Non-stationary Service Curves: Model and Estimation Method with Application to Cellular Sleep Scheduling

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

Doktor-Ingenieur

genehmigte

Dissertation

von

Dipl.-Math. Nico Becker, M. Sc.

geboren am 15. August 1984 in Guben

2021

1. Referent: Prof. Dr.-Ing. Markus Fidler
2. Referent: Prof. Dr.-Ing. Amr Rizk
Tag der Promotion: 12.02.2021

ABSTRACT

In today's computer networks, short-lived flows are predominant. Consequently, transient start-up effects such as the connection establishment in cellular networks have a significant impact on the performance. Although various solutions are derived in the fields of queuing theory, available bandwidths, and network calculus, the focus is, e.g., about the mean wake-up times, estimates of the available bandwidth, which consist either out of a single value or a stationary function and steady-state solutions for backlog and delay. Contrary, the analysis during transient phases presents fundamental challenges that have only been partially solved and is therefore understood to a much lesser extent.

To better comprehend systems with transient characteristics and to explain their behavior, this thesis contributes a concept of non-stationary service curves that belong to the framework of stochastic network calculus.

Thereby, we derive models of sleep scheduling including time-variant performance bounds for backlog and delay. We investigate the impact of arrival rates and different duration of wake-up times, where the metrics of interest are the transient overshoot and relaxation time. We compare a time-variant and a time-invariant description of the service with an exact solution. To avoid probabilistic and maybe unpredictable effects from random services, we first choose a deterministic description of the service and present results that illustrate that only the time-variant service curve can follow the progression of the exact solution. In contrast, the time-invariant service curve remains in the worst-case value.

Since in real cellular networks, it is well known that the service and sleep scheduling procedure is random, we extend the theory to the stochastic case and derive a model with a non-stationary service curve based on regenerative processes.

Further, the estimation of cellular network's capacity/ available bandwidth from measurements is an important topic that attracts research, and several works exist that obtain an estimate from measurements. Assuming a system without any knowledge about its internals, we investigate existing measurement methods such as the prevalent rate scanning and the burst response method. We find fundamental limitations to estimate the service accurately in a time-variant way, which can be explained by

the non-convexity of transient services and their super-additive network processes.

In order to overcome these limitations, we derive a novel two-phase probing technique. In the first step, the shape of a minimal probe is identified, which we then use to obtain an accurate estimate of the unknown service.

To demonstrate the minimal probing method's applicability, we perform a comprehensive measurement campaign in cellular networks with sleep scheduling (2G, 3G, and 4G). Here, we observe significant transient backlogs and delay overshoots that persist for long relaxation times by sending constant-bit-rate traffic, which matches the findings from our theoretical model. Contrary, the minimal probing method shows another strength: sending the minimal probe eliminates the transient overshoots and relaxation times.

Keywords: Cellular networks, sleep scheduling, DRX, network calculus, service curves, non-stationary service curves, transient backlog, transient delay

ZUSAMMENFASSUNG

In den heutigen Computernetzwerken sind kurzlebige Ströme vorherrschend. Folglich haben transiente Anlaufeffekte wie der Verbindungsaufbau in zellularen Netzwerken einen erheblichen Einfluss auf die Leistung. Obwohl verschiedene Lösungen in den Bereichen Warteschlangentheorie, verfügbare Bandbreiten und Netzwerkkalkül hergeleitet wurden, liegt das Hauptaugenmerk, z.B. auf den mittleren Aufwachzeiten, Schätzungen der verfügbaren Bandbreite, die sich entweder aus einem Einzelwert oder einer stationären Funktion zusammensetzen, und stationäre Lösungen für Puffer und Latenzen. Im Gegensatz dazu stellt die Analyse während transienter Phasen grundlegende Herausforderungen dar, die nur teilweise gelöst sind und daher in weitaus geringerem Maße verstanden werden.

Um Systeme mit transienten Eigenschaften besser zu verstehen und ihr Verhalten zu erklären, trägt diese Arbeit ein Konzept von nicht-stationären Dienstkurven bei, das in den Rahmen der stochastischen Netzwerkkalkülberechnung gehört.

Darin leiten wir Modelle der Schlafplanung ab, die zeitvariante Leistungsgrenzen für Pufferrückstände und Verzögerungen beinhalten. Wir untersuchen die Auswirkungen von Ankunftsraten und die unterschiedliche Dauer der Aufwachzeiten, wobei die Metriken über das vorübergehende Überschwingen und die Relaxationszeit von besonderem Interesse sind. Wir vergleichen eine zeitvariante und eine zeitinvariante Beschreibung des Dienstes mit einer exakten Lösung. Um probabilistische und möglicherweise unvorhersehbare Effekte durch zufällige Dienste zu vermeiden, wählen wir zunächst eine deterministische Beschreibung der Dienstkurve und präsentieren Ergebnisse, die veranschaulichen, dass nur die zeitvariante Dienstkurve dem Verlauf der exakten Lösung folgen kann. Im Gegensatz dazu verbleibt die zeitinvariante Dienstkurve im schlimmstmöglichen Wert.

Da es bekannt ist, dass in realen zellularen Netzwerken das Dienst- und Schlafplanungsverfahren zufällig ist, weiten wir die Theorie auf den stochastischen Fall aus und leiten ein Modell mit einer nicht-stationären Dienstkurve ab, die auf regenerativen Prozessen basiert.

Darüber hinaus ist die Schätzung aus Messungen der Kapazität/verfügbaren Bandbreite von zellularen Netzwerken ein wichtiges Thema, das die Forschung anzieht. Hier gibt es mehrere Arbeiten, die eine Schätzung aus Messungen erhalten. Ausgehend von einem Blackbox-System, untersuchen wir bestehende Messmethoden wie das vorherrschende Rate Scanning und die Burst Response-Methode. Wir haben grundlegende Einschränkungen bei der zeitvarianten Schätzung des Dienstes gefunden, die sich durch die Nicht-Konvexität der transienten Dienste und ihrer super-additiven Netzwerkprozesse begründen lassen.

Um diese Einschränkungen zu überwinden, leiten wir ein neuartiges, zweiphasiges Probe-Verfahren her. Im ersten Schritt wird die Form einer minimalen Probe iden-

tifiziert, die wir dann verwenden, um eine genaue Schätzung der unbekannten Dienste zu erhalten.

Um die Anwendbarkeit der Methode der minimalen Probe zu demonstrieren, haben wir eine umfassende Messkampagne in zellularen Netzwerken mit Schlafplanung durchgeführt (2G, 3G und 4G). Hier beobachten wir signifikante transiente Pufferrückstände und Verzögerungsüberschreitungen, die für lange Relaxationszeiten fortbestehen, indem wir Verkehr mit konstanter Bitrate senden, was den Erkenntnissen aus unserem theoretischen Modell entspricht. Im Gegensatz dazu zeigt die Methode der minimalen Probe eine andere Stärke: Das Senden der minimalen Probe eliminiert die transienten Überschwinger und Relaxationszeiten.

Schlagwörter: Zellulare Netzwerke, Schlafplanung, DRX, Netzwerkkalkül, Dienstkurven, nicht-stationäre Dienstkurven, transiente Puffer, transiente Latenzen

CONTENTS

LIST OF FIGURES

LIST OF TABLES

$\otimes$	min-plus convolution operator
$\wedge$	minimum operator
$\delta(t)$	burst function
$\Delta(\tau, t)$	deviation from additivity
$\lvert x\rvert_+$	$\max\{0, x\}$
α	probability of a packet arrival as a Bernoulli trial
β	probability of packet size as geometric random variable
ε, ξ	violation probability
ω, ψ	sample paths
Ω	set of sample paths
ϕ	minimal sample paths
Ψ_t	selected set of Ω
$\mathbb{R}$	real numbers
ρ_A	rate parameter
θ, ρ	free parameter of moment generating function (MGF)
σ, σ_A	burstiness parameter
Θ	maximum latency
X^ξ	quantile
a^j	arrival time of j-th packet
A	server A
$A(t)$	arrival process
$A^\varepsilon(t)$	statistical arrival envelope
$A_i(t)$	i-th arrival flows
$A_{mp}(t)$	arrivals minimal probe
$\tilde{A}_{mp}(t)$	estimate of $A_{mp}(t)$
$B(t)$	backlog
$B^\varepsilon(t)$	backlog quantile
$B^\xi(r)$	steady-state backlog for rate r
$B^\xi(r, t)$	backlog quantile for rate r

br	burst response
d^j	departure time of j-th packet
D	server D
$D(t)$	departures process
$D_\omega(t)$	departure sample path
$E(t)$	deterministic arrival envelope
$\mathcal{F}_0$	set of all non-negative, non-decreasing function that pass the origin
f, g	bivariate functions
h	bivariate or univariate functions
$\inf$	infimum
$K(t)$	state of the Markov chain
l^j	j-th packet length
M	moment generating function (MGF)
mp	minimal probe
n_{GoP}	number of frames in group of picture
q	stochastic service rate as Bernoulli trials
q_{env}	burst size of a leaky-bucket envelope
$\mathbf{Q}(t)$	transition matrix
p	service parameter for random wake-up times
$\mathbf{P}(t)$	state distribution
$P_i(t)$	i-th regeneration point
$\mathbb{P}$	set of regeneration points
r	rate for CBR traffic
$R(t)$	deterministic service rate
rs	rate scanning
$S(t)$	service process
$S^i(t)$	i-th service process
$S_i(t)$	i-th regeneration process
$S^{net}(t)$	network service process
$S^\varepsilon(t)$	statistical service envelope, such as ε-effective and non-stationary service curve
$S^\varepsilon_{br}(t)$	non-stationary service curve from burst response method
$S^\varepsilon_{mp}(t)$	non-stationary service curve from minimal probe method

$S^{\varepsilon}_{rs}(t)$	non-stationary service curve from rate scanning method
$S_{\omega}(t)$	service sample path
$sl(r)$	stationary latency (rate)
τ, t	time instances
$tl(r)$	transient latency (rate)
VFT^{j}	j-th Virtual Finishing Time
$W(t)$	delay

3GPP	3rd Generation Partnership Project
AMPS	Advanced Mobile Phone Service
br	burst response
CBR	Constant Bit Rate
CCDF	Complementary Cumulative Distribution Function
CSMA/ CA	Carrier Sense Multiple Access/Collision Avoidance
DRX	Discontinuous Reception Mode
EBB	Exponentially Bounded Burstiness
EDF	Earliest Deadline First
EDGE	Enhanced Data Rates for GSM Evolution
eNodeB	Evolved Node B (Base Station)
FCFS	First-Come-First-Serve
FIFO	First-In-First-Out
fps	frames per seconds
foi	flow of interest
Gbps	Gigabit per seconds
GoP	Group of Pictures
GPRS	General Packet Radio Service
GPS	Generalized Processor Sharing
GR	Guarantee Rate
gSBB	generalized Stochastically Bounded Burstiness
GSM	Global System for Mobile Communications
HARQ	Hybrid Automatic Repeat Request
HSPA	High Speed Packet Access
HTTP	Hypertext Transfer Protocol
ICMP	Internet Control Message Protocol
iid	independent and identically distributed
ISP	Internet Service Provider
kbps	Kilobit per seconds
LRD	Long Range Dependent
LTE	Long Term Evolution

MAC	Media Access Control
Mbps	Megabit per seconds
MGF	Moment Generating Function
MIMO	Multiple-Input-Multiple-Output
mp	minimal probe
MPEG	Moving Picture Experts Group
NTP	Network Time Protocol
OFDM	Orthogonal Frequency Division Multiplexing
OWD	One Way Delay
PDCCH	Physical Downlink Control Channel
pmf	probability mass function
PMOO	Pay Multiplexing Only Once
PSRG	Packet Scale Rate Guarantee
RRC	Radio Resource Control
rs	rate scanning
RTT	Round Trip Time
SBB	Stochastically Bounded Burstiness
SFA	separated flow analysis
SINR	Signal-To-Interference-Noise-Ratio
SP	Static Priority
SNR	Signal-to-Noise-Ratio
SOTAT	State-Of-The-Art Transient Bounds
TBS	Transport Block Size
TCP	Transmission Control Protocol
TPG	Train of Packet Groups
TTI	Transmission Time Interval
UE	User Equipment
VFT	Virtual Finishing Time
VoIP	Voice over IP
WTB	Wireless Transient Bounds

Part I

DISSERTATION

1 INTRODUCTION

In today's computer networks, most Internet connections are short-lived [109], such as TCP streams, which persist for less than five seconds in about 50% of cases [82]. Consequently, the TCP slow-start has a relevant influence on performance where the initial congestion window is discussed time and time again[63]. Apart from that, there are many other transient effects which have a significant impact on the performance, e.g., in wireless networks the time it takes to converge for routing protocols[75], using polling strategies to save energy, or in cellular networks the signaling procedures and discontinuous reception mode (DRX) to increase the battery lifetime of mobile devices [21].

Considering the cellular networks, we have, in the past 15 years, a tremendous increase of mobile devices or user equipments (UEs)[1], where more than five billion people, i.e., 65% [155] of the world population, possess a device these days. Moreover, we know that in 2017 almost 80% of the Internet users are online through UEs [117]. Hence, they particularly use and generate mobile data to communicate with each other and use applications, which leads to a huge data mix traffic. To name a few, we have HTTP web browsing traffic, telephony over VoIP, video streaming, the down- or upload of files such as music and videos, as well as periodic refresh messages, e.g., for news and messenger applications. For a detailed discussion of mobile applications see [117].

To be able to receive or transmit data in cellular networks, the UE must be connected to the base station (eNodeB). In the case of LTE (4G), the UE monitors the Physical Downlink Control Channel (PDCCH) to receive paging messages from the eNodeB. Doing this procedure continuously suffers the battery and reduces their lifetime. The longer no data can be received or transmitted, the more likely it is that further continuous polling of the PDCCH will lead to unnecessary battery consumption. The aim of the DRX mode is to extend the battery lifetime of UEs by entering different types of sleep states in which the UE turns off power-intensive parts, e.g., radio interfaces and displays. Generally, a mobile in an LTE network is in one of two radio resource control (RRC) states, i.e., `RRC_IDLE`

1 In the sequel, we refer to smartphones and tablets as user equipment (UE).

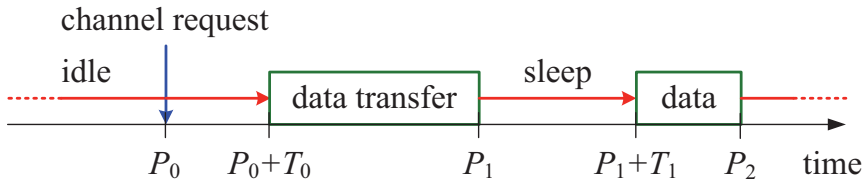


Figure 1.1: Sleep scheduling as implemented in discontinuous reception (DRX).

or RRC_CONNECTED [7]. In RRC_CONNECTED, the UE is connected to the eNodeB. Among other states in RRC_CONNECTED, the UE permanently monitors the channel draining the battery in the continuous reception mode. In RRC_IDLE, the UE does not possess an RRC connection [7]. Here, the UE has to establish one to the eNodeB, which consists of 16 to 19 signaling messages in uplink or downlink direction, respectively. This procedure leads to an additional delay of more than 100 ms, see Sec. 7.2.1. Consequently, data is buffered for later transmission while the mobile is asleep. This results in transient effects on backlog and delay that are non-negligible, see Sec. 7.4.1.

Fig. 1.1 illustrates a possible DRX sleep schedule process. In the beginning, the mobile is in an idle state. Due to a channel request at P_0, the mobile is awaking and takes T_0 units of time for activating. Subsequently, data can be transmitted. After the data transfer has taken place, the UE returns to sleep mode at P_1 and wakes up T_1 units later. The process is now repeated as often as required.

