)$. Multiplexing m statistically independent sources results in the addition of their respective effective bandwidths [66]. This is due to the fact that the MGF of the sum of independent random variables is given by the multiplication of the individual MGFs [[53] p. 182]. For m statistically independent Markov sources it can be shown using Chernoff's theorem that a stochastic envelope $E(t) = m\rho(\theta)t$ satisfies (2.15) with overflow profile $\varepsilon_p(b) = e^{-\theta b}$ [28].

A backlog bound for the arrivals of m EBB traffic sources at a constant rate server with capacity C is found in [26, 130] by invoking Boole's inequality, Chernoff's theorem and using a slack rate $C - m\rho(\theta) > 0$. We show the derivation of the

backlog bound using a sample path bound for gSBB traffic characterization with linear envelope $E(t)$ as follows

$$\begin{aligned}
\mathsf{P}[B > b] &\leq \mathsf{P}\left[\sup_{s\in[0,t]}\{A(s,t) - E(t-s)\} > b\right] \\
&\leq \sum_{s=0}^{t-1} \mathsf{P}\left[A(s,t) - E(t-s) > b\right] \\
&= \sum_{s=0}^{t-1} \mathsf{P}\left[A(s,t) - m\rho(\theta)(t-s) - (E(t-s) - m\rho(\theta)(t-s)) > b\right] \\
&\leq \sum_{u=1}^{\infty} e^{-\theta(b+(E-m\rho(\theta))u)} \\
&\leq \int_0^{\infty} e^{-\theta(b+(E-m\rho(\theta))u)} du \\
&= \frac{e^{-\theta b}}{\theta(E - m\rho(\theta))} = \varepsilon_s(b). \qquad (2.23)
\end{aligned}$$

for any $\theta > 0$ subject to the slack rate difference $E(t-s) > m\rho(\theta)(t-s)$. Substituting $E(t) = Ct$ yields the following backlog bound directly

$$\mathsf{P}[B > b] \leq \frac{e^{-\theta b}}{\theta(C - m\rho(\theta))}. \qquad (2.24)$$

In (2.23) we employed in the first line the sample path envelope with linear rate known from the gSBB model. In the second step we invoked Boole's inequality. In the third line we reformulated the expression to apply the statistical envelope formulation (2.15). From Chernoff's theorem we know that the envelope $m\rho(\theta)t$ satisfies (2.15) with overflow profile $\varepsilon_p(b) = e^{-\theta b}$ [28]. We employ a change of variables to adapt the point-wise violation probability as $\varepsilon_p(b) = e^{-\theta(b+(E-m\rho(\theta))u)}$ and find that the sum can be upper bounded by an integral due to the decreasing nature of the exponential function. We solve the integral to find a valid violation probability for the sample path envelope. Substituting $E(t) = Ct$ directly yields a valid backlog bound in (2.24). This result is also a locally-FIFO delay bound with $\mathsf{P}[W > b/C] \leq \varepsilon_s(b)$. Stochastic performance bounds that are based on integrability conditions for decaying point-wise envelope violation probabilities appeared in [22, 46, 121, 130].

The demonstrated performance bounds can be improved under additional assumptions such as independent traffic increments. Refined performance bounds can

then be attained by invoking Doob's maximal inequality following the construction of supermartingales [25, 26].

Counterparts to stochastic traffic envelopes are stochastic service curves. Both are fundamental to the derivation of stochastic performance bounds as will be shown in the sequel. Stochastic service curves describe service guarantees provided by queueing systems that are, however, violated at most with a given probability ε. A central definition of stochastic service curves is given in [28] and also in [34] as

$$\mathsf{P}\left[D(t) < \inf_{s\in[0,t]}\{A(s) + [S(t-s) - b]_+\}\right] \leq \varepsilon(b). \tag{2.25}$$

The deficit profile $\varepsilon(b)$ is a non-increasing function in b. Setting $\varepsilon(b) = 0$ for all $b \geq 0$ reduces the definition (2.25) to the deterministic case (2.1).

In order to describe traffic scheduling in queueing systems that comprise through as well as cross traffic as in Figure 2.3, we review so-called stochastic leftover service curves [28, 47, 74, 76, 103]. We restrict the exposition to the blind multiplexing model, i.e., with no assumptions on the scheduling algorithm. Further results on scheduling algorithms such as GPS and EDF can be found, for example, in [74, 103]. Note that the authors in [74] show that the gain of accurately characterizing scheduling algorithms instead of using blind multiplexing is modest compared to the leap from deterministic to stochastic service curves.

Consider the single node in Figure 2.3 with fixed capacity C, and incoming through traffic $A^{th}(t)$ and cross traffic $A^{cr}(t)$. Assume the cross traffic possesses an envelope $E^{cr}(t)$ satisfying (2.21) with sample path violation probability $\varepsilon_s(b)$. The through traffic $A^{th}(t)$ is, thus, provided with a leftover service curve $S^{lo}(t) = [Ct - E^{cr}(t)]_+$ with $\limsup_{t\to\infty} E^{cr}(t)/t < C$ to avoid triviality [28]. The leftover service curve $S^{lo}(t)$ satisfies (2.25) with deficit profile $\varepsilon_s(b)$.

Stochastic Performance Bounds and Concatenation Results

The stochastic network calculus framework enables the derivation of performance bounds based on the notion of stochastic service curves and sample path arrival envelopes. Stochastic backlog and delay bounds can be illustrated as vertical and

horizontal deviations of the service curve and the sample path arrival envelope that are subject to deficit and overflow profiles, respectively. Consider arrivals with envelope $E(t)$ according to (2.21) with overflow profile $\varepsilon^{th}(b)$ at the ingress of a system offering service curve $S(t)$ according to (2.25) with deficit profile $\varepsilon^{cr}(b)$. A backlog bound is found in [28, 34] as

$$\mathsf{P}\left[B > \sup_{s\geq 0}\{E(s) - S(s)\} + b\right] \leq \varepsilon(b), \tag{2.26}$$

where $\varepsilon(b) = \varepsilon^{th} \otimes \varepsilon^{cr}(b)$. For the scenario depicted in Figure 2.3 the term $\varepsilon^{th}(b)$ denotes the overflow profile for the *through* traffic. Further, the deficit profile of the service curve, i.e., $\varepsilon^{cr}(b)$ corresponds to the overflow profile of the *cross* traffic according to (2.21). Similarly, a locally-FIFO bound on the delay W is found as

$$\begin{aligned} &\mathsf{P}[W > w] \leq \varepsilon(b) = \varepsilon^{th} \otimes \varepsilon^{cr}(b), \\ \text{with} \quad &w = \inf\{s \geq 0 : S(t+s) \geq E(t) + b \ \ \forall t \geq 0\}. \end{aligned} \tag{2.27}$$

The derivation of single node performance bounds builds on the stochastic service curve definition (2.25), which has the following property: it uses a sample path formulation of the arrivals and makes a point-wise argument on the departures. Concatenation results in stochastic network calculus cannot be derived by straightforward recursion as in the deterministic case (2.13) [74]. The authors of [74] note the hard problem of constructing a stochastic network service curve that is expressed by the convolution of single node stochastic service curves. A successful approach to this problem was found in [26, 28] using an extended definition of a sample path service curve that permits the derivation of a stochastic network service curve [26, 28]. We recapitulate the definition from [28] with a minimal modification to discrete time, as we will rely on this formulation in the following chapters. A sample path service curve is given by

$$\mathsf{P}\left[\sup_{t\in[0,u]}\left\{\inf_{s\in[0,t]}\left\{A(s) + [S(t-s) - \varrho(u-t) - b]_+\right\} - D(t)\right\} > 0\right] \leq \varepsilon^{\varrho}(b). \tag{2.28}$$

with deficit profile

$$\varepsilon^{\varrho}(b) = \frac{1}{\varrho} \int_b^{\infty} \varepsilon(x) dx, \tag{2.29}$$

where $0 \leq s \leq t \leq u$ and $\varepsilon(x)$ corresponds to the deficit profile in (2.25). The deficit profile $\varepsilon^{\varrho}(b)$ for the sample path service curve is derived in [26, 28] by the application of Boole's inequality given the relaxation of the service curve by a slack rate parameter $\varrho > 0$ in (2.28). Equipped with (2.28) a stochastic network service curve is derived in [26, 28]. Thus, consider a tandem of n nodes as in Figure 2.2. Each node $i \in [1, n]$ provides a service curve $S_i(t)$ in the sense of (2.25) with deficit profile $\varepsilon_i(b)$. Using the notion $S^{-\varrho}(t) = S(t) - \varrho t$ a network service curve satisfying (2.25) is given in [26, 28] as

$$S^{net}(t) = S_1 \otimes S_2^{-\varrho} \otimes \cdots \otimes S_n^{-(n-1)\varrho}(t) \tag{2.30}$$

with deficit profile

$$\varepsilon^{net}(b) = \varepsilon_1^{\varrho} \otimes \varepsilon_2^{\varrho} \otimes \cdots \varepsilon_{n-1}^{\varrho} \otimes \varepsilon_n(b). \tag{2.31}$$

For the scenario in Figure 2.2 the formulation (2.25) holds with network arrivals $A_1(t)$, departures $D_n(t)$ and service curve $S^{net}(t)$ with deficit profile $\varepsilon^{net}(b)$.

We note that the derivation using (2.31) does not assume statistical independence of the cross traffic at the different nodes. End-to-end performance bounds that account for statistical independence of the individual service curves have been derived using an MGF based network calculus framework in [46]. Approaches to account for statistical independence in the formulation (2.30) can be found in [64, 111].

Next, we consider the scenario in Figure 2.4 comprising a network path of n nodes with single node persistent EBB cross traffic. We sketch the derivation of end-to-end performance bounds for the considered scenario [26]. First, derive leftover service curves $S_i^{lo}(t)$ for the individual nodes $i \in [1, n]$. Next, consolidate the network path into an equivalent system as in Figure 2.1 using the network service curve formulation (2.30). Finally, apply single node performance bounds formulated in

(2.26) and (2.27). End-to-end performance bounds for EBB cross traffic derived in [26, 28] from (2.30) are shown to scale in the number of traversed nodes n as $\Theta(n \log n)$ [19, 28]. This compares to the $\Theta(n)$ exact delay scaling known from queueing theory [68] and $\mathcal{O}(n)$ scaling for MGF based network calculus from [46], both derived under statistical independence conditions.

3 PROBLEM STATEMENT

In the previous chapters we reviewed the related work on features and modeling of Internet traffic and introduced the state-of-the-art in flow level performance evaluation of communication networks using the theory of network calculus. In the following sections we describe and summarize the research problems tackled in this thesis.

NETWORK MONITORING AND LRD TRAFFIC PARAMETER ESTIMATION

In the light of the analytical approaches to performance analysis that are discussed in Chapter 1 we note that it is important for the application of traffic engineering and QoS provisioning to understand the characteristics of the carried traffic and of the network. This encouraged researchers to investigate methods for inferring network and traffic properties [8, 39, 61, 62, 80, 81, 105, 122, 122, 123, 126]. These methods can be passive, i.e., based on non-intrusive monitoring, or active, i.e., based on injecting test traffic into the network.[1] Passive methods require direct access to the point of interest, e.g., a router. Active methods include injecting test packets (probes) into the network and measuring metrics related to packet interactions, e.g., send and receive timing information. One main difficulty of active methods is to design the test packets such that they reveal the desired metric. In some cases this has proven impossible [84]. Further, in practice it is not always possible to access the desired metric at all times, hence it may be accessed according to some sampling strategy [6, 84, 93, 115, 125, 129].

The approximate result in (1.2) indicates that long memory has a considerable impact on the queueing performance. Network performance monitoring applications require estimates of the relevant traffic properties. These estimates serve together with performance models to assess the provided QoS in operating networks. How-

1 In another definition passive methods use production traffic such as the self-clocking mechanism in TCP, opposed to injecting test traffic in active methods.

ever, we note that obtaining such traffic properties for monitoring purposes using traditional trace collection is challenging due to confidentiality issues and technical constraints on capture and storage speeds. In addition, standard traffic traces reflect statistics only at one vantage point. In contrast, in the following work we target a lightweight characterization of end-to-end connections without administrative support. In Chapter 4 we provide a sampling-based probing method for inferring LRD traffic characteristics in single and multi-node scenarios.

NON-ASYMPTOTIC PERFORMANCE EVALUATION WITH LRD TRAFFIC

An established model for LRD traffic in the literature [44, 66, 72, 95, 96] is a superposition of a linear process with a mean rate that we denote as λ, and an fBm process we denote as $V(t)$, i.e., the traffic arrivals are given by $A(t) = \lambda t + V(t)$. In Chapter 1 we reviewed the properties of the process $V(t)$, which is characterized by the variance parameter σ^2 and Hurst parameter $H \in (0.5, 1)$. One (unfriendly) property of the LRD process $V(t)$ is that it is heavily bursty with burst periods more likely to remain for long times. This property is based on the variance of $V(t)$ which grows super-linearly in time, i.e., $\sigma^2 t^{2H}$.

An envelope in the sense of (2.16) for fBm arrivals with mean rate λ, variance parameter σ^2 and LRD Hurst parameter $H \in (0.5, 1)$ is derived in [48, 74, 91] as

$$E(t) = \lambda t + \sqrt{-2 \log \varepsilon_p} \sigma t^H, \tag{3.1}$$

using the MGF for fBm traffic (1.3) and optimizing over θ to obtain the minimum in (2.16) at $\theta = \frac{1}{\sigma}\sqrt{-2 \log \varepsilon_p} t^{-H}$. Analyzing the expression in (3.1) it becomes apparent that for fBm with $H \in (0.5, 1)$ it is not possible to employ the sample path construction technique reviewed in Section 2.2. Recall that this technique is based on relaxing the statistical envelope $E(t)$ that satisfies (2.15) using a slack rate ϱ. Through the substitution of $b = b' + \varrho t$ in (2.15) it is achieved that the point-wise violation probability ε_p decays with t as $e^{-\theta \varrho t}$. This can be directly verified through the application of Chernoff's theorem. In Chapter 2 we showed that this technique permits the derivation of a statistical backlog bound for EBB traffic. However, inserting the

point-wise violation probability $e^{-\theta\varrho t}$ for ε_p in (3.1) we find that the designated envelope expression for fBm traffic grows faster than linearly with t due to $t^{\frac{1}{2}+H}$ with $H \in (0.5, 1)$. This violates the condition for finding a nontrivial statistical backlog bound at a constant rate server with capacity C that poses an upper bound on the growth of the envelope $E(t)$ such as $\limsup_{t\to\infty} E(t)/t < C$. Inserting $E(t)$ from (3.1) into the above inequality we find that the overflow probability ε_p decays with t at most with $e^{-t^{2-2H}}$.

A gSBB envelope for fBm traffic is proposed in [131] following the expression in (1.2). However, the formulation (1.2) does not provide a rigorous upper bound as in (2.21) but rather a lower bound through the approximation by the largest term. Based on the approximation (1.2) the authors of [132] analyze the behavior of LRD traffic under GPS scheduling. For an envelope rate $r > \lambda$ the corresponding approximate overflow profile is given by

$$\tilde{\varepsilon}_a = \exp\left(-\frac{1}{2\sigma^2}\left(\frac{r-\lambda}{H}\right)^{2H}\left(\frac{b}{1-H}\right)^{2-2H}\right). \tag{3.2}$$

The calculation of non-asymptotic upper performance bounds for LRD fBm traffic for single as well as multi-node scenarios has been an open problem in the field of stochastic network calculus for some time. In Chapters 5 and 6 we provide a novel approach to tackle the problem of finding backlog and delay bounds for queueing systems fed with LRD fBm traffic as well as for entire network paths with LRD fBm cross traffic. This is a central contribution of this thesis.

MAIN RESEARCH QUESTIONS

In the following, we summarize the main research questions that arise based on the literature review in Chapters 1 and 2. These questions identify the main research directions to which we contribute to with this thesis. We show the impact of answering the specified research questions on practical applications of quality of service monitoring and provisioning.

In this thesis we aim at a lightweight estimation method for LRD traffic parameters and at a non-asymptotic performance evaluation of networks carrying LRD traffic.

First, we seek to answer the following questions:

- Can LRD traffic properties be reliably estimated from sampled traffic subsets?
- Which sampling parameters govern the accuracy of the estimates?
- How can LRD traffic properties be estimated without administrative access?

Answering the research questions stated above enables efficient and online monitoring of LRD parameters from traffic samples without necessarily collecting entire traces. Further, optimizing the sampling parameters permits assessing and adjusting the measurement accuracy. Moreover, a traffic parameter estimator based on the interaction of packet probes and LRD traffic dispenses with administrative network support and paves the way for end-to-end traffic parameter estimation.

Estimating the parameters of LRD network traffic is the first step to analyze the provided QoS. It is of significant importance to deploy an analysis based on rigorous mathematical models as described in Chapter 2. Hence, the following research questions arise:

- Is it possible to derive non-asymptotic upper performance bounds for systems fed with LRD traffic?
- How do rigorous performance bounds derived using network calculus compare to state-of-the-art asymptotes?
- How does LRD traffic impact the service received by concurrent flows in resource sharing scenarios?

The answers to the questions above contribute a novel non-asymptotic analysis of queueing performance in the presence of LRD traffic. The QoS analysis of general scheduling models with LRD traffic lays the foundation for end-to-end performance evaluation under scheduling constraints. This leads to the following research questions:

- What is the impact of LRD traffic on the end-to-end network performance?
- Given a network path, how do performance bounds scale with the number of traversed nodes exhibiting LRD traffic?

- What are the implications of LRD on network dimensioning?

Answering these questions helps to understand the role of LRD traffic in data networks and its impact on the service provided for end-to-end connections. The corresponding results would carry the fundamental properties of the network calculus framework, i.e., scheduling formulations and end-to-end performance analysis, forward to the analysis of networks with LRD traffic such as the Internet.

4

LRD TRAFFIC PARAMETER ESTIMATION USING SAMPLING AND NETWORK PROBING

The inference of communication network characteristics, especially traffic properties, has been of great interest to communications engineers since the dawn of telephone networks. For example, using measurements of carried traffic load together with the queueing models developed by A. K. Erlang, it was possible to assess the performance of operating telephone networks. Today, in packet data networks the inference concept remains unchanged. Inferred metric estimates of, e.g., link capacity [39, 61], available bandwidth [62, 105, 122], service curves [78, 83], and traffic Hurst parameter [8, 123, 126] describe different characteristics of communication networks and carried traffic. These metrics serve together with mathematical models for performance evaluation to assess and design existing and future networks, respectively.

Commonly used metrics in modern data networks for inferring traffic and network properties are data volumes and packet timing information. Both metrics can be obtained from non-intrusive monitoring as well as dedicated active measurements. For example, a popular estimator of the average available bandwidth relies on the time dispersion of back-to-back packet pairs [39, 80, 81, 122]. Relevant for this thesis are procedures that acquire packet volume and timing information. This can be obtained on the one hand, through sampling the traffic. On the other hand, this information can be inferred by injecting packet probes of predefined pattern into the network and interpreting their interactions.

Sampling is typically deployed when continuous system monitoring is impractical either due to access restrictions or to significantly reduce processed data amounts. In the context of traffic monitoring it becomes clear that capturing traffic on links with speeds of 10 Gbps and more is challenging. Not only that the standard hard disk writing speeds are not sufficient but also the increasing storage space is costly. This establishes the need for methods that extract the desired metrics from data samples without bias, optimally, in an online manner without post-processing.

A fundamental result often employed in the sampling context is known as PASTA, Poisson Arrivals see Time Averages [129]. PASTA states that the portion of Poisson arrivals that see a system in a certain state corresponds, on average, to the portion of time the system spends in that state. This translates into the fact that packets arriving at a router as a Poisson process see on average the true mean queue length [129]. Similarly, monitoring the queue length at times according to a Poisson process reveals the true mean queue length [6].

PASTA has been generalized in [93] to find bias free estimates that are not limited to Poisson sampling. A recent powerful generalization of PASTA is provided in [6] and denoted as NIMASTA, i.e., Non-intrusive Mixing Arrivals See Time Averages. The NIMASTA theorem from [6] provides the basis for bias free estimates through an almost sure convergence of

$$\lim_{R\to\infty} \frac{1}{R} \sum_{i=1}^{R} \varphi(Y(\zeta_i)) = \mathsf{E}\left[\varphi(Y(0))\right], \tag{4.1}$$

where $Y(\zeta_i)$ is a sample of the process $Y(t)$ at time ζ_i and φ is a general positive function of Y. The sampling times ζ_i for $i \in \mathbb{N}$ are provided by a stochastic sampling process. The positive function φ enables adjusting the formulation to different metrics, e.g., different moments. The formulation (4.1) is proved in [6] for ergodic $Y(t)$ combined with stochastically mixing sampling processes. Note that the fGn process reviewed in Section 1.1 for modeling LRD traffic increments is proved to be ergodic in [30, 117]. The formulation (4.1) provides bias free estimates of certain traffic parameters, such as mean rate and variance.

Comparisons of different sampling processes are given, e.g., in [6, 102, 115, 125]. In [115] the authors compare Poisson and periodic sampling processes with respect to the variance of the sample mean. They show that Poisson or periodic sampling can be superior depending on the autocovariance of the sampled process. The authors of [6] present examples of zero bias sampling processes that possess different sample mean variances.

Investigations on LRD traffic parameter estimation that are related to this work were reported in [55, 102, 104]. For correlation lags tending to infinity, random sampling observations are shown in [102] to possess the long memory of the sampled

processes, as long as the distribution of the time between samples has a finite mean. The simulation studies [55, 104] investigated active probing techniques that could lead to LRD traffic estimates. The authors of [55] use numerical simulations to show the possibility of interpolating LRD traffic from probes that are based on back-to-back packet pairs described in [81, 122]. In [104] the authors build on a multi-fractal wavelet traffic model to derive a *deterministic* probing scheme that captures the multi-fractal wavelet coefficients of the traffic and show corresponding simulation results. Essential to the estimation in [104] is the assumption that the queue does not empty between the individual packets of a packet probing train.

The contribution of this chapter lies in the reliable estimation of LRD traffic correlations and Hurst parameter H, from *random* sampling observations. Our work differs substantially from [55, 102, 104] as we, for instance do not only regard the asymptotic regime, we do not rely on back-to-back packet probing and do not apply interpolation. We use diverse sampling processes and provide a general framework for reversing the impact of random sampling on the traffic correlations. This is illustrated in Section 4.1. Further, we examine in Section 4.2 the impact of limited sample sizes on the accuracy of the estimated traffic parameters. We formulate an asymptotically unbiased estimator based on the samples. Contrary to traditional trace-driven approaches that reflect the traffic statistics at *one* network node, we design in Section 4.3 a probing scheme that enables the estimation of the dominant traffic correlations over entire network paths. Finally, we present experimental evaluation results obtained from lab measurements as well as results from an extensive Internet measurement campaign. Results presented in this chapter are joint work with Z. Bozakov and M. Fidler [112].

4.1 TRAFFIC SAMPLING AND LRD TRAFFIC PARAMETER ESTIMATION

In this section we present a model for sampling long memory traffic. Our goal is to extract the LRD characteristics from observed samples. The main findings in this section are that the sampling process impacts the observations, and that this impact

can be reversed under certain conditions to estimate the autocovariance of the long memory traffic.

We define a sampling model on the basis of three stationary discrete time processes, i.e., for $t \in \mathbb{N}_0$. First, we consider the LRD traffic increment process $Y(t)$ that is characterized by a Hurst parameter H and slowly decaying autocovariance given in (1.1). The increment process is related to the LRD cumulative arrival process $A(t)$ through $Y(t) = A(t+1) - A(t)$. Second, we consider the sampling process $X(t)$ and the observed process $Z(t)$, which stems from the interaction of $X(t)$ and $Y(t)$.

We define $X(t)$ as a point process that takes the value of one when a sample is taken and zero otherwise. Formally, $X(t)$ is a train of Kronecker deltas, where a Kronecker delta is defined in [[97] p. 420] as

$$\delta(t) = \begin{cases} 1 & \text{for } t = 0 \\ 0 & \text{otherwise} \end{cases}$$

The times between two consecutive Kronecker deltas, i.e., the inter-sample time, are independent and identically distributed (iid) random variables with a given probability distribution that characterizes the sampling process $X(t)$. We define the sampling intensity μ_X as $\mathsf{E}[X(t)]$ with $0 \leq \mu_X \leq 1$.[1]

We consider the observed stochastic process $Z(t)$ that is generated by random samples $X(t)$ of the increment process $Y(t)$ and is given by

$$Z(t) = X(t)Y(t). \tag{4.2}$$

Our goal is to infer the LRD characteristics of the traffic increment process $Y(t)$ given the observed process $Z(t)$. The long memory of $Y(t)$ is resembled in the slow decay of its autocovariance $c_Y(\tau)$ given in (1.1). The following lemma relates the autocovariance of the observed process $Z(t)$ to the autocovariance of $Y(t)$.

1 Throughout this chapter we use $\mu_{(\cdot)}$ to denote the expected value $\mathsf{E}[(\cdot)]$.

Lemma 4.1 *Given stationary and independent stochastic processes $X(t)$ and $Y(t)$ and let $Z(t) = X(t)Y(t)$. The autocovariance of $Z(t)$ can be decomposed into*

$$c_Z(\tau) = \left(c_X(\tau) + \mu_X^2\right) c_Y(\tau) + c_X(\tau)\mu_Y^2.$$

Proof of Lemma 4.1. For independent and stationary processes $X(t)$ and $Y(t)$ and for $Z(t)$ given by (4.2), it follows from the definition of the autocovariance function in Chapter 1 that

$$\begin{aligned} c_Z(\tau) &= \mathsf{E}\left[Z(t)Z(t+\tau)\right] - \mu_Z^2 \\ &= \mathsf{E}\left[X(t)Y(t)X(t+\tau)Y(t+\tau)\right] - \mu_X^2\mu_Y^2 \\ &= \mathsf{E}\left[X(t)X(t+\tau)\right]\mathsf{E}\left[Y(t)Y(t+\tau)\right] - \mu_X^2\mu_Y^2 \\ &= \left(c_X(\tau) + \mu_X^2\right)\left(c_Y(\tau) + \mu_Y^2\right) - \mu_X^2\mu_Y^2 \\ &= \left(c_X(\tau) + \mu_X^2\right) c_Y(\tau) + c_X(\tau)\mu_Y^2 \end{aligned}$$

where $c_{(\cdot)}(\tau)$ denotes the autocovariance of process $(\cdot)$ at lag τ. ∎

Lemma 4.1 demonstrates the impact of the chosen sampling distribution on the observed autocovariance. The observed statistics are determined by the properties of the sampling process $X(t)$, specifically, its sampling intensity μ_X and autocovariance $c_X(\tau)$. For sampling processes with autocovariances that decay fast enough in τ, the observed $c_Z(\tau)$ has the same asymptotic tail decay as $c_Y(\tau)$ for $\tau \to \infty$. Specifically, for $\lim_{\tau\to\infty} \frac{c_X(\tau)}{c_Y(\tau)} = 0$ we compute from Lemma 4.1

$$\lim_{\tau\to\infty} \frac{c_Z(\tau)}{c_Y(\tau)} = \lim_{\tau\to\infty} \left(c_X(\tau) + \mu_X^2\right) + \frac{c_X(\tau)}{c_Y(\tau)}\mu_Y^2 = \mu_X^2, \tag{4.3}$$

i.e., the asymptotic tail decay of the observed autocovariance $c_Z(\tau)$ is the same as for $c_Y(\tau)$ except for a rescaling factor μ_X^2. The condition above holds, for example, for SRD sampling processes and for Markovian sampling processes with exponentially decaying $c_X(\tau)$. On a log-log scale $c_Z(\tau)$ decays linearly with slope $2H - 2$ for $\tau \to \infty$. This observation matches the asymptotic result from [102].

Next, we investigate the structure of the observed $c_Z(\tau)$ at some positive finite τ for four different sampling processes defined by corresponding inter-sample distri-

butions. We consider the inter-sample distributions: geometric, periodic, Gamma, and uniform given in Table 4.1. The first two distributions are characterized by probability mass functions (pmf), while the second pair is characterized by probability density functions (pdf). In the following, we use $f(\tau)$ for pdf and pmf for notational simplicity, yet we precisely make clear in the description of the derived formulae, whether pmf or pdf is considered. Next, we derive analytical expressions for the autocovariance function $c_X(\tau)$ and for μ_X to assess the impact of the different sampling processes $X(t)$ on the observed autocovariance $c_Z(\tau)$ through Lemma 4.1. To this end, we rephrase a standard result from the framework of [31], in particular, its equation Eq. (4.6.1).