We model the sleep duration T_i in the next chapters, where we assume either deterministic or random wake-up times in Sec. 2.2.2, 3.1 and in Chap.5, 6, respectively.

However, the current literature mainly investigate the obvious trade-off for the DRX mode, i.e., between battery lifetime and wake-up delays where a mobile during sleep mode cannot be paged for incoming packets. Besides the time needed to establish a connection, it has to wait until it wakes up to monitor the PDCCH. Therefore, a packet intended for a sleeping UE has to be buffered in between, which in turn leads to an increased delay.

The trade-off for similar mechanisms in 2G and 3G is analyzed, e.g., in [116, 130]. For 4G, a lot of research is done by modeling the DRX states and mean wake-up time by semi-Markov chains [23, 144, 148, 156, 157]. Note that we consider the DRX mode in more detail in Sec. 7.2.1.

In addition to optimizing DRX cycles [60, 138], there are applications such as video and music streaming, where a continuous connection to the base station causes a permanent battery discharge at the UE that affects the user experience. In [134], the authors investigated that traffic shaping can

save up to 60% of the energy. The idea is to send bursts such that the mobile can switch off the radio interface in between. Although this method can save energy, it may lead to a massive increase in packet delays. For example, in one of our earlier works [21], we observed that sending traffic at a higher rate than the LTE system can serve, leads to a massive blowing up of packet delays, e.g., from 10 ms to more than 300 ms. Thus, for delay-sensitive applications, one should be careful by sending burst traffic and adapt the traffic according to the service a network provides.

To find the service of a network, measurement-based approaches are the preferred choice. The methods assume a system, as described by Lübben et.al. in [103].

Here, the author interprets a general system from the classical system theory as a computer network, where the inputs become arrivals, the system's responses are the service of the network and the outputs departures. Depending on what knowledge we have about the internals of the system, we divide the methods into grey- [15, 27] or black-box [103] models. In grey-box models, we already have certain information about the service, such as that it has a constant rate and a specific one-way delay (OWD). So from measurements, only these two parameters need to be determined, whereas a black box model does not provide any information at all.

Further classifications can be made by the general choice of arrival traffic. Do we estimate the service from existing traffic, or do we actively inject test traffic into the network? Methods using the first case are called passive [15, 32, 101] and the second case is called active measurement methods, where popular test traffic are packet pairs [136], i.e., two successive packets with a predefined spacing, packet trains [108], i.e., sending a series of packets with constant rate and packet chirps [120], which are packet trains where the rate increases within the train. The goal is to find the available bandwidth, i.e., the long-term average unused capacity of a system. An overview of the different methods can be found in [133, 136].

Even though the measurement methods used to estimate the available bandwidth in wired networks perform well, the accuracy of wireless transmission technologies is found to be poor. In [99], the authors described difficulties encountered in IEEE 802.11 wireless networks. More specifically, tools such as Spruce [136], TOPP [108], PathChirp [120], and many others are impacted by wireless networks where the mechanisms are affected by contention for channel access from other traffic. Although a new method

called WBest is being introduced to overcome these problems, we know that additional challenges need to be dealt with for cellular networks.

For example, in LTE, the resource allocation is done by the base station, where several packets are in one transport block due to concatenation or segmentation. As a consequence, methods, e.g., based on packet pairs, are not suitable anymore.

Generally, the known methods try to estimate the available bandwidth or maximum capacity, which results in a single value. Hence, it is naturally not possible to represent time-dependent characteristics. So they are not suitable to illustrate the influence of DRX mode on the service in mobile networks such as in LTE. We take a closer look on that in Sec.3.2.

Therefore, we find that a time-dependent service estimation to investigate, e.g., transient effects in cellular networks is not present so far. Moreover, the analysis of transient effects in computer networks is sparse. In queuing, theory results are mainly derived as steady-state solutions. As an example, we have an M|M|1 system where the solution follows from linear balance equations [126]. To investigate transient behavior, we have to solve a set of differential equations. Here, mainly numerical solutions exist [153], or the solutions are highly specific ones such as for the TCP congestion control [109].

In order to address these problems, we choose the theory of network calculus that provides the necessary foundations.

1.1 THE THEORY OF NETWORK CALCULUS

The network calculus is a framework established at the beginning of the 1990s. The seminal work of Cruz [53, 54] laid a foundation to model computer networks where several networks can be concatenated, and performance bound , e.g., backlog and delay, are derived. Generally, we consider a system as in Fig. 1.2. Here, several endpoints, such as servers, personal computers, printers, smartphones, and so on, are connected via a network. The network consists out of nodes that are, e.g., routers and switches, which are then modeled as a queuing system with a queue and a server. We marked a possible route where a server A sends data to a server D through the network. Due to the complexity of such networks, we simplify it for a moment and assume a system as in Fig. 1.3, where all possible nodes are combined into one system with a corresponding queue and a specific ser-

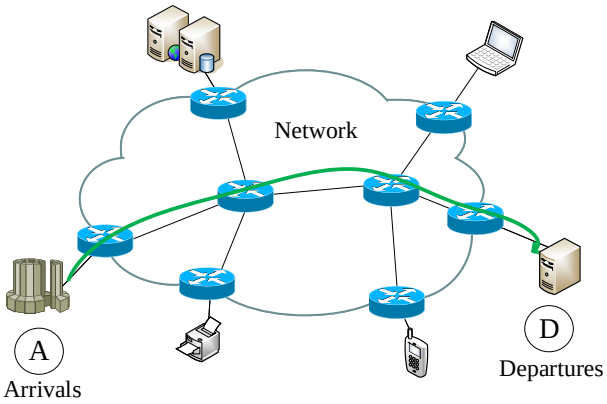


Figure 1.2: Simplification of the Internet

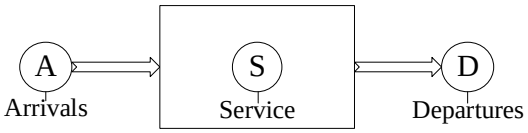


Figure 1.3: System Model

vice rate. Note that we use the terms network and system interchangeably during the whole work.

The idea is that data at the sender A in the form of, e.g., bits enter the network. We name this input arrivals A. Then, the arrivals are processed with a service S through the network and leave the system as output traffic at the server D, which we denote as departures D.

With network calculus, we are then able to derive performance bounds such as backlog and delay. Moreover, the deterministic network calculus[37, 52, 96] is a theory that can explore transient effects. Because time-variant and non-linear systems are replaced by time-invariant and linear bounds, only worst-case results are obtained, where e.g.the maximal backlog and delay is computed, as we will show in Sec. 2.1.

In order to take advantage of the statistical nature of data flows, the stochastic network calculus [31, 37, 47, 48, 68, 70, 73, 86, 98] is derived. Here, the knowledge about certain traffic characteristics such as the scheduling discipline and statistical multiplexing is used to obtain statistical guarantees of the type $P[backlog > x] \leqslant \varepsilon$ and $P[delay > x] \leqslant \varepsilon$. Although the stochastic network calculus takes time-variant systems into account, it

typically either assumes stationary random processes or uses stationary bounds. Hence, a complete analysis of transient phases is not possible.

However, apart from the derivation of performance bounds, the network calculus also provides a suitable framework for the estimation of the networks service. Depending on the assumptions such as linearity and stationarity, we can classify the existing work. Researches that deal with work-conserving systems, are [15, 27, 32, 140]. Assuming the system is min-plus linear we have deterministic, time-invariant models [10, 15, 27, 79, 101] or stochastic models in stationary systems [32, 105, 103, 140]. Note that these works, either assume time-invariance or stationarity where univariate functions are used to describe the system. Hence, a time-dependent analysis of the network or characterization of the service is not possible.

Anyway, in [11, 40], a time-variant system is derived that replaces the univariate functions with bivariate ones. The use of the notion of bivariate functions gives us a basis and enables us to extend the current literature to model changes over time accurately and to find time-variant or non-stationary service characteristics which occur, e.g., during the connection establishment in cellular networks such as LTE.

1.2 THESIS CONTRIBUTIONS

In this thesis, we contribute a non-stationary service curve model in the framework of network calculus. It enables us to analyze the transient behavior of computer networks and systems such as cellular networks. We derive solutions for systems with sleep scheduling and present results of the time-dependent formulation of performance bounds backlog and delay. Moreover, we estimate the service of unknown systems where we show the limitations of existing methods regarding a transient description of the service. A new two-phase measurement method is developed, which can estimate a non-stationary service within a defined accuracy. Further, we substantiate our findings from simulations by performing an extensive measurement campaign in cellular networks. The evaluation provides insights into transient service characteristics of 2G, 3G, and 4G networks. Next, we describe the contributions in more detail.

In the first place, we use the notion of time-variant systems as in [11, 40], i.e., to model changes over time we use bivariate instead of univariate functions. Then, we model systems with deterministic sleep scheduling, i.e., with fixed wake-up times and compare a time-invariant formulation

of the service [37] with a time-variant one. Here we can illustrate that both service descriptions show the same growth of performance bounds, such as the backlog during the sleep period. How the backlog is cleared after the start of service can only be explained by the time-variant model, whereas the time-invariant model remains in the worst-case. Due to our choice of a Poisson arrival process, we are able to derive statistical arrival envelopes, for example, by applying Chernoff's bound and compare the results with an exact solution. We observe that only the time-variant service follows the progression of the exact solution where a closer bound is achieved by using Doob's Martingale inequality. In addition to that, we investigate the influence of the arrival rate, service rate, and the duration of sleep cycles where the measures of interests are, e.g., the transient overshoot and the relaxation time, i.e., the time it takes to reach steady-state.

Secondly, we derive a non-stationary service curve based on the time-variant concept and the regenerative service processes followed by the time-dependent description of statistical performance bounds backlog and delay. It allows us to model systems with sleep scheduling, where the wake-up times and services are random. It lays the theoretical basis to investigate the transient and stationary behavior of systems with random wake-up times, such as the DRX mode in cellular networks. From our work in [21], we know that the transition from `RRC_IDLE` to `RRC_CONNECTED` state takes a random amount of time. Thus, this step from the deterministic to the stochastic network calculus is necessary to analyze, e.g., the specific implementation of sleep scheduling in real networks.

The third main contribution is about unknown systems and the estimate of their service curves from measurements of probe traffic. More precisely, our goal is to estimate a general service model of a linear system, with a time-varying, regenerative service that avoids specific assumptions about the network internals. Thereby, we consider known measurement methods, i.e., the rate scanning [103, 105] and burst response method [101] and refine them to estimate non-stationary service curves. In doing so, we come across additional difficulties due to the non-convexity and super-additivity of the service. To overcome these limitations, we are developing a novel method which we call minimal probing that estimates a non-stationary service curve and provides a measure of accuracy. The method consists of two steps. In the first step, it computes a probe that is minimal under certain conditions. The second step then consists of using this minimal probe and estimating a non-stationary service curve with a defined accuracy. Our method can reveal

the actual service progression of the network, that includes various effects such as transient delays due to sleep scheduling, stationary OWDs, time-dependent service rates, and service outages due to wireless transmission characteristics, see Sec. 7.4.2.

To analyze transient and stationary latencies, we explore a model with a server with vacations from the literature [37] and elaborate on the effects of transient and stationary latencies on the (super)-additivity of the service in detail. Our findings demonstrate that the service is additive in case of a transient latency, whereas it is super-additive in case of stationary latency, which is comparable to OWDs. Additionally, we show the deviation from additivity for networks of n systems with random sleep scheduling in series. Since real networks have non-zero OWDs, this result is essential for our measurements of production networks.

The fourth contribution is about a measurement-based evaluation of the DRX mode in cellular networks and includes the first practical validation of our new method in several real production networks. The campaign produces results in 2G, 3G, and 4G networks for two different providers at day and night. In order to provide statistically relevant results, we have fixed the position for each provider, the transmission technology, and the time of day and assume stationary channel conditions for each setup. Then, we measured for each setting the non-stationary service curve by performing up to 2000 measurements, each. As an example, for 4G and the provider ISP1 we estimated the service curves over ten nights, as described in Sec. 7.4.2, i.e., we send 100 bursts and estimate the minimal probe for $\varepsilon = 0.05$. Subsequently, we send the minimal probe 100 times, take the backlog $B(t)$ at $t = 1$ sec and compute the 0.95-quantile $B^{\varepsilon}(t)$. Out of this, we obtain the non-stationary service curve, as described in Sec.6.4. We repeat this procedure ten times and take the average over all estimates. Thus, we have 1000 measurements for the burst response and 1000 measurements for minimal probing. Taking the 0.95 confidence interval of the burst response estimates confirms that our new approach yields stable results. Moreover, from our results, we are able to explain the behavior of the DRX mode, i.e., the specific implementation of sleep scheduling in cellular networks. Out of the non-stationary service curve obtained from minimal probing, we extract a set of information, including the accuracy of the method, transient and stationary latencies, service outages, and capacity limits. Hence, in comparison to other methods that estimate stationary results or estimate

only a single value, e.g., for the available bandwidth, our method produces much more information.

We sum up that by taking into account, the measurement results from Chap. 7, we believe that the measurement results provide good evidence that our model of non-stationary service curves and the method for estimating the shape of the service curve is suitable for characterizing key aspects of mobile network service, such as stationary and transient delays and rate limitations.

1.3 THESIS OUTLINE

The remainder of this thesis is structured as follows. In Chap. 2 we introduce the network calculus, where we explain the main ideas and concepts of this theory. We describe the basics for the deterministic and stochastic network calculus. It lays the foundation for the next chapters.

Next, we consider the related work in Chap. 3. There, we divide the chapter into two parts. The first part considers the state-of-the-art of time-variant systems in network calculus, whereas the second part, is about measurement-based methods to estimate the service of unknown systems.

It illustrates the main difficulties of the current literature. To better formalise the limitations, we state the problems in Chap. 4.

In Chap. 5 we define the non-stationary service curve. We present a method of construction and derive performance bounds for backlog and delay. Also, we present models of systems with sleep scheduling.

To estimate the non-stationary service of unknown systems, we refine known methods in Chap. 6 and explain their limitations, which are due to the convexity and (super)-additivity of the service. We introduce a new and novel method, the minimal probing method that is able to overcome the limitations as shown, e.g., in simulations.

In Chap. 7, we do the first practical validation of the minimal probing method. There, we perform a massive measurement campaign in the cellular networks of 2G, 3G, and 4G for two different providers. From each service curve, we extract all possible information, such as the transient and stationary delays, the accuracy of the method, capacity limit, and outages.

Finally, we conclude the thesis in Chap. 8 and present further ideas and topics for future work.

2

SYSTEM MODELS IN NETWORK CALCULUS

In this chapter, we lay the foundation for deriving a time-dependent model, enabling us to analyze transient phases in cellular networks caused by, e.g., the DRX-cycles in LTE. In order to do so, we choose the network calculus framework. We summarize the theory and present the main results briefly. An extended elaboration of network calculus can be found in [37, 52, 96].

Recalling the system model as in Fig. 1.3 from Sec.1.1 the basic point of view is to consider data at time t that enters a system in form of an arrival function $A(t)$ which is processed with a service $S(t)$ and leave the system as departures $D(t)$. Thereby, we use the cumulative form of these functions. Thus, the cumulative arrival function $A(t)$ represents all data, e.g., the number of bits in the time interval $(0, t]$. Further we define no arrivals for $t \leqslant 0$ and consider only positive time-scales $t \geqslant 0$ with $A(0) = 0$. For reasons of causality the number of bits cannot be negative, and so it is straightforward that $A(t)$ for $t \geqslant 0$ is a non-negative, non-decreasing function .