From [31] we derive the autocorrelation density

$$\mathsf{E}\left[X(t)X(t+\tau)\right] = \mu_X \sum_{k=1}^{\infty} f^{(*k)}(\tau), \tag{4.4}$$

where $f^{(*k)}(\tau)$ is the k-fold self-convolution of the probability density function of the inter-sample times $f(\tau)$. The intuition behind this formula is that starting from one Kronecker delta at $X(t)$, another Kronecker delta at $X(t+\tau)$ can be the first, the second, the third, etc. Kronecker delta to come after the one at $X(t)$. The k-fold self-convolution of $f(\tau)$ expresses the density function for the sum of k independent random variables each described by $f(\tau)$. The derivation in [31] uses a small time interval of length $\Delta t \to 0$, such that $\mu_X \Delta t$ is the probability that a Kronecker delta occurred in $[t, t+\Delta t)$, see Eq. (4.5.9) in [31]. To calculate the correlations of $X(t)$ at lag τ the author of [31] deduces the conditional probability that a Kronecker delta occurred at $[t+\tau, t+\tau+\Delta t)$ given a Kronecker delta at $[t, t+\Delta t)$. This is given by $\sum_{k=1}^{\infty} f^{(*k)}(\tau)\Delta t$ in Eq. (4.5.11) in [31]. It follows that $(\Delta t)^2 \mu_X \sum_{k=1}^{\infty} f^{(*k)}(\tau)$ is the autocorrelation of Kronecker deltas, that are τ apart, observed in time slots of length Δt. We exploit the property that $f^{(*k)}$ can be formulated as a power series for the considered distributions and that its sum in (4.4) converges.

A discrete-time extension of the correlation calculation from [31] with probability mass functions is straightforward. To this end we replace the probability density functions with probability mass functions and consider a time slot $\Delta t = 1$ such that we obtain correlation functions instead of densities. For the continuous time

distributions considered we regard the correlations on a fixed time slot basis. We use a discretization with a time slot of unit size.

GEOMETRIC SAMPLING. First, we consider inter-sample times drawn from a geometric distribution with pmf $f(\tau) = p(1-p)^{\tau-1}$ for $\tau \in \mathbb{N}$, i.e., the geometric sampling process is a Bernoulli process with parameter p. Hence, the probability of drawing a sample at one time slot is p and the sampling intensity is $\mu_X = p$. We calculate the autocovariance for $\tau > 0$ of the geometric sampling process as

$$c_X(\tau) = 0. \tag{4.5}$$

Proof The k-fold self-convolution of $f(\tau)$ is the probability mass function (pmf) of the sum of k geometrically distributed random variables, i.e., negative binomial distributed with parameters p and k [[53] p. 61]. We insert the negative binomial pmf for $f^{(*k)}(\tau)$ into (4.4) to find

$$\begin{aligned}
\mathsf{E}\left[X(t)X(t+\tau)\right] &= \mu_X \sum_{k=1}^{\tau} \binom{\tau-1}{k-1} p^k (1-p)^{\tau-k} \\
&= \mu_X \sum_{s=0}^{l} \binom{l}{s} p^s (1-p)^{l-s} p \\
&= \mu_X p \\
&= \mu_X^2.
\end{aligned}$$

In the first line we used the support of the pmf $f^{(*k)}(\tau)$ to bound $1 \leq k \leq \tau$. In the second line we substituted $\tau = l+1$ and $k = s+1$ and rearranged the sum. In the third line we used the binomial identity $\sum_{s=0}^{l} \binom{l}{s} x^s y^{l-s} = (x+y)^l$. Finally, we inserted $\mu_X = p$. The autocovariance is given by $c_X(\tau) = \mathsf{E}\left[A(t)A(t+\tau)\right] - \mu_X^2$ which completes the proof. ∎

In fact, the independence of the increments of the geometric sampling process implies that $c_X(\tau) = 0$. Plugging (4.5) into Lemma 4.1 yields for the observations

$$c_Z(\tau) = c_Y(\tau)\mu_X^2. \tag{4.6}$$

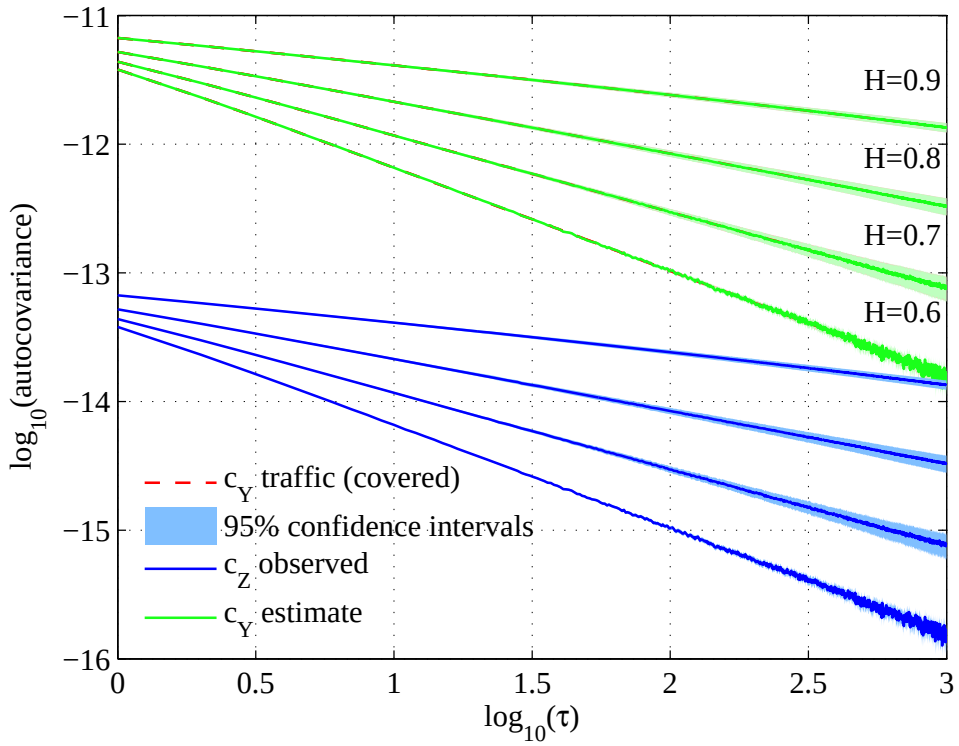


Figure 4.1: Autocovariances of LRD traffic processes and its geometric sampled counterparts. The observed autocovariance "c_Z observed" maintains the autocovariance structure of the traffic process autocovariance. The reconstructed "c_Y estimate" precisely covers the autocovariance of the traffic process "c_Y traffic".

Geometric sampling or, generally, sampling processes with uncorrelated increments, only rescale the autocovariance of the sampled process $Y(t)$ by μ_X^2 as shown in (4.6). The structure of the traffic autocovariance $c_Y(\tau)$, i.e., the linear decay on a log-log scale, remains unchanged after sampling. Figure 4.1 depicts autocovariance estimates from observations $Z(t)$ denoted "c_Z observed". These are obtained by applying geometric sampling to LRD traffic processes with Hurst parameters $H \in \{0.6, 0.7, 0.8, 0.9\}$.[2] Complying with (4.6) Figure 4.1 shows that the observations from geometric sampling exhibit the linear autocovariance decay of $c_Y(\tau)$ with slope $2H - 2$. Geometric sampling enables the reconstruction of the traffic autocovariance by solving for $c_Y(\tau) = c_Z(\tau)/\mu_X^2$. Additional traffic parameters need not to be estimated for reconstructing $c_Y(\tau)$. We reconstruct the traffic autocovariance to find that the reconstructed autocovariance denoted "c_Y estimate" exactly covers the traffic autocovariance "c_Y traffic" in Figure 4.1.

2 We used synthetic LRD traffic traces of length 2.5×10^8 time slots for the simulations. For each considered H the simulation was repeated 25 times. We fixed these parameters for all considered sampling processes in the sequel.

PERIODIC SAMPLING. Periodic sampling is practical because of its simplicity. The sampling process $X(t)$ is given by a comb of Kronecker deltas with period ω. Thus, the inter-sample times are drawn from a degenerate distribution with pmf $f(\tau) = \delta(\tau - \omega)$. The intensity of the periodic sampling process $X(t)$ is given as $\mu_X = 1/\omega$. A moment's consideration reveals that the autocorrelation of $X(t)$ is found as $\mathsf{E}[X(t)X(t+\tau)] = 1/\omega$ for $\tau = i\omega$, $i \in \mathbb{N}$ and zero otherwise. Given μ_X we calculate the autocovariance of $X(t)$ for $\tau > 0$ as

$$c_X(\tau) = \begin{cases} 1/\omega - 1/\omega^2 & \text{for } \tau = i\omega,\ i \in \mathbb{N} \\ -1/\omega^2 & \text{otherwise} \end{cases} \tag{4.7}$$

Combining (4.7) and Lemma 4.1 we find the observed autocovariance

$$c_Z(\tau) = \mu_X \left(c_Y(\tau) + \mu_Y^2 (1 - \mu_X) \right) \tag{4.8}$$

at the sampling times $\tau = i\omega$, $i \in \mathbb{N}$. In contrast to geometric sampling, periodic sampling observations do *not* converge to a rescaled version of the traffic autocovariance $c_Y(\tau)$ for $\tau \to \infty$. Instead, $c_Z(\tau)$ converges to $c_Z(\tau) = \mu_X \mu_Y^2 (1 - \mu_X)$ for $\tau \to \infty$. This distortion can be observed in Figure 4.2a.

The reconstruction of the traffic autocovariance from the observed $c_Z(\tau)$ is possible at $\tau = i\omega$, $i \in \mathbb{N}$ by solving (4.8) for $c_Y(\tau)$. However, this requires the knowledge of the traffic mean rate μ_Y. Note that (4.1) does not hold for the periodic sampling process as it is not mixing. The traffic mean rate estimator μ_Z/μ_X is not an unbiased estimator in case of periodic sampling [6]. In general, the fixed structure of the sampling process may lead to a sampling period that coincides with periodicities in the sampled process, thus distorting the result.

For known μ_Y, Figure 4.2a shows the successful reconstruction of the traffic autocovariance $c_Y(\tau)$ at $\tau = i\omega$, $i \in \mathbb{N}$. Note that "c_Y estimate" exactly covers the original traffic autocovariance "c_Y traffic".

Next, we analyze Gamma and uniform sampling processes. For mathematical tractability, we use continuous time for the derivation of the autocorrelation of the following sampling process $X(t)$. For discretization we use a time slot of unit size.

Note that the discretization error diminishes for autocorrelation lags much larger than the discretization time slot.

GAMMA SAMPLING. The inter-sample times of a Gamma sampling process $X(t)$ are drawn from a Gamma distribution with probability density function

$$f_{v,\iota}(\tau) = \frac{\iota^v}{\Gamma(v)}\tau^{v-1}e^{-\iota\tau} \tag{4.9}$$

with parameters $v > 0, \iota > 0$ and support $\tau > 0$. The sampling intensity is given by $\mu_X = \iota/v$. The choice of the parameter v gives the Gamma distribution a great flexibility. The Gamma sampling process degenerates to periodic sampling for $v \to \infty$ subject to constant intensity μ_X [5]. Moreover, it is generally known that for $v = 1$ the Gamma distribution degenerates to the exponential distribution, i.e., the continuous time memoryless equivalent of the aforementioned geometric distribution. In this case the sampling process is a Poisson process with intensity ι. Thus, for $v = 1$ we recover the results for the geometric sampling process.

Next, we show Gamma sampling results for selected values for v. We calculate the autocovariance of the Gamma sampling process for $v = 2$ as

$$c_X(\tau) = -\mu_X^2 e^{4\mu_X\tau}, \tag{4.10}$$

and for $v = 4$ as

$$c_X(\tau) = -\mu_X^2 e^{8\mu_X\tau} - 2\mu_X^2 \sin(4\mu_X\tau)\, e^{-4\mu_X\tau}. \tag{4.11}$$

Proof Given the time between two Kronecker deltas is Gamma distributed as (4.9) with parameters v, ι. The sum in (4.4) reduces to the following expression

$$\sum_{k=1}^{\infty} f^{(*k)}(\tau) = \frac{e^{-\iota\tau}}{\tau}\sum_{k=1}^{\infty}\frac{(\iota\tau)^{kv}}{\Gamma(kv)}. \tag{4.12}$$

We substitute $v = 2$ into (4.12), use that $\Gamma(kv) = (kv-1)!$, and evaluate the sum in (4.12) as

$$\sum_{k=1}^{\infty} \frac{(\iota\tau)^{2k}}{(2k-1)!} = \iota\tau \sinh(\iota\tau),$$

using the series expansion for $\sinh(x)$ function from [1] Eq. (4.5.62). We then exploit the identity $\sinh(x) = (e^x - e^{-x})/2$ and substitute $\iota = 2\mu_X$ to evaluate (4.4) as function of μ_X and τ as

$$\mathsf{E}\left[X(t)X(t+\tau)\right] = \mu_X^2 \left(1 - e^{-4\mu_X\tau}\right).$$

For $v = 4$ we evaluate the sum in (4.12) as

$$\sum_{k=1}^{\infty} \frac{(\iota\tau)^{4k}}{(4k-1)!} = \iota\tau \left(\frac{\sinh(\iota\tau) - \sin(\iota\tau)}{2}\right), \tag{4.13}$$

using the series expansion for $\sinh(x)$ and $\sin(x)$ functions from [1] in Eq. (4.5.62) and Eq. (4.3.65), respectively. We insert (4.13) into (4.12) and finally into (4.4) substituting the identity $\sinh(x) = (e^x - e^{-x})/2$ and $\iota = 4\mu_X$ to find

$$\mathsf{E}\left[X(t)X(t+\tau)\right] = \mu_X^2 \left(1 - e^{-8\mu_X\tau} - 2\sin(4\mu_X\tau)\, e^{-4\mu_X\tau}\right).$$

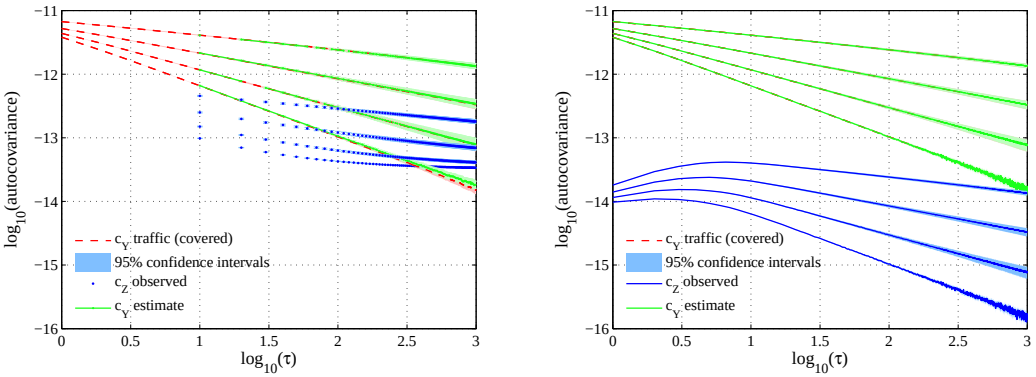
In the last step, the autocovariance of the sampling process $X(t)$ is found as $c_X(\tau) = \mathsf{E}\left[X(t)X(t+\tau)\right] - \mu_X^2$. ∎

Combining (4.10) with Lemma 4.1 yields the following expression for the observed autocovariance after Gamma sampling with $v = 2$,

$$c_Z(\tau) = \mu_X^2 \left(1 - e^{-4\mu_X\tau}\right) c_Y(\tau) - \mu_X^2\mu_Y^2 e^{-4\mu_X\tau}. \tag{4.14}$$

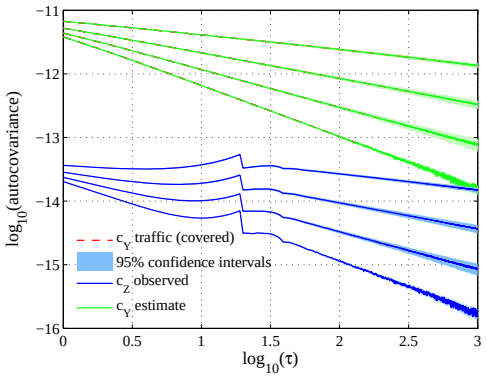
Similarly, the observed autocovariance for $v = 4$ can be deduced by combining (4.11) with Lemma 4.1 as

$$\begin{aligned} c_Z(\tau) = \mu_X^2 \left(1 - e^{-8\mu_X\tau} - 2\sin(4\mu_X\tau)\, e^{-4\mu_X\tau}\right) c_Y(\tau) \\ - \mu_X^2\mu_Y^2 \left(e^{-8\mu_X\tau} + 2\sin(4\mu_X\tau)\, e^{-4\mu_X\tau}\right). \end{aligned} \tag{4.15}$$



(a) Periodic sampling

(b) Gamma sampling



(c) Uniform sampling

Figure 4.2: Autocovariance of LRD processes for different stochastic sampling processes. Note that "c_Z observed" is obviously distorted and that "c_Y traffic" is covered by the reconstructed "c_Y estimate".

From (4.14) and (4.15) it is obvious that we are able to extract $c_Y(\tau)$ from the observed autocovariance $c_Z(\tau)$. However, this requires an estimate of the traffic intensity μ_Y that can be obtained, owing to the NIMASTA theorem by insertion of the identity function φ into (4.1) as exemplified in [6]. Further, the formulations (4.14) and (4.15) satisfy (4.3) such that for $\tau \to \infty$ the observed autocovariance $c_Z(\tau)$ converges to a rescaled version of the traffic covariance, i.e., $\mu_X^2 c_Y(\tau)$.

The ability to estimate $c_Y(\tau)$ is not limited to the considered cases of $v \in \{2, 4\}$. Lemma 4.1 can be used to estimate $c_Y(\tau)$ for Gamma sampling processes with arbitrary parameters as long as the autocovariance $c_X(\tau)$ is computable.

Figure 4.2b shows an example of the reconstruction of the traffic autocovariance $c_Y(\tau)$ from observed autocovariances using Gamma sampling with $v = 2$. The observed autocovariance denoted "c_Z observed" are obviously distorted. Note that

the reconstructed autocovariance "c_Y estimate" exactly covers the original traffic autocovariance "c_Y traffic".

UNIFORM SAMPLING. Uniform sampling is especially attractive owing to the finite support of $f(\tau)$, which is convenient for practical measurements. In contrast to geometric or Gamma sampling the inter-sample times cannot become arbitrarily large. Consider inter-sample times drawn from a uniform distribution with the density function $f(\tau) = 1/v$ for $\tau \in [0, v]$. The sampling intensity of the uniform sampling process $X(t)$ is given by $\mu_X = 2/v$. We calculate the autocovariance of the uniform sampling process $X(t)$ for $\tau \in [0, v]$ as

$$c_X(\tau) = \mu_X^2 \left(\frac{1}{2} e^{\frac{1}{2}\mu_X \tau} - 1 \right). \tag{4.16}$$

Proof Given the time between two Kronecker deltas is uniformly distributed with $f(\tau) = 1/v$ for $\tau \in [0, v]$ and zero otherwise. First, consider the sum in (4.4) for $\tau \in [0, v]$. The sum $\sum_{k=1}^{\infty} f^{(*k)}(\tau)$ from (4.4) can be expanded as

$$\begin{aligned} \sum_{k=1}^{\infty} f^{(*k)}(\tau) = f(\tau) &+ \int_0^\tau f(x_1) f(\tau - x_1) dx_1 \\ &+ \int_0^\tau \int_0^{x_1} f(\tau - x_1) f(x_1 - x_2) f(x_2) dx_2 dx_1 \\ &+ \int_0^\tau \int_0^{x_1} \int_0^{x_2} f(\tau - x_1) f(x_1 - x_2) f(x_2 - x_3) f(x_3) dx_3 dx_2 dx_1 \\ &+ \cdots, \end{aligned} \tag{4.17}$$

where the kth summand contains the pdf of the sum of k independent uniform random variables expressed through the convolution of pdfs. All arguments of the pdfs in (4.17) are in the range $[0, \tau]$ such that we can replace the pdfs in (4.17)

with $1/v$. Equation (4.17) evaluates then to the series expansion of the exponential function, i.e.,

$$\begin{aligned}\sum_{k=1}^{\infty} f^{(*k)}(\tau) &= \frac{1}{v} + \frac{\tau}{v^2} + \frac{\tau^2}{2!v^3} + \frac{\tau^3}{3!v^4} + \cdots \\ &= \frac{1}{v}\sum_{k=0}^{\infty} \frac{(\tau/v)^k}{k!} \\ &= \frac{1}{v} e^{\tau/v}. \end{aligned} \tag{4.18}$$

Finally, we use (4.18) in conjunction with (4.4), and $v = 2/\mu_X$ to derive

$$\mathsf{E}\left[X(t)X(t+\tau)\right] = \frac{1}{2}\mu_X^2 e^{\mu_X \tau/2},$$

for $\tau \in [0, v]$. The autocovariance of the process $X(t)$ is found as $c_X(\tau) = \mathsf{E}\left[X(t)X(t+\tau)\right] - \mu_X^2$, which completes the proof ∎

With (4.16) and Lemma 4.1 the observed autocovariance is expressed as

$$c_Z(\tau) = \mu_X^2 \frac{1}{2} e^{\frac{1}{2}\mu_X \tau} c_Y(\tau) + \mu_Y^2 \mu_X^2 \left(\frac{1}{2} e^{\frac{1}{2}\mu_X \tau} - 1\right) \tag{4.19}$$

for $\tau \in [0, v]$. Since the uniform sampling process $X(t)$ is mixing [6], we conclude for $\tau > v$ that the autocorrelation $\mathsf{E}\left[A(t)A(t+\tau)\right]$ converges to μ_X^2 [37, 117]. Thus, the observed autocovariance $c_Z(\tau)$ is given as $\mu_X^2 c_Y(\tau)$ for $\tau \to \infty$.

The formulation (4.19) enables the extraction of the traffic autocovariance $c_Y(\tau)$ from the observed counterpart $c_Z(\tau)$ using an estimate for μ_Y, that can be obtained from (4.1) based on the NIMASTA property [6].

Figure 4.2c shows that the observed autocovariance $c_Z(\tau)$ is significantly distorted for $\tau \in [1, v]$. Note that lower sampling intensities μ_X imply larger v, i.e., a larger span with obvious distortion in $c_Z(\tau)$. Figure 4.2c also presents the reconstruction of the traffic autocovariance $c_Y(\tau)$ from the observed autocovariance for uniform sampling. Note that the reconstructed autocovariance "c_Y estimate" exactly covers the original traffic autocovariance.

We summarize the results for the discussed sampling processes and the corresponding parameterizations in Table 4.1. It provides expressions for reconstructing

the traffic autocovariance $c_Y(\tau)$ after geometric, periodic, Gamma and uniform sampling with intensity μ_X. The sampling intensity μ_X is directly related to the respective distribution parameters. The reconstruction of the traffic autocovariance $c_Y(\tau)$ from periodic, Gamma and uniform sampling requires knowledge of the mean traffic rate μ_Y.

In the following, we assess the advantages and disadvantages of the presented sampling distributions. Periodic and uniform sampling are practically appealing due to the finite support of the inter-sample distribution. Note that, generally, periodic sampling may yield misleading results if the sampling period coincides with periodicities in the target process. In general, we showed that it is possible to extract the traffic autocovariance $c_Y(\tau)$ from sampling observations. However, periodic, Gamma and uniform sampling require knowledge of the mean traffic intensity μ_Y for the inference. Otherwise the observed autocovariance $c_Z(\tau)$ is strongly distorted as depicted in Figure 4.2.

We find that a major advantage of geometric sampling, i.e., memoryless, is that the autocovariance structure of $c_Y(\tau)$ is preserved in the observations as given in (4.6). This result holds generally for sampling processes with uncorrelated increments. This stands in contrast to the distortion observed for periodic, Gamma and uniform sampling. For example, we can infer the Hurst parameter H directly from the slope of $c_Z(\tau)$. Further, a reconstruction of $c_Y(\tau)$ is easily done as only the known sampling intensity μ_X is required. Note that the Internet Engineering Task Force (IETF) proposes memoryless sampling for network probing purposes [101]. As a result of this discussion, we focus in the following analysis on geometric sampling.

Table 4.1: Parametrization of the considered sampling distributions and traffic autocovariance reconstruction formulae

	inter-sample distribution $f(\tau)$	autocovariance $c_X(\tau) = \mathsf{E}[X(t)X(t+\tau)] - \mu_X^2$ for $\tau > 0$	reconstructed traffic autocovariance $c_Y(\tau)$	remarks
Geometric	$p(1-p)^{\tau-1}$	0	$\frac{c_Z(\tau)}{\mu_X^2}$	$\mu_X = p$
Periodic	$\delta(\tau-\omega)$	$1/\omega - 1/\omega^2$ for $\tau = i\omega$, $i \in \mathbb{N}$ $-1/\omega^2$ otherwise	$\frac{c_Z(\tau)-\mu_X\mu_Y^2(1-\mu_X)}{\mu_X}$	$c_Y(\tau)$ at $\tau = i\omega$, $\mu_X = 1/\omega$
Gamma	$\frac{\iota^\upsilon}{\Gamma(\upsilon)}\tau^{\upsilon-1}e^{-\iota\tau}$	$-\mu_X^2 e^{-4\mu_X\tau}$	$\frac{c_Z(\tau)+\mu_X^2\mu_Y^2 e^{-4\mu_X\tau}}{\mu_X^2\left(1-e^{-4\mu_X\tau}\right)}$	for $\upsilon = 2$, $\mu_X = \iota/\upsilon$
Uniform	$1/\upsilon$ for $0 \le \tau \le \upsilon$	$\mu_X^2\left(\frac{1}{2}e^{\frac{1}{2}\mu_X\tau} - 1\right)$	$\frac{2c_Z(\tau)-\mu_X^2\mu_Y^2\left(e^{\frac{1}{2}\mu_X\tau}-2\right)}{\mu_X^2 e^{\frac{1}{2}\mu_X\tau}}$	$c_Y(\tau)$ for $\tau \le \upsilon$, $\mu_X = 2/\upsilon$

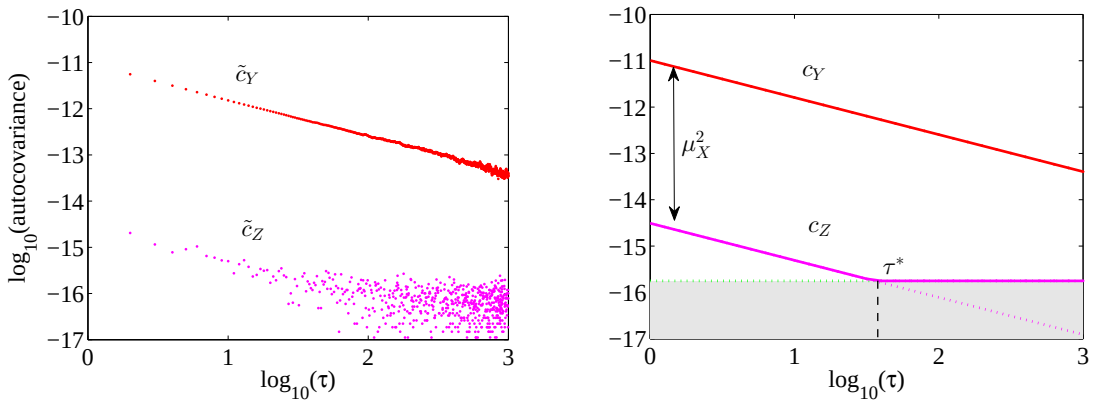
4.2 ACCURACY AND BIAS UNDER FINITE SAMPLING

In the sequel, we analyze the impact of finite sample sizes and sampling parameters on the observed autocovariance. We focus on geometric sampling because of its major advantages, as discussed at the end of Section 4.1. We show the impact of finite sample sizes on the accuracy of the used statistical estimators. The main findings in this section are that the estimation error increases with autocovariance lag τ as τ^{2-2H} with traffic Hurst parameter H and that there is a nonlinear trade-off between the sampling intensity μ_X and the sample size T with respect to the accuracy.

In the sequel we denote the sample autocovariance of a given process $(\cdot)$ by $\tilde{c}_{(\cdot)}(\tau)$ at lag τ, which is an estimator of the population autocovariance $c_{(\cdot)}(\tau)$. Similarly, we consider the sample mean $\tilde{\mu}_{(\cdot)}$ to estimate the population mean $\mu_{(\cdot)}$.

We analyze different aspects of the impact of finite sample sizes on the observations. First, we analyze a measure we denote observation range. Here, we focus on the observability of the autocovariance of sampled traffic. This implies that the variability that is introduced by the limited sample size should not conceal the target autocovariance. Second, we analyze the estimation accuracy. Hence, we analyze the impact of the finite sample size on the autocovariance of the geometric sampling process and consequently on the observed autocovariance. We construct a metric that captures the relative estimation error and show the impact of sampling parameters on this metric. Third, we analyze the bias of the used statistical autocovariance estimator when fed with finite sample sizes. We show that the deployed estimator is asymptotically unbiased and that the absolute bias depends on H.