We define $\mathcal{F}_0$ as the set of all non-negative, non-decreasing functions that pass the origin, i.e.,

$$\mathcal{F}_0 = \{f : f(t) \geqslant f(\tau) \geqslant 0 \quad \forall t \geqslant \tau, f(0) = 0\}, \tag{2.1}$$

such that $A(t) \in \mathcal{F}_0$. Here, $A(t)$ is the short-handed form to denote all bits in the interval $(0, t]$.

If we consider the interval $(\tau, t]$, where $t \geqslant \tau \geqslant 0$ we distinguish between univariate and bivariate functions. For the arrivals and the bivariate case we take all data into account which arrives in the interval $(\tau, t]$ such that for this concrete time-interval we have that $A(\tau, t) = A(t) - A(\tau)$, where we assume $A(\tau, t)$ to be additive. Note that the set $\mathcal{F}_0$ has to be slightly adapted to bivariate functions , i.e., $\mathcal{F}_0 = \{f : f(\tau, t) \geqslant f(\tau, \nu) \geqslant 0 \; \forall t \geqslant \nu \geqslant \tau, f(t, t) = 0 \; \forall t\}$ see [51, 70]. In the univariate case we do not consider the time instances τ and t explicitly, where $t \geqslant \tau \geqslant 0$. We are only interested in the length of the interval. Hence, we get the arrivals by $A(t - \tau)$. Generally, we consider univariate functions as the time-invariant and bivariate functions as the time-variant case. It might look like a small change, but as we will see, it has significant implications for the analysis of transient effects.

In the same way, we define the cumulative departures $D(t)$, $D(t-\tau)$, $D(\tau,t)$ and services $S(t)$, $D(t-\tau)$, $S(\tau,t)$ provided by the network for $t \geqslant \tau \geqslant 0$ at the appropriate intervals.

Next, we can compute the departures $D(t)$ from the arrivals $A(t)$ and the service $S(\tau,t)$ with the following assumptions. First, we assume a lossless system. This corresponds to a buffer with sufficient space to store the incoming data. An extension to lossy systems can be found in [51], where the author introduced a traffic clipper that discards non-conforming data. Another way is shown in [38], where packets are dropped when delays or buffering constraints are violated. An overview of lossy systems can be found in [37, 96]. Throughout this work, we assume that time is discrete. Continuous-time requires an additional discretization step, see [47]. Apart from that it is reasonable to assume $A(t) \geqslant D(t)$ for all $t \geqslant 0$. Last but not least, arrivals from several flows $A_i(t)$ for $i = 1, \ldots, m$ are multiplexed to one flow $A(t)$ by summing up the individual flows, i.e., $A(t) = \sum_{i=1}^{m} A_i(t)$.

We then use the concept of service curves as in [12, 13, 36, 55, 57, 95, 127] to obtain the departure guarantee [11, 37, 38, 40] for the service process $S(\tau,t)$. Hence, we get

$$D(t) \geqslant \inf_{\tau\in[0,t]} \{A(\tau) + S(\tau,t)\} =: A \otimes S(t). \tag{2.2}$$

In Eq. (2.2) we defined the operator $\otimes$ which is known as convolution under min-plus algebra [37, 96].

In case we choose a time-invariant service process, i.e., $S(\tau,t) = S(t-\tau)$, Eq. (2.2) becomes

$$D(t) \geqslant \inf_{\tau\in[0,t]} \{A(\tau) + S(t-\tau)\}. \tag{2.3}$$

Performance Bounds

The service guarantee of Eq. (2.2) enables us to derive performance bounds such as backlog and delay. The backlog can be considered as the amount of data in the system, which includes the data in the queue as well as all in transmission. It can be computed as the vertical deviation of arrivals and departures at a given time $t \geqslant 0$. Thus,

$$B(t) = A(t) - D(t) \tag{2.4}$$

at some time $t \geqslant 0$. An upper bound of the backlog follows immediately by Eq. (2.2) see [96], i.e.,

$$B(t) \leqslant \sup_{\tau \in [0,t]} \{A(\tau, t) - S(\tau, t)\}. \tag{2.5}$$

Similarly, the first-come-first-serve (FCFS) delay at time t can be viewed as the horizontal deviation. It is defined as

$$W(t) = \inf\{w \geqslant 0 : A(t) \leqslant D(t+w)\}. \tag{2.6}$$

Again, an upper bound follows from Eq. (2.2) as

$$W(t) \leqslant \inf\left\{w \geqslant 0 : \sup_{\tau \in [0,t]} \{A(\tau, t) - S(\tau, t+w)\} \leqslant 0\right\}. \tag{2.7}$$

Assuming time-invariant functions simplify the computation of backlog and delay as in Eq.(2.14).

Comparisons to the system theory and properties of the min-plus convolution

Despite the inequality, which will be an equal sign later on, at this point, the attentive reader may recognize similarities of Eq. (2.3) to the system theory. There, we map the incoming, arriving signals to the outgoing departures by the impulse response function. Most commonly, the system is linear and time-invariant, see 3 for the definitions. As before, a time-invariant impulse response means that it does not depend on the explicit time-instances itself, but to the amount of time. Furthermore, if the system is linear, then the impulse response describes the system completely. In the discrete case, the mapping is done by

$$D(t) = \sum_{\tau} A(\tau) \cdot \mathbb{S}(t-\tau) =: (A * \mathbb{S})(t), \tag{2.8}$$

where $*$ is the convolution operator and $\mathbb{S}(t)$ is the system response to the Dirac unity impulse $\delta(t)$. In the discrete case Dirac impulse function becomes the Kronecker δ-function [78], which means that $\delta(t) = 1$ for $t = 0$ and zero otherwise, such that it is the neutral element of the convolution. For continuous functions the sum becomes an integral [115] and $\int_{-\infty}^{\infty} \delta(t)dt = 1$, where $\delta(0) = \infty$ and zero otherwise.

By comparing the convolution in system theory, the sum becomes an infimum and the multiplication a plus sign, respectively, under min-plus

convolution. Despite these differences, there are also similarities. As in the case of system theory, we have an exact service curve in min-plus algebra if and only if the system is time-invariant and linear. Again, these definitions can be found in the appendix, see definition 5. Shortly spoken, min-plus linearity means that any (min-plus linear) combination of input signals $c_1 + A_1 \wedge c_2 + A_2$ results in the output signal $c_1 + D_1 \wedge c_2 + D_2$, where $\wedge$ denotes the minimum operator and c_1 and c_2 are constants. So, under the min-plus algebra, the service curve is the response of the system. Similarly to the Dirac impulse function we can define a neutral element of the min-plus convolution, i.e., $\delta \otimes S(t) = S(t)$, where

$$\delta(t) = \begin{cases} 0 & \text{for } t = 0, \\ \infty & \text{for } t > 0. \end{cases} \tag{2.9}$$

It can be seen as a burst impulse such that $A(t) = \delta(t)$. It reveals the service $S(0, t)$ for all $t \geqslant 0$ as the system's burst response. We will use this property in a further course to develop a new measurement method.

From a mathematical point of view the algebraic structures $(\mathbb{R} \cup \infty, \wedge, +)$ and $(\mathcal{F}_0, \wedge, \otimes)$ are commutative dioids. They are not rings because there exists no inverse element for the minimum ($\wedge$)-operator. A list of properties can be found in the appendix 6 or for deeper insights into this algebra see [17, 37, 96]. For our purposes, the most important properties are associativity, distributivity, and commutativity. The operators $\wedge, \otimes, +$ are all associative and commutative, apart from $\otimes$ that is commutative not in general but only in the univariate case. Moreover, $+, \otimes$ are distributive with repect to $\wedge$. Particularly noteworthy is the associativity of $\otimes$ as it facilitates the concatenation of systems in series. Further properties follow in the corresponding sections. For more information about similarities from system theory to network calculus, see [37, 96].

Hereinafter, we follow the classification from [46] where the author subclassifies into deterministic and stochastic network calculus. The classical deterministic network calculus replaces possible time-variant functions, which are crucial for a time-dependent analysis of transient phases, by time-invariant functions, e.g., to express the service curve in Eq. (2.2). As a consequence, we are able to consider transient phases, but as we will show below, the analysis remains in the worst case due to invariant functions. In contrast, stochastic network calculus includes time-variant services and, as described in [46] by using stationary random processes or stationary

bounds specified by invariant functions, again. Please note that we present among others results from our own works [19].

2.1 DETERMINISTIC NETWORK CALCULUS

In the following, we describe the basic principles for deterministic network calculus and give examples for service curves, performance bounds, scheduling disciplines, end-to-end analysis of several systems in a row, and envelope functions for the arrivals and service curves. A good overview of the deterministic network calculus we get from the work of Le Boudec [94, 96, 97].

Recalling the statement that the deterministic network calculus uses time-invariant functions we start with Eq. (2.3), where the system offers a lower service curve $S(t)$ to an arrival process $A(t)$ to obtain the departure guarantee $D(t)$. The lower service curve is a lower bound on the amount of service the arrivals receive. Note, that Eq. (2.3) becomes equal, if $S(t-\tau)$ is linear. Then, the lower bound on the departures is also an upper bound and the service curve is called exact. Hence,

$$D(t) = \inf_{s\in[0,t]} \{A(s) + S(t-s)\}. \tag{2.10}$$

Examples for exact service curves are, e.g., constant rate server where $S(t) = R \cdot t$ with capacity R, and a leaky-bucket-shaper, see Eq. (2.15).Further results of service curves are formulated in [12, 13, 36, 55, 58, 95, 127].

Apart from the different types of service curves mentioned above, the concept even allows the characterization of different scheduling algorithms. Besides the fact that we do not make any assumptions regarding a scheduling discipline, we give a brief list of some well-known types. To be precise, we find more information about first-in-first-out (FIFO) in [37, 57, 96], for earliest-deadline first (EDF) see [102], Generalized Processor Sharing (GPS) in [37, 91] and Static Priority Scheduling (SP) is explored in [37, 96].

2.1.1 *End-to-End Analysis*

As in system theory, one of the main strengths of min-plus convolution is the concatenation of tandem systems along a network path, where the individual service curves of the systems can be easily concatenated by convolution, resulting in a network service curve that specifies the available

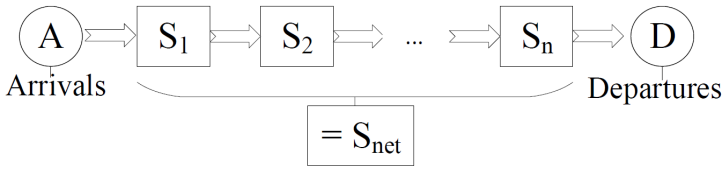


Figure 2.1: A network path with n systems and corresponding service curves $S_i(t)$ for $i \in [1, n]$ in tandem are simplified to one end-to-end service curve S_{net}

end-to-end service. In this way, a general framework for the analysis of entire networks is created. Let's consider a system as in Fig. 2.1. Due to the associativity of the min-plus convolution, we can compute an end-to-end service curve for the whole network. In particular, we have

$$\begin{aligned} D(t) &\geqslant (A \otimes S_1) \otimes S_2) \otimes ...) \otimes S_n \\ &= A \otimes \underbrace{(S_1 \otimes S_2 ... \otimes S_n)}_{:=S_{net}}(t), \end{aligned} \tag{2.11}$$

where we used associativity in the second line and convolved all individual service curves $S_i(t)$ for $i \in [1, n]$ to obtain the network service curve S_{net} [37].

2.1.2 *Envelope Functions*

Generally, we cannot expect to know the total traffic in advance. An example is a video encoder that changes the video's quality based on some criteria that might be unknown. For more information about the effect of changing the quantization parameter *qp* of the H-264 encoder on the signal-to-noise ratio (SNR) and delay, see on our own work [71]. Consequently, the traffic might be highly bursty, and certain delays, backlog, and loss constraints could be violated. However, the network calculus provides some ideas and models to deal with such problems. More precisely, we look for some upper limits of arrival traffic
(arrival envelope) or, if the service is not fully known, for a lower limit of the service curve (service envelope). These envelope functions could be stated as some strict bounds (deterministic) or some bounds that allow a small probability of violation (probabilistic), and can be used for the

provisioning of deterministic or statistical guarantees and performance bounds.

In this section, we consider the deterministic case. So, we call $E(t)$ a deterministic upper bound of the arrivals if the function is non-decreasing and non-negative, i.e., $E(t) \in \mathcal{F}_0$, see Eq. (2.1) and it holds that the cumulative traffic is bounded as follows

$$A(t) - A(\tau) \leqslant E(t-\tau) \tag{2.12}$$

$\forall t \geqslant \tau \geqslant 0$. We can rewrite Eq. (2.12) by adding $A(\tau)$ on both sides and taking the infimum on the right hand side

$$A(t) \leqslant A \otimes E(t). \tag{2.13}$$

The backlog bound from Eq (2.5) becomes

$$B(t) \leqslant \sup_{\tau \in [0,t]} \{E(t-\tau) - S(t-\tau)\} = \sup_{u \in [0,t]} \{E(u) - S(u)\}, \tag{2.14}$$

where we used the univariate service description from Eq. (2.3) and substitute $u = t - \tau$. Note that we obtain a bound for the delay similarly.

An example of traffic envelopes is a leaky-bucket-shaper. It allows a specific, immediate amount of arrivals and delays other traffic, such that, e.g., the output can be handled by the network. A leaky-bucket-shaper has the form

$$E(t) = q_{env} + \rho_{env} t, \tag{2.15}$$

where q_{env} is the max instantaneous burst of arrivals permitted, and ρ_{env} is the envelope rate, which is an upper bound on the mean rate of the arrivals [52, 53]. Note that we already got the conditions for a lower bound of the service from Eq. (2.3). A good overview of traffic arrival envelopes we find in [106] and for service curves in [70].