OBSERVATION RANGE. In a first consideration, we assume no deviations of the sample statistics from the population statistics. This assumption will be relaxed in the following considerations. First, we characterize the impact of the variability that is introduced through the finite sample size. It is known that finite length realizations of iid sequences yield non-zero autocovariance values for lags $\tau > 0$ [13]. Remember that theoretically (for infinite sequences) the autocovariance of iid sequences equals



(a) Sampling (b) Schematic description

Figure 4.3: Distorted observations due to finite sampling.

zero for positive lags. The autocovariance values obtained for finite sample sizes possess a Gaussian distribution around the expected value of the autocovariance [13]. To characterize the impact of the finite sample size we calculate a range of lags where the autocovariance of the finitely observed LRD process is significantly different from the autocovariance observed for finite sampled iid sequences. Beyond this range of lags we consider the observed autocovariance to be strongly distorted by the variability introduced through the finite sample size. For this comparison we use a standard technique from [8, 13].

From (1.1) and (4.6) it is obvious that the autocovariance $c_Z(\tau)$ of the geometrically sampled LRD traffic process is a slowly decaying function of the lag τ. We determine a range of lags $\tau \in [0, \tau^*]$ denoted *observation range* where $\tilde{c}_Z(\tau)$ is not significantly distorted. We define τ^* as the intersection of $c_Z(\tau)$ from (4.6) and the 0.95 confidence interval for geometrically sampled iid sequences of finite size T with mean μ_Y and variance σ_Y^2. Prior to τ^* we reject the hypothesis that there is no correlation at the significance level 0.95. In other words, the observed autocovariance is significantly different from the autocovariance of sampled iid sequences. This range ensures that the variability due to the finite sample size does not conceal the target traffic autocovariance.

In Figure 4.3a we exemplarily depict a sample autocovariance of the traffic process $\tilde{c}_Y(\tau)$ and the corresponding autocovariance $\tilde{c}_Z(\tau)$ of the geometrically sampled observations. It is obvious that $\tilde{c}_Z(\tau)$ is distorted for increasing lags τ.

Figure 4.3b schematically depicts the previously explained observation range τ^* as intersection of $c_Z(\tau)$ and the 0.95 confidence interval for geometrically sampled finite Gaussian iid sequences (shaded), that is calculated in the sequel. Remember that we are interested in the correlations prior to τ^*.

We consider a sample path $z(t)$ of the observation process $Z(t)$ from geometrically sampling a Gaussian iid sequence $Y(t)$ as described by (4.2). The mean and variance of $Y(t)$ are given by μ_Y and σ_Y^2. The sample size is T such that $t \in [1, T]$. The mean of the observations is given by $\mu_Z = \mu_X \mu_Y$ with sampling intensity μ_X. The variance of the observations is given by the independence of $X(t)$ and $Y(t)$ as

$$\sigma_Z^2 = \sigma_X^2 \mu_Y^2 + \sigma_Y^2 \mu_X^2 + \sigma_X^2 \sigma_Y^2, \tag{4.20}$$

and for geometric sampling as

$$\sigma_Z^2 = \sigma_X^2 \mu_Y^2 + \sigma_Y^2 \mu_X. \tag{4.21}$$

since for geometric sampling it holds $\sigma_X^2 + \mu_X^2 = \mu_X$. We consider the following unbiased estimator of the observed autocovariance $\tilde{c}_Z(\tau)$

$$\tilde{c}_Z(\tau) = \frac{1}{T-\tau} \sum_{t=1}^{T-\tau} \left(z(t) - \mu_Z\right) \left(z(t+\tau) - \mu_Z\right). \tag{4.22}$$

We expand the product as $z(t)z(t+\tau) - z(t)\mu_Z - z(t+\tau)\mu_Z + \mu_Z^2$ and apply the central limit theorem to approximate the distribution of the individual terms for large $T - \tau$. We sketch the derivation for the first term only as the rest follows straightforwardly. Since $z(t)$ and $z(t+\tau)$ are independent we calculate the expected value of the product as μ_Z^2 and the variance of the product as $\sigma_Z^4 + 2\mu_Z^2\sigma_Z^2$. Applying the CLT to the sum $\frac{1}{T-\tau}\sum_{t=1}^{T-\tau} z(t)z(t+\tau)$ we find that it is approximately normal distributed with mean μ_Z^2 and variance $\frac{\sigma_Z^4 + 2\mu_Z^2\sigma_Z^2}{T-\tau}$. We apply the CLT to the remaining terms to find

$$\begin{aligned} \tilde{c}_Z(\tau) &\approx \mathcal{N}\left(\mu_Z^2, \tfrac{\sigma_Z^4 + 2\mu_Z^2\sigma_Z^2}{T-\tau}\right) - \mathcal{N}\left(2\mu_Z^2, \tfrac{2\mu_Z^2\sigma_Z^2}{T-\tau}\right) + \mu_Z^2 \\ &= \mathcal{N}\left(0, \tfrac{\sigma_Z^4 + 4\mu_Z^2\sigma_Z^2}{T-\tau}\right), \end{aligned} \tag{4.23}$$

where we use the notion $\mathcal{N}(\mu, \sigma^2)$ for normal random variables with mean μ and variance σ^2. The 0.95 confidence interval for $\tilde{c}_Z(\tau)$ is approximated by $\pm 2\sigma_Z \sqrt{\sigma_Z^2 + 4\mu_Z^2} / \sqrt{(T-\tau)}$.[3] Plugging (4.20) into (4.23) we obtain a general expression for the 0.95 confidence interval of $\tilde{c}_Z(\tau)$. For geometric sampling processes we substitute (4.21) into (4.23) and assume $T \gg \tau$ to find the confidence interval $\pm 2\sqrt{(\mu_X(\mu_Y^2+\sigma_Y^2)-\mu_X^2\mu_Y^2)^2 + 4\mu_X^2\mu_Y^2(\mu_X(\mu_Y^2+\sigma_Y^2)-\mu_X^2\mu_Y^2)} / \sqrt{T}$, where we substituted $\sigma_X^2 = \mu_X - \mu_X^2$. Recall that only the observed autocovariance that is larger than the given confidence interval is significant.

We calculate the observation range τ^* for LRD traffic with covariance $c_Y(\tau) = K\sigma_Y^2\tau^{2H-2}$, with some positive constant K, as the intersection of $c_Z(\tau) = c_Y(\tau)\mu_X^2$ and the positive border of the confidence interval given above. We find that

$$K\sigma_Y^2\tau^{2H-2}\mu_X^2 = 2\sqrt{(\mu_X(\mu_Y^2+\sigma_Y^2)-\mu_X^2\mu_Y^2)^2 + 4\mu_X^2\mu_Y^2(\mu_X(\mu_Y^2+\sigma_Y^2)-\mu_X^2\mu_Y^2)} / \sqrt{T}$$

$$\Downarrow$$

$$\tau^* = \left[\frac{K\sigma_Y^2\mu_X^2\sqrt{T}}{2\sqrt{(\mu_X(\mu_Y^2+\sigma_Y^2)-\mu_X^2\mu_Y^2)^2 + 4\mu_X^2\mu_Y^2(\mu_X(\mu_Y^2+\sigma_Y^2)-\mu_X^2\mu_Y^2)}} \right]^{\frac{1}{2-2H}} . \quad (4.24)$$

The metric τ^* expresses the impact of the sampling parameters on the observability of the traffic autocovariance after sampling. From τ^* we can infer that higher H, i.e., stronger LRD, delivers a higher observation range. The observation range goes to infinity $\tau^* \to \infty$ for an infinite sample size $T \to \infty$.

ESTIMATION ACCURACY. Next, we evaluate the impact of limited sample sizes on the estimation accuracy. First, we investigate the sample autocovariance $\tilde{c}_X(\tau)$ for limited sample sizes before analyzing the impact of the limited sample size on $\tilde{c}_Z(\tau)$ of the observations. Finally, we evaluate the impact of the finite sample size on the estimates of the traffic autocovariance obtained from $\tilde{c}_Z(\tau)$. In the sequel we restrict the deviation of sample statistics from population statistics to the deviation in $\tilde{c}_X(\tau)$ from $c_X(\tau)$, i.e., we assume $\tilde{c}_Y(\tau) = c_Y(\tau)$, $\tilde{\mu}_X = \mu_X$ and $\tilde{\mu}_Y = \mu_Y$.

We derive an approximation of the distribution of the sample autocovariance $\tilde{c}_X(\tau)$ using the CLT for large $T - \tau$. Consider a sample path $x(t)$ of the geometric sampling

3 For ease of notation we use 2 times the standard deviation instead of the known rule 1.967 times the standard deviation [97].

process $X(t)$ with known intensity μ_X. Next, we employ the following unbiased estimator to obtain the autocovariance $\tilde{c}_X(\tau)$

$$\tilde{c}_X(\tau) = \frac{1}{T-\tau} \sum_{t=1}^{T-\tau} \left(x(t) - \mu_X\right)\left(x(t+\tau) - \mu_X\right). \tag{4.25}$$

We apply the same steps used for the derivation of (4.23). After expanding the product in (4.25) we apply the CLT for large $T - \tau$ to approximate the distribution of the individual terms such that

$$\begin{aligned}\tilde{c}_X(\tau) &\approx \mathcal{N}\left(\mu_X^2, \tfrac{\sigma_X^4 + 2\mu_X^2\sigma_X^2}{T-\tau}\right) - \mathcal{N}\left(2\mu_X^2, \tfrac{2\mu_X^2\sigma_X^2}{T-\tau}\right) + \mu_X^2 \\ &= \mathcal{N}\left(0, \tfrac{\sigma_X^4 + 4\mu_X^2\sigma_X^2}{T-\tau}\right).\end{aligned}$$

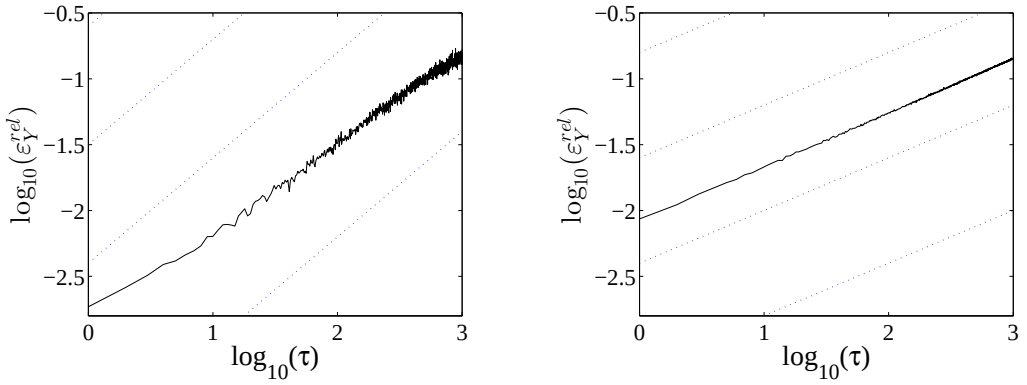
We use this result to calculate the 0.95 confidence interval for the distribution of the sample autocovariance $\tilde{c}_X(\tau)$, that is denoted c_X^{95}, as $\pm 2\mu_X\sqrt{1 + 2\mu_X - 3\mu_X^2}/\sqrt{T}$ with $\sigma_X^2 = \mu_X - \mu_X^2$ and $T \gg \tau$.

Next, we use c_X^{95} to investigate the impact of the limited sample size T on the observed autocovariance $\tilde{c}_Z(\tau)$. Remember that we consider only the variability that stems from the finite length sampling process $X(t)$ and assume $\tilde{c}_Y(\tau) = c_Y(\tau)$, $\tilde{\mu}_X = \mu_X$ and $\tilde{\mu}_Y = \mu_Y$. Given these assumptions and the formulation in Lemma 4.1 we calculate the confidence interval for $\tilde{c}_Z(\tau)$ as $c_Z^{95}(\tau) = c_X^{95}(c_Y(\tau) + \mu_Y^2)$.

We consider $\tilde{c}_Z(\tau)/\mu_X^2$ as estimator for $c_Y(\tau)$ and deduce the confidence interval $c_Y^{95}(\tau)$ for this estimator as $c_Y^{95}(\tau) = c_X^{95}(c_Y(\tau) + \mu_Y^2)/\mu_X^2$. To assess the impact of the variability due to the limited size of the sampling process $X(t)$ on the traffic autocovariance estimator we define the relative estimation error $\epsilon_Y^{rel}(\tau) = \frac{|c_Y^{95}(\tau)|}{c_Y(\tau)}$. We calculate $\epsilon_Y^{rel}(\tau)$ as

$$\begin{aligned}\epsilon_Y^{rel}(\tau) &= \frac{|c_Y^{95}(\tau)|}{c_Y(\tau)} \\ &= \frac{|c_X^{95}|}{\mu_X^2}\left(1 + \frac{\mu_Y^2}{c_Y(\tau)}\right) \\ &\approx \frac{2\sqrt{1 + 2\mu_X - 3\mu_X^2}}{\sqrt{T}\mu_X}\left(1 + \frac{\mu_Y^2}{c_Y(\tau)}\right).\end{aligned} \tag{4.26}$$

The prefactor in (4.26) is only dependent on the sampling parameters, i.e., the sampling intensity μ_X and the sample size T. Optimizing the sampling parameters



(a) $H = 0.6$ (b) $H = 0.8$

Figure 4.4: Estimation error under finite sampling depends on H.

aims at reducing the prefactor of the relative estimation error. The second term depends solely on the traffic properties, i.e., its intensity μ_Y and its autocovariance $c_Y(\tau)$. Given that the traffic autocovariance $c_Y(\tau)$ decreases naturally with lag τ the relative estimation error increases with τ. From (4.26) we observe that the relative estimation error introduced through $\tilde{c}_X(\tau)$ decays with increasing sample size T or with increasing sampling intensity μ_X. For fixed τ and practical sampling intensities, i.e., $\mu_X \leq 0.1$, we find a nonlinear trade-off between sampling intensity μ_X and sampling size T. The prefactor in (4.26) behaves as $2/\sqrt{T}\mu_X$. From this result we conclude that given finite sample sizes, sampling intensity has a stronger impact on accuracy than sample size.

Next, we investigate the impact of the Hurst parameter $H \in (0.5, 1)$ on the relative estimation error $\epsilon_Y^{rel}(\tau)$. To this end, we consider (4.26) for increasing lags τ. As τ increases $c_Y(\tau)$ decays leading to $\mu_Y^2/c_Y(\tau) \gg 1$ in (4.26). Then, the relative estimation error $\epsilon_Y^{rel}(\tau)$ becomes

$$\begin{aligned} \epsilon_Y^{rel}(\tau) &\approx \frac{2\sqrt{1+2\mu_X-3\mu_X^2}}{\sqrt{T}\mu_X}\left(\frac{\mu_Y^2}{c_Y(\tau)}\right) \\ &\approx \frac{2\sqrt{1+2\mu_X-3\mu_X^2}\mu_Y^2}{\sqrt{T}\mu_X\sigma_Y^2}\tau^{2-2H}, \end{aligned} \tag{4.27}$$

where we substituted $\sigma_Y^2\tau^{2H-2}$ for $c_Y(\tau)$. From (4.27) it is obvious that the relative estimation error increases with the considered lag τ. For LRD traffic with high H, $\epsilon_Y^{rel}(\tau)$ increases slower with τ compared to traffic with low H.

Figure 4.4 depicts the increase of $\epsilon_Y^{rel}(\tau)$ with τ for two different H values.[4] The figure includes auxiliary lines with slope $2-2H$ to show that the estimation error closely matches the behavior described by (4.27).

Finally, we deduce a sampling dimensioning rule using (4.27) to infer the needed sample size T in order to retain fixed estimation error $\epsilon_Y^{rel}(\tau)$. We find from (4.27) that the sample size has to increase by $T \simeq \max\{\tau^{4-4H}, \tau\}$ to achieve constant $\epsilon_Y^{rel}(\tau)$ for increasing τ. The term τ in the maximum stems from the need to increase T at least by τ to retain the same number of points in the autocovariance calculation. Note that for $H \in (0.5, 0.75)$ the sample size T must increase faster than linearly with τ to achieve constant $\epsilon_Y^{rel}(\tau)$.

BIAS OF AUTOCOVARIANCE ESTIMATOR. Next, we investigate the impact of finite sample sizes on the bias of the deployed statistical autocovariance estimators. We consider the cases of directly observing the entire traffic process $Y(t)$ as well as after sampling. We show that the deployed estimator is asymptotically unbiased as the sample size T tends to infinity. First, we consider the direct observation of a sample path $y(t)$ of the traffic process of finite length T. We consider the autocovariance estimator

$$\tilde{c}_Y(\tau) = \frac{1}{T-\tau} \sum_{t=1}^{T-\tau} \left(y(t) - \tilde{\mu}_{Y_0}\right)\left(y(t+\tau) - \tilde{\mu}_{Y_\tau}\right) \tag{4.28}$$

with means $\tilde{\mu}_{Y_0} = \frac{1}{(T-\tau)} \sum_{t=1}^{T-\tau} y(t)$ and $\tilde{\mu}_{Y_\tau} = \frac{1}{(T-\tau)} \sum_{t=1}^{T-\tau} y(t+\tau)$. The means $\tilde{\mu}_{Y_0}$ and $\tilde{\mu}_{Y_\tau}$ are sample means over the *first* and *last* $T-\tau$ points of the sample, respectively. Any autocovariance estimator (4.28) is unbiased iff $\mathsf{E}\left[\tilde{c}_Y(\tau)\right] = c_Y(\tau)$ [[97] p. 327].

Theorem 4.1 (Bias of the Autocovariance Estimator for the Traffic Process $Y(t)$) *Given a sample path $y(t)$ of length T of the LRD traffic increment process $Y(t)$ characterized by σ_Y^2 and $H \in (0.5, 1)$. The expected value of the autocovariance estimator* (4.28) *is approximated by*

$$\mathsf{E}\left[\tilde{c}_Y(\tau)\right] \approx c_Y(\tau) - \frac{\sigma_Y^2}{(T-\tau)^{2-2H}}.$$

4 We used 100 synthetic LRD traffic traces each of length $T = 2 \times 10^8$ time slots for the simulations.

From Theorem 4.1 we conclude that the autocovariance estimator (4.28) is asymptotically unbiased for $T \to \infty$ and $T \gg \tau$. The impact of H is apparent as higher H requires larger sample sizes T to achieve negligible bias.

Proof of Theorem 4.1. Given a sample path $y(t)$ with sample means $\tilde{\mu}_{Y_0} = \frac{1}{(T-\tau)} \sum_{t=1}^{T-\tau} y(t)$ and $\tilde{\mu}_{Y_\tau} = \frac{1}{(T-\tau)} \sum_{t=1}^{T-\tau} y(t+\tau)$. Fix σ_Y^2, $H \in (0.5, 1)$ and T. Consider the autocovariance estimator (4.28) and expand the product as $y(t)y(t+\tau) - y(t)\tilde{\mu}_{Y_\tau} - y(t+\tau)\tilde{\mu}_{Y_0} + \tilde{\mu}_{Y_0}\tilde{\mu}_{Y_\tau}$. We compute the expected values for the individual terms as follows

$$\mathsf{E}\left[\frac{1}{T-\tau}\sum_{t=1}^{T-\tau} y(t)y(t+\tau)\right] = c_Y(\tau) + \mu_Y^2,$$

where $c_Y(\tau)$ and μ_Y are the population parameters. Exchanging the expected value and the sum above directly yields the result. We also calculate

$$\mathsf{E}\left[\frac{1}{T-\tau}\sum_{t=1}^{T-\tau} y(t)\tilde{\mu}_{Y_\tau}\right] = \mathsf{E}[\tilde{\mu}_{Y_0}\tilde{\mu}_{Y_\tau}],$$

where we used that $\frac{1}{T-\tau}\sum_{t=1}^{T-\tau} y(t) = \tilde{\mu}_{Y_0}$. The same argument applies for the product $y(t+\tau)\tilde{\mu}_{Y_0}$. We rephrase and estimate the last term as

$$\begin{aligned} \mathsf{E}\left[\tilde{\mu}_{Y_0}\tilde{\mu}_{Y_\tau}\right] &= \mathrm{Cov}\left[\tilde{\mu}_{Y_0}, \tilde{\mu}_{Y_\tau}\right] + \mathsf{E}\left[\tilde{\mu}_{Y_0}\right]\mathsf{E}\left[\tilde{\mu}_{Y_\tau}\right] \\ &\leq \mathrm{Var}\left[\tilde{\mu}_{Y_0}\right] + \mu_Y^2, \end{aligned} \tag{4.29}$$

where we use $\mathrm{Cov}\left[(\cdot)_i, (\cdot)_j\right]$ to denote the covariance of the random variables $(\cdot)_i$ and $(\cdot)_j$. Note that $\mathsf{E}\left[\tilde{\mu}_{Y_0}\right] = \mu_Y$ and $\mathsf{E}\left[\tilde{\mu}_{Y_\tau}\right] = \mu_Y$, as these are unbiased estimators of the population mean. Observe that the sample ranges for $\tilde{\mu}_{Y_0}$ and $\tilde{\mu}_{Y_\tau}$, i.e., $[1, T-\tau]$ and $[\tau+1, T]$ overlap by $T-2\tau$. For $T \gg \tau$ we have $\mathrm{Cov}\left[\tilde{\mu}_{Y_0}, \tilde{\mu}_{Y_\tau}\right] \approx \mathrm{Var}\left[\tilde{\mu}_{Y_0}\right]$, i.e., the bound in (4.29) is good. Finally, we use from [[8] p. 54] that the variance of the mean of $T-\tau$ samples of an LRD process decays as $\sigma_Y^2/(T-\tau)^{2-2H}$ to derive

$$\mathsf{E}\left[\tilde{\mu}_{Y_0}\tilde{\mu}_{Y_\tau}\right] \approx \frac{\sigma_Y^2}{(T-\tau)^{2-2H}} + \mu_Y^2.$$

Putting all the pieces together we obtain

$$\begin{aligned} \mathsf{E}\left[\tilde{c}_Y(\tau)\right] &= c_Y(\tau) - \mathrm{Cov}[\tilde{\mu}_{Y_0}, \tilde{\mu}_{Y_\tau}] \\ &\approx c_Y(\tau) - \frac{\sigma_Y^2}{(T-\tau)^{2-2H}}, \end{aligned} \tag{4.30}$$

i.e., the estimator $\tilde{c}_Y(\tau)$ underestimates the covariance, where the bias diminishes with increasing $T - \tau$. ∎

The derivation above shows that for a traffic process $y(t)$ that is directly accessible the autocovariance estimator is asymptotically unbiased. The bias in (4.30) originates from the fact that we do not know the true traffic intensity μ_Y. We resort to the unbiased estimators $\tilde{\mu}_{Y_0}$ and $\tilde{\mu}_{Y_\tau}$ that have a variance that depends on H and decays with the sample size T. Thus, the bias diminishes with increasing sample size with a convergence speed that depends on H.

Next we consider the bias of the presented autocovariance estimator when applied to a sample path $z(t)$ of the observed process $Z(t)$ with finite size T such as

$$\tilde{c}_Z(\tau) = \frac{1}{T-\tau} \sum_{t=1}^{T-\tau} \left(z(t) - \tilde{\mu}_{Z_0}\right)\left(z(t+\tau) - \tilde{\mu}_{Z_\tau}\right). \tag{4.31}$$

The sample means $\tilde{\mu}_{Z_0}$ and $\tilde{\mu}_{Z_\tau}$ are given by $\tilde{\mu}_{Z_0} = \frac{1}{(T-\tau)} \sum_{t=1}^{T-\tau} z(t)$ and $\tilde{\mu}_{Z_\tau} = \frac{1}{(T-\tau)} \sum_{t=1}^{T-\tau} z(t+\tau)$.

Theorem 4.2 (Bias of the Autocovariance Estimator for the Observed Process $Z(t)$) *Given a sample path $z(t)$ of length T of the observed process $Z(t)$. The expected value of the autocovariance estimator* (4.31) *is approximated by*

$$\mathsf{E}[\tilde{c}_Z(\tau)] \approx c_Z(\tau) - \frac{c_Z(0)}{T-\tau} - \frac{2}{(T-\tau)^2} \sum_{t=1}^{T-\tau-1} (T-\tau-t) c_Z(t).$$

From Theorem 4.2 we conclude that the autocovariance estimator (4.31) is asymptotically unbiased for $T \to \infty$ and $T \gg \tau$.

Proof of Theorem 4.2. The proof is structured similarly to the proof of Theorem 4.1. Given a sample path $z(t)$ with sample means $\tilde{\mu}_{Z_0}$ and $\tilde{\mu}_{Z_\tau}$. We calculate the expected value $\mathsf{E}[\tilde{c}_Z(\tau)]$ by expanding the product in (4.31) and computing the expected

values of the individual terms, i.e., $z(t)z(t+\tau) - z(t)\tilde{\mu}_{Z_\tau} - z(t+\tau)\tilde{\mu}_{Z_0} + \tilde{\mu}_{Z_0}\tilde{\mu}_{Z_\tau}$. We find

$$\mathsf{E}\left[\frac{1}{T-\tau}\sum_{t=1}^{T-\tau} z(t)z(t+\tau)\right] = c_Z(\tau) + \mu_Z^2,$$

where $c_Z(\tau)$ and μ_Z are the population parameters and

$$\mathsf{E}\left[\frac{1}{T-\tau}\sum_{t=1}^{T-\tau} z(t)\tilde{\mu}_{Z_\tau}\right] = \mathsf{E}[\tilde{\mu}_{Z_0}\tilde{\mu}_{Z_\tau}].$$

The same argument of the second term applies for the product $z(t+\tau)\tilde{\mu}_{Z_0}$. The last term is approximated as

$$\begin{aligned}\mathsf{E}\left[\tilde{\mu}_{Z_0}\tilde{\mu}_{Z_\tau}\right] &= \mathrm{Cov}\left[\tilde{\mu}_{Z_0}, \tilde{\mu}_{Z_\tau}\right] + \mathsf{E}\left[\tilde{\mu}_{Z_0}\right]\mathsf{E}\left[\tilde{\mu}_{Z_\tau}\right] \\ &\leq \mathrm{Var}\left[\tilde{\mu}_{Z_0}\right] + \mu_Z^2.\end{aligned}$$

We obtain the expected value of the covariance estimator as

$$\mathsf{E}\left[\tilde{c}_Z(\tau)\right] = c_Z(\tau) - \mathrm{Cov}[\tilde{\mu}_{Z_0}, \tilde{\mu}_{Z_\tau}],$$

where $c_Z(\tau)$ is the population parameter. The sample ranges for $\tilde{\mu}_{Z_0}$ and $\tilde{\mu}_{Z_\tau}$, i.e., $[1, T-\tau]$ and $[\tau+1, T]$ respectively, overlap by $T-2\tau$. For $T \gg \tau$ we estimate $\mathrm{Cov}[\tilde{\mu}_{Z_0}, \tilde{\mu}_{Z_\tau}] \approx \mathrm{Var}[\tilde{\mu}_{Z_0}]$. We invoke the identity from [[113] p. 54] that

$$\mathrm{Var}\left[\sum_{i=1}^{k} N_i\right] = \sum_{i=1}^{k}\sum_{j=1}^{k}\mathrm{Cov}\left[N_i, N_j\right],$$

where $\mathrm{Cov}\left[N_i, N_j\right]$ denotes the covariance of the random variables N_i and N_j for $i, j \in [1, k]$. We rearrange $\mathrm{Var}\left[\tilde{\mu}_{Z_0}\right]$ as follows

$$\begin{aligned}\mathrm{Var}\left[\tilde{\mu}_{Z_0}\right] &= \mathrm{Var}\left[\tfrac{1}{(T-\tau)}\sum_{t=1}^{T-\tau} z(t)\right] \\ &= \frac{1}{(T-\tau)^2}\sum_{i=1}^{T-\tau}\sum_{j=1}^{T-\tau}\mathrm{Cov}[z(i), z(j)] \\ &= \frac{c_Z(0)}{T-\tau} + \frac{2}{(T-\tau)^2}\sum_{t=1}^{T-\tau-1}(T-\tau-t)c_Z(t).\end{aligned}$$

In the last line we rearranged the double sum and used the notion $\mathrm{Cov}[z(t), z(t+\tau)] = c_Z(\tau)$. The expected value of the sample covariance follows as

$$\mathsf{E}[\tilde{c}_Z(\tau)] \approx c_Z(\tau) - \frac{c_Z(0)}{T-\tau} - \frac{2}{(T-\tau)^2} \sum_{t=1}^{T-\tau-1} (T-\tau-t) c_Z(t).$$

∎

Under geometric sampling described by (4.6), the estimator $\tilde{c}_Z(\tau)$ is asymptotically unbiased since for $T \gg \tau$ the following limits exist

$$\begin{aligned} &\lim_{T\to\infty} \frac{c_Z(0)}{T-\tau} = 0, \\ &\lim_{T\to\infty} \frac{2}{(T-\tau)^2} \sum_{t=1}^{T-\tau-1} (T-\tau-t) c_Z(t) = 0. \end{aligned} \tag{4.32}$$

To obtain the second limit we use (4.6) to substitute for $c_Z(\tau)$ in (4.32). We upper bound the term on the left hand side of (4.32) using

$$\frac{2\sigma_Y^2 \mu_X^2}{(T-\tau)} \sum_{t=1}^{T-\tau-1} t^{2H-2},$$

where we omitted the subtraction of t in the brackets in (4.32) and inserted $\sigma_Y^2 t^{2H-2}$ for $c_Y(t)$. As t^{2H-2} is a decreasing function in t, with $H \in (0.5, 1)$, we upper bound the sum $\sum_{t=1}^{T-\tau-1} t^{2H-2}$ by the integral $\int_{t=0}^{T-\tau-1} t^{2H-2} dt = \frac{(T-\tau-1)^{2H-1}}{2H-1}$. Thus, we find the limit

$$\lim_{T\to\infty} \frac{2\sigma_Y^2 \mu_X^2 (T-\tau-1)^{2H-1}}{(2H-1)(T-\tau)} = 0, \tag{4.33}$$

yielding that the bias goes to zero for $T \to \infty$.