Next, we consider the first example with an MPEG video source. Here, we assume that the video consists of a set of pictures/ frames, where each frame is either an I-frame or a B-frame. For simplicity, we do not use P-frames. Thereby, the order of the frames, i.e., the group of pictures (GoP) is $(\underbrace{\text{IBBB}\ldots\text{BBB}}_{n_{\text{GoP}}\ \text{frames}}\text{I})$. To learn more about MPEG and the GoP see [74]. Further, we assume that the frames arrive in a uniform order. Hence, time is discretised with time-slots $t \in [0, 1, 2, \ldots]$. In each time-slot, we assume

that a packet arrives with probability $\alpha = 1$, which corresponds to the deterministic case. The packet size of the B-frames B_{frame} are equal to two in all cases, whereas we change the size of the I-frames I_{frame}, i.e., $I_{frame} \in [10, 20, 30, 40, 50]$. The length of a GoP is $n_{GoP} = 50$. In order to compute performance bounds such as the backlog, we choose a leaky-bucket arrival envelope from Eq. (2.15). For sure, the permitted burst size q_{env} is equivalent to the size of the I-frames. We adjust the envelope rate ρ_{env} according to the mean rates for all I-frame sizes, which are between two and three. We increase all rates by 10% to get an upper bound of the mean rates.
The service curve is a latency rate function of the form

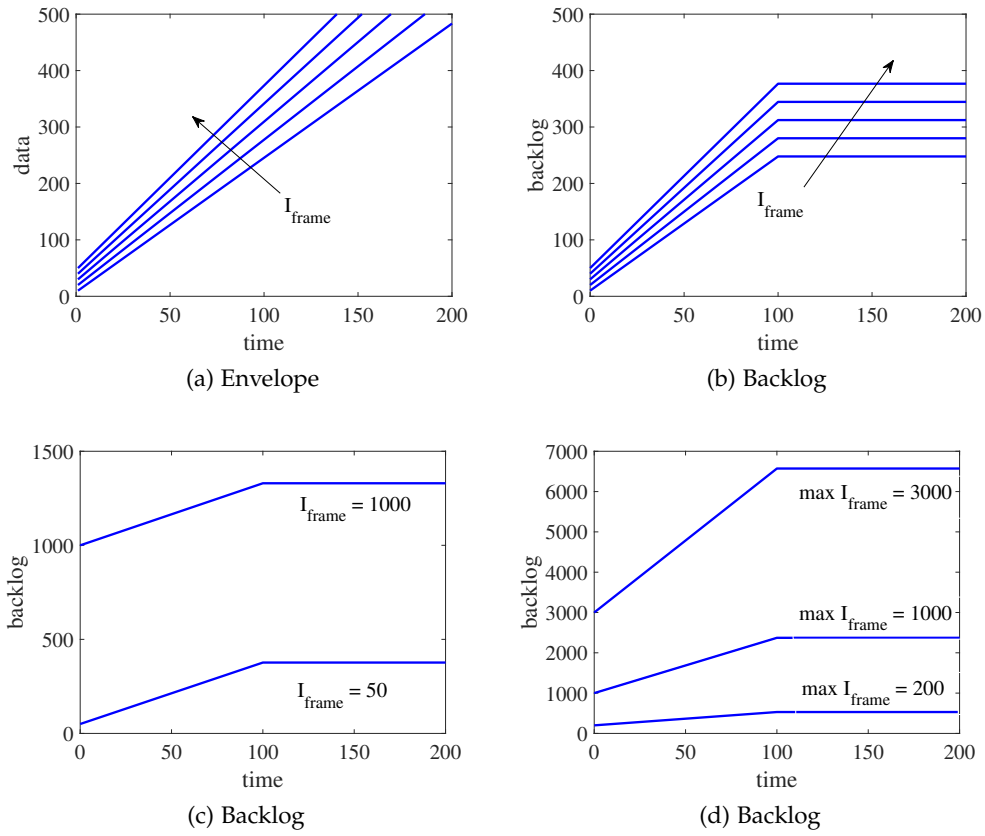
$$S^{slr}(\tau, t) = R[t - \tau - T]_+ = S^{slr}(t - \tau) \tag{2.16}$$

$\forall t \geqslant \tau \geqslant 0$, where the superscript slr indicates that the latency is stationary, i.e., time-invariant. Note that a stationary latency conforms to a propagation delay. However, the parameters in Eq.(2.16) are $R = 4$ for the rate and $T = 100$ for the the latency.

Fig. 2.2 shows the first indications of the influence of the I-frame's size. Clearly, the burst size q_{env} of the leaky-bucket envelope depends on the I-frame size. In consequence the mean rate increases, too, see Fig. 2.2a. The influence on the backlog we illustrate in Fig. 2.2b. Here, the backlog bound increases until the service starts at $T = 100$. Afterwards, the backlog bound remains at this worst-case value. The reason is that the time-invariant model is generally non-decreasing in t due to the sup in Eq.(2.14) if S is invariant. Thus, we conclude that the I-frame sizes have a high impact on the backlog bounds.

To substantiate this, we modify the example. Therefore, we increase the I-frame size to 1000 and set $n_{GoP} = 1000$. This way, we have for every 1000 frames one I-frame. The mean rate is around three, i.e., similar as in the above case, where $I_{frame} = 50$. Although the average rates are similar, we see in Fig. 2.2c a significant difference in the backlog bounds. This is due to the enormous burst of 1000.

The cases so far considered a fixed I-frame size. We now choose a case where the size of the I-frame increases continuously. More precisely, we set $n_{GoP} = 50$, again, and $I_{frame} = t$. The results we see in Fig. 2.2d where we plotted the backlog for a maximum amount of time $t_{max} \in [200, 1000, 3000]$. This yields a maximum I-frame of 200 for $t_{max} = 200$ and similarly of 3000 for $t_{max} = 3000$. Consequently, the backlog in Fig. 2.2d increases with t.



(a) Envelope (b) Backlog

(c) Backlog (d) Backlog

Figure 2.2: Deterministic leaky-bucket envelope

Note, we adjusted the service rate R such that $R > \rho_{env}$ for every t_{max}. Otherwise, the backlog would grow continuously.

This example illustrates what can happen if the video gets longer and so the number of frames is getting higher and higher. It means that with every additional frame, there is a chance that this new frame is larger than all the other ones before, such that the arrival envelope increases.

In this scenario, we only have 3000 as the maximum amount of frames. However, it is not unusual that videos have 60 frames per second (fps), today. This means that watching a movie of 2 hours and longer results in approximately 500.000 frames.

Thus, despite some outstanding qualities that include the system modeling and the concatenation of the tandem system, the worst-case view results in an overly pessimistic performance bounds and is a significant shortcoming of the deterministic network calculus. This becomes even more obvious by considering, e.g., audio and video applications that tolerate violations with a small probability of the delay, backlog, and loss constraints. A discussion of client requirements can be found in [67].

Apart from that, it is not guaranteed that there exist some deterministic envelopes for arrival traffic and service. An example is radio channels, where the service is random due to fading, interference, and non-deterministic medium access control. Consequently, deterministic worst-case bounds are not trivial and may not even exist in random systems.

An extension of the deterministic network calculus that overcomes these limitations is the stochastic network calculus. It guarantees some statistical bounds that hold with a certain probability $\varepsilon > 0$ for some arrival and service functions [25].

2.2 STOCHASTIC NETWORK CALCULUS

The deterministic network calculus is a framework with outstanding qualities, for example, by modeling systems and concatenation of systems in series. However, in the following, we want to overcome the described drawbacks from the previous section by introducing the stochastic network calculus and presenting the main results. A broad overview can be found in [17, 31, 37, 46, 70, 86].

For repetition, a major disadvantage of the deterministic network calculus is that a few rare events could lead to an extremely weak estimation of performance bounds. The stochastic network calculus is a framework that

excludes these rare events. Thereby, the objective is to find or bound the probability that performance measures, such as delay and backlog, exceed a given threshold.

Let ε be a small violation probability, which could be in the order of 10^{-5}, it leads to the following bounds for delay and backlog [25, 46]

$$P[delay > q] \leqslant \varepsilon \qquad \text{or} \qquad P[B > b] \leqslant \varepsilon. \tag{2.17}$$

Note that for $\varepsilon = 0$ we have the deterministic case. Moreover, it has been shown that stochastic service curves can efficiently model effects such as the variability of the service provided by radio channels [69, 143] or the CSMA/CA random access control [29].

Although stochastic network calculus has a certain strength by exploiting the knowledge of scheduling algorithms and achieving a multiplexing gain, it is still modest in comparison to the leap from deterministic to stochastic network calculus [98]. Therefore, we use a blind multiplexing model. Thus, no assumptions are made about the scheduling algorithms such as GPS, EDF, and FIFO. Further scheduling traffic, and service models have been investigated, for example, in [25, 30, 31, 37, 46, 47, 48, 49, 52, 56, 70, 86, 89, 96, 98, 106, 128, 135, 149, 151].

In the following, we classify the stochastic network calculus according to the random systems description in [46]. There, Ciucu describes a random system and their statistical envelopes either as random processes, see [34, 92, 152] or as non-random functions as in [25, 149].

In the first case, time-variance is embedded in bivariate functions, whereby these functions are interpreted as random processes.

In the second case, invariant functions are used to encounter the variability of random systems as a probabilistic bound.

2.2.1 *Random Processes*

We start the consideration of random processes in stochastic network calculation by recalling equation Eq. (2.2), where $D(t) \geqslant \inf_{\tau \in [0,t]} \{A(\tau) + S(\tau, t)\}$. Here, the arrivals $A(\tau, t)$, service $S(\tau, t)$ and departures $D(\tau, t)$ are bivariate deterministic functions with the properties as seen so far and as stated, e.g., in [11, 40]. There, the authors Chang and Cruz noted that the time-varying system with the bivariate deterministic functions could be extended to the stochastic case as random processes where statistical envelopes of arrivals

and services are derived. The envelopes are bounds that may be violated with a defined probability. In [37, 40], the author used the concept of a dynamic server where the service is a random process to derive the departure bound from Eq.(2.2). To achieve equality, the system requires linearity [96]. An example that satisfies the definition of a dynamic server and leads to equality is a loss-less, work-conserving server with time-varying capacity. In comparison, a non-linear system is obtained by a first-in-first-out (FIFO) scheduler [101], i.e., it satisfies only $\geqslant$. However, results for services with bivariate random processes are shown in [35, 56]. Here, random processes are analysed, and stationary bounds are derived. In order to classify the results, we first need a better understanding of what stationarity is.

Definition 1 (Stationarity). *A process* $S(t)$ *is stationary if it holds that*

$$P[S(\tau, t) \leqslant x] = P[S(\tau + \delta_s, t + \delta_s) \leqslant x], \tag{2.18}$$

for any $\tau, t, \delta_s \geqslant 0$. In other words: If the probability of observing a certain amount of service in an interval does not depend on the time instances themselves, but only on the length of the interval.

The stationary bounds thus lead to a situation similar to the deterministic case, although time-dependent processes are investigated.

In the next chapter 3, we will present some first results for time-dependent services in a deterministic scenario. The extension to the stochastic case is shown in the chapter 5 where non-stationary systems are shown.

2.2.2 *Non-Random Functions*

In the section for deterministic network calculus, we have already seen that the derivation of deterministic envelopes in section 2.1.2 could lead to an overly pessimistic estimate, especially when the number of arrivals is not known or tends to infinity. Then some single packets can lead to an extremely bad envelope function for Eq. (2.12). To overcome this, we introduce now statistical envelopes as non-random functions, a stochastic extension of Eq. (2.12) and present a method of construction. We call $E(t)$ a statistical envelope of the arrivals $A(t)$ if it holds $\forall t \geqslant \tau \geqslant 0$ that

$$P[A(\tau, t) > E(t - \tau) + \sigma_A] \leqslant \varepsilon(\sigma_A) \tag{2.19}$$

where $\varepsilon(\sigma_A)$ is the violation probability, also referred as overflow profile with parameter σ_A. Depending on the choice of $E(t)$, σ_A and $\varepsilon(\sigma_A)$ we

distinguish between several models, such as the Exponentially Bounded Burstiness (EBB) model by Yaron and Sidi [149, 150], the Stochastically Bounded Burstiness (SBB) model by Starobinski and Sidi [135], the generalized Stochastically Bounded Burstiness (gSBB) [87, 151] and the effective envelopes by Boorstyn et.al [25]. To get deeper insights see [46, 86]. For $E(t-\tau) = \rho_A(t-\tau)$ we have an extension of the affine leaky-bucket function from Eq. (2.15).

Note that the formulation in Eq. (2.19) holds for a fixed τ, only. Therefore, it is also referred as a point-wise violation probability. In order to obtain guarantees for the complete path we formulate Eq. (2.19) as follows

$$P[\exists\tau \in [0,t] : A(\tau,t) > E(t-\tau) + \sigma_A] \leqslant \varepsilon(\sigma_A) \tag{2.20}$$

and call it a sample path envelope. Note, that $\varepsilon(\sigma_A)$ in Eq. (2.19) and Eq. (2.20) are different. Alternatively, we can rewrite Eq. (2.20) as

$$P[A(\tau,t) \leqslant E(t-\tau) + \sigma_A,\ \forall\tau \in [0,t]] \geqslant 1-\varepsilon(\sigma_A). \tag{2.21}$$

The advantage is that we can use an equivalent formulation, which can be estimated for all and not just a specific value of τ. More precisely, we have

$$P\left[\sup_{\tau\in[0,t]} \{A(\tau,t) - E(t-\tau)\} > \sigma_A\right] \leqslant \sum_{\tau=0}^{t} P\left[A(\tau,t) - E(t-\tau) > \sigma_A\right] \tag{2.22}$$

where we used the union bound to find an upper bound of the sample path envelope. Note the similarities of the left hand side in Eq. (2.20) to the backlog computation in Eq. (2.14). To get more information about point-wise versus sample path envelopes and how to handle the sum in Eq.(2.22) that might tend to infinity for $t \to \infty$ see [70, 73].

To construct such envelopes, we refer to Mao [106] and Fidler [70] where two methods of construction are mentioned. The first one defines rate-variance envelopes with the help of the central limit theorem, see [44, 89]. The second method uses moment generating functions (MGF) to construct an envelope and is our method of choice.

In the following, we choose $\sigma = 0$ which corresponds to the effective envelopes from Boorstyn et.al [25]. Moreover, to clarify that the statistical

arrival envelope has a relation to the violation probability ε we denote the statistical arrival envelope as $A^\varepsilon(t)$ from now on. Hence Eq. (2.21) becomes

$$P[A(\tau, t) \leqslant \mathcal{A}^\varepsilon(t-\tau),\ \forall \tau \in [0, t]] \geqslant 1-\varepsilon. \tag{2.23}$$

For the arrivals we define the MGF as

$$M_A(\theta, t) = E[e^{\theta(A(\tau,\tau+t))}] \tag{2.24}$$

where $\theta > 0$ is a free parameter. Using Chernoff's bound, see in the appendix and theorem 1 we obtain for $E(t-\tau)$ in Eq.(2.22) and $\theta > 0$ the following statistical envelope

$$P[A(\tau, \tau+t) > A^\varepsilon(t)] \leqslant e^{-\theta A^\varepsilon(t)} M_A(\theta, t) \tag{2.25}$$

where it holds for every $\tau \geqslant 0$ by assuming stationarity.

A method of construction for the statistical envelope $A^\varepsilon(t)$ we find in [47, 98]. Following similar steps for the arrival envelope, we have that

$$A^\varepsilon(t) = \frac{1}{\theta(t)} \left(\ln M_A(\theta, t) + \rho t - \ln(\rho\varepsilon)\right) \tag{2.26}$$

is a statistical envelope function of $A(t)$ in Eq. (2.25) that provides the sample path guarantee from Eq. (2.20) and Eq.(2.21) for all $t \geqslant 0$. Above, $\varepsilon \in (0, 1]$ is a probability of overflow, and $\theta(t) > 0$ and $\rho \in (0, 1/\varepsilon]$ are free parameters where ρ is referred to as the slack rate. The derivation is similar to Eq.(5.6). Note, that for a computationally efficient implementation $\theta(t)$ is a time-variant parameter, which allows to optimize $\theta(t)$ for each t, individually. A backlog bound follows immediately by substituting $A^\varepsilon(t-\tau)$ for $A(\tau, t)$. Thus, Eq.(2.5) becomes

$$P[B(t) \leqslant \sup_{\tau\in[0,t]} \{A^\varepsilon(t-\tau) - S(\tau, t)\}] \geqslant 1-\varepsilon. \tag{2.27}$$

Next, we give a concrete example for a statistical arrival envelope and present results for the backlog. Let us reconsider the example from the section about deterministic network calculus, where we used a deterministic pattern for the packet sizes in a video that arises in each time-slot with probability one. As already mentioned, some applications tolerate losses, e.g., video streaming. Hence, we relax the assumption that in a discrete interval of length t, the probability of a successful packet arrival is one. In other words, in each time-slot, the probability of a packet arrival is $\alpha \in [0, 1]$,

which is an independent Bernoulli trial. Consequently, the number of arrivals $N(t)$ in the complete interval is binomially distributed with parameter α. So, we can interpret α as an average arrival rate. Further, we do not have deterministic packet sizes anymore, which is reasonable, for example, for radio channels with changing characteristics. We assume independent and identically distributed (iid) packet size $Y(i)$ for $i = 0, 1, 2, \ldots$ which follow a geometric distribution with parameter $\beta \in (0, 1]$. Using basic stochastic, we obtain $1/\beta$ as the average packet size. Thus, this example is a discrete-time M|M|1-queue equivalent of a Poisson process. We will use it throughout the theoretical part of this work as our stochastic arrival process, where the quotient α/β has the interpretation of the system's utilization if we assume a constant rate service $R = 1$.