In general we find from inserting Lemma 4.1 into (4.32) that sampling processes with autocovariances that decay to zero, i.e., satisfying a weakly-mixing condition as defined in [30, 117], deliver asymptotically unbiased estimates using the autocovariance estimator (4.31).

Examples for sampling processes that yield asymptotically unbiased results are geometric, Gamma and uniform sampling. For periodic sampling, inserting $c_Z(\tau)$

from (4.8) into Theorem 4.2 yields that the autocovariance estimator (4.31) is biased for $T \to \infty$. It is obvious that periodic sampling clearly violates the mixing condition.

In this section we showed the impact of finite sample sizes and sampling parameters on the accuracy of the estimated traffic autocovariance. We capitalized on geometric sampling due to its advantages that are explored in Section 4.1. First, we analyzed the observability of the autocovariance of sampled traffic to find a range of lags where the variability which is introduced through the limited sample size does not conceal the target autocovariance. Second, we quantified the impact of the finite sample size on the autocovariance of the geometric sampling process and constructed a metric that describes the arising relative estimation error of the traffic autocovariance. One main finding at this point is that the estimation error increases with autocovariance lag τ as τ^{2-2H}. Concerning the impact of the sampling parameters on the relative estimation error we discovered a nonlinear trade-off between the sampling intensity μ_X and the sample size T. Lastly, we analyzed the impact of the finite sample size on the deployed statistical autocovariance estimators which are shown to be asymptotically unbiased with an absolute bias that depends on the traffic Hurst parameter H.

4.3 PROBING METHODOLOGY AND INTERNET MEASUREMENT CAMPAIGN

In Sections 4.1 and 4.2 we inferred characteristics of LRD traffic from observed samples and investigated the accuracy of the deployed statistical estimators under finite sample sizes. Results from Sections 4.1 and 4.2 can be directly applied in monitoring scenarios, i.e., if traffic is directly accessible. In the sequel we investigate the problem of inferring LRD traffic characteristics without direct monitoring capability. Equipped with the derivations from the previous sections we present a method for inferring LRD traffic characteristics from the interaction of injected packet probes with the traffic. We presented our method in [112]. In contrast to traffic traces that reflect the characteristics at a *single* observation point, i.e., a single node, we extend the view to infer dominant characteristics over entire paths.

Network Probing Methodology for Inferring LRD Traffic Characteristics

We consider a single node scenario with a basic queueing model as shown in Figure 2.3. The cross traffic is LRD whereas the through traffic is the probing traffic flow. The proposed probing technique uses the delays of single packet probes to detect the busy periods at the router output link. Assuming non-preemptive FIFO packet scheduling, a packet probe would experience the minimum one way delay d_{min} only if the router is idle. If the router is busy with transmitting cross traffic, the probing packet will have to wait for its turn. Hence, if there is concurrent cross traffic the delay d_p experienced by a probe packet would be larger than d_{min}.

The measured one way delay is given by $d_p = t_r - t_s$, where t_r and t_s denote the receive and send times respectively. Each probe yields a sample of the egress link state, that is either busy or not. We denote the process that describes the egress link state as $Y(t)$. The values of $Y(t) \in \{0, 1\}$ denote the states (i) outgoing link not busy, i.e., $Y(t) = 0$ and (ii) busy, i.e., $Y(t) = 1$, respectively, at time t. The process $Y(t)$ captures the utilization of the node egress. The observed process $Z(t)$ arises by sampling of the egress link state process $Y(t)$ using the sampling process $X(t)$. The sampling process $X(t) \in \{0, 1\}$ takes the value one when a probe is sent, i.e., a sample is taken, and zero otherwise. The time between two probes follows a distribution as discussed in Section 4.1, e.g., $X(t)$ is a geometric sampling process. If a probe encounters a busy outgoing link, i.e., $d_p > d_{min}$, we set the observed process $Z(t)$ to one, otherwise $Z(t)$ is set to zero. We construct the observed process according to (4.2) as follows

$$Z(t) = \begin{cases} 1 & \text{if } d_p > d_{min} \text{ and } X(t) = 1 \\ 0 & \text{otherwise.} \end{cases} \tag{4.34}$$

From [44, 49] it is known that the autocovariance decay of LRD traffic is preserved at the output of a queue or traffic shaper, i.e., the process $Y(t)$ has the same autocovariance decay from (1.1) as the LRD traffic input. From (4.6) we deduce that for the geometric sampling process $X(t)$ the observed process $Z(t)$ has the

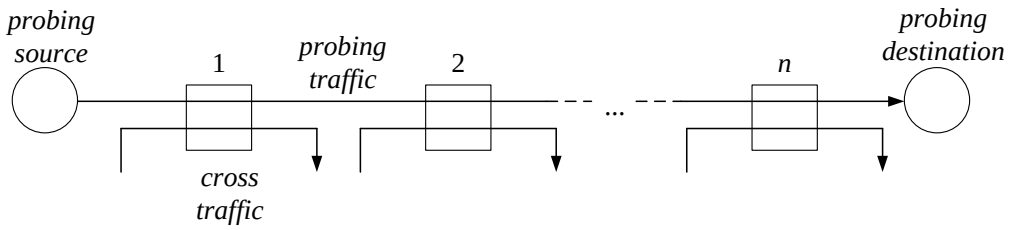


Figure 4.5: n node topology with probing traffic and LRD cross traffic at nodes $i \in [1, n]$.

same autocovariance as $Y(t)$ except for a multiplicative rescaling factor. Thus, the observed process has the same autocovariance decay as the LRD traffic input.

Next, we extend the previous probing method to cover an end-to-end path with n nodes each having fresh multiplexed/demultiplexed LRD traffic as depicted in Figure 4.5. We consider the observations at the egress of the path and show that these observations also exhibit LRD. In particular, if different cross traffic flows at nodes $i \in [1, n]$ possess different Hurst parameter values $H_i \in (0.5, 1)$, we show that the largest H_i dominates the path observations. This complements a result on end-to-end performance domination by the maximum Hurst parameter value along a path that is presented in Chapter 6. A similar result on maximum Hurst parameter domination in multiplexing scenarios is derived in [49].

Consider the n node path depicted in Figure 4.5 with independent LRD cross traffic at each node. The busy state process $Y_i(t) \in \{0, 1\}$ at the egress of node i has the autocovariance given by (1.1) with Hurst parameter value H_i. Consider the end-to-end delay of a probe sent over the n node path after subtracting the minimum delay. The (excess) delay will only be positive if *any* of the routers on the path was busy when the probe arrives at its ingress, else the excess delay equals zero. This relation can be expressed as logical OR operation of the processes $Y_i(t)$ characterizing the busy states at the egress of the routers $i \in [1, n]$. For a transmitted probe at time t, i.e., $X(t) = 1$, we can formulate the OR operation using $Y_i(t) \in \{0, 1\}$ to find $Z_i(t)$ at the egress of node i as

$$Z_i(t) = \begin{cases} Y_1(t) & \text{for } i = 1 \\ Z_{i-1}(t) + Y_i(t) - Z_{i-1}(t)Y_i(t) & \text{for } i \in [2, n]. \end{cases} \tag{4.35}$$

For ease of exposition we assume in (4.35) that an idle node forwards probe packets instantaneously to the next node. Hence, a probe packet at time t observes $Y_i(t)$ for

all $i \in [1, n]$. We may relax this assumption at the cost of extra notation. We then formulate (4.35) for probes that observe $Y_i(t)$ at time $t_i \geq t_{i-1}$ for $i \in [2, n]$.

Next, we use a two node example calculation to motivate the derivation of (4.35). The egress of nodes 1 and 2 is characterized by $Y_1(t)$ and $Y_2(t)$, respectively. A packet probe at time t, i.e., $X(t) = 1$, that observes a busy path with $d_p(t) > d_{min}$, hence $Z_2(t) = 1$, implies that $Y_1(t) = 1$ (logically) OR $Y_2(t) = 1$. This translates to the consideration that the excess delay comes from the first OR (logically) from the second node. Otherwise, no excess delay is measured and $Z_2(t) = 0$. Since $Y_1(t)$ and $Y_2(t) \in \{0, 1\}$ we can simply formulate the OR operation as

$$Z_2(t) = Y_1(t) + Y_2(t) - Y_1(t)Y_2(t). \tag{4.36}$$

Recall that for the single node case we find that for a transmitted probe at time t, i.e., $X(t) = 1$ it holds that $Z_1(t) = Y_1(t)$.

Next, we derive the autocovariance $c_{Z_2}(\tau)$ as a function of $c_{Y_i}(\tau)$ and μ_{Y_i} for $i \in \{1, 2\}$. First, we express the autocovariance $c_{Z_2}(\tau)$ as

$$c_{Z_2}(\tau) = \mathsf{E}\left[Z_2(t)Z_2(t+\tau)\right] - \mu_{Z_2}^2. \tag{4.37}$$

Then, we calculate the expected value of Z_2 from (4.36) using the independence of $Y_1(t)$ and $Y_2(t)$ as $\mu_{Z_2} = \mu_{Y_1} + \mu_{Y_2} - \mu_{Y_1}\mu_{Y_2}$ where $\mu_{Y_i} = \mathsf{E}\left[Y_i\right]$ for $i \in [1, 2]$. We insert (4.36) into the autocorrelation expression $\mathsf{E}\left[Z_2(t)Z_2(t+\tau)\right]$ to find

$$\begin{aligned}\mathsf{E}\left[Z_2(t)Z_2(t+\tau)\right] = {} & c_{Y_1}(\tau) + \mu_{Y_1}^2 + c_{Y_2}(\tau) + \mu_{Y_2}^2 + 2\mu_{Y_1}\mu_{Y_2} \\ & - 2\left(\mu_{Y_2}\left(c_{Y_1}(\tau) + \mu_{Y_1}^2\right)\right) - 2\left(\mu_{Y_1}\left(c_{Y_2}(\tau) + \mu_{Y_2}^2\right)\right) \\ & + \left(c_{Y_1}(\tau) + \mu_{Y_1}^2\right)\left(c_{Y_2}(\tau) + \mu_{Y_2}^2\right),\end{aligned} \tag{4.38}$$

where we first expanded the terms, then applied the linearity of the expectation operator, and used the independence of $Y_1(t)$ and $Y_2(t)$. We insert (4.38) and μ_{Z_2} from above into (4.37) to find

$$c_{Z_2}(\tau) = c_{Y_1}(\tau)\left(1 - \mu_{Y_2}\right)^2 + c_{Y_2}(\tau)\left(1 - \mu_{Y_1}\right)^2 + c_{Y_1}(\tau)c_{Y_2}(\tau). \tag{4.39}$$

From the autocovariance decay described by (1.1) with different $H_i \in (0.5, 1)$ for $i \in [1, 2]$ we can directly show that the autocovariance at the egress of the second node $c_{Z_2}(\tau)$ is governed by the slower decaying autocovariance $c_{Y_i}(\tau)$, i.e., the larger Hurst parameter H_i for $i \in [1, 2]$.

The two node example described in (4.36) is readily extended to cover n tandem nodes as in Figure 4.5 using the formulation of $Z_n(t)$ as the result of the OR operation of $Y_i(t)$ for $i \in [1, n]$. For example, for $n = 3$ we find $Z_3(t)$ as the result of the OR operation of $Y_1(t)$, $Y_2(t)$ and $Y_3(t)$. Equivalently, $Z_3(t)$ is the result of the OR operation of $Z_2(t)$ and $Y_3(t)$. Thus, we can iteratively formulate $Z_i(t)$ for node $i \in [2, n]$ as the result of the OR operation of the busy state process $Y_i(t)$ at node i and $Z_{i-1}(t)$, i.e., the outcome of the OR operations up to the previous node $i - 1$ as given in (4.35).

Repeating the steps (4.36) to (4.39) for the recursive formulation in (4.35) yields

$$c_{Z_i}(\tau) = c_{Z_{i-1}}(\tau)c_{Y_i}(\tau) + c_{Z_{i-1}}(\tau)(1 - \mu_{Y_i})^2 + c_{Y_i}(\tau)(1 - \mu_{Z_{i-1}})^2. \tag{4.40}$$

Through recursive insertion and reordering we reformulate (4.40) for an n node path as

$$c_{Z_n}(\tau) = \sum_{i=1}^{n} c_{Y_i}(\tau)K_i + \mathcal{O}\left(\left(\tau^{2\max_i\{H_i\}-2}\right)^2\right), \tag{4.41}$$

with constants K_i for $i \in [1, n]$. For the n-th node we find that the second term in (4.40) keeps the sum of autocovariances of the past $n - 1$ iterations as $\sum_i^{n-1} K_i c_{Y_i}(\tau)$ with different prefactors K_i, while the last term adds the current autocovariance $K_n c_{Y_n}(\tau)$ with the prefactor K_n. We find that the first term of (4.40) is of higher order, i.e., it involves the multiplication of at least two autocovariance functions $c_{Y_i}(\tau)c_{Y_j}(\tau)$ for $i, j \in [1, n]$ which leads to a faster decay in τ. This is captured by the last term in (4.41). Next, we show an example calculation for (4.41) for $n = 3$.

We insert (4.39) into (4.40) to find the following autocovariance at the egress of the third node

$$\begin{aligned} c_{Z_3}(\tau) &= k_1 c_{Y_1}(\tau) + k_2 c_{Y_2}(\tau) + k_3 c_{Y_3}(\tau) \\ &\quad + k_4 c_{Y_1}(\tau) c_{Y_3}(\tau) + k_5 c_{Y_1}(\tau) c_{Y_2}(\tau) + k_6 c_{Y_2}(\tau) c_{Y_3}(\tau) \\ &\quad + k_7 c_{Y_1}(\tau) c_{Y_2}(\tau) c_{Y_3}(\tau), \end{aligned} \tag{4.42}$$

using the constants k_i for $i \in [1, 7]$.

The expressions (4.41) and (4.40) reveal that for a given n node path the autocovariance of the end-to-end observations is governed by the strongest LRD on the path. In Chapter 6 we show that traffic with the highest H along a network path has the strongest impact on the derived non-asymptotic end-to-end performance bounds.

Experimental Testbed Evaluation

We validate the introduced probing methodology in a controlled Emulab-based testbed environment called FILab.[5] The testbed environment allows reconfigurable topologies with exclusive node use. Our testbed topology depicted in Figure 4.6 shows probe as well as cross traffic senders. The cross traffic senders S_1 and S_2 transmit synthesized LRD traffic traces with given Hurst parameter H to the receiver nodes R_1 and R_2. The traffic traces were created by superpositioning 10^5 heavy tailed on-off traffic sources. The relation between the tail index ξ of the on-off sources and the resulting Hurst parameter H is given by $H = (3 - \xi)/2$ [128]. Each sender transmits traffic with a mean rate of 50 Mbps and a constant packet size of 1500 Bytes. We set the capacity of all links to $C = 1$ Gbps.

The probe sender S_p transmits Internet Control Massage Protocol (ICMP) echo request probes to the receiver R_p and measures the round trip time (RTT) using *libpcap*. We choose round trip times for reasons of practical deployment, i.e., to circumvent clock synchronization problems at probe sender and receiver. We sub-

5 Within the framework of FILab [58] we deploy nodes with Supermicro X8DTU server mainboards with 2.2 GHz Intel E5520 Xeon processors, quad port Intel 82576EB Gigabit Ethernet Controllers, and Ubuntu 10.04 LTS with kernel 2.6.32-24, FIFO scheduling and buffers for 5000 packets.

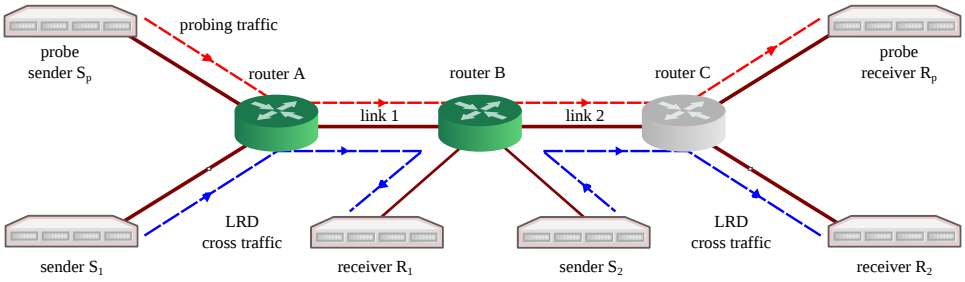


Figure 4.6: Testbed for empirical evaluation of the LRD probing methodology

Table 4.2: Probing Traffic Parameters.

distribution	parameter p	discretization slot	# of probes	mean rate
geometric	0.1	1 ms	10^6	70 kbps

stantiated first results by measuring one way delays (not shown here) using Endace DAG packet capture cards and network taps at S_p and R_p. Note that measuring RTTs corresponds to characterizing the round trip path such that the probing source and destination in Figure 4.5 are in fact located on the same physical machine. The developed probing software *H-probe* is available at [14].

We assume that dropped probes are due to a busy router, hence we set $Z(t) = 1$ for dropped samples. Further, we assume negligible perturbation due to probing as we choose the smallest probe size possible and a probing rate of less than 0.01 per mill compared to the capacity. To reduce the impact of non-queueing jitter in routers we employ a heuristic that substitutes d_{min} in (4.34) by the average $\mathsf{E}[d]$. This heuristic masks small bursts but it does not alter the long tail of the burst length distribution, which is fundamental for LRD as described in [128].

In the first experiment we consider only the single node case, i.e., S_1 transmits LRD traffic with Hurst parameter $H \in \{0.6, 0.7, 0.8, 0.9\}$ to R_1 on link 1, while S_2 idles. We repeat each experiment 25 times. We estimate the autocovariances for the traffic traces using (4.28) and for the sampling as well as probing observations using (4.31). We use an observation lag range of $\tau \in [1, 10^3]$ and measurement parameters that are given in Table 4.2.

Figure 4.7 presents inferred Hurst parameter estimates. Here, we distinguish between the following cases: (i) accessing entire traffic traces, (ii) samples from traffic traces, and (iii) probing results. The sampling results are from random

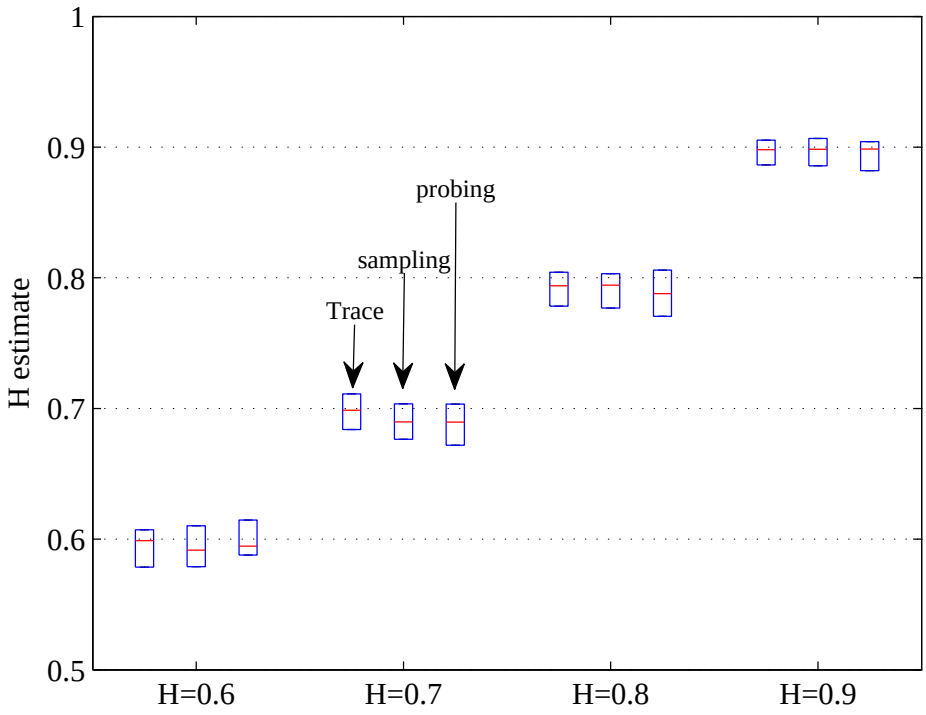


Figure 4.7: Hurst parameter estimates from direct and indirect access, i.e., (a) trace analysis, (b) trace sampling, and (c) probing in the FILab testbed.

samples as described in Section 4.1 with parameters from Table 4.2. The probing results stem from observations that are based on the interaction of probes with replayed LRD traffic traces at shared testbed nodes (router A) as described at the beginning of this section. The boxplots in Figure 4.7 show successful estimation of H with reliable accuracy using packet probes without administrative access to network internal nodes. We show corresponding autocovariance estimates in Figure 4.8.

In a second experiment we investigate the end-to-end measurement scenario. We inject LRD traffic on both links as depicted in Figure 4.6. First, we deploy LRD traffic of clearly distinguishable $H_i \in \{0.6, 0.9\}$, i.e., $\{H_1 = 0.6, H_2 = 0.9\}$ and vice versa, on the respective links $i \in \{1, 2\}$. As predicted by (4.39) end-to-end measurements given in Table 4.3 reflect the autocovariance decay with the *highest* H_i on the path. Note that we know from (4.41) that the order is not important such that the constellations $\{H_1 = 0.6, H_2 = 0.9\}$ and $\{H_1 = 0.9, H_2 = 0.6\}$ show the same end-to-end behavior, i.e., an end-to-end estimate of $H = 0.9$. For comparison we

Table 4.3: End-to-end Hurst parameter estimates in a 2 node scenario. The *highest H* along a path prevails.

	estimated H on run #				
	1	2	3	4	5
$\{H_1 = 0.6, H_2 = 0.9\}$	0.87	0.89	0.89	0.90	0.90
$\{H_1 = 0.9, H_2 = 0.6\}$	0.87	0.88	0.88	0.90	0.90
$\{H_1 = 0.6, H_2 = 0.6\}$	0.59	0.62	0.64	0.63	0.63
$\{H_1 = 0.9, H_2 = 0.9\}$	0.92	0.92	0.89	0.92	0.89

show in Table 4.3 end-to-end measurement results for paths with homogeneous LRD traffic ($H_1 = H_2$) that show congruent estimates.

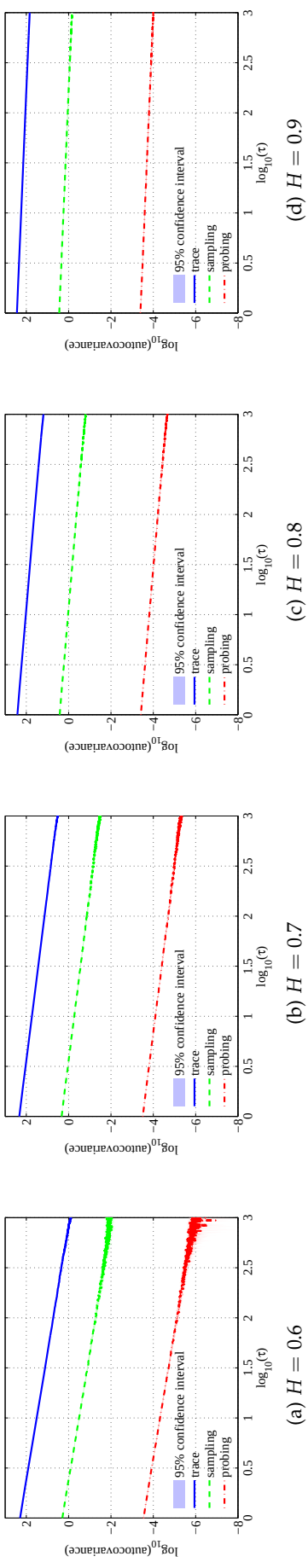
Internet Measurement Campaign

We deploy the developed LRD probing software [14] in an extensive Internet measurement study. The probes are transmitted from a university lab server in Germany [58] targeting different Internet destinations across the world. We maintain the probing parameters from Table 4.2. We set the measurement duration to 3 hours, which is a duration up to which piecewise stationarity has been monitored in [54, 116]. In addition, we deploy the Leybourne-McCabe stationarity test to filter out measured non-stationary time series [73].

In Figure 4.9 we show a snapshot of the autocovariance estimation results for PlanetLab targets that are located at different universities, specifically the universities of Pennsylvania, Virginia and Massachusetts. We observe that paths to the targets are fixed using *Traceroute*. We recognize that the autocovariance varies across different Internet paths and at different times of day. We observe typical LRD decay of the autocovariance in Figures 4.9a and 4.9b. For some target paths, e.g., Figure 4.9c, we observe periodicities on various time scales. This phenomenon was reported in offline traces in [16] and can be partially traced back to protocol specifications. We also note that the estimated Hurst parameter varies significantly with time. In Figure 4.10 we depict H estimates from continued measurements over an excerpt of 1 month. The estimated autocovariances show strong correlation, thus the Hurst

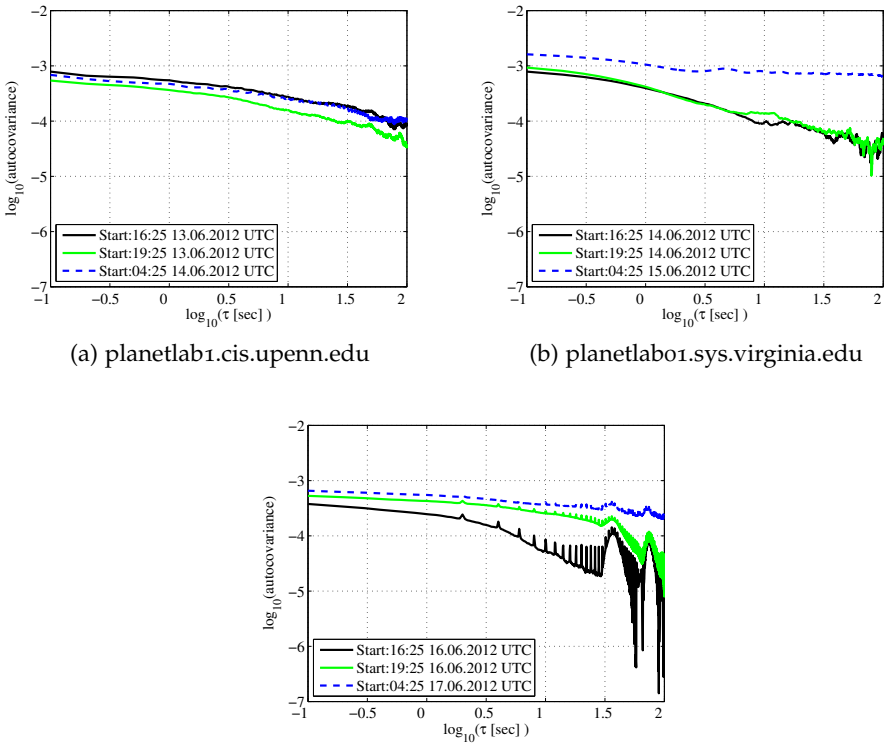
parameter estimates mostly lie in the range $[0.8, 0.9]$. We show extended results from our Internet measurement study in Appendix A.

In this chapter we provided a lightweight method for inferring the characteristics of LRD traffic from random samples. We recovered LRD traffic properties from observations made by different stochastic sampling processes and quantified the impact of finite sample sizes. We provided expressions for the accuracy of the estimates as a function of the sampling parameters. Our method lends itself to network performance monitoring scenarios. To this end, we provide a practical tool to explore the structure of traffic correlations along connection paths without administrative network support. In the next chapter we derive upper performance bounds for queueing systems fed with LRD traffic. The provided estimation method for LRD traffic parameters enables together with a rigorous mathematical analysis the assessment of the performance of operating networks as well as a better design and control for future networks.