Now, to compute the statistical arrival envelope $\mathcal{A}^\varepsilon(t)$ from Eq. (2.26) we need the moment generating functions from Eq. (2.24). From [125] we can derive the corresponding MGFs for $N(t)$ and $Y(i)$, i.e.,

$$M_N(\vartheta, t) = (\alpha e^\vartheta + 1 - \alpha)^t$$

and

$$M_Y(\theta) = \beta e^\theta / (1 - (1 - \beta) e^\theta)$$

for $\theta \in [0, -\ln(1-\beta))$. Then, we get the cumulative arrivals for an interval $(\tau, t]$ from $A(\tau, t) = \sum_{i=N(\tau)+1}^{N(t)} Y(i)$. Thus, it is a doubly stochastic process with MGF

$$M_A(\theta, t-\tau) = M_N(\ln M_Y(\theta), t-\tau) \tag{2.28}$$

[73, 125] so that by insertion of $M_N(\vartheta, t)$ and $M_Y(\theta)$ we obtain

$$M_A(\theta, t) = \left(\frac{\alpha\beta e^\theta}{1 - (1-\beta)e^\theta} + 1 - \alpha \right)^t . \tag{2.29}$$

Clearly, $M_A(\theta, t) = (M_A(\theta, 1))^t$ as $A(t)$ has iid increments.

In Fig. 2.3 we show the backlog changes over time for Eq. (2.27) and the time-invariant service curve from (2.16) with parameter $T = 100$ and $R = 1$. For the arrival process we choose $\alpha = 0.09$ and $\beta = 0.3$ such that we obtain a utilization of 0.3. We optimize θ and $\rho \in (0, 1/\varepsilon]$ numerically, where the overflow probability is set to $\varepsilon = 10^{-9}$.

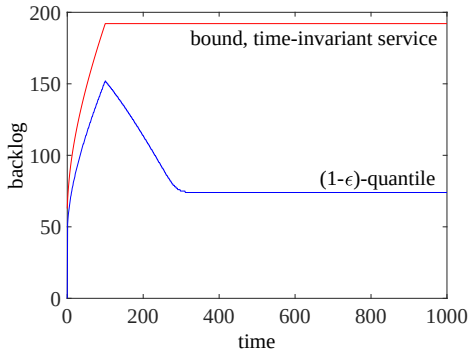


Figure 2.3: Progression of the transient backlog over time. The time-invariant service model remains in the worst-case bound whereas the shape of the $(1-\varepsilon)$-quantile decreases.

For the time-invariant service curve from the deterministic network calculus Eq. (2.16) we see that the backlog bound increases until the service starts at $t = 100$. Afterwards it remains at this value, although the service rate with $R = 1$ is significantly larger than the utilization $a/\beta = 0.3$. This is because the univariate function depends only on the width of the interval $(\tau, t]$ and not to the corresponding time instances such that the sup in Eq. (2.27) is non-decreasing in time t. Consequently, the backlog considers the transient phase at the beginning, but remains at the worst-case due to the univariate service function as in the example of the deterministic network calculus and Fig.(2.2).

Due to our choice of a discrete Poisson arrival process, we can compare the results with an exact solution. This exact solution is obtained from a discrete Markov chain. Here, the arrivals at time t are represented by the state of the Markov chain $K(t) \geqslant 0$ and starts in state $K(0) = 0$. Hence, the initial state distribution $\mathbf{P}(0)$ is the column vector $(1, 0, 0, \ldots)$. For all other states $t > 0$ the state distribution follows from the repeated insertion of $\mathbf{P}(t) = \mathbf{Q}(t)\mathbf{P}(t-1)$, where $\mathbf{Q}$ is the transition matrix. For all $t \leqslant T$ no service is available and the state depends only on the arrivals which follow a binomial distribution, such that $\mathbf{Q}$ is assembled with the probabilities $q_{i,i} = 1-\alpha$, $q_{i,i+1} = \alpha$, and all other $q_{i,j} = 0$. For $t > T$ the service started, i.e., the chain makes a transition to $q_{i,i-1} = (1-\alpha)\beta$, $q_{i,i} = (1-\alpha)(1-\beta)+\alpha\beta$, $q_{i,i+1} = \alpha(1-\beta)$, and $q_{i,j} = 0$ else. Letting $t \to \infty$ it tends to the geometric state distribution [16], i.e.,

$$P[K(\infty) = k] = \frac{\beta - \alpha}{\beta(1-\alpha)} \left(\frac{\alpha(1-\beta)}{(1-\alpha)\beta} \right)^k .$$

Now, we can compute the backlog distribution from

$$P[B(t) = b] = \sum_{k=1}^{\infty} P[B(t) = b|K(t) = k]\, P[K(t) = k], \tag{2.30}$$

where $b > 0$. Note that the sum starts at $k = 1$, since for the case that $b = 0$ $P[B(t) = 0] = P[K(t) = 0]$.

As seen from Eq. (2.30) the backlog in state k is conditional and follows from the sum of k geometric random variables that is negative binomial, i.e., for $k, b > 0$

$$P[B(t) = b|K(t) = k] = \binom{b-1}{k-1} \beta^{k-1} (1-\beta)^{b-k}.$$

We included the backlog progression for a $(1-\varepsilon)$-quantile of $B(t)$ in Fig. 2.3. As before, the backlog increases until $t = 100$. In contrast to the time-invariant service, the backlog declines and converges towards a stationary backlog bound that is significantly lower in comparison to the worst-case bound. The deviation of the two curves, e.g., for $t = 100$ in Fig. [**?**] is due to inequalities such as the union and the Chernoff bound, which are used to derive the Poisson envelope. Especially for the Poisson process, the deviation can be reduced using tighter martingale bounds [45]. We will introduce and compare them in Sec. 5.2.2. Results and their envelopes for other non-trivial arrival processes can be obtained, e.g., for self-similar, long-range dependent [100, 122], and heavy-tailed processes [100]. The advantage of comparing bounds with an exact solution is that we get an estimate of the accuracy for the bounds. This knowledge helps us to evaluate thresholds for cases where no exact solution is available.

However, as seen from the backlog progression of the exact solution in Fig. 2.3, the performance bounds for statistical arrival envelopes and deterministic, time-invariant service curves does not match this pattern and remains in the worst-case. Therefore, we present some of the state-of-the-art results in the next chapter. Among other things, we show the extension of the network calculus to time-variant services, and additionally, the ε-effective service curve, which is able to handle random services.

In the cases described so far, we know the services in the system. Scenarios in which they are unknown, we have not yet investigated. To close this gap, we also present measurement methods, e.g., in the field of network calculus, that can estimate the service for unknown systems. We also explain their disadvantages in the calculation of time-variant service curves.

3

RELATED WORK

In this chapter, we present, among other things, well-chosen parts of the current literature to analyze systems and estimate their services in a time-dependent way.

So far, in the previous chapter, we have introduced basic principles in network calculus. As can be seen in Chap. 2, the standard concepts, such as the deterministic network calculus in Sec. 2.1, take transient phases into account, but they remain in the worst-case and therefore, cannot accurately predict system progress due to time-invariant functions.

The extension to the stochastic network calculus in Sec.2.2 either assumes stationarity as in Sec.2.2.2 or uses stationary bounds to handle time-variant random processes, see Sec. 2.2.1. Further references are [31, 37, 47, 48, 68, 70, 73, 86, 98]. So, the currently shown theory is not able to take a similar course as the exact solution from the example in Fig. 2.3. We therefore present in Sec. 3.1 a time-dependent extension of the service curves in network calculus, which follows the desired progression. For the stochastic network calculus, this extension is accompanied by the removal of the assumption of stationarity, which is also dealt with in Chap. 5. In Sec.3.1.3 we also introduce the ε-effective service curve. It is a statistical service curve that is able to consider time-variant, random processes, but under the assumption of stationarity.

All these considerations have one major thing in common, namely that we already know the service in the network. If we relax this assumption, i.e., if we have only a little or no knowledge at all about the service of the system, the question arises how to find an adequate service curve. In Sec. 3.2, we introduce methods, which are state-of-the-art in estimating random services in (cellular) networks. Further, we show limitations of these methods in measuring the service in a time-dependent way. Please note that we present among others results from [19].

3.1 TIME-VARIANT NETWORK CALCULUS

To find time-dependent performance bounds, we relax the assumptions of time-invariance in the network calculus as in [11, 38, 39, 40], where a

time-varying min-plus system theory, is developed. Note that the service is not necessarily deterministic and can also be a stochastic process. For a summary see [37].

Following the description from the introduction of Chap. 2, we have non-negative, non-decreasing, bivariate functions to model arrivals, departures, and services, such that for a time-varying lower service curve $S(\tau, t)$ Eq. (2.2) holds. The service curve is called exact, if and only if the system is min-plus linear. Thus Eq. (2.2) holds with equality. Examples for a min-plus linear system for bivariate functions are work-conserving links with a time-varying capacity, see [37, 38, 40].

Generally, the system description by bivariate functions is less intensively studied than for an univariate system. An overview of algebraic properties we find, e.g., in [37, 38, 40]. Taking a closer look at the properties of time-invariant functions , see Appendix, we find many similarities. For sure, the backlog is still the vertical and the delay the horizontal deviation of arrivals and departures. Hence, the performance bounds from Chap. 2 are still valid. Further, if a system has a certain constraint on the delay and backlog Chang [37] solved the problem for time-varying services. The upper envelope functions for the arrivals have to be adjusted to bivariate functions. In this sense, $E(\tau, t)$ is an upper arrival envelope of the arrivals if it holds for all $t \geqslant \tau$

$$A(\tau, t) \leqslant E(\tau, t). \tag{3.1}$$

Clearly, due to the equal sign $A(\tau, t)$ itself is an upper envelope [68]. Similarly, it holds for the service. Then, we still have the property of associativity. As in equation Eq. (2.11) it is allowed, for example, to convolve two systems in series with the service curve $S_1(\tau, t)$ and $S_2(\tau, t)$ into a single system $S(\tau, t)$, i.e.

$$S(\tau, t) = \inf_{\nu \in [\tau, t]} \{S_1(\tau, \nu) + S_2(\nu, t)\} = S_1 \otimes S_2(\tau, t).$$

Having n systems in series the iterative convolution yields $S_{net}(\tau, t) = S_1 \otimes S_2 \otimes ... \otimes S_n(\tau, t)$. Apart from these and other similarities, there are also characteristics that no longer apply, e.g., that the convolution of bivariate functions is no longer commutative, i.e.,

$$S_1 \otimes S_2(\tau, t) \neq S_2 \otimes S_1(\tau, t).$$

Furthermore, we will see later on that the assumption of additivity, i.e. $S(t) = S(\tau) + S(\tau, t)\ \forall t \geqslant \tau \geqslant 0$, is generally not correct. In Chap. 6, we will examine the effects of the assumption of the additivity and how it influences the measurement methods.

3.1.1 *Deterministic Network Calculus*

For the moment we extend the theory of deterministic network calculus and consider deterministic, time-dependent, bivariate service curves. In Sec. 2.1.2 we introduced the latency rate service function in Eq. 2.16, that is a univariate, and so time-invariant function. We are able to extent these functions to time-variance, where the latency is transient. This is done as described in our works [19], by the following bivariate function, where for all $t \geqslant \tau \geqslant 0$ we have

$$S^{tlr}(\tau, t) = \begin{cases} 0, & t \leqslant T \\ R(t-T), & t > T, \tau \leqslant T \\ R(t-\tau), & t > T, \tau > T \end{cases}$$

or equivalently

$$S^{tlr}(\tau, t) = R[t - \max\{\tau, T\}]_+ \tag{3.2}$$

where $[x]_+ = \max\{0, x\}$ is the non-negative part of x. To emphasize the transient latency, we abbreviated by superscript tlr.

For numerical evaluation we use the same example as in Sec. 2.2.2. Hence, we use the same Poisson arrival process, with $\alpha = 0.09$, $\beta = 0.3$ and $\varepsilon = 10^{-9}$. It enables us to compare the results with the exact solution from Fig. 2.3. For the service, we choose again a latency of $T = 100$ and a rate of $R = 1$. This applies both to the time-invariant S^{slr} and the time-variant service S^{tlr}.

In Fig.3.1 we show the results for the time-variant, time-invariant and the $(1-\varepsilon)$-quantile of the Markov-chain, where the progression for the exact $(1-\varepsilon)$-quantile and the time-invariant bounds are the same as in Fig.2.3. As discussed earlier, the time-invariant, stationary latency function cannot recover the shape of the exact solution, whereas the function with the transient latency does. Again, the deviation is due to the bounds which are used to derive the envelope function of the Poisson process.

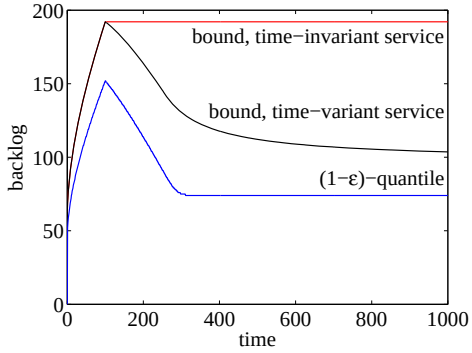
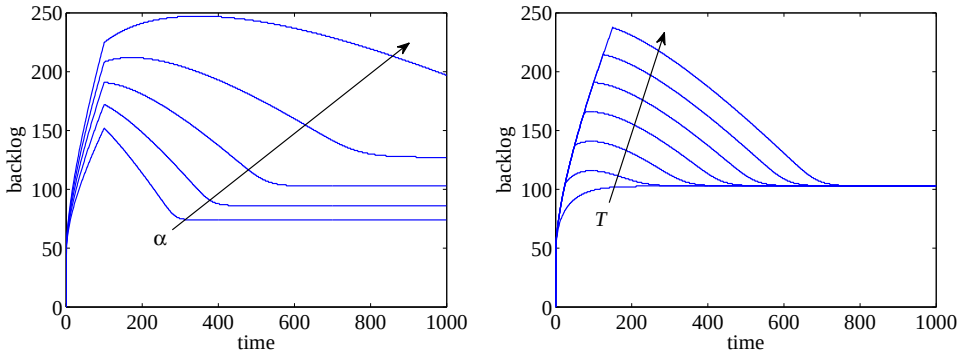


Figure 3.1: Progression of the transient and stationary backlog over time. The time-variant service model correctly estimates the shape of the $(1-\varepsilon)$-quantile.



(a) $\alpha = \{0.09, 0.12, \ldots, 0.21\}$ (b) $T = \{0, 25, \ldots, 150\}$

Figure 3.2: Impact of arrival rate α and sleep cycle T on the backlog quantile.

The importance of dealing with such transient effects is shown in Fig. 3.2. Here, we use the same parameters as in Fig. 2.3 and Fig. 3.1 and illustrate the influence of the arrival rates α for $T = 100$ and the influence of the sleep cycle T for $\alpha = 0.15$ on the backlog quantile.

We are particularly interested in the maximum overshoot compared to the steady-state solution and the time needed to reach the steady-state backlog within a defined range, i.e., the relaxation time, see [141].