(a) $H = 0.6$ (b) $H = 0.7$ (c) $H = 0.8$ (d) $H = 0.9$

Figure 4.8: Autocovariance estimates of LRD traffic traces of different H given from (i) analyzing entire traffic traces, (ii) traffic trace sampling and (iii) packet probing in the FILab testbed. We show mean autocovariances of 25 traces including (narrow) confidence intervals. Observe that Hurst parameter estimation is possible with reliable accuracy. The shift in the y-axis between the trace autocovariance and the sampling autocovariance is given by square of the intensity of the sampling process $\mu_X = 0.1$ as in (4.6). The shift in the y-axis between the trace autocovariance and the probing autocovariance is due to the calculation of the probing autocovariance over the utilization process at the node egress as given in (4.34).



(a) planetlab1.cis.upenn.edu

(b) planetlab01.sys.virginia.edu

(c) planetlab1.cs.umass.edu

Figure 4.9: Autocovariance estimates for end-to-end Internet paths obtained through measurements from FILab [58] to the specified targets. Each measurement is 3 hours long. For different paths and across different times of day we obtain varying autocovariance structures. For certain targets as (c) we observe obvious periodicities on different time scales.

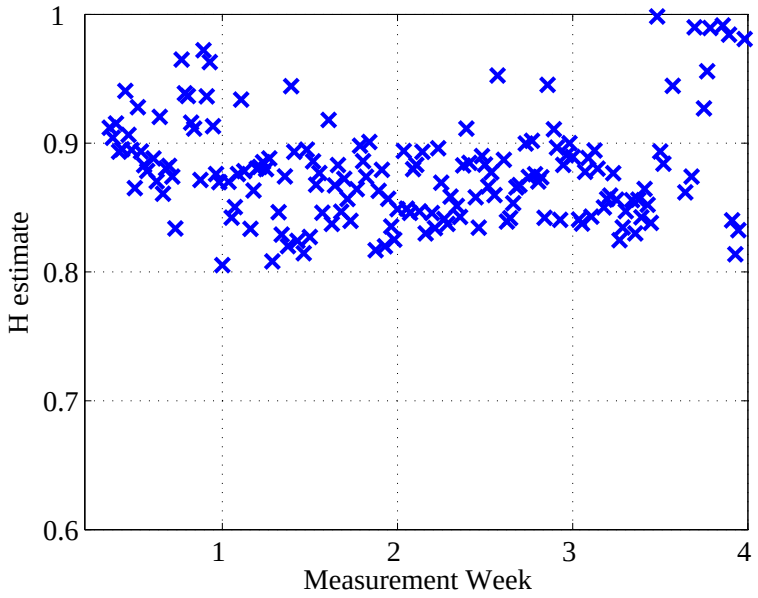


Figure 4.10: Hurst parameter estimates from continued measurements over one month (1.9.2012-30.9.2012). Target is planetlab1.cis.upenn.edu. One data point equals a measurement over three hours.

5

SINGLE NODE QUEUEING ANALYSIS WITH LONG MEMORY TRAFFIC

In this chapter we derive upper performance bounds for long memory traffic arrivals at a single queueing system. The long memory arrivals are modeled by an fBm process with Hurst parameter $H \in (0.5, 1)$ as described in Chapter 1. We use a non-asymptotic analysis to show that long memory has a significant impact on the queueing performance. Compared to the approximate, respectively, asymptotic results reviewed in Chapter 1, the derived results are non-asymptotic upper bounds. The performance bounds are based on an original sample path envelope for long memory traffic. We show that systems fed with long memory traffic behave fundamentally different than systems exhibiting light tailed EBB traffic. Performance bounds for EBB traffic known from the literature are reviewed in Chapter 2.

In addition, we consider in this chapter resource sharing scenarios as depicted in Fig. 2.3 with through and cross traffic. Approaches that belong to effective bandwidths or to large deviations theory that are reviewed in Chapter 1 do not lend themselves easily to the analysis of scheduling. In this chapter we analyze the service received by a through traffic flow after scheduling long memory cross traffic. To this end, we utilize the concept of leftover service curves from the stochastic network calculus framework reviewed in Chapter 2. We show the impact of long memory cross traffic on the queueing performance of through traffic. The results in this chapter support earlier conclusions on the efficiency of buffering long memory traffic [96] as well as general conclusions on buffer sizing in modern data networks [3]. Results presented in this chapter were developed in a joint work with M. Fidler [109, 110].

Unless stated otherwise, the definitions and properties reviewed in Chapter 2 apply to the functions utilized in the sequel. This chapter provides the foundation for Chapter 6, where we derive scaling properties of end-to-end delay bounds in multi-node networks with long memory traffic.

This chapter is structured as follows. We first derive a sample path envelope for fBm traffic and deduce backlog and delay bounds for a server with fBm input. We then inspect the asymptotic properties of the derived performance bounds revealing the tail behavior of the related overflow probabilities. After a comparison with performance bounds for EBB traffic, that show fundamentally different queueing behavior, we finally present a leftover service curve formulation under fBm cross traffic.

5.1 SAMPLE PATH ANALYSIS

Consider fBm traffic that is characterized by the following parameters: mean rate $\lambda > 0$, variance parameter $\sigma^2 > 0$ and Hurst parameter $H \in (0.5, 1)$. We use a discrete time model, where time is divided into slots of fixed size, thus time is a dimensionless counter of slots and the mean rate λ is given in bits per time slot.

The target of this section is to derive a sample path envelope for fBm traffic that satisfies (2.20). To this end, we consider the point-wise envelope (3.1) with time dependent overflow probability $\varepsilon_p(t)$. The point-wise envelope is then given as

$$E(t) = \lambda t + \sqrt{-2\log \varepsilon_p(t)}\sigma t^H. \tag{5.1}$$

We introduce the free parameters $\beta \in (0, 1-H)$ and $\eta \in (0,1)$ to express the point-wise violation probability as $\varepsilon_p(t) = \eta^{t^{2\beta}}$. The following theorem provides the construction of fBm sample path envelopes.

Theorem 5.1 (FBM Sample Path Envelope) *Given fBm traffic with mean rate λ, variance parameter σ^2, and LRD Hurst parameter $H \in (0.5, 1)$. A sample path envelope satisfying* (2.20) *is given by*

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

with overflow probability

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

where $\beta \in (0, 1-H)$ *and* $\eta \in (0,1)$ *are free parameters.*

The proof of Theorem 5.1 uses the following result on the Gamma function.

Lemma 5.1 (Gamma Function) *For* $x \in (0,1)$ *and* $\nu > 0$ *it holds that*

$$\int_0^\infty x^{t^\nu} dt = \frac{\Gamma\left(\frac{1}{\nu}\right)}{\nu(-\ln x)^{\frac{1}{\nu}}}.$$

Proof of Lemma 5.1. The definition of the Gamma function by Euler states that for all $y > 0$

$$\Gamma(y) = \int_0^1 (-\ln t)^{y-1} dt.$$

We substitute $t = e^{-zk^\nu}$ where $z > 0$ and $\nu > 0$. It follows that $dt = -z\nu e^{-zk^\nu} k^{\nu-1} dk$ and

$$\Gamma(y) = z^y \nu \int_0^\infty k^{\nu y - 1} e^{-zk^\nu} dk.$$

Assuming $\nu > 0$ and letting $y = 1/\nu$ yields

$$\Gamma\left(\frac{1}{\nu}\right) = z^{\frac{1}{\nu}} \nu \int_0^\infty e^{-zk^\nu} dk.$$

Finally, we substitute $z = -\ln x$ where $x \in (0,1)$. ∎

Proof of Theorem 5.1. Given the free parameters $\eta \in (0,1), \beta \in (0, 1-H)$. We use s, t as time indexes and rephrase the definition of sample path envelopes from (2.20) as

$$\mathsf{P}\left[\sup_{s\in[0,t]} \{A(t-s,t) - E(s)\} > 0\right] \leq \varepsilon_s, \qquad \forall t \geq 0. \tag{5.2}$$

We follow the steps reviewed in Section 2.2 for finding a rigorous upper bound on the overflow probability by applying Boole's inequality, such that we can write

$$\begin{aligned}\mathsf{P}\left[\sup_{s\in[0,t]}\{A(t-s,t)-E(s)\}>0\right] &= \mathsf{P}\left[\bigcup_{s=0}^{t}\{A(t-s,t)-E(s)>0\}\right]\\ &\leq \sum_{s=1}^{t}\mathsf{P}[A(t-s,t)>E(s)]. \end{aligned} \tag{5.3}$$

Above we used the fact that the overflow probability at $s=0$ is trivially zero since by definition $A(t,t)=E(0)=0$.

We apply Chernoff's theorem to evaluate the expression

$$\begin{aligned}\mathsf{P}[A(t-s,t)>E(s)] &\leq e^{-\theta E(s)}\mathsf{E}\left[e^{\theta A(t-s,t)}\right]\\ &= e^{-\theta E(s)}e^{\theta\lambda s+\frac{\theta^2\sigma^2}{2}s^{2H}}, \end{aligned} \tag{5.4}$$

for any $\theta>0$, where we used the moment generating function (1.3) and the stationarity of the arrival increments in the second step.

We insert the envelope $E(s)=\lambda s+\sqrt{-2\log\eta}\sigma s^{H+\beta}$ into (5.4) and minimize the right hand side over $\theta>0$ to attain the minimum at $\theta=\frac{1}{\sigma}\sqrt{-2\log\eta}s^{\beta-H}$ such that (5.4) becomes

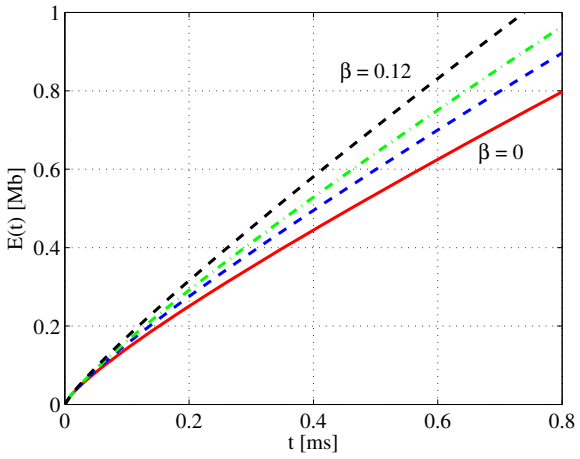
$$\mathsf{P}[A(t-s,t)>E(s)]\leq\eta^{s^{2\beta}}.$$

Equipped with the fact that $\eta^{s^{2\beta}}$ decreases monotonically with increasing s, due to $\eta\in(0,1)$, we upper bound the sum on the right hand side of (5.3) by upper bounding each summand with index s by an integral over $(s-1,s]$ and finally let $t\to\infty$ to find

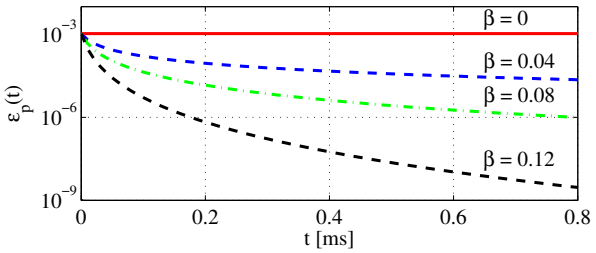
$$\sum_{s=1}^{\infty}\eta^{s^{2\beta}}\leq\int_0^{\infty}\eta^{s^{2\beta}}ds=\frac{\Gamma(\frac{1}{2\beta})}{2\beta(-\log\eta)^{\frac{1}{2\beta}}},$$

where we applied Lemma 5.1 in the last step to solve the integral. ∎

The slack parameter $\beta>0$ relaxes the envelope in Theorem 5.1 compared to the formulation (5.1) as shown in Figure 5.1a for different β values. Note that the long term envelope rate is $\lim_{t\to\infty}E(t)/t=\lambda$. For the special case $\beta=0$ the envelope



(a) The relaxation parameter $\beta > 0$ tilts the envelope $E(t)$. For $\beta = 0$ we obtain the envelope from (5.1).

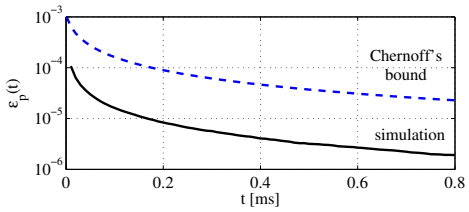


(b) For $\beta \in (0, 1-H)$ the point-wise overflow probability $\varepsilon_p(t) = \eta^{t^{2\beta}}$ decays with t, however, slower than exponentially. For $\beta = 0$ the point-wise overflow probability $\varepsilon_p(t)$ remains constant.

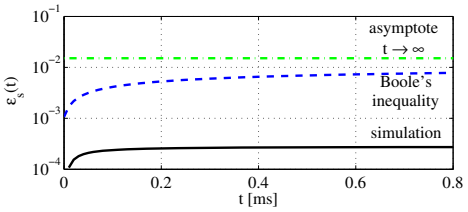
Figure 5.1: Envelope relaxation with $\beta \in (0, 1-H)$ produces decaying point-wise overflow probabilities $\varepsilon_p(t) = \eta^{t^{2\beta}}$.

in Theorem 5.1 reduces to the formulation in (5.1) and $\varepsilon_p(t)$ remains constant, thus the sample path probability would stay unbounded. For $\beta > 0$ the point-wise overflow probability $\varepsilon_p(t)$ decays slower than exponentially due to $\beta \in (0, 1-H)$ and $H \in (0.5, 1)$. For Figure 5.1a, as well as the following figures, we consider a link with capacity $C = 1$ Gbps and use a time slot of 10 μs which is the transmission time of a 10 kb packet. We set the traffic parameters for the numerical example in Figure 5.1a as measured in Internet traces in [45, 72, 100, 128] scaled to the considered link capacity. This results for λ and σ to amount to 5 kb per time slot, which translates physically to a utilization of 0.5. We set the Hurst parameter to $H = 0.7$.

In Figure 5.2 we examine the main ingredients of Theorem 5.1, i.e., Chernoff's theorem and Boole's inequality. In Figure 5.2a we compare the point-wise overflow



(a) Simulation results compared to the point-wise overflow probability $\varepsilon_p(t) = \eta^{t^{2\beta}}$, that are conservative due to Chernoff's theorem.



(b) Simulation results for length t sample paths vs. sample path overflow probability ε_s that are strict upper bounds due to Boole's inequality. The sample path overflow probabilities converge to the given finite value from Theorem 5.1 for $t \to \infty$.

Figure 5.2: Overflow probabilities for envelopes from Theorem. 5.1 with $\beta = 0.04$.

probability $\varepsilon_p(t)$ with simulation results for 10^9 fBm sample paths generated in Matlab. The results are conservative due to Chernoff's theorem. In Figure 5.2b we depict overflow probabilities for sample paths of length t. Thus, we mark the sample paths that violate the envelope at least one time in $[0, t]$ in the simulation and denote the corresponding curve "simulation" in Figure 5.2b. We plot the analytical counterpart denoted "Boole's inequality" as the cumulative sum of the point-wise violation probability up to t. This sum is upper bounded by the sample path probability from Theorem 5.1 which is depicted in Figure 5.2b as a horizontal line. The results are conservative upper bounds due to Boole's inequality.

5.2 BACKLOG AND DELAY BOUNDS

In the sequel we deduce performance bounds on backlog and delay at a single server with fBm input. We utilize the sample path envelope given in Theorem 5.1.

Theorem 5.2 (Backlog and Delay Bound) *Consider fBm arrivals characterized by λ, σ^2 and $H \in (0.5, 1)$ at the ingress of a constant rate lossless work-conserving server with capacity C. For $C > \lambda$ the steady-state backlog B is upper bounded by b subject to the overflow probability ε_s such that*

$$\mathsf{P}[B > b] \leq \varepsilon_s = \frac{\Gamma(\frac{1}{2\beta})}{2\beta(-\log \eta)^{\frac{1}{2\beta}}},$$

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

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

Under FIFO scheduling the steady-state delay W is upper bounded by $\mathsf{P}[W > b/C] \leq \varepsilon_s$.

Proof of Theorem 5.2. We use s, t as time indexes and rephrase the steady-state backlog from (2.17) as

$$\mathsf{P}[B > b] = \mathsf{P}\left[\sup_{s\in[0,t]} \{A(t-s,t) - Cs\} > b\right],$$

to directly find that it is a special case of the sample path envelope (5.2), where the envelope is substituted by $b + Cs$.

Consider $E(t)$ as given in Theorem 5.1. If $E(t) \leq b + Ct$ for all $t \geq 0$ it follows for the steady-state backlog that

$$\mathsf{P}[B > b] \leq \frac{\Gamma(\frac{1}{2\beta})}{2\beta(-\log\eta)^{\frac{1}{2\beta}}}.$$

Fix b and C. We seek the largest envelope $E(t) \leq b + Ct$ for all $t \geq 0$. The envelope $E(t)$ given in Theorem 5.1 is a concave function in t since the first derivative $\partial E(t)/\partial t$ is monotonically decreasing and the second derivative $\partial^2 E(t)/\partial t^2$ is negative. To fit $E(t)$ under the affine curve $b + Ct$ we seek the point t^* where $E(t)$ and $b + Ct$ are tangent, that is where t^* minimizes the vertical deviation between $E(t)$ and $b + Ct$. We find t^* as solution for $\partial E(t)/\partial t = C$ as

$$\begin{aligned} & \lambda + (H+\beta)\sqrt{-2\log\eta}\sigma t^{H+\beta-1} = C \\ \Rightarrow \quad & t = \left(\frac{\sqrt{-2\log\eta}\sigma(H+\beta)}{C-\lambda}\right)^{\frac{1}{1-(H+\beta)}} := t^*. \end{aligned} \tag{5.5}$$

Inserting t^* into $E(t^*) = b + Ct^*$ and solving for η yields

$$\begin{aligned}
& E(t^*) = b + Ct^* \\
\Leftrightarrow \quad & (\lambda - C)\left(\frac{C-\lambda}{\sqrt{2}\sigma(H+\beta)}\right)^{\frac{1}{H+\beta-1}} (\sqrt{-\log\eta})^{-\frac{1}{H+\beta-1}} \\
& \quad + \sqrt{2}\sigma(\sqrt{-\log\eta})^{1-\frac{H+\beta}{H+\beta-1}}\left(\frac{C-\lambda}{\sqrt{2}\sigma(H+\beta)}\right)^{\frac{H+\beta}{H+\beta-1}} = b \\
\Leftrightarrow \quad & (\sqrt{-\log\eta})^{\frac{1}{H+\beta-1}} = b^{-1}\left[\sqrt{2}\sigma\left(\frac{C-\lambda}{\sqrt{2}\sigma(H+\beta)}\right)^{\frac{H+\beta}{H+\beta-1}}(1-(H+\beta))\right] \\
\Leftrightarrow \quad & \sqrt{-\log\eta} = \left(\frac{C-\lambda}{\sqrt{2}\sigma(H+\beta)}\right)^{H+\beta}\left(\frac{b}{\sqrt{2}\sigma(1-(H+\beta))}\right)^{1-(H+\beta)} \\
\Leftrightarrow \quad & \eta = \exp\left(-\frac{1}{2\sigma^2}\left(\frac{C-\lambda}{H+\beta}\right)^{2(H+\beta)}\left(\frac{b}{1-(H+\beta)}\right)^{2-2(H+\beta)}\right)
\end{aligned}$$

The delay bound is deduced as the maximum horizontal distance of $E(t)$ and Ct. It can be inferred from the backlog bound at a constant rate server with capacity C as $\mathsf{P}[W > b/C] \leq \varepsilon_s$. ∎

In Figure 5.3 we plot backlog bounds obtained from Theorem 5.2 as well as the approximate result from (1.2). The backlog bounds are calculated for the following scenario. A constant rate server with capacity $C = 1$ Gbps is fed with fBm input with parameters $\lambda = 0.5$ Gbps, $\sigma = 0.25$ Gbps, and $H = 0.75$. Both overflow probabilities, from Theorem 5.2 and from (1.2), decay slower than exponentially due to LRD. The relative difference of the two backlog bounds decreases for smaller violation probabilities ε. In Figure 5.4 we show the impact of the traffic parameters on the backlog bounds from Theorem 5.2. For a fixed capacity C, increasing λ, σ, or H leads to an increase in the backlog bound. Note the strong impact of H, hence Figure 5.4d is on log scale.

Note that for the case $\beta = 0$ the parameter η from Theorem 5.2 reduces to the overflow probability ε_a of the approximation (1.2). Inserting η from Theorem 5.2 with $\beta = 0$ into (5.5) we find

$$t^* = \left(\frac{b}{C-\lambda}\right)\left(\frac{H}{1-H}\right),$$

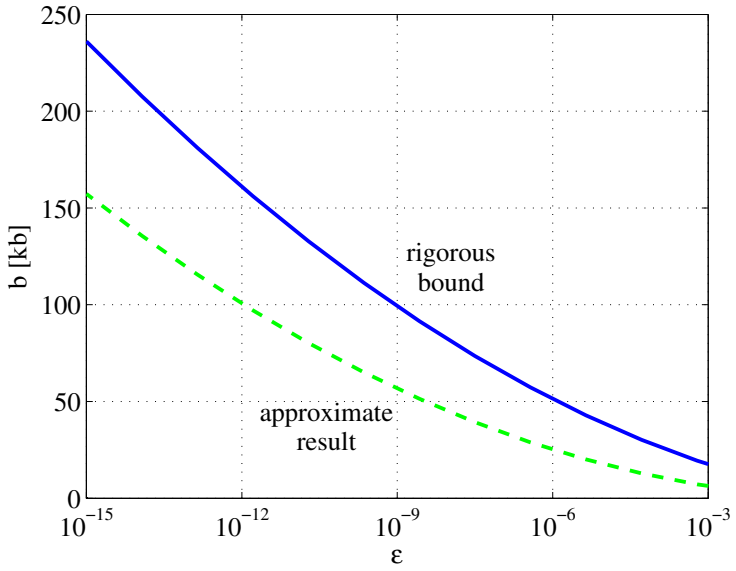
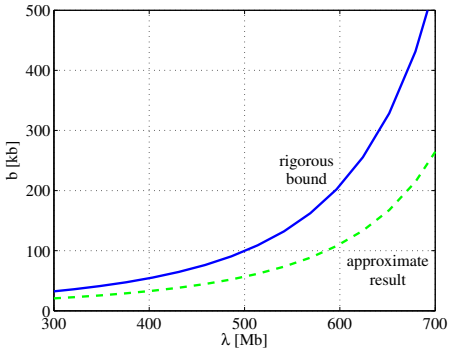
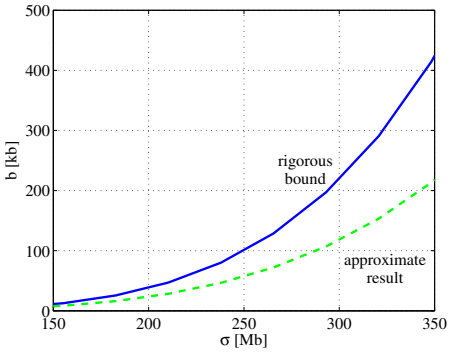


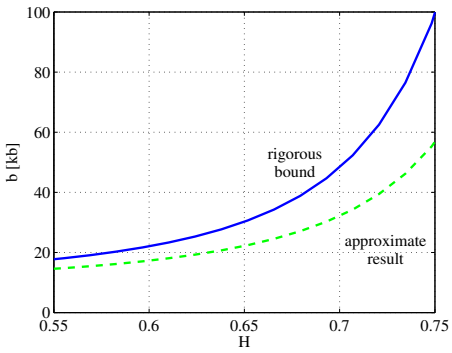
Figure 5.3: Backlog bound from Theorem 5.2 compared to the approximation from (1.2) for fBm arrival traffic with parameters $\lambda = 0.5$ Gbps, $\sigma = 0.25$ Gbps, and $H = 0.75$ at a server with capacity $C = 1$ Gbps. Observe the slower than exponential decay for both bounds.



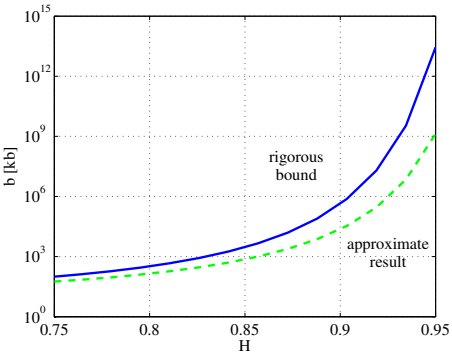
(a) Impact of mean rate λ



(b) Impact of std. deviation σ



(c) Impact of H, moderate H



(d) Impact of H, large H

Figure 5.4: Backlog bounds from Theorem 5.2 and from the approximation (1.2) for different traffic parameters subject to $\varepsilon = 10^{-9}$. The Hurst parameter has an immense impact, thus Figure 5.4d is on log scale.

recovering the critical overflow time scale from [95]. Observe the strong impact of LRD, that is expressed by $H \in (0.5, 1)$, on the critical time scale.

Next, we describe the construction of affine fBm sample path envelopes. We will resort to the affine fBm envelope in Section 5.4 to characterize the service leftover to concurrent traffic by fBm cross traffic in resource sharing scenarios. The following corollary builds on the gSBB characterization reviewed in Section 2.2.

Corollary 5.1 (Affine FBm Envelopes) *Given fBm traffic characterized by mean rate λ, variance parameter σ^2 and $H \in (0.5, 1)$. The envelope function $E(t) = rt$ is a sample path envelope satisfying* (2.21) *for $r > \lambda$ subject to the overflow profile*

$$\varepsilon_s(b) = \frac{\Gamma(\frac{1}{2\beta})}{2\beta\vartheta^{\frac{1}{2\beta}}} b^{-\frac{1-(H+\beta)}{\beta}},$$

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

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

We use the backlog bound from Theorem 5.2 to derive the affine sample path envelope in Corollary 5.1. The affine fBm sample path envelope $E(t) = rt$ is equal or greater than the envelope in Theorem 5.1 for all t. The derivation steps are identical to the proof of Theorem 5.2.

5.3 LARGE BUFFER ASYMPTOTE

In the following, we minimize the sample path violation probability ε_s from Theorem 5.2 over β analytically, which enables drawing conclusions on the large buffer asymptotic behavior of systems fed with long memory traffic. Note that we optimized the free parameter β numerically in Figures 5.3 and 5.4. While Theorem 5.2 provides rigorous non-asymptotic bounds that hold for every positive b, the following theorem concerns the tail of the overflow probabilities.

Theorem 5.3 (Weibull Tail of Overflow Probabilities) *The sample path bound in Theorem 5.2 exhibits a Weibull tail decay. The log-asymptotic decay of ε_s in b is given by*

$$\log \varepsilon_s \simeq \log \varepsilon_a \simeq -Kb^{2-2H},$$

where ε_a is given by the largest term approximation (1.2) *and K is a positive constant.*

Proof of Theorem 5.3. To obtain the log-asymptotic decay in Theorem 5.3 we minimize ε_s from Theorems 5.1 and 5.2 over β. To this end, we use Stirling's formula $\Gamma(x) \simeq \sqrt{2\pi/x}\,(x/e)^x$ for $x \gg 1$, which is given in [[1] p. 257], to approximate ε_s from Theorems 5.1 and 5.2 by

$$\tilde{\varepsilon}_s = \frac{\sqrt{\pi}}{\sqrt{\beta}(2e\beta(-\log \eta))^{\frac{1}{2\beta}}} \quad \text{for} \quad \beta \ll 1. \tag{5.6}$$

The approximation using Stirling's formula is exact in the limit $\beta \to 0$.

We simplify (5.6) for small β to derive a near optimal solution β^*. We assume $\beta \ll (1-H)$ and approximate $(1-H-\beta)$ by $(1-H)$ to compute the derivative $\partial\tilde{\varepsilon}_s/\partial\beta$ of (5.6). We solve $\partial\tilde{\varepsilon}_s/\partial\beta = 0$ for β and find that the minimum of (5.6) is attained at $\beta = -\mathcal{W}(1/(2\log \varepsilon_a))$ with the shorthand notation ε_a from (1.2). $\mathcal{W}(z)$ denotes Lambert's W function that is the inverse of $z = xe^x$. For $\varepsilon_a < e^{-e/2}$ it is real-valued. Since β is assumed to be small we find a good approximation of the optimal solution as

$$\beta^* = \frac{1}{2(-\log \varepsilon_a)},$$

where we used a linear segment to estimate Lambert's W function. Note that β^* decreases in b as $-\log \varepsilon_a$ increases with b, which improves the estimate accuracy at the tail. Remember that Theorem 5.1 and 5.2 hold for any $\beta \in (0, 1-H)$. Hence, inserting β^* into the previous theorems yields a rigorous upper bound that is near to the minimal solution.

In the sequel we use the definition

$$\chi := \left(\frac{H^H(1-H)^{1-H}}{(H+\beta)^{H+\beta}(1-(H+\beta))^{1-(H+\beta)}}\right)^2, \tag{5.7}$$

that is in $(\frac{1}{4}, 1)$ for $H \in (0.5, 1)$ and $\beta \in (0, 1-H)$ and approaches 1 for decreasing β. We insert β^* into (5.6) to simplify the terms algebraically yielding

$$\tilde{\varepsilon}_s = \sqrt{\pi}\varepsilon_a \left[\frac{\log \eta}{\log \varepsilon_a}\right]^{\log \varepsilon_a} \sqrt{-2 \log \varepsilon_a}.$$

Next, we substitute η from Theorem 5.2 and use χ from (5.7) to collect terms. We find the closed-form

$$\begin{aligned} \tilde{\varepsilon}_s &= \sqrt{\pi}\varepsilon_a \frac{b}{C-\lambda} \chi^{\log \varepsilon_a} \sqrt{-2 \log \varepsilon_a} \\ &= \frac{b}{C-\lambda}\, \varepsilon_a^{1+\log \chi} \sqrt{2\pi(-\log \varepsilon_a)} \end{aligned} \tag{5.8}$$

as a near optimal solution of Theorem 5.2 where we used $\chi^{\log \varepsilon_a} = \varepsilon_a^{\log \chi}$ with $\varepsilon_a, \chi > 0$. For $b \to \infty$ it follows that $\beta^* \to 0$, $\log \chi \to 0$, and (5.8) becomes exact from Stirling's formula. Finally, we evaluate the limit

$$\lim_{b \to \infty} \frac{\log \tilde{\varepsilon}_s}{\log \varepsilon_a} = 1,$$

which completes the proof. ∎

Theorem 5.3 shows that both ε_s and ε_a have the same log-asymptotic behavior in b, precisely, a Weibull tail. The formulation (5.8) enables drawing conclusions on buffer dimensioning based on the rigorous sample path analysis from Theorem 5.1. This strengthens previous conclusions from [96] that are based on the approximation given in (1.2). Examining ε_a reveals that the impact of spare capacity $C - \lambda$ and the buffer size b on the overflow probability is mainly determined by the Hurst parameter H. For $H = 0.5$ the arrivals degenerate to EBB standard Brownian motion. Here, the impact of $C - \lambda$ and b on ε_a is equal such that, e.g., halving the spare capacity is compensated by doubling the buffer size. This trade-off deteriorates for increasing $H \in (0.5, 1)$. Equation (5.8) confirms this behavior using sample path analysis. This result underlines the importance of spare capacity in systems with LRD traffic and simultaneously supports modern arguments on buffer size reduction [3].

Table 5.1: Traffic parameters used for Figure 5.5.

Flow type	# of sources	mean rate	variability	burstiness/LRD
EBB	$m = 100$	$\lambda = 5$ Mbps	$P = 5\lambda$	U
fBm	$m = 1$	$\lambda = 500$ Mbps	$\sigma = \lambda/2$	H

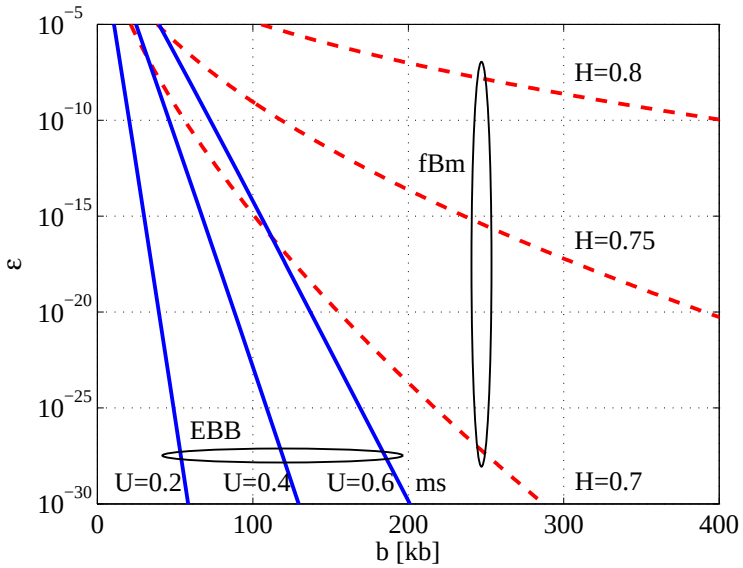


Figure 5.5: Inefficiency of buffering LRD traffic: Weibullian decay of the overflow probability for fBm traffic where $\log \varepsilon \simeq -K_1 b^{2-2H}$ compared to the exponential decay $\log \varepsilon \simeq -K_2 b$ for EBB traffic with some positive constants K_1, K_2.

Figure 5.5 compares the tail behavior of the buffer overflow probabilities for fBm and EBB traffic input to a server with capacity $C = 1$ Gbps. The EBB traffic is generated by Markov on-off sources described in Section 2.2. The overflow probability of the backlog bound for EBB traffic is shown in (2.24). EBB and fBm traffic parameters are given in Table 5.1. The EBB traffic consists of an aggregate of 100 flows. FBm is typically employed to model aggregate traffic [44, 72, 95, 96]. Hence, we plot results for a single flow using (5.8). Note that (5.8) perfectly matches the numerically optimized results from Theorem 5.2 depicted in Figures 5.3 and 5.4.

Figure 5.5 shows the decay of the buffer overflow probability for fBm traffic for different Hurst parameter values. The slower than exponential (Weibull) decay is strongly affected by the Hurst parameter H. In case of EBB traffic the burstiness parameter U affects the slope of the decay. The efficiency of buffering in case of exponential decay is obvious as it may achieve very small ε_s through slight

buffer increase. Note that the key to evaluating buffering efficiency is the form of the depicted tail decay function. From Figure 5.5 it is clear that the queueing behavior with LRD traffic is fundamentally different from that with EBB traffic. For strong LRD, e.g., $H = 0.8$ very large buffer fillings could occur with non-negligible probability. From the formulation (5.8) we conclude that buffering is inefficient for LRD traffic due to the over-proportional price paid to reduce the buffer overflow probability.

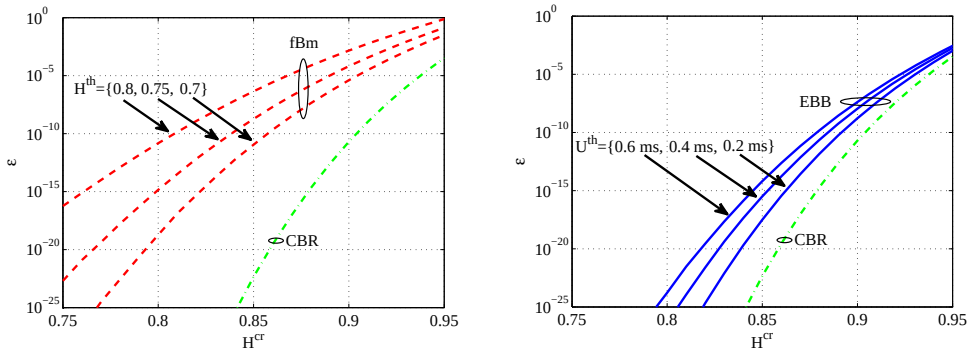
5.4 LEFTOVER SERVICE UNDER FBM CROSS TRAFFIC

We regard a system such as in Figure 2.3 where resources on the egress link are shared between through and cross traffic. Given the affine sample path envelope from Corollary 5.1 we are able to derive a leftover service curve in accordance to Section 2.2. The leftover service curve presented in the following corollary characterizes the service remaining for a through flow after scheduling fBm cross traffic at a server with capacity C.

Corollary 5.2 (FBm Leftover Service Curve) *Given a constant rate server with capacity C serving through traffic and fBm cross traffic with envelope $E(t) = rt$ that satisfies Corollary 5.1. The function $S(t) = (C - r)t$ is a valid service curve satisfying* (2.25) *that is offered to the through traffic subject to a deficit profile $\varepsilon(b)$. The deficit profile $\varepsilon(b)$ equals the overflow profile $\varepsilon_s(b)$ of the affine fBm envelope from Corollary 5.1.*

The proof of Corollary 5.2 follows directly from (2.25), (2.21) and Corollary 5.1.

Equipped with the leftover service curve formulation from Corollary 5.2 and sample path arrival envelopes for different types of through traffic, we derive performance bounds for the configuration in Figure 2.3 using (2.26) and (2.27). We consider fBm cross traffic and three basically different types of through traffic, namely constant bit rate (CBR), EBB, and fBm. We fix the through traffic mean rate to $m\lambda$, with m denoting the number of flows each possessing mean rate λ. For the calculations we use affine sample path envelopes satisfying (2.21). The first scenario with CBR traffic is a degenerate case since it is deterministic traffic. Hence, the envelope for CBR traffic is simply given by $E(t) = rt$ with $r = m\lambda$ and overflow



(a) FBm through traffic

(b) EBB through traffic

Figure 5.6: Violation probability of a delay bound of 1 ms for (a) FBm and (b) EBB through traffic at a constant rate server with fBm cross traffic. We compare the two scenarios of fBm and EBB traffic to the deterministic CBR traffic case, respectively. Observe that the increasing Hurst parameter of the cross traffic H^{cr} has considerable impact on the performance of the through traffic regardless of the through traffic correlations.

profile $\varepsilon_s(b) = 0$ for all $b \geq 0$. For EBB traffic we use the Markov on-off model reviewed in Section 2.2. We use a sample path envelope $E(t) = rt$ obtained from the backlog bound (2.24) as gSBB traffic characterization as discussed in Section 2.2. The corresponding overflow profile is given by $\varepsilon_s(b) = e^{-\theta b}/(\theta(r - m\rho(\theta)))$ for any $\theta > 0$ satisfying $m\rho(\theta) < r$ with $\rho(\theta)$ from (2.22). Since fBm is a model for aggregate traffic we set $m = 1$ for fBm through traffic. The affine envelope for the fBm through traffic is constructed as described by Corollary 5.1.

Next, we derive a delay bound for through traffic as a function of through and cross traffic parameters denoted by the superscripts *th* and *cr* respectively. Through traffic is characterized by the envelope rate r^{th} and respective overflow profile $\varepsilon^{th}(b^{th})$. The leftover service curve is determined by the cross traffic envelope with respective parameters r^{cr} and $\varepsilon^{cr}(b^{cr})$. A delay bound for the through traffic follows from (2.27) as

$$\mathsf{P}\left[W > \frac{b^{th} + b^{cr}}{C - r^{cr}}\right] \leq \varepsilon^{th}(b^{th}) + \varepsilon^{cr}(b^{cr}),$$

subject to $r^{th} + r^{cr} \leq C$. We numerically optimize over the free parameters of the envelopes, i.e., b^{th}, r^{th} and b^{cr}, r^{cr} and over β for fBm traffic and θ for EBB traffic.

We consider a server with capacity $C = 1$ Gbps and present in Figure 5.6 the violation probability of a delay bound of 1 ms for fBm and EBB through traffic under

Table 5.2: Traffic parameters used for Figure 5.6.

traffic	type	# sources	mean rate	variability	correlation
through	CBR	$m = 100$	$\lambda = 2.5$ Mb/s	–	–
	EBB	$m = 100$	$\lambda = 2.5$ Mb/s	$P = 5\lambda$	U^{th}
	fBm	$m = 1$	$\lambda = 250$ Mb/s	$\sigma = \lambda/2$	H^{th}
cross	fBm	$m = 1$	$\lambda = 250$ Mb/s	$\sigma = \lambda/2$	H^{cr}

fBm cross traffic. Each case is compared to the baseline scenario of deterministic CBR through traffic. The traffic parameters are summarized in Table 5.2. Figure 5.6 clearly shows the impact of the Hurst parameter H^{cr} of the cross traffic on the performance of through flows. The case in Figure 5.6a shows that the influence of H^{th} becomes much less pronounced for increasing cross traffic H^{cr}, i.e., cross traffic LRD becomes the dominating effect.

In this chapter we derived a sample path envelope for fBm traffic that enabled us to infer rigorous upper performance bounds for single nodes fed with LRD traffic. We investigated the respective queueing behavior to show the inefficiency of buffering in the presence of LRD. After inspecting the asymptotic tail behavior of the derived performance bounds we presented a leftover service curve that describes the service provided to concurrent flows after scheduling LRD traffic. Based on the leftover service curve formulation derived above we extend the analysis in Chapter 6 to cover entire network paths with LRD cross traffic.

6 NETWORK PATH ANALYSIS WITH LONG MEMORY TRAFFIC

In this chapter, we derive upper performance bounds for networks with long memory traffic. To this end, we deduce a network service curve in the sense of (2.30) for end-to-end paths as depicted in Figure 2.4. The network service curve is composed of the convolution of tandem node service curves that exhibit each long memory cross traffic. We capitalize on single node results, especially the leftover service curve, derived in Chapter 5 to express the service provided by the tandem nodes. The derivation of a network service curve allows the calculation of rigorous end-to-end performance bounds under long memory traffic. The results derived in this chapter were developed in a joint work with M. Fidler [107, 108, 110]. The scaling result presented in this chapter appeared first in [107]. Since the discovery of self-similarity and LRD in Internet traffic the evaluation of end-to-end network performance under long memory traffic has been an open problem. The scaling result presented in this chapter was recovered independently in the context of performance evaluation with heavy tailed traffic in [77, 79].

Considerable research has been dedicated to the growth of performance bounds in the number of traversed path nodes n subject to fixed violation probabilities. For EBB traffic a scaling result of $\Theta(n \log n)$ was proven in [19, 28]. This compares to a scaling result of $\mathcal{O}(n^3)$ that is shown in [26] to arise from adding per node bounds as suggested by [130]. The root of this improvement lies in the network service curve formulation [28]. For $(\sigma(\theta), \rho(\theta))$ cross traffic an end-to-end scaling result of $\mathcal{O}(n)$ is derived in [46] under statistical independence of the cross traffic. From queueing theory it is known for queueing networks with tandem $M|M|1$ queues that end-to-end delays scale in $\Theta(n)$ [68].

The contributions of this chapter are end-to-end performance bounds for networks with LRD traffic. We first present a formulation for sample path leftover service under fBm cross traffic that is essential to the evaluation of the end-to-end network performance. Equipped with this result we infer the end-to-end service of a network path under fBm cross traffic and derive corresponding performance

bounds. We show that end-to-end delays scale with $\mathcal{O}\left(n(\log n)^{\frac{1}{2-2H}}\right)$ for network paths of length n under LRD cross traffic with Hurst parameter $H \in (0.5, 1)$. This result makes the substantial impact of LRD on the network performance explicit and enables drawing conclusions on designing and dimensioning networks. Our scaling recovers the EBB result of $\mathcal{O}(n \log n)$ from [28] for the EBB standard Brownian motion case of $H = 0.5$. Further, we establish a relation in the asymptotic regime between end-to-end and single node performance bounds as a function of the network path length n. Finally, we make a case for spare capacity as we show its considerable impact on the network performance in the presence of LRD traffic.

6.1 NETWORK PERFORMANCE ANALYSIS

In Chapter 5, we derived a service curve representation for single nodes with fBm cross traffic. Stochastic network calculus enables the derivation of network service curves that describe end-to-end paths by composition of the service curves of the individual tandem nodes. A general construction method of network service curves has been reviewed in Section 2.2. Network service curves permit the derivation of end-to-end performance bounds by collapsing a network path into a single equivalent system and substituting in single node expressions for backlog or delay such as (2.26), respectively, (2.27).

Consider a network path as depicted in Figure 2.4, where fBm cross traffic is multiplexed and de-multiplexed at each tandem node. In the first step, we derive a leftover service curve that satisfies (2.28) with a deficit profile that follows from (2.29). Consequently, we combine the leftover service curves of the individual nodes as in (2.30) to obtain a network service curve $S^{net}(t)$. The deficit profile for the network service curve follows from (2.31). The following corollary provides a sample path leftover service curve under fBm cross traffic.

Corollary 6.1 (Sample Path fBm Leftover Service Curve) *Consider through traffic under blind multiplexing with fBm cross traffic that is characterized by λ, σ^2 and $H \in (0.5, 1)$ at a constant rate server with capacity C. Assume the fBm cross traffic possesses an affine*

envelope $E(t) = rt$ *according to Corollary 5.1. The through traffic is offered a service curve* $S(t) = (C - r)t$ *defined by* (2.28) *with a deficit profile*

$$\varepsilon^{\varrho}(b) = \frac{\Gamma(\frac{1}{2\beta})}{2\varrho\vartheta^{\frac{1}{2\beta}}(1-(H+2\beta))} b^{-\frac{1-(H+2\beta)}{\beta}}$$

according to (2.29)*, where* $\varrho \in (0, C - r)$ *and* $\beta \in \left(0, \frac{1-H}{2}\right)$ *are free parameters and* ϑ *is defined in Corollary 5.1.*

The sample path deficit profile in Corollary 6.1 is obtained by plugging the deficit profile from Corollary 5.1 into (2.29) as

$$\varepsilon^{\varrho}(b) = \frac{\Gamma(\frac{1}{2\beta})}{2\beta\varrho\vartheta^{\frac{1}{2\beta}}} \int_b^{\infty} x^{-\frac{1-(H+\beta)}{\beta}} dx,$$

that has the finite solution given in Corollary 6.1 for $\beta < \frac{1-H}{2}$.

Using Corollary 6.1 we are able to derive an end-to-end service curve that satisfies (2.30) with a deficit profile as given in (2.31). We consider through traffic with affine envelope rate r^{th} that traverses a network path of n nodes as depicted in Figure 2.4 with fBm cross traffic. For ease of notation we assume n homogeneous nodes, each with capacity C and fBm cross traffic possessing identical parameters λ, σ^2, and $H^{cr} \in (0.5, 1)$. We adopt the convention of the superscripts th and cr for through and cross traffic respectively from Section 5.4. At each node $i \in [1, n]$ we obtain a leftover service curve $S_i(t) = (C - r^{cr})t$, where we use r^{cr} to denote the envelope rate for the fBm cross traffic. Plugging $S_i(t)$ into the min-plus convolution in (2.30) yields an end-to-end service curve $S^{net}(t) = (C - r^{cr} - \Delta)t$ with constant $\Delta := (n-1)\varrho$ and $\Delta \in (0, C - r^{cr})$ for stability. We calculate the deficit profile for the network service curve $S^{net}(t)$ by inserting $\varepsilon^{\varrho}(b)$ from Corollary 6.1 for nodes $i \in [1, n-1]$ and $\varepsilon(b)$ from Corollary 5.2 for node n into the min-plus convolution (2.31). After algebraic simplification we obtain the following deficit profile for the network service curve

$$\varepsilon^{net}(b) = \frac{\Gamma(\frac{1}{2\beta})}{2\vartheta^{\frac{1}{2\beta}}} \inf_x \left\{ \frac{(n-1)^2 (\frac{b-x}{n-1})^{-\frac{1-(H+2\beta)}{\beta}}}{\Delta(1-(H+2\beta))} + \frac{x^{-\frac{1-(H+\beta)}{\beta}}}{\beta} \right\}, \tag{6.1}$$

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

Table 6.1: Traffic parameters used for Figures 6.1, 6.2, 6.3, and 6.4.

traffic	type	# sources	mean rate	variability	burstiness/LRD
through	CBR	$m = 100$	$\lambda = 2.5$ Mb/s	–	–
cross	EBB	$m = 100$	$\lambda = 2.5$ Mb/s	$P = 5\lambda$	U^{cr}
	fBm	$m = 1$	$\lambda = 250$ Mb/s	$\sigma = \lambda/2$	H^{cr}

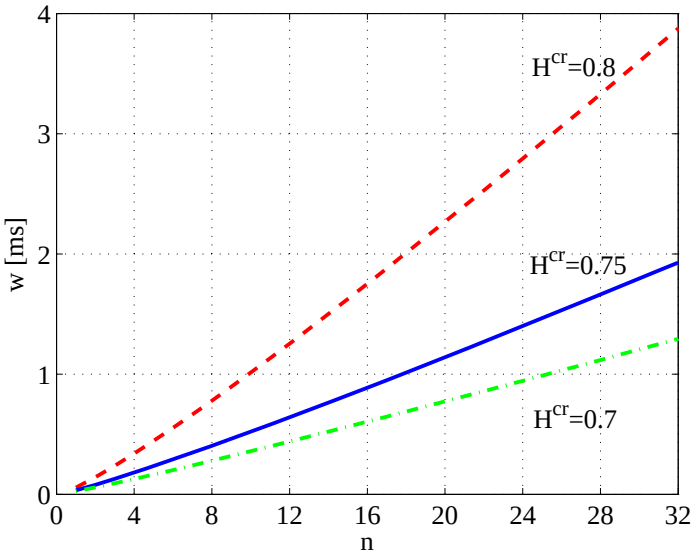


Figure 6.1: End-to-end delay bounds for through traffic at a tandem of n constant rate nodes with LRD cross traffic subject to fixed violation probability $\varepsilon = 10^{-12}$. The delay bounds that are shown for different cross traffic Hurst parameter values H^{cr}, grow super-linearly in n.

Using (6.1) we are able to derive end-to-end performance bounds by insertion into expressions for single node performance bounds such as (2.26) and (2.27). For example, an end-to-end delay bound for CBR through traffic with rate r^{th} can be computed using (2.27) as

$$\mathsf{P}\left[W > \frac{b}{(C - r^{cr} - \Delta)}\right] \leq \varepsilon^{net}(b), \tag{6.2}$$

with stability condition $r^{th} + r^{cr} + \Delta \leq C$. A straightforward extension for through traffic with overflow profile ε^{th} implies a delay bound $\mathsf{P}[W > b/(C - r^{cr} - \Delta)] \leq \varepsilon^{net} \otimes \varepsilon^{th}(b)$. We depict the end-to-end delay bounds for a fixed violation probability and CBR through traffic in Figure 6.1 as a function of the number of traversed nodes n. Traffic parameters are given in Table 6.1 where each node possess a capacity of $C = 1$ Gbps. We numerically optimize the free parameters of the

envelopes. Figure 6.1 shows the impact of LRD cross traffic on end-to-end network performance. Observe that the end-to-end delay bounds for through traffic grow super-linearly with the number of traversed nodes n.

Next, we briefly examine the scenario of inhomogeneous LRD cross traffic, i.e., the individual cross traffic flows at nodes $i \in [1, n]$ are characterized by different Hurst parameter values H_i. From the min-plus convolution in (2.31) and the Weibull tail decay shown in (5.8), it is obvious that the cross traffic with the highest H_i, that leads to the slowest decay of $\varepsilon_i^o(b)$ in (2.31), has the strongest impact on the deficit profile of the network service curve $\varepsilon^{net}(b)$.

In the following section, we analytically inspect the scaling of end-to-end performance bounds for networks with homogeneous LRD traffic.

6.2 SCALING OF END-TO-END BOUNDS IN THE NUMBER OF TANDEM NODES

Next, we inspect the growth of end-to-end performance bounds with the number of traversed nodes n under LRD cross traffic. This scaling expresses the additional delay arising on longer network paths, subject to unchanged violation probability.

Theorem 6.1 (Scaling Analysis in the Number of Traversed Nodes) *Given n homogeneous tandem nodes, each exhibiting fBm cross traffic with Hurst parameter $H \in (0.5, 1)$. End-to-end backlog and delay bounds for through traffic scale in the number of nodes n as*

$$\mathcal{O}\left(n\,(\log n)^{\frac{1}{2-2H}}\right)$$

subject to a fixed violation probability.

Proof of Theorem 6.1. The proof is similar to the proof of Theorem 5.3 and uses some of its basic steps. First, we neglect the irregularity in (2.31) that is due to the last hop and upper bound ε_n by ε_n^o to obtain the simplified deficit profile from (6.1)

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

We assume small $\beta \ll (1-H)/2$ and use Stirling's formula $\Gamma(x) \simeq \sqrt{2\pi/x}\,(x/e)^x$ for $x \gg 1$ from in [1] to find

$$\tilde{\varepsilon}^{net}(b) = \frac{\sqrt{\pi\beta}n(n-1)(\frac{b}{n})^{-\frac{1-(H+2\beta)}{\beta}}}{(2e\beta\vartheta)^{\frac{1}{2\beta}}\Delta(1-(H+2\beta))}. \tag{6.3}$$

Next, we optimize (6.3) over β and attain the minimum at $\beta = \mathcal{W}\left(n^{2-2H}/2(-\log\tilde{\varepsilon}_a)\right)$, where we used the shorthand notation $\tilde{\varepsilon}_a$ that is defined in (3.2). $\mathcal{W}$ denotes Lambert's W function deployed in the proof of Theorem 5.3. Since β is assumed to be small, a good approximation of the optimal solution, respectively, of Lambert's W function is given by a linear segment such as

$$\beta^* = \frac{n^{2-2H}}{2(-\log\tilde{\varepsilon}_a)}. \tag{6.4}$$

We define $\psi = (1-H)/(1-(H+2\beta))$ and use the definition of χ from (5.7), which is in $\left[\frac{4}{3\sqrt{3}},1\right]$ for $H \in (0.5,1)$ and $\beta \in (0, \frac{1-H}{2})$. We insert β^* into (6.3) and obtain after some algebraic manipulations

$$\tilde{\varepsilon}^{net}(b) = \frac{n-1}{n^H}\left(\frac{H}{1-H}\right)^H \frac{\sqrt{\pi}\psi\sigma b^{1+H}}{\Delta(r^{cr}-\lambda)^{1+H}}\tilde{\varepsilon}_a^{(1+\log\chi)n^{2H-2}}. \tag{6.5}$$

Note that for $\beta \to 0$ both χ and ψ tend to 1. Assuming $\beta^* \ll (1-H)/2$ we can generally bound χ and ψ from below, respectively, from above using constants. Next, we expand the shorthand notation $\tilde{\varepsilon}_a$ given by (3.2) and use positive constants k_i with $i \in \{0,1,2\}$ to collect terms that do not depend on n, b, and H to obtain

$$\tilde{\varepsilon}^{net}(b) \leq n^{1-H}\,k_1 b^{1+H} e^{-k_2 b^{2-2H} n^{2H-2}}.$$

We aim at expressing b as a function of n subject to the violation probability $\tilde{\varepsilon}^{net}$ that is bounded from above by a constant for all n. We choose

$$b = n(k_0 \log n)^{\frac{1}{2-2H}}, \tag{6.6}$$

for $n \geq 2$ and obtain by insertion

$$\tilde{\varepsilon}^{net} \leq (k_0 \log n)^{\frac{1+H}{2-2H}} k_1 n^{2-k_0 k_2}.$$

Generally, there exists $k_0 > 2/k_2$ such that $\tilde{\varepsilon}^{net}$ is upper bounded by a constant for all n. This is due to the fast decay of the second factor $k_1 n^{2-k_0 k_2}$ that eliminates the growth of the first factor $(k_0 \log n)^{\frac{1+H}{2-2H}}$. Note that (6.6) is the function of the smallest order that satisfies this constraint. For fixed $\tilde{\varepsilon}^{net}$ the formulation in (6.6) provides the scaling stated in Theorem 6.1. This scaling holds for the backlog as well as for the delay bound that is given in (6.2).

Finally, through inserting (6.6) into (6.4) we verify that β^* decreases with n where the decay is proportional to $1/\log(n)$. This confirms the assumption that starting from a small β^* it remains small and decreases for increasing n. Recall that Stirling's approximation is exact in the limit for $\beta \to 0$. ∎

Theorem 6.1 characterizes the impact of LRD cross traffic on end-to-end performance bounds. For a given violation probability end-to-end performance bounds grow super-linearly in the number of traversed nodes n. The growth is determined by the tail decay of the overflow profile from Corollary 6.1 that is invoked in the min-plus convolution (2.31). In essence, the Hurst parameter H impacts the poly-logarithmic scaling component in Theorem 6.1. The impact of LRD on the scaling of end-to-end delay bounds is deemed moderate due to the poly-logarithmic component in Theorem 6.1. The provided scaling specifies the additional cost in terms of delay for using longer network paths while retaining a given violation probability.