As can be seen in Fig. 3.2a, the impact of α is significant on the transient overshoot and the relaxation time. Interestingly, the maximum overshoot occurs after T for large α, i.e., after the service has already started, or strictly speaking during the transition from binomial to geometric state distribution. Intuitively, this can be explained by the fact that after time T there is a certain probability that the service is smaller than the arrivals, see Sec. 2.2.2. By increasing T for fixed α, we investigate the relaxation time, where Fig.3.2b illustrates that it can reach values that are an order of magnitude greater than T.

3.1.2 *Stochastic Network Calculus*

Considering the time-variance with stochastic processes, we find several interesting works in recent years [33, 66, 112, 113]. Similarly as our example for the deterministic network calculus it is also based on [37, 38, 40].

For example, in [112] Nikolaus et al. derived transient end-to-end delay bounds where the underlying topology is a sink tree with one flow of interest (*foi*), and n cross traffic flows. After simplifying the topology into a tandem system, arrival bounds for each cross flow are computed at the point where it crosses the *foi*. Then, with MGFs and Hölders inequality delay bounds for separated flow analysis (SFA) and pay multiplexing only once (PMOO) are derived. Comparing SFA and PMOO for arrivals of the form of fractional Brownian motion (fBm) yields that PMOO outperforms SFA in case of sink tree topologies since SFA applies more the Hölder inequality. Further, the authors showed that for long-range dependent (LRD) traffic, i.e., the Hurst parameter is between $(0.5, 1)$, and by increasing the stochastic dependence of the flows, the delay bounds get worse, e.g., by using the Hölder inequality.

Another example of time-dependent end-to-end delay bounds is presented in [33]. Here, Champati et al. analyzed the transient behavior of arrival traffic traversing multi-hop wireless networks. Novel wireless transient bounds (WTB) are compared with state-of-the-art transient bounds (SOTAT). As a result, WTB provides tighter bounds than SOTAT, which cannot adequately deal with the short-term variability of wireless service. Additionally, a new theoretical model for queuing is introduced. It can handle an initial buffer greater than zero and random services, such as the Rayleigh-block fading channel model.

In comparison to this thesis, these papers consider the violation probability for end-to-end delay bounds to find closer bounds, whereas in our work, the transient behavior over the entire time is investigated to find new insights, e.g., into the implementation of the DRX mode in cellular networks and the consequences for backlog and delay. Note that in [18] the author extended our results from [19, 20] to an analysis of backlog and delays, where the service and additionally the arrivals are time-variant. It is the first work that considers time-dependent arrivals and services at once. It substantiates our idea to investigate processes in a time-variant way.

3.1.3 *ε-effective Service Curve*

As described earlier in Sec. 2.2.2, one way to handle time-variant services in stochastic network calculus is done by the use of stationary functions. In [31] the authors derived the so called ε-effective service curve. It belongs to the effective envelopes by Boorstyn et al. [25] where we set $\sigma = 0$ on the right hand side in Eq.(2.20). We define a lower service envelope of the system $\mathcal{S}^{\varepsilon}(t)$ that satisfies

$$P[S(\tau, t) \geqslant \mathcal{S}^{\varepsilon}(t-\tau) \forall \tau \in [0, t]] \leqslant 1 - \varepsilon \tag{3.3}$$

for $t \geqslant 0$, where $\varepsilon \geqslant 0$ is a probability of underflow. Generally, a system has an ε-effective service curve if a service envelope $\mathcal{S}^{\varepsilon}(t) \geqslant 0$ such as in Eq. (3.3) provides for all $t \geqslant 0$ the service guarantee

$$P\left[D(t) \geqslant \inf_{\tau \in [0,t]} \{A(\tau) + \mathcal{S}^{\varepsilon}(t-\tau)\}\right] \geqslant 1 - \varepsilon, \tag{3.4}$$

where ε is the violation probability [31]. In [101], the authors Liebeherr, Fidler, and Valaee have accomplished to transfer the estimate of the available bandwidth into the min-plus theory, where the service curve expresses the available bandwidth. Furthermore, a measurement based rate scanning method is introduced. It can represent the available bandwidth for different time-scales, which is usually given as a single value.

Even though Eq. (3.4) can handle the stochastic nature of systems, it cannot represent in a time-variant way. The reason is that the ε-effective service curve only considers the time interval's length and not the time instances itself. Regarding our earlier example, the backlog's progression would be similar to the deterministic case, due to the same arguments. In order to obtain a behavior as for Eq. (3.2) we will extend the definition of Eq. (3.4) in Chap. 5.

3.2 MEASUREMENT-BASED ESTIMATION METHODS

So far, we have used a model-based approach to analyze transient effects in networks. In the current literature, it is possible in some places to specify bounds, e.g., for backlog and delay in networks with respect to time-variant perspectives. An example is given in Sec 3.1, where we have shown a time-variant service curve of sleep planning.

However, we assumed that we are completely informed about the internals of the system. To identify the service curve of a system without knowledge of its internals, we now investigate measurement methods that estimate them. The resulting service curve can be used to evaluate sleep scheduling implementations in cellular networks, as in Sec. 7.4.2.

To emphasize that this topic has become more important in recent years, we refer to [10, 15, 27, 32, 79, 101, 103, 105, 140].

Generally, we estimate the service curve of an unknown system by analyzing the arrivals and departures. We distinguish between grey-box [15, 27] and black-box [103] approaches. A black-box contains no information at all about the system and usually assumes only linearity and stationarity. Contrary, a grey-box contains further information, this could be the assumption that the service curve has the shape of a latency-rate function so that only the two parameters, latency, and rate have to be determined from measurements.

Further classifications can be undertaken, such as the distinction between deterministic and time-invariant models as in [10, 15, 26, 27, 79, 101, 121, 139] and stochastic models of stationary systems, see [32, 103, 105, 140].

As an example, in [139], the authors introduced a deterministic and time-invariant method to estimate a service curve which is similar to a latency rate function consisting out of a rate and an error term. In this grey-box the parameters rate and error term are determined for the *Guaranteed Rate* (GR) [77] and *Packet Scale Rate Guarantee* (PSRG) model [22]. For the GR model, we obtain the j-the departure packet d^j by

$$d^j \leqslant \mathrm{VFT}^j + E_{error}, \tag{3.5}$$

where E_{error} is an error term and VFT^j the j-th Virtual Finishing Time which is defined for $j \geqslant 1$ as

$$\mathrm{VFT}^j = \max\{a^j, \mathrm{VFT}^{j-1}\} + \frac{l^j}{r}. \tag{3.6}$$

Here, a^j is the arrival time of packet j with the corresponding packet length l^j while the server has a rate r. Note that initially, $\mathrm{VFT}^0 = 0$. For a server with a constant rate of r, the error term E_{error} is interpreted as how late the server is. Knowing the arrivals a^j and the departures d^j, the rate r is then computed as the maximum throughput over backlog periods. The error term E_{error} is calculated by $E_{error} = \Theta - \frac{L^{min}}{r}$, where Θ is the maximum latency in each burst period and L^{min} the minimum packet length.

Contrary to that, Cetinkaya et al. [32] allocated resources with real-time performance requirements in a stochastic, black-box model using admission control criteria at the egress to compute mean service curves.

Besides the service identification in network calculus much research has been done in a familiar topic named the available bandwidth estimation. Here, the aim is to find the long-term average of the unused capacity in networks.

This can be done by passive or active measurements. Passive measurements [32, 101] use the given traffic to observe the departures and find an estimate. Therefore, all such methods highly depend on a parameter that is not under control. Active measurements has the advantage of choosing the probe traffic on its own. Typical probes are packet pairs that measure the gap of the two packets at the ingress and compare it at the egress [136]. Packet trains, where a sequence of constant bit rate traffic is send [10, 27, 83, 101, 103, 105, 108, 111], and packet chirps that use one packet train and increase the rate over time [101, 120]. We will discuss the question according to the right probing traffic in Chap. 6 and present a minimal probe, that is the optimal probe under certain conditions.

Anyway, the authors Bredel and Fidler analyzed in [28] the different types of probing methods and illustrated that in wireless networks rather, the fair share is measured than the available bandwidth.

For cellular networks, the authors in [132] explained that in, e.g., LTE networks, the resource allocation is done by the base station where several data packets can be concatenated or segmented to a transport block size (TBS) which is determined from the radio signal condition by the eNodeB. Thereby, the base station computes the TBS in a predefined transmission time interval (TTI) with a length of 1 ms. Hence, the inter-packet gaps highly depend on the resource allocation at the base station and can be in the order of magnitude of 10^{-6} seconds or additionally, the time of one or more TTIs. As a consequence, the available bandwidth estimation out of packet pairs by computing the dispersion of the inter-packet gap is inaccurate.

The authors then describe an active measurement method called "train of packet groups"(TPG), which is based on pathQuick [114]. In pathQuick, packets are sent in a packet train with a fixed inter-packet gap while increasing the packet size within this train. The available bandwidth is then calculated for the first packet size, where the end-to-end delay increases by using this packet size and dividing it by the inter-packet gap. In [132], the method is adapted to LTE networks, i.e., several packet groups are sent with

an inter packet-group gap of TTI. Starting with a low number of packets and small packet sizes in each group, the end-to-end delay is measured for the last packet and compared to the first one. If the difference is less than a TTI, so 1 ms, then the size of the packet group is less than the TBS, and another packet is added until the maximum number of packets per group is reached. Then, the packet sizes are successively increased, again, until the end-to-end delay for the last packet is larger than a TTI in comparison to the first packet. Finally, the available bandwidth is computed for the first size of the packet group where the delay criteria is triggered, i.e., adding all packet sizes in one group together and dividing it by one TTI.

Another method that can handle specific properties in mobile networks is presented in [110]. Here, the authors describe a passive method that estimates the maximum capacity that a UE could receive. Similarly, as in [132], the packets are assigned to certain transport blocks by comparing the inter-packet gaps. Unlike before, this algorithm averages the transmitted data over two transmission blocks by carefully chosen the number of packets and dividing it by the time of the last in comparison to the first packet of this selected group of packets. At the end, the maximum overall estimation is taken as the maximum capacity a user can get.

Comparing all these methods, they have in common. They seek to find a single value for the available bandwidth or the maximum capacity or obtain an univariate estimate of the form $S^{\varepsilon}(t-\tau)$. Thus, a time-variant estimate is clearly not possible.

In order to find a time-variant service description we consider the min-plus algebra in network calculus, again. Here a service identification tries to solve $D(t) = \inf_{\tau\in[0,t]}\{A(\tau) + S(\tau,t)\}$ for $S(\tau,t)$. Because an inverse of the infimum operator does only exists in special cases [101], but not in general, it is not trivial to get an adequate solution for $S(\tau,t)$. To find time-variant service curves, we present two known methods, rate scanning (Sec. 3.2.1) and burst response (Sec. 3.2.2), for a measurement-based estimation of the system's service.

We explain the fundamental limitations of these methods to measure transient effects accurately, which motivates us to build up a new theory and measurement method to overcome these limitations in the next chapters.

3.2.1 *Rate Scanning*

First, we consider the rate scanning method from [101, 105] that fulfills the definition of the ε-effective service curve (3.4). It is derived in min-plus [104] and max-plus [103, 105] algebra.

In [104] Luebben, Fidler and Liebeherr extended the rate scanning method from [101] and formulated it in max-plus algebra. In max-plus algebra, the sum (or integral in the continuous case) becomes a maximum and the multiplication a plus sign. More information about the max-plus algebra can be found in [17, 37, 88, 97]. In comparison to the min-plus formulation of the rate scanning method, the advantage is that the estimate can directly be obtained by packets timestamps at the sender and receiver.

From the measurement-based approach, the authors are able to estimate an ε-effective service curve out of the steady-state delay percentiles of the used probing traffic. A min-plus formulation of the max-plus ε-effective service curve from [104] is shown in [105]. Further, the authors proved a connection from the available bandwidth to the left-over service process in network calculus. It lays the foundation for a bandwidth estimation or system identification of networks with random service, e.g., in wireless and cellular networks.

Here, we will concentrate on the min-plus formulation of the method. For the system identification it uses constant rate probes $A(t) = rt$ for a set of rates $r \in \mathbb{R}$. We obtain the backlog $B(r,t)$ at time t for the rate r from the departures $D(t)$, i.e., $B(r,t) = rt - D(t)$. Due to a random service $B(r,t)$ becomes a random variable, too. Repeating the measurements for every arrival rate $r \in \mathbb{R}$ we are able to compute quantiles of the backlog distribution. Generally, the quantile is defined as

$$X^{\xi} = \inf\{x \geqslant 0 : P[X \leqslant x] \geqslant 1 - \xi\}. \tag{3.7}$$

Hence, we get the backlog quantiles by

$$B^{\xi}(r,t) = \inf\{x \geqslant 0 : P[B(r,t) \leqslant x] \geqslant 1 - \xi\} \tag{3.8}$$

where ξ is the quantile's violation probability. By letting $t \to \infty$ the backlog for the rate r becomes independent to the time t and tends towards the steady-state distribution $B^{\xi}(r)$. Finally, the ε-effective service curve from

the rate scanning method is obtained by taking the maximum out of all probing rates $r \in \mathbb{R}$, such that it holds that

$$S^{\varepsilon}_{rs}(t) = \max_{r \in R}\{rt - B^{\xi}(r)\}, \tag{3.9}$$

where we used the union bound and get the violation probability $\varepsilon = \sum_{r \in R} \xi$. The subscript rs indicates the relation of the ε-effective service curve to the rate scanning method. Due to the union bound and the sum of the single violation probability ξ, ε increases by the number of rates. Therefore, the set of probing rates $\mathbb{R}$ has to be selected wisely.

With the help of tests, we find a maximum probing rate as in [101, 105], the spacing between the rates can be, e.g. linear or geometrical.

Note that Eq. (3.9) has the form of a Legendre-Fenchel transform of the backlog [101] and has various useful properties, also in the network calculus [72, 79]. Since the results lead generally towards convex function [123] it is also known as convex conjugate. The convexity has the property that the Legendre-Fenchel transform is its own inverse. In contrast, for non-convex functions a convex hull of the function can be obtained [123].

However, due to the stationarity of the rate scanning method, we are not able to follow transient effects, because of the same reasons as seen before. In Sec. 6.1 we adapt the method and extend it to a time-variant description and show further limitations which are based on the convex form of the service curve.

3.2.2 *Burst Response*

Next, we introduce the burst response method. The basic idea, we already stated in Chap. 2 where we compared the min-plus with the system theory. We formulated that under min-plus algebra the service curve is the response of the system and that sending the neutral element $\delta(t)$ of min-plus convolution results in the service $S(t)$ [96], where $\delta(t)$ is defined as in Eq.(2.9), i.e.,

$$\delta(t) = \begin{cases} 0 & \text{for } t = 0, \\ \infty & \text{for } t > 0. \end{cases}$$

Sending $\delta(t)$ as probe, i.e., $A(\tau) = \delta(t)$ corresponds to a burst probe and reveals the service $S(0,t)$ for all $t \geqslant 0$ since

$$D(t) = \inf_{\tau \in [0,t]} \{\delta(\tau) + S(\tau, t)\} = S(0,t). \tag{3.10}$$

Under the assumption of additive service processes as defined in [86, p. 6,7] and Appendix Def.3 a time-variant service $S(\tau, t)$ can be obtained for all $t \geqslant \tau \geqslant 0$ as

$$S(\tau, t) = S(0,t) - S(0,\tau). \tag{3.11}$$

The advantage of this method is that we can identify the service $S(t)$ out of the departures $D(t)$.

Apart from these useful features, the burst probing method has limitations. For certain systems, such as FCFS multiplexer, burst probes cause non-linear behavior. Further, the burst traffic preempts other traffic, which leads to an overly optimistic estimate of the service [101]. Moreover, the intuitive and basic assumption of additive service processes [86], which is crucial for this method, is not generally justifiable.

Despite these limitations of burst probing, we will carry out further investigations and present results showing the deviation from additivity. It can be demonstrated that, e.g., a stationary latency leads to non-additive services. This is an essential point since a stationary latency can be viewed as a one-way delay. Because of the fact that every real system has a OWD larger than zero, we conclude that the estimate from burst probing always leads to an overestimation of the service.

However, we are able to overcome these problems and present a new measurement-based estimation method consisting out of two phases in which burst probes play an important role.

4

PROBLEM STATEMENT

In order to analyze, optimize, and to further develop computer networks, much research is done. By doing so, a lot of effects occur that influence the performance. Some of them have a temporary impact. Examples are the TCP slow-start and, in particular, the choice of the initial congestion window [109], convergence of routing protocols [75], and the DRX mode in cellular networks to save battery power of mobile devices [21].

Regarding the DRX mode, the research focuses on topics such as the optimization of the DRX cycle lengths versus battery drain [23, 119, 144, 156, 157]. To model the DRX states, one way is to use semi-Markov chains to analyze the stationary mean wake-up times [23, 144, 148, 156, 157]. Further, it is possible to obtain stationary queuing delays for the DRX by using an M|G|1 model with vacations [147]. Similar to these examples, many works assume stationarity to obtain steady-state solutions.

A framework that works without the assumption of stationarity is the deterministic network calculus [37, 52, 96]. It uses the min-plus theory from Sec. 2.1. Here, the worst-case scenarios are considered by replacing time-variant and non-stationary systems with time-invariant linear bounds. This allows an analysis of transient phases, with the disadvantage that it stays in the maximum backlog and delay. An extension to time-variant systems is possible, as presented in Sec. 3.1. It already shows the advantage of investigating systems in a time-variant way. More precisely, in Fig. 3.1 we compare time-variant with time-invariant systems. By having deterministic service curves, we demonstrate that only the time-variant description of the service curve follows the exact solution, whereas the time-invariant remains in the worst-case backlog.

Depending on the parameters for the arrivals and service curve, the difference between these curves may even increase. Fig. 3.2 illustrates this and shows the impact of, e.g., the arrival rate α on the maximum overshoot after the wake-up time T and that the relaxation can reach values that are a magnitude larger than T.

Although these results look promising, in practice, a constant service rate with deterministic wake-up times is not realistic, e.g., due to the random nature of wireless channels and random wake-up times [21].

So, the stochastic network calculus takes time-variant systems also into account, but either assumes stationary random processes or uses stationary bounds [46], see Sec 2.2. The extension to time-dependent stochastic processes is also considered in the current literature where, e.g., end-to-end delay bounds for wireless multi-node networks and sink tree topolo-

gies [33, 112, 113] are derived. The progression of the backlog and delay over time is not considered.

Therefore, we address the problem of:

1.) Analysis of transient phases in systems with random wake-up times and random services.

To do this we extend the theory from deterministic to stochastic network calculus and in particular the ε-effective service curve from Sec. 3.1.3 to a bivariate and therefore time-dependent formulation. Due to the fact that certain properties of the min-plus algebra, which hold in the univariate case, are not valid anymore with bivariate functions, such as the commutativity, makes the derivation and analysis of these systems more complex.

By finding a solution for this challenge, we are able to follow accurately backlog and delay progressions over time, even in the stochastic case. This way, we can, e.g., model and simulate non-stationary service characteristics that help us to investigate DRX wake-up times in cellular networks, which enables us to find new insights into the behavior of such networks. Knowing the stochastic features of the service enables us to illustrate the impact of random wake-up times on the transient overshoot and relaxation times for backlog and delay. As we will see in section 7.5, the relaxation time can take several seconds for transmission technologies such as EDGE and HSPA. Therefore, possible video applications are likely to be lagged, which reduces the user experience.

Since we cannot assume to have knowledge about the service every time, we also consider methods that estimate the service curve from measurements out of probe traffic.

On the one hand, we have methods that seek to find the available bandwidth [132] or maximum capacity [110] of, e.g., cellular networks. Even though these methods provide promising results, they are not able to estimate non-stationary service characteristics, since we only get a single value for the available bandwidth and maximum capacity, respectively.

On the other hand we know methods from network calculus, such as the rate scanning method that can represents the available bandwidth as a service curve in network calculus [101] and estimates the unknown service for systems with random service [103, 105]. Although estimates over time-intervals are obtained, a major drawback to find a non-stationary service curve is because of, e.g., the assumption of stationarity and as we will see in Sec. 5.2.1 transient phases can be analyzed by the non-convex part of service curves, whereas the rate scanning always provides convex functions. Similarly, we have the burst response method as in Sec. 3.2.2 that yields time-variant service descriptions but overestimates the service in case of non-zero OWDs and preempts other traffic. Therefore, we have that a measurement-based estimation of non-stationary service curves is not present in the current literature such that we address the problem of:

2.) Conservative estimate of non-stationary service curve in systems with unknown service.

To meet this challenge, we have to solve the problem of the overestimation and find a function that is non-convex by construction. Thereby, we have to find a suitable probe traffic. Intuitively, a probe that is too small will provide little information about the service, since the observed departures are limited by the arrivals. On the other hand, a probe that is too large will deteriorate the estimate, e.g., in the extreme case of a burst probe [101].

We present our solution by introducing the novel method of minimal probing. This method consists of two steps. In the first step, a minimal probe is estimated, which is optimal under certain conditions. In the second part, we use the minimal probe to estimate a conservative and time-variant service curve.

We show the advantages in Sec. 7.4, where we perform a first proof-of-concept study in cellular networks to demonstrate the applicability. We show maximum overshoots and massive relaxation times of up to several seconds in these networks. When examining the question of the optimal traffic to be sent into the network, we come up with the minimal probe that eliminates the transient effects so that stationary buffers and latencies are achieved as quickly as possible, taking into account the characteristics of, e.g., the DRX implementations.

5 THE NON-STATIONARY SERVICE CURVE

As seen so far, dealing with time variance is not trivial. Despite the fact that new results have occasionally been produced in recent years, in stochastic network computation, time-dependent formulations are usually either replaced by stationary envelopes or modeled by random processes including stationary bounds, see [46].
One example for the use of stationarity we have already seen in Sec. 3.1.3 where we introduced the ε-effective service curve, which uses a univariate description of the service process.
In the following, we extend the ε-effective service curve to a formulation that takes transient changes into account. As in [11, 40], we consider time-variant systems using bivariate instead of univariate functions and obtain non-stationary characteristics of the service. Based on this, we are able to model systems with non-deterministic sleep scheduling, such as in cellular networks with their stochastic wake-up times of UE's to establish a connection after entering an idle state, see [21]. Due to this non-stationary service curve model, we get insights into transient phases and can analyze the occurring effects. It is an excellent basis to investigate the impact of sleep scheduling on the service quality, where we are especially interested in measures such as the transient overshoot and the relaxation time until the steady-state is approached, see Fig. 3.2.

In this chapter, we introduce regenerative processes and the derivation of the non-stationary service curve that includes a method of construction. We introduce an example for systems with random wake-up times and show results for the delay and backlog. We present a representation of the service curve that explains how we are able to analyze transient changes. It allows us later to analyse the effects of sleep scheduling in cellular networks from measurements. Please note that we present among others results from [19, 20].

5.1 REGENERATIVE SERVICE PROCESSES

Contrary to the current literature, that makes the assumption of stationary service processes, we assume throughout this work that the service is a

regenerative process [126] with regeneration points $\mathbb{P} = \{P_0, P_1, P_2, \dots\}$ where $P_0 = 0$ and $P_i < P_{i+1}$ for all $i \geqslant 0$. We divide the service $S(\tau, t)$ into segments separated by the regeneration points, i.e., for all $0 \leqslant \tau \leqslant t \leqslant P_{i+1} - P_i$ and $i \in \mathcal{N}_0$ we have

$$S_i(\tau, t) = S(\tau + P_i, t + P_i). \tag{5.1}$$

Thus, between the ith and $(i+1)$th regeneration point the service is described by $S_i(\tau, t)$. So, for all $i, j, x \geqslant 0$ and $0 \leqslant \tau \leqslant t \leqslant \min\{P_{i+1} - P_i, P_{j+1} - P_j\}$ we have that $S_i(\tau, t)$ are statistical replicas in the sense that

$$P[S_i(\tau, t) \leqslant x] = P[S_j(\tau, t) \leqslant x]. \tag{5.2}$$

Hereinafter we omit the index i due to Eq. (5.1) where we assume that $t \leqslant P_i$, which means that the next regeneration point is spaced sufficiently apart.

5.2 NON-STATIONARY SERVICE CURVES

Now, we consider $S(\tau, t)$ as a non-stationary random service process and derive a time-variant, i.e., bivariate lower service envelope function $\mathcal{S}^\varepsilon(\tau, t)$ that conforms to

$$P[S(\tau, t) \geqslant \mathcal{S}^\varepsilon(\tau, t),\ \forall \tau \in [0, t]] \geqslant 1 - \varepsilon, \tag{5.3}$$

for all $t \geqslant 0$, where $\varepsilon \in (0, 1]$ is a probability of underflow. Here, the essential difference to the state-of-the-art envelope functions [31], such as the ε-effective envelope from Eq. (3.4) and the arrival envelope from Eq. (2.25) is the use of bivariate instead of univariate functions, which is crucial for modeling transient phases, see Sec. 2.1.2 and 2.2.2.
By adding $A(\tau)$ on both sides, we obtain

$$P[A(\tau) + S(\tau, t) \geqslant A(\tau) + \mathcal{S}^\varepsilon(\tau, t),\ \forall \tau \in [0, t]] \geqslant 1 - \varepsilon, \tag{5.4}$$

where Eq. (5.4) makes a sample path argument for all $\tau \in [0, t]$. Thus, it holds that

$$P\left[\inf_{\tau \in [0,t]} \{A(\tau) + S(\tau, t)\} \geqslant \inf_{\tau \in [0,t]} \{A(\tau) + \mathcal{S}^\varepsilon(\tau, t)\}\right] \geqslant 1 - \varepsilon.$$

Substituting the definition of the min-plus convolution $D(t) = \inf_{\tau\in[0,t]}\{A(\tau) + S(\tau,t)\}$ leads to

$$P[D(t) \geqslant A \otimes \mathcal{S}^\varepsilon(t)] \geqslant 1-\varepsilon, \tag{5.5}$$

for all $t \geqslant 0$. We refer to $\mathcal{S}^\varepsilon(\tau,t)$ as *non-stationary service curve*. It extends [31] and has the capability to analyze transient changes over time. Note that for univariate functions the convolution is commutative, whereas for bivariate functions the convolution is bivariate , again, but not commutative, i.e., $S_1 \otimes S_2(\tau,t) \neq S_2 \otimes S_1(\tau,t)$.

In order to derive a non-stationary service curve $\mathcal{S}^\varepsilon(\tau,t)$ we use similar steps as for the arrival envelope from Eq. (2.26) and obtain

$$\mathcal{S}^\varepsilon(\tau,t) = -\frac{1}{\theta(\tau,t)}(\ln M_S(-\theta,\tau,t) + \rho(t-\tau) - \ln(\rho\varepsilon)), \tag{5.6}$$

where $\theta(\tau,t) > 0$ and $\rho \in (0, 1/\varepsilon]$ are free parameters and $M_S(-\theta,\tau,t) = E[e^{-\theta S(\tau,t)}]$ is the negative MGF, respectively, Laplace transform.

Derivation of Eq. (5.6)

For the derivation of Eq. (5.6) we start with Eq. (5.3) and use the complementary formulation

$$\xi := P[\exists\tau \in [0,t] : S(\tau,t) < \mathcal{S}^\varepsilon(\tau,t)] \leqslant \varepsilon.$$

We show with basic steps from the stochastic network calculus [47, 98] that $\xi \leqslant \varepsilon$, which proves that $\mathcal{S}^\varepsilon(\tau,t)$ from (5.6) satisfies Eq. (5.3).

It holds for a set of free parameters $\theta(\tau,t) \geqslant 0$, that

$$\xi \leqslant \sum_{\tau=0}^{t-1} P[S(\tau,t) < \mathcal{S}^\varepsilon(\tau,t)] \leqslant \sum_{\tau=0}^{t-1} e^{\theta(\tau,t)\mathcal{S}^\varepsilon(\tau,t)} M_S(-\theta,\tau,t),$$

where we used the union bound and additionally the Chernoff bound, i.e., for a random variable X and every $\theta \geqslant 0$, it follows that $P[X \leqslant x] \leqslant e^{\theta x} M_X(-\theta)$. Since we have $S(t,t) = 0$ and $\mathcal{S}^\varepsilon(t,t) \leqslant 0$ by definition the case $\tau = t$ is omitted, here. Now, by insertion of $\mathcal{S}^\varepsilon(\tau,t)$ from Eq. (5.6) it follows that

$$\xi \leqslant \rho\varepsilon \sum_{\tau=0}^{t-1} e^{-\rho(t-\tau)} = \rho\varepsilon \sum_{v=1}^{t} e^{-\rho v} \leqslant \rho\varepsilon \int_0^\infty e^{-\rho y}\,dy = \varepsilon,$$

where each summand is bounded by $e^{-\rho v} \leqslant \int_{v-1}^{v} e^{-\rho y} dy$ since $e^{-\rho v}$ is decreasing. Finally, letting $t \to \infty$ and solving the integral completes the proof that $\xi \leqslant \varepsilon$ for all $t \geqslant 0$.

5.2.1 *Random Sleep Scheduling*

The consideration of the deterministic cases from Chap. 2 and Chap. 3 in addition to the fact that the amount of time it takes for a UE to leave the idle state and establish a connection in LTE's DRX-mode is random [21], substantiates the need of a time-variant theory with a random wake-up time T in the network calculus. Note that it is also possible to use Semi-Markov models as in [23, 144, 148, 156, 157] to analyze mean wake-up delays.

Using the concept of non-stationary service curve, we consider a work-conserving system with random sleep scheduling and random service increments $Z(t)$ for $t \geqslant 0$. We get the service process for all $t > \tau \geqslant 0$ as $S(\tau, t) = \sum_{v=\tau+1}^{t} Z(v)$. As usual, $S(t, t) = 0$ for all $t \geqslant 0$.