Next, we compare the scaling of the end-to-end delay bounds from Theorem 6.1 to a corresponding scaling for light tailed EBB cross traffic from [28]. To this end, we consider the scenario in Figure 2.4 with CBR through traffic and homogeneous EBB cross traffic that is given by aggregation of m on-off sources. We use the EBB sample path envelope from Section 5.4 to derive a leftover service curve $S_i(t) = (C - r^{cr})t$ with deficit profile $\varepsilon(b) = e^{-\theta b}/(\theta(r^{cr} - m\rho(\theta)))$ for nodes $i \in [1, n]$. A sample path deficit profile follows from (2.29) after relaxing the service curve by a slack rate $\varrho > 0$ according to (2.28). By insertion of $S_i(t)$ into (2.30) we derive an end-to-end service

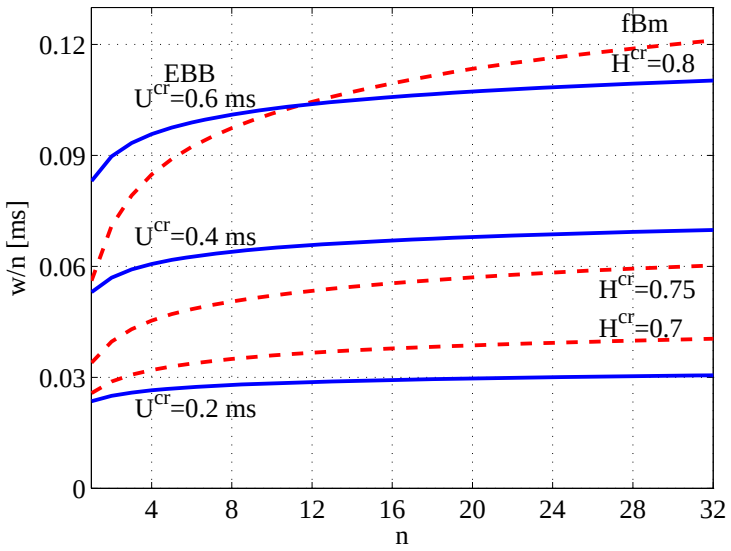


Figure 6.2: Normalized end-to-end delay bounds w/n for a path of n nodes with CBR through traffic and fBm or EBB cross traffic, respectively. Observe the poly-logarithmic growth for LRD traffic in comparison to the logarithmic growth with n for EBB traffic.

curve $S^{net}(t) = (C - r^{cr} - \Delta)t$ with $\Delta = (n-1)\varrho$ and a deficit profile according to (2.31) as

$$\varepsilon^{net}(b) = \frac{1}{r^{cr} - m\rho(\theta)} \inf_x \left\{ \frac{(n-1)^2 e^{-\theta \frac{b-x}{n-1}}}{\Delta\theta^2} + \frac{e^{-\theta x}}{\theta} \right\}, \tag{6.7}$$

where $x \in (0, b)$ and $\theta > 0$ is given such that $m\rho(\theta) < r^{cr}$. Solving $\varepsilon^{net}(b)$ for b yields that end-to-end performance bounds under EBB cross traffic scale in $\mathcal{O}(n \log n)$ [28]. Theorem 6.1 recovers the scaling of $\mathcal{O}(n \log n)$ for $H = 0.5$ which implies EBB class standard Brownian motion.

Figure 6.2 depicts end-to-end delay bounds for CBR through traffic and EBB or LRD cross traffic, respectively. We depict delay bounds normalized to the path length, i.e., w/n versus the number of nodes n. Traffic parameters are given in Table 6.1 where the capacity at each node amounts to $C = 1$ Gbps. We optimize the parameters of the traffic envelopes numerically. Figure 6.2 shows the logarithmic and poly-logarithmic scaling of the normalized delay bounds under EBB and LRD cross traffic, respectively. In the case of EBB, increasing the burstiness of the cross traffic using the parameter U^{cr} reproduces tiered versions of w/n. However, for the

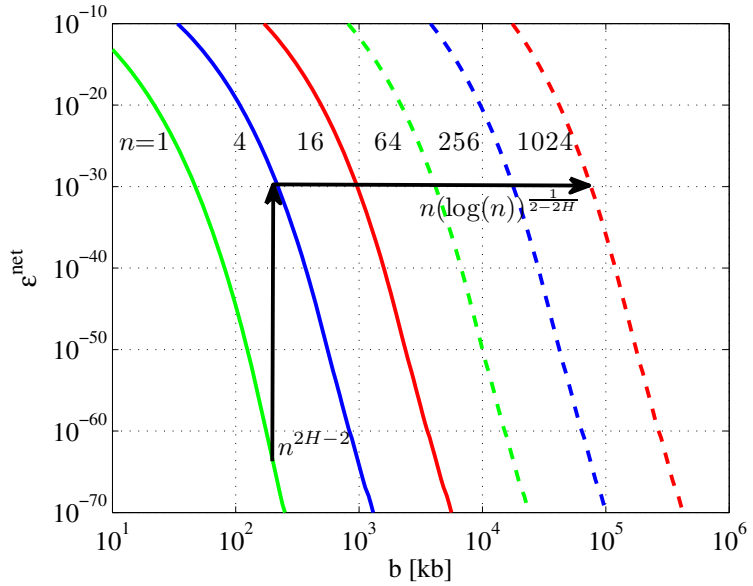


Figure 6.3: Scaling of the deficit profile ε^{net} of the network service curve (6.5). In the horizontal direction we depict the scaling from Theorem 6.1, while in the vertical direction we illustrate the scaling of (6.8). Traffic parameters are taken from Table 6.1, with fixed $H = 0.75$.

LRD case, increasing the Hurst parameter H^{cr} impacts the poly-logarithmic scaling component, thus it changes the shape of w/n.

6.3 ASYMPTOTIC BEHAVIOR OF END-TO-END PERFORMANCE BOUNDS

We examine the asymptotic behavior of end-to-end multi-node performance bounds in comparison to single node performance bounds. We consider CBR through traffic and fBm cross traffic and compare the corresponding deficit profiles of the network service curve from (6.5) to the single node formulation from (5.8). In the large buffer asymptotic regime we find the following behavior

$$\lim_{b\to\infty} \frac{\log \tilde{\varepsilon}^{net}(b)}{\log \tilde{\varepsilon}_s(b)} = n^{2H-2}. \tag{6.8}$$

The end-to-end deficit profile $\tilde{\varepsilon}^{net}$ possesses the same log-asymptotic decay in b as the single node expression $\tilde{\varepsilon}_s$ from (5.8) except for a factor that depends on the network path length n and the Hurst parameter H. For $H = 0.5$ the formulation in (6.8) recovers the ratio n^{-1} that is straightforwardly derived from (2.24) and (6.7), respectively from the original EBB results in [26].

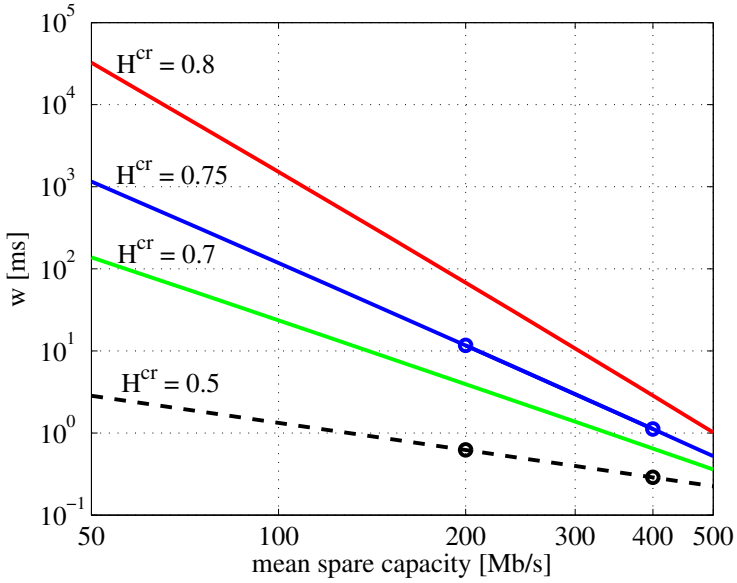


Figure 6.4: End-to-end delay bounds for a network path of $n = 10$ nodes, cross traffic with Hurst parameter H^{cr} and varying mean spare capacity. For uncorrelated cross traffic, i.e., $H^{cr} = 0.5$, cutting the spare capacity by half doubles the delay bound w. For LRD cross traffic spare capacity becomes essential owing to a stronger impact on w. For example, halving the spare capacity leads to a tenfold increase of the delay bound for $H^{cr} = 0.75$.

Figure 6.3 shows the tail decay of $\varepsilon^{net}(b)$ for network paths of length n. All nodes have capacity $C = 1$ Gbps while the traffic possesses the parameters in Table 6.1. We illustrate the scalings from Theorem 6.1 and expression (6.8) using a logarithmic scale for $\varepsilon^{net}(b)$ as well as for b. We fix $H = 0.75$ such that the scaling in vertical direction reduces to $1/\sqrt{n}$ on the depicted logarithmic scale. From Theorem 6.1 we find that the scaling in n , i.e., in the horizontal direction, is given by $n(\log n)^2$ on a linear scale.

6.4 NETWORK DIMENSIONING: THE NEED FOR SPARE CAPACITY

In this section we illustrate the impact of the average spare capacity provided by the network on the derived end-to-end performance bounds. This is similar to investigations known from queueing theory on the impact of the utilization on single node queueing measures such as the average waiting time [10].

We consider a homogeneous network as depicted in Figure 2.4 and define the average spare capacity as the residual capacity after subtracting the mean rates of

both through and cross traffic from the node capacity C. We regard the average spare capacity similar to single node utilization as a network operating point. From the view of network operators a low spare capacity is cost efficient. In the sequel, we present an argument relating end-to-end performance bounds to the average spare capacity, hence, to the corresponding network operating points.

We build on the single node argument on buffering efficiency in the presence of LRD from Section 5.3 and from the approximation (1.2), respectively. We regard the end-to-end deficit profile $\tilde{\varepsilon}^{net}$ in (6.5) that is formulated as multiple of the shorthand notation $\tilde{\varepsilon}_a$ that is given in (3.2). The expression for $\tilde{\varepsilon}_a$ describes how the Hurst parameter H governs the relationship of $r - \lambda$ to b. Note that $r - \lambda$ in (3.2) is the difference between the affine envelope rate and the mean rate of the fBm traffic. To achieve a fixed violation probability $\tilde{\varepsilon}_a$, halving $r - \lambda$ requires doubling b for $H = 0.5$. However, this relationship deteriorates for increasing H and becomes, e.g., an eightfold increase of b in case of $H = 0.75$.

Next, we consider an example of a network path with $n = 10$ tandem nodes, CBR through traffic and fBm cross traffic parameterized as in Table 6.1. In Figure 6.4 we illustrate end-to-end delay bounds from (6.5) for varying mean spare capacity subject to a fixed violation probability $\varepsilon = 10^{-12}$. We optimize all free parameters numerically. Halving the spare capacity leads to doubling the delay bound for $H^{cr} = 0.5$ as indicated by the markers. For $H^{cr} = 0.75$ the delay bound increases more than eightfold. Figure 6.4 shows the fundamental need for spare capacity to obtain good performance in the presence of LRD network traffic. In the context of network dimensioning this result evidently demonstrates the importance of over-provisioning. Moreover, in the light of congestion control this result implies that source rate throttling can effectively improve network performance. The improvement depends, however, on the network load as indicated by the logarithmic axis scale in Figure 6.4. The impact on heavily loaded networks is stronger than on lightly loaded ones, as in the first case slight changes to the mean traffic rates have a high relative impact on the spare capacity compared to lightly loaded networks.

In this chapter we derived upper end-to-end performance bounds for networks carrying LRD traffic. First, we deduced a sample path leftover service curve to describe individual nodes each exhibiting LRD cross traffic. Equipped with this

result we derived a network service curve that captures the service provided by a tandem of such nodes. The network service curve allowed us to deduce end-to-end performance bounds for network paths with LRD cross traffic. We found that end-to-end delays scale as $\mathcal{O}\left(n(\log n)^{\frac{1}{2-2H}}\right)$ in the number of traversed node n. This result reveals the impact of LRD on the network performance. Finally, we showed how the derived results establish the significance of spare capacity in networks carrying LRD traffic.

7

CONCLUSIONS AND FUTURE WORK

In this thesis we contributed a non-asymptotic end-to-end performance evaluation of communication networks carrying long memory traffic. Further, we provided a lightweight sampling framework to estimate the statistical properties of long memory traffic in single and multi-node scenarios. In the following, we summarize the conclusions of our work and provide pointers for possible future investigations with corresponding applications.

We presented a lightweight method to extract long memory traffic properties from random traffic samples. After proving the applicability of our approach to estimate the traffic autocovariance for different stochastic sampling processes we identified the limitations of using rigid periodic sampling. To enable practical deployment, we quantified the impact of limited sample sizes on the accuracy of the estimates and provided asymptotically unbiased estimators. Complementary to trace driven approaches that analyze traffic properties at a single point of interest, we provide a network probing framework that reflects the dominant characteristics on end-to-end network paths without administrative support.

We derived a formulation for sample path envelopes for long memory traffic with corresponding violation probabilities. Equipped with this result we obtained upper performance bounds for single nodes fed with long memory traffic. We find that the rigorous performance bounds possess a Weibullian tail decay. Our bound recovers an approximate result known from literature at the most probable time scale for buffer overflow. For long memory traffic we observed a fundamentally different queueing behavior compared to light tailed EBB traffic. For resource sharing scenarios we used an affine sample path envelope to deduce a formulation for the service leftover for concurrent through traffic by long memory cross traffic. This formulation allows an end-to-end performance evaluation with long memory traffic using the framework of stochastic network calculus.

We capitalized on expressions for convolution form networks known from stochastic network calculus to derive non-asymptotic end-to-end performance bounds un-

der long memory cross traffic. We first deduced an equivalent system formulation to the investigated network path by deriving a network service curve through composition of tandem node service curves each under long memory traffic. Consequently, by applying single node techniques to the equivalent system we find end-to-end performance bounds for through traffic traversing networks with long memory cross traffic. Our results culminated in a scaling for end-to-end performance bounds in the number of traversed nodes n that is given as $\mathcal{O}\left(n(\log n)^{\frac{1}{2-2H}}\right)$ with long memory Hurst parameter $H \in (0.5, 1)$. This scaling characterizes the impact of long memory on end-to-end performance bounds in communication networks. Our scaling result recovers a prominent finding for EBB traffic for the case of $H = 0.5$. A fundamental implication of our end-to-end performance results is that spare capacity, hence over-provisioning, is decisive for reliable network performance in the presence of long memory traffic.

The results presented in this thesis lead to some open research questions and possible future work directions. We sketch some of the remaining open questions and point out adequate approaches. One potential research direction builds on the presented end-to-end probing framework. A possible extension relates the presented end-to-end autocovariance inference problem to the broad class of network tomography problems. One approach to this challenge is based on multicast trees that are investigated in [17] in the context of loss rate identification on inner network links.

Another potential research direction comprises the investigation of the impact of statistical independence on the end-to-end performance bounds under long memory cross traffic. A possible approach to this problem builds on the statistical independence considerations that are presented in different calculation domains within the stochastic network calculus framework such as in [46] or in [64].

Part II

APPENDIX

A

DATA EXCERPT FROM THE INTERNET MEASUREMENT CAMPAIGN

We performed measurements for multiple months using the probing software *H-probe* [14] starting from our lab in Germany and targeting a number of PlanetLab nodes around the world. Next, we briefly describe the measurement setup:

- Discretization slot length $\delta = 1$ ms.
- Geometrically distributed inter-sample times with parameter $p = 0.1$.
- Number of probes collected 10^6 (approx. 3 hours)
- ICMP probing packets of size 64 Bytes
- Probing rate 100 pkt/s approx. 70 kbps

In the following, we present a representative set of the measurement results, where the target is planetlab1.cis.upenn.edu. We show end-to-end autocovariance estimates from continued measurements at two different days. In addition, we show estimates from daily measurements starting at a fixed time, that is 10:45 UTC for the time span of 17-24.7.2012.

We deduce Hurst parameter estimates from the slope of the autocovariance function that is theoretically given by $2H - 2$. Slope estimates are obtained through least square regression. The H estimates from Internet measurements have a moderately higher variance compared to active probing results from Figure 4.7. The autocovariance decay functions exhibit a complex structure including LRD decay and periodicities.

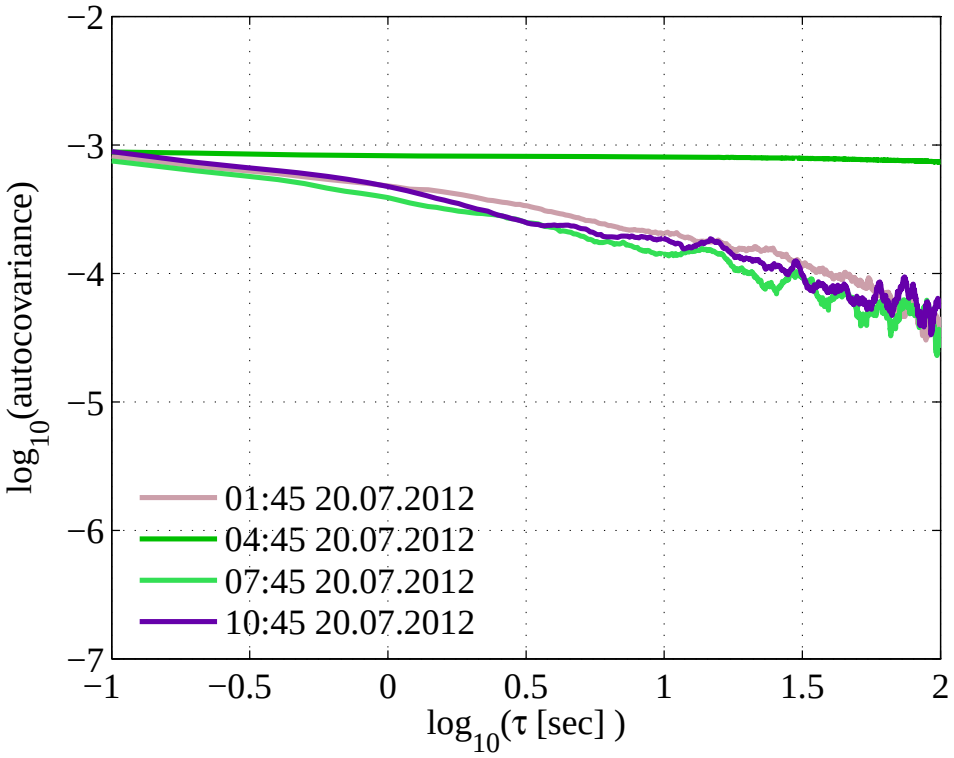


Figure A.1: Autocovariance estimates 20.7.2012 UTC

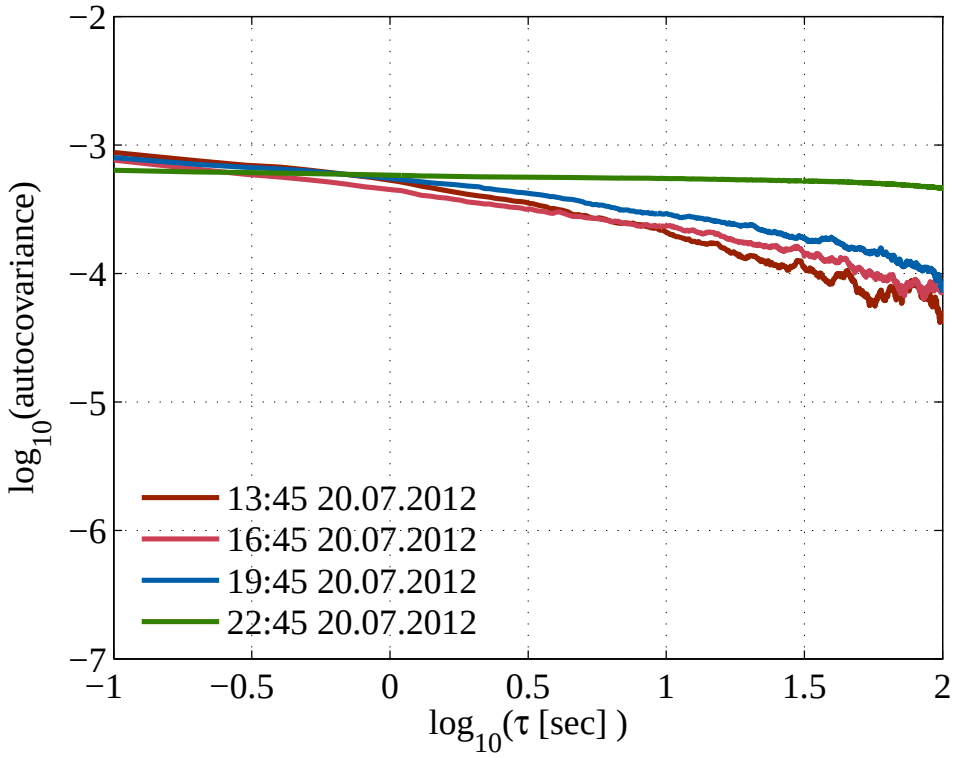


Figure A.2: Autocovariance estimates 20.7.2012 UTC

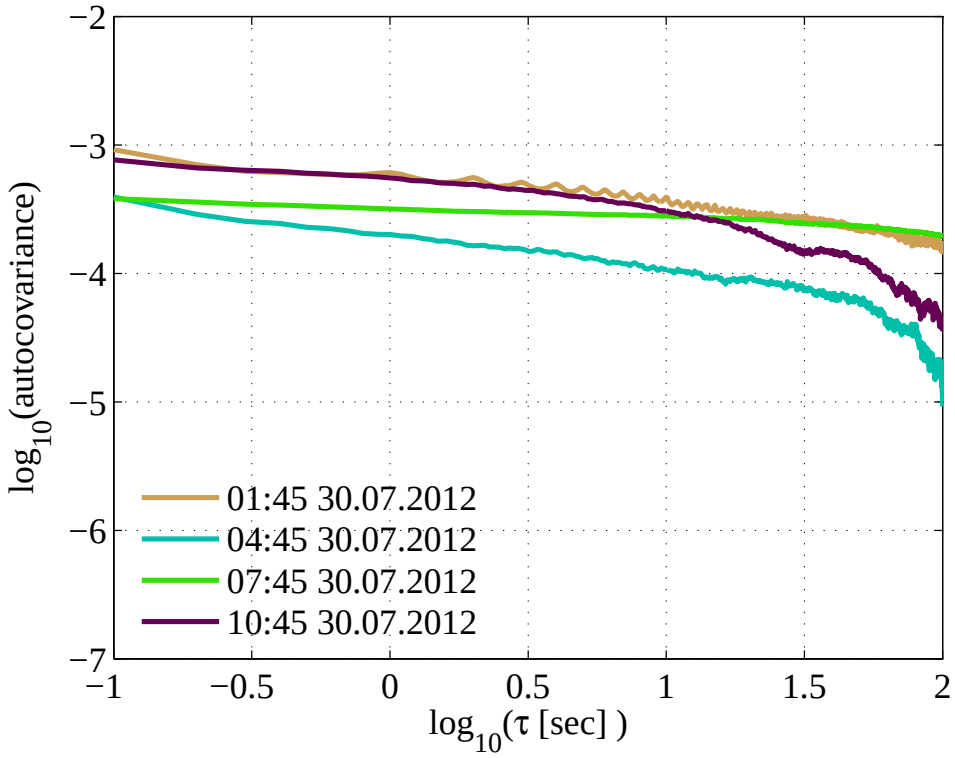


Figure A.3: Autocovariance estimates 30.7.2012 UTC

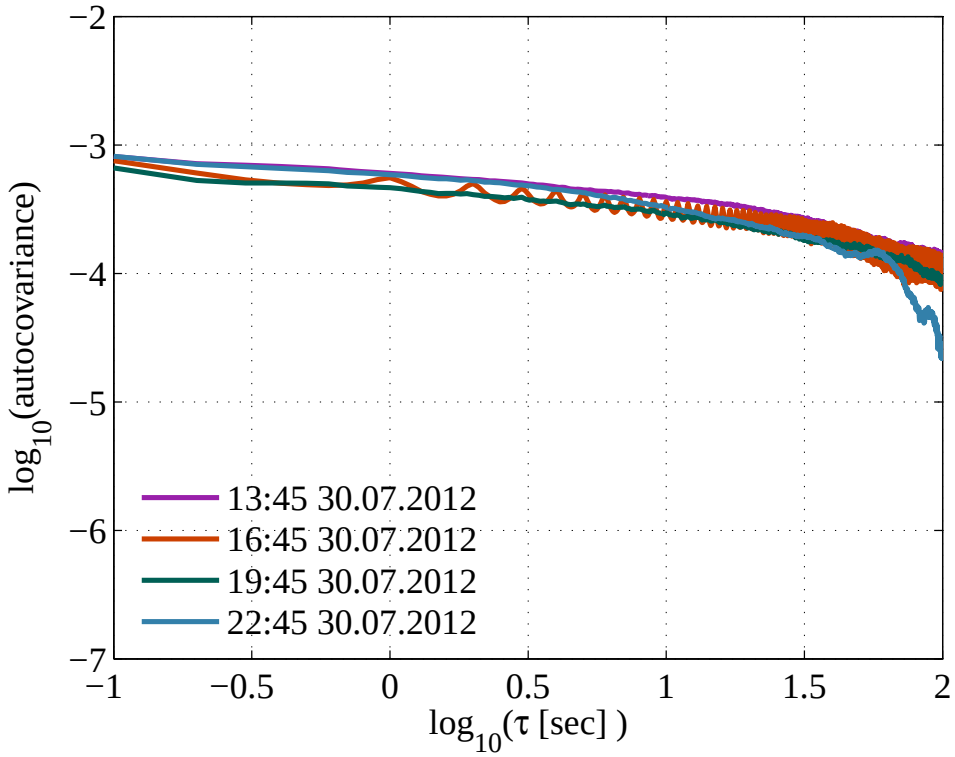


Figure A.4: Autocovariance estimates 30.7.2012 UTC

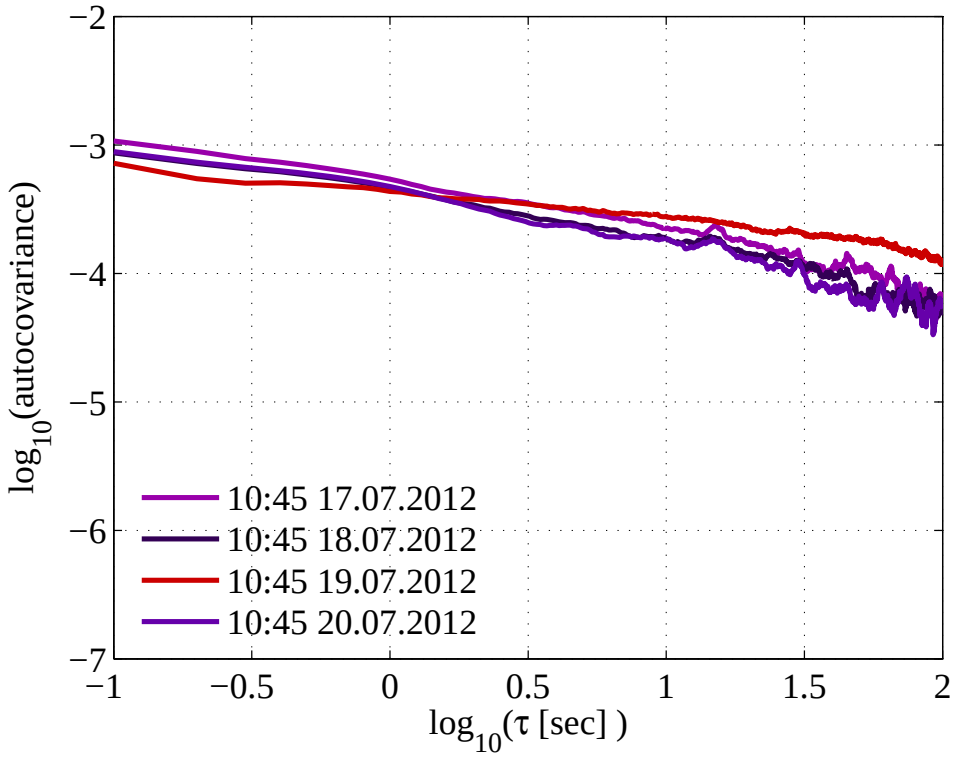


Figure A.5: Autocovariance estimates at 10:45 UTC

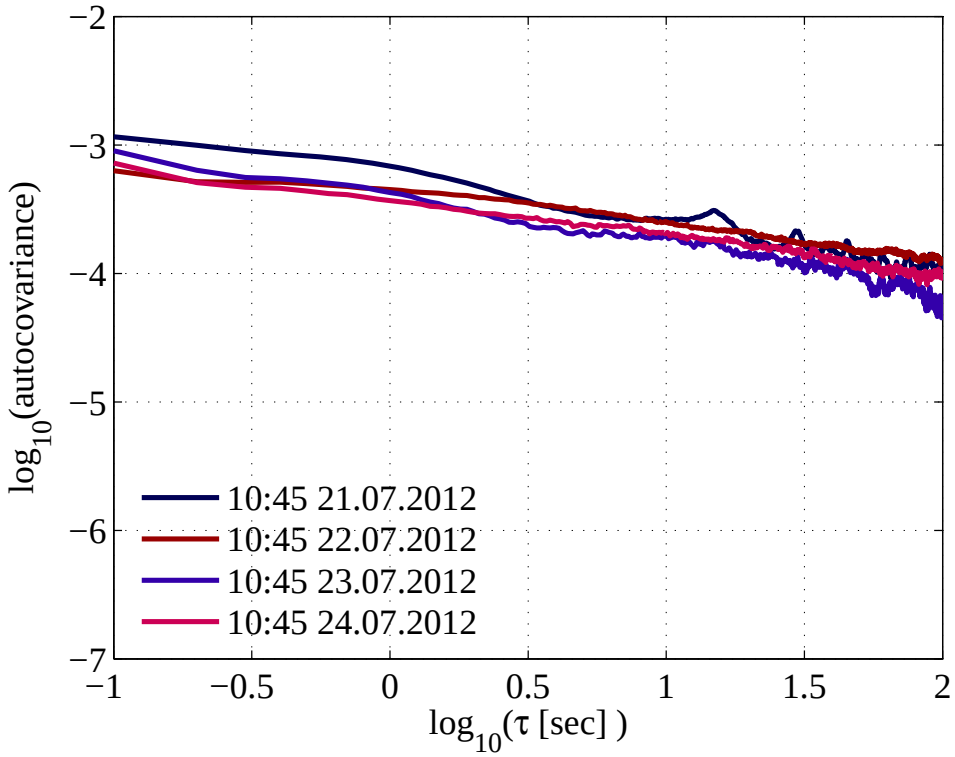


Figure A.6: Autocovariance estimates at 10:45 UTC

BIBLIOGRAPHY

[1] M. Abramowitz and I. Stegun. *Handbook of Mathematical Functions*. Dover, December 1964.

[2] R. Agrawal, R. L. Cruz, C. M. Okino, and R. Rajan. A Framework for Adaptive Service Guarantees. In *Proc. of Allerton Conf. on Communications, Control and Computing*, pages 693–702, 1998.

[3] G. Appenzeller, I. Keslassy, and N. McKeown. Sizing Router Buffers. *SIGCOMM Comput. Commun. Rev.*, 34(4):281–292, August 2004.

[4] F. Baccelli, G. Cohen, G. J. Olsder, and J.-P. Quadrat. *Synchronization and Linearity: An Algebra for Discrete Event Systems.* John Wiley & Sons, 1992.

[5] F. Baccelli, S. Machiraju, D. Veitch, and J. Bolot. On Optimal Probing for Delay and Loss Measurement. In *Proc. of ACM IMC*, pages 291–302, October 2007.

[6] F. Baccelli, S. Machiraju, D. Veitch, and J. Bolot. The Role of PASTA in Network Measurement. *IEEE/ACM Trans. Networking*, 17(4):1340–1353, August 2009.

[7] N. Basher, A. Mahanti, A. Mahanti, C. Williamson, and M. Arlitt. A Comparative Analysis of Web and Peer-to-peer Traffic. In *Proc. of WWW*, pages 287–296, 2008.

[8] J. Beran. *Statistics for Long-Memory Processes*. Chapman & Hall/CRC, October 1994.

[9] R. Birke, M. Mellia, M. Petracca, and D. Rossi. Understanding VoIP from Backbone Measurements. In *Proc. of IEEE INFOCOM*, pages 2027–2035, May 2007.

[10] G. Bolch, S. Greiner, H. de Meer, and K. Trivedi. *Queueing Networks and Markov Chains: Modeling and Performance Evaluation with Computer Science Applications*. John Wiley & Sons, May 2006.

[11] R.-R. Boorstyn, A. Burchard, J. Liebeherr, and C. Oottamakorn. Statistical Service Assurances for Traffic Scheduling Algorithms. *IEEE J. Select. Areas Commun.*, 18(12):2651–2664, December 2000.

[12] J.-Y. Le Boudec and P. Thiran. *Network Calculus: A Theory of Deterministic Queuing Systems for the Internet*. Number 2050 in Lecture Notes in Computer Science. Springer, 2001.

[13] G. E. P. Box, G. M. Jenkins, and G. C. Reinsel. *Time Series Analysis: Forecasting and Control*. Wiley Series in Probability and Statistics. Wiley, 2008.

[14] Z. Bozakov, A. Rizk, and M. Fidler. H-Probe Software, 2012. Available at: `http://www.ikt.uni-hannover.de/h-probe`.

[15] R. Bringhurst. *The Elements of Typographic Style*. Version 2.5. Hartley & Marks, Publishers, 2002.

[16] A. Broido, R. King, E. Nemeth, and Kc Claffy. Radon Spectroscopy of Interpacket Delay. In *Proc. of IEEE High Speed Networking Workshop*, March 2003.

[17] T. Bu, N. Duffield, F. Lo Presti, and D. Towsley. Network Tomography on General Topologies. *ACM SIGMETRICS Perform. Eval. Rev.*, 30(1):21–30, June 2002.

[18] A. Burchard, J. Liebeherr, and S. D. Patek. A Min-Plus Calculus for End-to-End Statistical Service Guarantees. *IEEE Trans. Inform. Theory*, 52(9):4105–4114, September 2006.

[19] A. Burchard, J. Liebeherr, and F. Ciucu. On Superlinear Scaling of Network Delays. *IEEE/ACM Trans. Networking*, 19(4):1043–1056, August 2011.

[20] J. Cao and K. Ramanan. A Poisson Limit for Buffer Overflow Probabilities. In *Proc. of IEEE INFOCOM*, pages 994–1003, June 2002.

[21] C.-S. Chang. Stability, Queue Length and Delay. II. Stochastic Queueing Networks . In *Proc. of IEEE Conference on Decision and Control*, pages 1005–1010, December 1992.

[22] C.-S. Chang. Stability, Queue Length, and Delay of Deterministic and Stochastic Queueing Networks. *IEEE Transactions on Automatic Control*, 39(5):913–931, May 1994.

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

[24] J. Choe and N. B. Shroff. Queueing Analysis of High-Speed Multiplexers including Long-Range Dependent Arrival Processes. In *Proc. of IEEE INFOCOM*, pages 617–624, March 1999.

[25] F. Ciucu. Exponential Supermartingales for Evaluating End-to-end Backlog Bounds. *ACM SIGMETRICS Perform. Eval. Rev.*, 35(2):21–23, September 2007.

[26] F. Ciucu. *Scaling Properties in the Stochastic Network Calculus*. PhD thesis, Univ. of Virginia, August 2007.

[27] F. Ciucu and J. B. Schmitt. Perspectives on Network Calculus: No Free Lunch, but Still Good Value. *SIGCOMM Comput. Commun. Rev.*, 42(4):311–322, August 2012.

[28] F. Ciucu, A. Burchard, and J. Liebeherr. Scaling Properties of Statistical End-to-end Bounds in the Network Calculus. *IEEE/ACM Trans. Networking*, 14(6): 2300–2312, June 2006.

[29] F. Ciucu, J. B. Schmitt, and H. Wang. On Expressing Networks with Flow Transformations in Convolution-form. In *Proc. of IEEE INFOCOM*, pages 1979–1987, April 2011.

[30] I. Cornfeld, Y. Sinai, and S. Fomin. *Ergodic Theory*. Springer, 1982.

[31] D. Cox and P. Lewis. *The Statistical Analysis of Series of Events*. Methuen's Statistical Monographs, 1966.

[32] M. Crovella and A. Bestavros. Self-similarity in World Wide Web traffic: Evidence and Possible Causes. *IEEE/ACM Trans. Networking*, 5(6):835–846, December 1997.

[33] R. L. Cruz. A Calculus for Network Delay, Part I and II: Network Elements in Isolation and Network Analysis. *IEEE Trans. Inform. Theory*, 37(1):114–141, January 1991.

[34] R. L. Cruz. Quality of Service Management in Integrated Services Networks. In *Proc. of 1st Research Review, Center of Wireless Communication, UCSD*, June 1996.

[35] R. L. Cruz. SCED+: Efficient Management of Quality of Service Guarantees. In *Proc. of IEEE INFOCOM*, pages 625–634, March 1998.

[36] R. L. Cruz and C. M. Okino. Service Guarantees for Window Flow Control. In *Proc. of Allerton Conf. on Communications, Control and Computing*, October 1996.

[37] P. Doukhan. *Mixing: Properties and Examples*. Lecture Notes in Statistics. Springer, 1994.

[38] P. Doukhan, G. Oppenheim, and M. Taqqu. *Theory and Applications of Long-Range Dependence*. Birkhäuser Boston, 2003.

[39] C. Dovrolis, P. Ramanathan, and D. Moore. Packet-dispersion Techniques and a Capacity-estimation Methodology. *IEEE/ACM Trans. Networking*, 12(6): 963–977, December 2004.

[40] N. G. Duffield and N. O'Connell. Large Deviations and Overflow Probabilities for the General Single-Server Queue, with Applications. *Math. Proc. Camb. Phil. Soc.*, 118(2):363–375, September 1995.

[41] A. Elwalid and D. Mitra. Effective Bandwidth of General Markovian Traffic Sources and Admission Control of High Speed Networks. *IEEE/ACM Trans. Networking*, 1(3):329–343, June 1993.

[42] A. K. Erlang. The Theory of Probabilities and Telephone Conversations. *Nyt Tidsskrift for Matematik*, 20(B):33–39, 1909.

[43] A. K. Erlang. Solution of some Problems in the Theory of Probabilities of Significance in Automatic Telephone Exchanges. *Elektrotkeknikeren*, 13:138–155, 1917.

[44] A. Erramilli, O. Narayan, and W. Willinger. Experimental Queueing Analysis with Long-range Dependent Packet Traffic. *IEEE/ACM Trans. Networking*, 4(2): 209–223, April 1996.

[45] A. Feldmann, A. C. Gilbert, P. Huang, and W. Willinger. Dynamics of IP traffic: A Study of the Role of Variability and the Impact of Control. *SIGCOMM Comput. Commun. Rev.*, 29(4):301–313, August 1999.

[46] M. Fidler. An End-to-End Probabilistic Network Calculus with Moment Generating Functions. In *Proc. of IEEE/ACM of IWQoS*, pages 261–270, June 2006.

[47] M. Fidler. A Survey of Deterministic and Stochastic Service Curve Models in the Network Calculus. *IEEE Communications Surveys and Tutorials*, 12(1):59–86, 2010.

[48] N. Fonseca, G. Mayor, and C. Neto. On the Equivalent Bandwidth of Self-similar Sources. *ACM Trans. Model. Comput. Simul.*, 10(2):104–124, April 2000.

[49] A. Ganesh, N. O'Connell, and D. Wischik. *Big Queues*. Springer, 2004.

[50] R. J. Gibbens and P. J. Hunt. Effective Bandwidths for the Multi-type UAS Channel. *Queueing Systems*, 9:17–28, October 1991.

[51] P. Gill, M. Arlitt, Z. Li, and A. Mahanti. Youtube Traffic Characterization: A View from the Edge. In *Proc. of ACM IMC*, pages 15–28, October 2007.

[52] P. W. Glynn and W. Whitt. Logarithmic Asymptotics For Steady-State Tail Probabilities in a Single-Server Queue. pages 131–156, January 1994.

[53] G. Grimmet and D. Stirzaker. *Probability and Random Processes*. Oxford University Press, 2001.

[54] H. Gupta, A. Mahanti, and V. Ribeiro. Revisting Coexistance of Poissonity and Self-similarity in Internet Traffic. In *Proc. of IEEE/ACM MASCOTS*, September 2009.

[55] G. He and J. Hou. On Exploiting Long Range Dependency of Network Traffic in Measuring Cross Traffic on an End-to-end Basis. In *Proc. of IEEE INFOCOM*, pages 1858–1868, March 2003.

[56] H. E. Hurst. Long-term Storage Capacity of Reservoirs. *Transactions of the American Society of Civil Engineers*, 116:770–799, 1951.

[57] J. Hüsler and V. Piterbarg. Extremes of a Certain Class of Gaussian Processes. *Stochastic Processes and their Applications*, 83:257–271, October 1999.

[58] Future Internet Lab. URL `http://www.filab.uni-hannover.de/`.

[59] ISO/IEC Standard 11172-2 - Information Technology. Coding of Moving Pictures and Associated Audio for Digital Storage Media at up to about 1,5 Mbit/s – Part 2: Video, 1993.

[60] J. R. Jackson. Networks of Waiting Lines. *Operations Research*, 5(4):518–521, August 1957.

[61] V. Jacobson. Pathchar: A Tool to Infer Characteristics of Internet Paths., April 1997. URL `ftp://ftp.ee.lbl.gov/pathchar/`.

[62] M. Jain and C. Dovrolis. End-to-end Available Bandwidth: Measurement Methodology, Dynamics, and Relation with TCP Throughput. *IEEE/ACM Trans. Networking*, 11:537–549, August 2003.

[63] Y. Jiang and P. J. Emstad. Analysis of Stochastic Service Guarantees in Communication Networks: A Traffic Model. In *Proc. of International Teletraffic Congress (ITC)*, pages 2337–2346, August 2005.

[64] Y. Jiang and Y. Liu. *Stochastic Network Calculus*. Springer, September 2008.

[65] F. P. Kelly. Effective Bandwidths at Multi-class Queues. *Queueing Systems*, 9: 5–16, October 1991.

[66] F. P. Kelly. Notes on Effective Bandwidths. Number 4 in Royal Statistical Society Lecture Notes, pages 141–168. Oxford University, 1996.

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

[68] L. Kleinrock. *Queueing Systems, Volume 1: Theory*. John Wiley & Sons, 1975.

[69] E. W. Knightly. Second Moment Resource Allocation in Multi-service Networks. *ACM SIGMETRICS Perform. Eval. Rev.*, 25(1):181–191, June 1997.

[70] A. Kumar, D. Manjunath, and J. Kuri. *Communication Networking: An Analytical Approach*. Morgan Kaufmann Publishers Inc., 2004.

[71] J. Kurose. On Computing Per-Session Performance Bounds in High-Speed Multi-Hop Computer Networks. *ACM SIGMETRICS Perform. Eval. Rev.*, 20(1): 128–139, June 1992.

[72] W. Leland, M. Taqqu, W. Willinger, and D. Wilson. On the Self-similar Nature of Ethernet Traffic (Extended Version). *IEEE/ACM Trans. Networking*, 2(1):1–15, February 1994.

[73] S. J. Leybourne and B. P. M. McCabe. A Consistent Test for a Unit Root. *Journal of Business and Economic Statistics*, 12(2):157–166, April 1994.

[74] C. Li, A. Burchard, and J. Liebeherr. A Network Calculus with Effective Bandwidth. *IEEE/ACM Trans. Networking*, 15(6):1442–1453, December 2007.

[75] J. Liebeherr, D. E. Wrege, and D. Ferrari. Exact Admission Control for Networks with a Bounded Delay Service. *IEEE/ACM Trans. Networking*, 4(6): 885–901, December 1996.

[76] J. Liebeherr, S. D. Patek, and A. Burchard. Statistical Per-Flow Service Bounds in a Network with Aggregate Provisioning. In *Proc. of IEEE INFOCOM*, pages 1680–1690, March 2003.

[77] J. Liebeherr, A. Burchard, and F. Ciucu. Delay Bounds for Networks with Heavy-Tailed and Self-Similar Traffic. Technical Report arXiv:0911.3856v1, November 2009.

[78] J. Liebeherr, M. Fidler, and S. Valaee. A System-Theoretic Approach to Bandwidth Estimation. *IEEE/ACM Trans. Networking*, 18(4):1040–1053, August 2010.

[79] J. Liebeherr, A. Burchard, and F. Ciucu. Delay Bounds in Communication Networks With Heavy-Tailed and Self-Similar Traffic. *IEEE Trans. Inform. Theory*, 58(2):1010–1024, February 2012.

[80] X. Liu, K. Ravindran, and D. Loguinov. What Signals Do Packet-pair Dispersions Carry? In *Proc. of IEEE INFOCOM*, pages 281–292, March 2005.

[81] X. Liu, K. Ravindran, and D. Loguinov. A Queueing-Theoretic Foundation of Available Bandwidth Estimation: Single-Hop Analysis. *IEEE/ACM Trans. Networking*, 15(4):918–931, August 2007.

[82] P. Loiseau, P. Goncalves, G. Dewaele, P. Borgnat, P. Abry, and P. Primet. Investigating Self-Similarity and Heavy-Tailed Distributions on a Large-Scale Experimental Facility. *IEEE/ACM Trans. Networking*, 18(4):1261–1274, August 2010.

[83] R. Lübben, M. Fidler, and J. Liebeherr. A Foundation for Stochastic Bandwidth Estimation of Networks with Random Service. In *Proc. of IEEE INFOCOM*, pages 1817–1825, April 2011.

[84] S. Machiraju, D. Veitch, F. Baccelli, and J. Bolot. Adding Definition to Active Probing. *SIGCOMM Comput. Commun. Rev.*, 37(2):17–28, March 2007.

[85] B. B. Mandelbrot. *Fractals: Form, Chance, and Dimension*. W. H. Freeman and Company, 1977.

[86] B. B. Mandelbrot and J. Van Ness. Fractional Brownian Motions, Fractional Noises and Applications. *SIAM Review*, 10(4):422–437, October 1968.

[87] B. B. Mandelbrot and J. R. Wallis. Noah, Joseph and Operational Hydrology. *Water Resources Research*, 4(5):909–918, 1968.

[88] M. Mandjes. *Large Deviations for Gaussian Queues*. John Wiley & Sons, 2007.

[89] S. Mao and S. Panwar. A Survey of Envelope Processes and their Applications in Quality of Service Provisioning. *IEEE Communications Surveys Tutorials*, 8 (3):2–20, 2006.

[90] L. Massoulie and A. Simonian. Large Buffer Asymptotics for the Queue with FBM Input. *Applied Probability*, 36:894–906, 1999.

[91] G. Mayor and J. Silvester. Time Scale Analysis of an ATM Queueing System with Long-Range Dependent Traffic. In *Proc. of IEEE INFOCOM*, pages 205–212, April 1997.

[92] R. Mazumdar. *Performance Modeling, Loss Networks, and Statistical Multiplexing*. Morgan & Claypool, 2010.

[93] B. Melamed and W. Whitt. On Arrivals that See Time Averages. *Operations Research*, 38(1):156–172, February 1990.

[94] O. Narayan. Exact Asymptotic Queue Length Distribution for Fractional Brownian Traffic. *Advances in Performance Analysis*, 1:39–64, 1998.

[95] I. Norros. A Storage Model with Self-similar Input. *Queueing Systems*, 16(3): 387–396, September 1994.

[96] I. Norros. On the Use of Fractional Brownian Motion in the Theory of Connectionless Networks. *IEEE J. Select. Areas Commun.*, 13(6):953–962, August 1995.

[97] A. Papoulis and S. Pillai. *Probability, Random Variables and Stochastic Processes*. McGraw-Hill, 2002.

[98] A. K. Parekh and R. G. Gallager. A Generalized Processor Sharing Approach to Flow Control in Integrated Services Networks: The Single-Node Case. *IEEE/ACM Trans. Networking*, 1(3):344–357, June 1993.

[99] K. Park and W. Willinger. *Self-similar Network Traffic and Performance Evaluation*. Wiley, 2000.

[100] V. Paxson and S. Floyd. Wide-Area Traffic: The Failure of Poisson Modeling. *IEEE/ACM Trans. Networking*, 3(3):226–244, June 1995.

[101] V. Paxson, G. Almes, J. Mahdavi, and M. Mathis. RFC2330 - Framework for IP Performance Metrics, 1998. Available at: `http://www.tools.ietf.org/html/rfc2330`.

[102] A. Philippe and M. Viano. Random Sampling of Long-memory Stationary Processes. *Journal of Statistical Planning and Inference*, 140:1110–1124, May 2010.

[103] J.-Y. Qiu and E. W. Knightly. Inter-class Resource Sharing using Statistical Service Envelopes . In *Proc. of IEEE INFOCOM*, pages 1404–1411, March 1999.

[104] V. Ribeiro, M. Coates, R. Riedi, S. Sarvotham, B. Hendricks, and R. Baraniuk. Multifractal Cross-Traffic Estimation. In *Proc. of ITC Conference on IP Traffic, Modeling and Management*, September 2000.

[105] V. Ribeiro, R. Riedi, R. Baraniuk, J. Navratil, and L. Cottrell. pathChirp: Efficient Available Bandwidth Estimation for Network Paths. In *Proc. of Passive and Active Measurement Workshop (PAM)*, April 2003.

[106] V. Ribeiro, R. Riedi, and R. Baraniuk. Multiscale Queuing Analysis. *IEEE/ACM Trans. Networking*, 14(5):1005–1018, October 2006.

[107] A. Rizk and M. Fidler. Statistical End-to-end Performance Bounds for Networks under Long Memory FBM Cross Traffic. Technical Report arXiv:0909.0633v1, September 2009.

[108] A. Rizk and M. Fidler. Statistical End-to-end Performance Bounds for Networks under Long Memory FBM Cross Traffic. In *Proc. of IEEE/ACM IWQoS*, June 2010.

[109] A. Rizk and M. Fidler. Sample Path Bounds for Long Memory FBM Traffic. In *Proc. of IEEE INFOCOM MC*, March 2010.

[110] A. Rizk and M. Fidler. Non-asymptotic End-to-end Performance Bounds for Networks with Long Range Dependent FBM Cross Traffic. *Computer Networks*, 56(1):127–141, 2012.

[111] A. Rizk and M. Fidler. On Multiplexing Models for Independent Traffic Flows in Single- and Multi-Node Networks. *IEEE Transactions on Network and Service Management*, 10(1):1–14, March 2013.

[112] A. Rizk, Z. Bozakov, and M. Fidler. Estimating Traffic Correlations from Sampling and Active Network Probing. In *Proc. of IFIP Networking*, 2013. accepted for publication.

[113] S. M. Ross. *Introduction to Probability Models*. Academic Press, Inc., 2006.

[114] S. M. Ross and E. A. Peköz. *A Second Course in Probability*. ProbabilityBookstore.com, 2007.

[115] M. Roughan. A Comparison of Poisson and Uniform Sampling for Active Measurements. *IEEE J. Select. Areas Commun.*, 24(12):2299–2312, December 2006.

[116] M. Roughan, D. Veitch, and P. Abry. Real-time Estimation of the Parameters of Long-range Dependence. *IEEE/ACM Trans. Networking*, 8(4):467–478, August 2000.

[117] G. Samorodnitsky. *Long Range Dependence*. Foundations and Trends in Stochastic Systems. Now Publishers Inc., 2007.

[118] H. Sariowan, R. L. Cruz, and G. C. Polyzos. Scheduling for Quality of Service Guarantees via Service Curves. In *Proc. of IEEE ICCCN*, pages 512–520, September 1995.

[119] J. B. Schmitt and U. Roedig. Sensor Network Calculus - A Framework for Worst Case Analysis. In *Proc. of DCOSS*, pages 141–154, June 2005.

[120] J. B. Schmitt, F. A. Zdarsky, and M. Fidler. Delay Bounds under Arbitrary Multiplexing: When Network Calculus Leaves You in the Lurch... In *Proc. of IEEE INFOCOM*, pages 1669–1677, April 2008.

[121] D. Starobinski and M. Sidi. Stochastically Bounded Burstiness for Communication Networks. *IEEE Trans. Inform. Theory*, 46(1):206–212, January 2000.

[122] J. Strauss, D. Katabi, and F. Kaashoek. A Measurement Study of Available Bandwidth Estimation Tools. In *Proc. of ACM IMC*, pages 39–44, October 2003.

[123] M. Taqqu, V. Teverovsky, and W. Willinger. Estimators for Long-Range Dependence: An Empirical Study. *Fractals*, 3(4):785–798, 1995.

[124] M. Taqqu, W. Willinger, and R. Sherman. Proof of a Fundamental Result in Self-similar Traffic Modeling. *SIGCOMM Comput. Commun. Rev.*, 27(2):5–23, April 1997.

[125] M. Bin Tariq, A. Dhamdhere, C. Dovrolis, and M. Ammar. Poisson versus Periodic Path Probing (or, does PASTA Matter?). In *Proc. of ACM IMC*, pages 119–124, October 2005.

[126] D. Veitch and P. Abry. A Wavelet-Based Joint Estimator of the Parameters of Long-Range Dependence. *IEEE Trans. Inform. Theory*, 45(2):878–897, April 1999.

[127] W. Whitt. Tail Probabilities with Statistical Multiplexing and Effective Bandwidths in Multi-class Queues. *Telecommunication Systems*, 2(1):71–107, August 1993.

[128] W. Willinger, M. Taqqu, R. Sherman, and D. Wilson. Self-similarity through High-variability: Statistical Analysis of Ethernet LAN Traffic at the Source Level. *IEEE/ACM Trans. Networking*, 5(1):71–86, February 1997.

[129] R. Wolff. Poisson Arrivals See Time Averages. *Operations Research*, 30(2): 223–231, March 1982.

[130] O. Yaron and M. Sidi. Performance and Stability of Communication Networks via Robust Exponential Bounds. *IEEE/ACM Trans. Networking*, 1(3):372–385, June 1993.

[131] Q. Yin, Y. Jiang, S. Jiang, and P. Y. Kong. Analysis of Generalized Stochastically Bounded Bursty Traffic for Communication Networks. In *Proc. of IEEE LCN*, pages 141–149, November 2002.

[132] X. Yu, I. L.-J. Thng, Y. Jiang, and C. Qiao. Queueing Processes in GPS and PGPS with LRD Traffic Inputs. *IEEE/ACM Trans. Networking*, 13(3):676–689, June 2005.

OWN PUBLICATIONS

D. Dietrich, A. Rizk, and P. Papadimitriou. Multi-Domain Virtual Network Embedding with Limited Information Disclosure. In *Proc. of IFIP Networking*, May 2013. accepted for publication.

K. Mahmood, A. Rizk, and Y. Jiang. On the Flow-Level Delay of a Spatial Multiplexing MIMO Wireless Channel. In *Proc. of ICC*, pages 1–6, June 2011.

A. Rizk and M. Fidler. On the Identifiability of Link Service Curves from End-Host Measurements. In *Proc. of NET-COOP*, pages 53–61, September 2008.

A. Rizk and M. Fidler. Statistical End-to-end Performance Bounds for Networks under Long Memory FBM Cross Traffic. In *Proc. of IEEE/ACM IWQoS*, June 2010.

A. Rizk and M. Fidler. Sample Path Bounds for Long Memory FBM Traffic. In *Proc. of IEEE INFOCOM MC*, March 2010.

A. Rizk and M. Fidler. Leveraging Statistical Multiplexing Gains in Single- and Multi-hop Networks. In *Proc. of IEEE/ACM IWQoS*, pages 1–9, June 2011.

A. Rizk and M. Fidler. Non-asymptotic End-to-end Performance Bounds for Networks with Long Range Dependent FBM Cross Traffic. *Computer Networks*, 56(1): 127–141, 2012.

A. Rizk and M. Fidler. On Multiplexing Models for Independent Traffic Flows in Single- and Multi-Node Networks. *IEEE Transactions on Network and Service Management*, 10(1):1–14, March 2013.

A. Rizk, Z. Bozakov, and M. Fidler. Estimating Traffic Correlations from Sampling and Active Network Probing. In *Proc. of IFIP Networking*, May 2013. accepted for publication.

CURRICULUM VITAE

Name	Amr Rizk
Day of birth	24.08.1983
Education	
since 03/2009	Research assistant and PhD student: Institute of Communications Technology at Leibniz Universität Hannover.
	Research in the field of stochastic performance evaluation of communication networks. Member of the Emmy-Noether Research Group ProPerBounds - a DFG funded research project.
05/2008 – 02/2009	Research assistant and PhD student: Institute of Multimedia Communications at Technische Universität Darmstadt.
	Research in the field of stochastic performance evaluation of communication networks. Member of the Emmy-Noether Research Group ProPerBounds - a DFG funded research project.
10/2001 – 03/08	Studies: Electrical Engineering, Information Technology and Business Administration at Technische Universität Darmstadt.
	Diploma thesis at the Institute of Multimedia Communications at Technische Universität Darmstadt, titled *"Network Tomography using Min-Plus System Theory"*

	Student project at the Institute of Communication Technology at Technische Universität Darmstadt, titled *"Analysis of OFDM with and without Guard Interval as a Promising Candidate for Future Mobile Radio Systems"*
06/2001	High school: Deutsche Evangelische Oberschule, Cairo

Work Experience

04/2007 – 10/2007	Student assistant at the Multimedia Communications at Technische Universität Darmstadt
08/2006 – 03/2007	Student trainee at Robert Bosch GmbH, Automotive Aftermarket, Karlsruhe, Germany
11/2005 – 07/2006	Student assistant at the Multimedia Communications at Technische Universität Darmstadt

Teaching Experience

since 03/2009	Exercises on "Teletraffic Theory"
since 03/2009	Lab exercises on "Congestion Control"
since 05/2008	Co-advising several BA and MSc theses
05/2008 – 02/2009	Exercises "Communication Networks"

INDEX

COLOPHON

This thesis was typeset with LaTeX 2ε using Hermann Zapf's *Palatino* and *Euler* type faces. The listings are typeset in *Bera Mono*, originally developed by Bitstream, Inc. as "Bitstream Vera".

The typographic style is provided by André Miede. It is inspired by Bringhurst's genius as presented in *The Elements of Typographic Style* [Bringhurst, 2002]. It is available for LaTeX via CTAN as "`classicthesis`".

Final Version as of September 12, 2013.