The system regenerates in the moment it enters a sleep state and wakes up after a random amount of time $T \geqslant 0$. Hence, it follows for every time instances $t \leqslant T$ that $Z(t) = 0$. For the interval $(\tau, t]$ and $t > T$ we have to distinguish whether $\tau \leqslant T$ or $\tau > T$. Thus, we have to count the number of usable time-slots after T, leading to

$$U(\tau, t) = [t - \max\{\tau, T\}]_+.$$

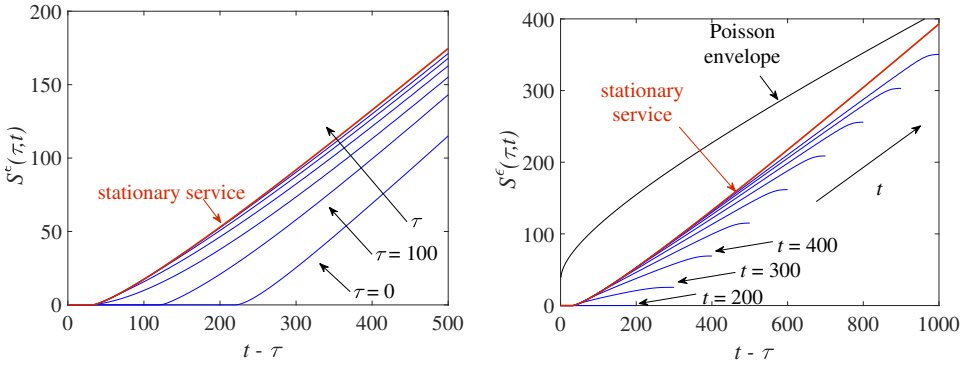
To get the non-stationary service curve from Eq. (5.6) the MGF of $U(\tau, t)$ is needed. In our case, it consists out of three terms

$$M_U(\theta, \tau, t) = e^{\theta(t-\tau)} P[T \leqslant \tau] + \sum_{v=\tau+1}^{t} e^{\theta(t-v)} P[T = v] + P[T > t], \quad (5.7)$$

and correspond to the cases described above. Let's assume iid service increments $Z(t)$ for $t > T$ with MGF $M_Z(\theta)$. Then, the MGF of the service process is

$$M_S(\theta, \tau, t) = E\left[(M_Z(\theta))^{U(\tau,t)}\right] = M_U(\ln M_Z(\theta), \tau, t). \quad (5.8)$$

In order to present numerical results, we model T as a geometric random variable with parameter p, where $P[T = v] = p(1-p)^v$, and $Z(t)$ for $t > T$



(a) $\tau \in \{0, 100, 200, \dots\}$ (b) $t \in \{100, 200, \dots\}$

Figure 5.1: Non-stationary service curves of random sleep scheduling.

following a basic wireless outage model [73] as iid Bernoulli trials with parameter q. Due to the memorylessness of the processes, solutions for this specific case may also be derived, e.g., from a Markov model. We note that the service curve in Eq. (5.6) is not limited to memoryless processes. However, it enables us to compute certain reference results in Chap. 6.

Now, consider the three terms in Eq. (5.7). The probability for the first term follows from the geometric random variable, immediately. So, $P[T \leqslant \tau] = 1 - (1-p)^{\tau+1}$. The second one is

$$\sum_{\upsilon=\tau+1}^{t} e^{\theta(t-\upsilon)} \mathcal{P}[T = \upsilon] = e^{\theta t} p \sum_{\upsilon=\tau+1}^{t} \left(e^{-\theta}(1-p)\right)^{\upsilon},$$

where we substitute $y = e^{-\theta}(1-p)$ and compute the geometric series as

$$\sum_{\upsilon=\tau+1}^{t} y^{\upsilon} = \frac{y^{\tau+1} - y^{t+1}}{1-y}.$$

Next, we need the MGF of the Bernoulli service increments, which is $M_Z(\theta) = qe^{\theta} + 1 - q$. Inserting the terms into Eq. (5.8) yields the MGF for the service process $M_S(\theta, \tau, t)$. Finally, putting it into Eq. (5.6) results in a valid representation of a non-stationary service curve.

In the following, we present first results for the non-stationary service curve. The parameters are $p = 0.1$ and $q = 0.5$. Hence, the mean stationary service rate is $E[Z] = q = 0.5$ and the mean transient latency is $E[T] = (1-p)/p = 9$. For the service curves we choose $\rho = 10^{-4}$ and $\varepsilon = 10^{-6}$ as the probability of underflow, where $\theta(\tau, t)$ is optimized numerically.

In Fig. 5.1 $\mathcal{S}^{\varepsilon}(\tau, t)$ is shown to illustrate the effects of sleep scheduling on the service curve by considering two different perspectives on the representation of $\mathcal{S}^{\varepsilon}(\tau, t)$. In Fig. 5.1a, we fix τ and compute the corresponding service curve. For small τ, we are close to the last regeneration point. Therefore, the probability that the service has not started yet is immense. For sure, by increasing the width of the interval $t - \tau$, the service increases, as well. Moreover, the influence of the transient phase that occurs at the beginning of each regeneration point decreases with increasing τ. That makes sense since we extend the distance to the last regeneration point and reduces so the impact of the wake-up time T. Further, we observe that for large τ, the service $\mathcal{S}^{\varepsilon}(\tau, t)$ converges towards a stationary service curve. It is marked as the red curve. It is the same service increment process as before but without sleep scheduling.

Fig. 5.1b shows the service curve, again. Here, we change the perspective, i.e., we do not fix τ but t, instead. In this case, τ is variable. A small width of the intervals means that we increase τ for a fixed t. It corresponds to a decreasing influence of the transient sleep phase. Interestingly, we observe a non-negligible part at the beginning of the curve where the service is zero, e.g., for $t = 100$ and $t = 200$. This is due to the outages of the Bernoulli service increments process. The fact that the stationary service curve has the same progress confirms the statement.

Besides the outages on the left-hand side of the service curve, we observe a non-convex shape on the right-hand side, where τ decreases and tends to zero. It leads to a higher impact of the sleep scheduling and, therefore, of the transient phase. We emphasize that the non-convex part is essential to analyze transient phases. Thus, from here on, we prefer the presentation from Fig. 5.1b to display a non-stationary service curve.

5.2.2 *Performance Bounds & improved Arrival Envelopes*

Apart from the fact that we are able to show the transient part of the service curve in Fig. 5.1b, it additionally matches with the way we compute statistical performance bounds by fixing t and evaluating, e.g., the backlog for all other $\tau \in [0, t]$ and taking the maximum out of it, see Eq. (2.27). To present performance bounds for the non-stationary service curve we choose the Poisson arrivals from Sec. 2.2.2 and Sec.3.1.

There, we derived an arrival envelope based on the union bound and Chernoff's theorem and compared the backlog for time-variant and invari-

ant services with an exact solution. We observed that only the time-variant service could follow the progress of the exact solution in Fig. 3.1. Additionally, we noticed a non-negligible difference between these two curves, which is due to the use of union and Chernoff bound. To find a better approximation, we use a technique from [85, 118] applying Doob's martingale inequality [61, Theorem 3.2, p. 314]. In particular we have for a martingale, (see Appendix Def.7) X_n and $\eta > 0$ that

$$P\left[\sup_n X_n \geqslant \eta\right] \leqslant E[X_0]\eta^{-1}. \tag{5.9}$$

Then, an arrival envelope for processes with iid increments is

$$\mathcal{A}^\varepsilon(t) = \frac{1}{\theta}(\ln M_A(\theta, t) - \ln \varepsilon), \tag{5.10}$$

where $M_A(\theta, t) = E[e^{\theta A(\tau,\tau+t)}] = (M_A(\theta, 1))^t$ for $\tau, t \geqslant 0$ is the MGF of $A(t)$ and $\theta > 0$ is a free parameter.

Derivation of Eq. (5.10)

Lets take the envelope

$$\mathcal{A}^\varepsilon(t) = \rho_A t + \sigma_A.$$

where $\rho_A > 0$ is the rate and $\sigma_A \geqslant 0$ the burstiness parameter. In order to get ρ_A and σ_A, we insert $\mathcal{A}^\varepsilon(t)$ into Eq. (2.23) and derive for $\theta > 0$

$$\begin{aligned}
&1 - P[A(\tau, t) \leqslant \rho_A(t-\tau) + \sigma_A, \forall \tau \in [0, t]] \\
&= P\left[\max_{\tau\in[0,t-1]} \{A(\tau, t) - \rho_A(t-\tau)\} > \sigma_A\right] \\
&= P\left[\max_{\tau\in[1,t]} \left\{e^{\theta(A(t-\tau,t)-\rho_A\tau)}\right\} > e^{\theta\sigma_A}\right],
\end{aligned}$$

where $A(t, t) = 0$. Now, we consider for fixed $t > 0$ the process $U(\tau) = e^{\theta(A(t-\tau,t)-\rho_A\tau)}$, for $\tau \in [0, t]$. Then, it follows that

$$\begin{aligned}
U(\tau + 1) &= U(\tau)e^{\theta(A(t-\tau-1,t-\tau)-\rho_A)} \\
&= U(\tau)e^{\theta A(t-\tau-1,t-\tau)}e^{-\theta\rho_A}.
\end{aligned}$$

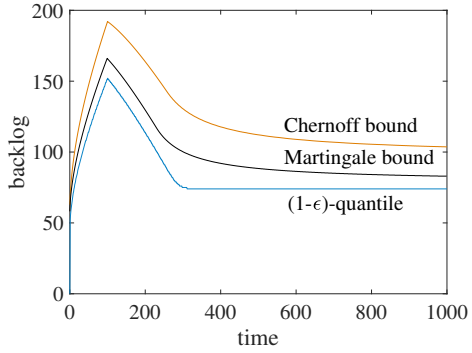


Figure 5.2: Progression of the transient backlog over time. Comparison of the Poisson arrival process obtained by Chernoff's vs. Martingale bound in relation to the $(1-\varepsilon)$-quantile of the exact solution. The Martingale bound clearly improves the estimate.

By using the property of independence the conditional expectation becomes

$$\begin{aligned}&\mathsf{E}[U(\tau+1)|U(\tau),U(\tau-1),\dots,U(1)]\\=&U(\tau)\mathsf{E}[e^{\theta A(t-\tau-1,t-\tau)}]e^{-\theta\rho_A}.\end{aligned}$$

Moreover, we have for iid increments $\mathsf{E}[e^{\theta A(t-\tau-1,t-\tau)}] = M_A(\theta,1)$, that $\mathsf{E}[U(\tau+1)|U(\tau),U(\tau-1),\dots U(1)] = U(\tau)$ is a martingale if $e^{\theta\rho_A} = M_A(\theta,1)$. Therefore, we get for the parameter $\rho_A = \ln M_A(\theta,1)/\theta$ such that for $t > 0$ we have $\rho_A t = \ln M_A(\theta,t)/\theta$. By using Doob's martingale inequality from Eq. (A.2) and the reformulation from [85] it follows for non-negative martingales $U(\tau)$ for $\tau \geqslant 1$ that

$$x\mathsf{P}\left[\max_{\tau\in[1,t]}\{U(\tau\} \geqslant x\right] \leqslant \mathsf{E}[U(1)].$$

With $\mathsf{E}[U(1)] = 1$ and $x = e^{\theta\sigma_A}$ we have that

$$\mathsf{P}\left[\max_{\tau\in[0,t]}\{A(\tau,t) - \rho_A(\theta)(t-\tau)\} > \sigma_A\right] \leqslant e^{-\theta\sigma_A}.$$

Finally, we let $\varepsilon = e^{-\theta\sigma_A}$ such that we obtain $\sigma_A = -\ln\varepsilon/\theta$.

For a comparison of the two different arrival envelopes, i.e. the first envelope which uses the union bound and Chernoff's theorem and the second one that is based on Doob's martingale inequality, we use the same example as in Sec. 3.1 and Fig. 3.1. More precisely, we have the same Poisson arrival process, with parameter $\alpha = 0.09$, $\beta = 0.3$ and $\varepsilon = 10^{-9}$ and the deterministic transient service curve (3.2) with latency $T = 100$ and a rate of $R = 1$.

As seen from Fig. 5.2, both bounds provide good estimates and follow the shape of the $(1-\varepsilon)$-quantile for the time-variant service very well. The deviation to the exact solution is due to the inequalities that are invoked in the derivation of the envelopes. We observe that the use of Doob's martingale inequality is much closer at the exact solution and, therefore, clearly improves the bound from Chernoff's theorem.

As a consequence, in the following we use the arrival envelope obtained from Doob's inequality rather than Chernoff's bound.

Performance Bounds

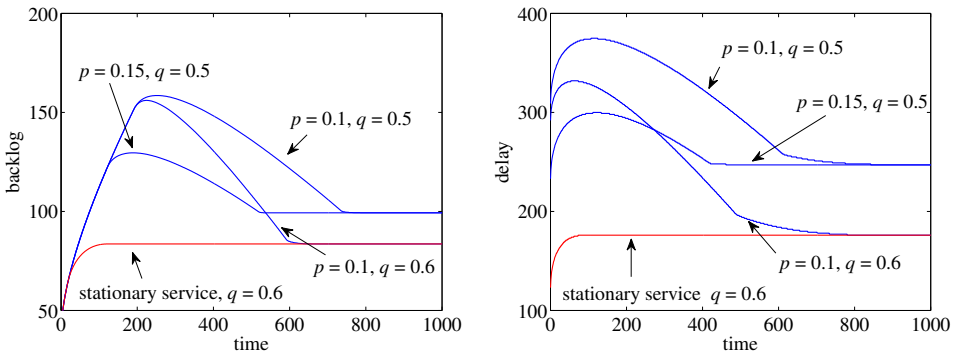
Next, we present performance bounds for non-stationary service curves $\mathcal{S}^\varepsilon(\tau,t)$ and statistical arrival envelopes $\mathcal{A}^\varepsilon(t-\tau)\ \forall t \geqslant \tau \geqslant 0$. A statistical backlog bound follows with Eqs. (5.5) and (2.21) as

$$\mathcal{P}\Big[B(t) \leqslant \sup_{\tau\in[0,t]} \{\mathcal{A}^\varepsilon(t-\tau) - \mathcal{S}^\varepsilon(\tau,t)\}\Big] \geqslant 1-2\varepsilon \tag{5.11}$$

and a first-come-first-served delay bound as

$$\mathcal{P}\Big[W(t) \leqslant \inf\Big\{w \geqslant 0 : \sup_{\tau\in[0,t]} \{\mathcal{A}^\varepsilon(t-\tau) - \mathcal{S}^\varepsilon(\tau,t+w)\} \leqslant 0\Big\}\Big] \geqslant 1-2\varepsilon, \tag{5.12}$$

where the ε for $\mathcal{A}^\varepsilon(t-\tau)$ and $\mathcal{S}^\varepsilon(\tau,t)$ is the same, such that we obtain an error probability of 2ε. Intuitively, the backlog and delay bound are the maximal vertical and horizontal deviation of $\mathcal{A}^\varepsilon(t-\tau)$ and $\mathcal{S}^\varepsilon(\tau,t)$, respectively. To visualize this, we added the Poisson arrivals to Fig. 5.1b.



(a) backlog (b) delay

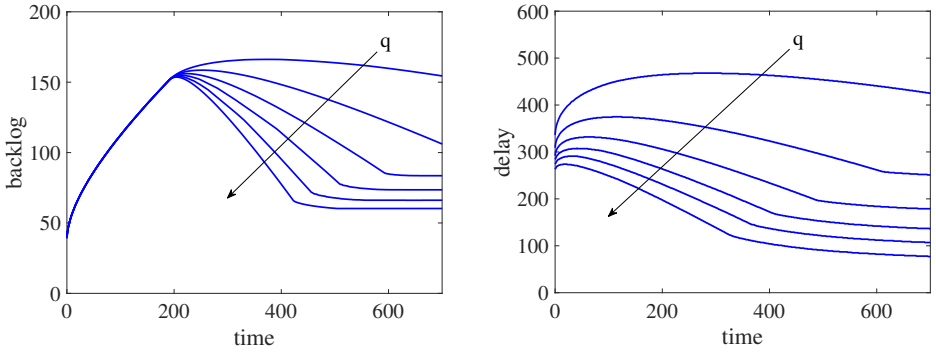
Figure 5.3: Transient backlog and delay of random sleep scheduling.

Now, we choose for the arrival process $\alpha = 0.06$, $\beta = 0.3$, and $\varepsilon = 10^{-6}$. Then, in Fig. 5.3, we show for different parameters of the service p and q the backlog and delay for this doubly stochastic system, where the long-term utilization can be computed as $\alpha/(\beta q)$. As before, the free parameters θ and ρ are optimized numerically. First of all, we observe that the transient overshoot and relaxation time, i.e., the time it takes to reach steady-state, is immense in comparison to the stationary case, which is the same stochastic process but without any wake-up time and is marked as the red curve. In this example the mean transient latency is computed by $\mathsf{E}[T] = (1-p)/p$ and only depends on p. Since $p \in (0, 1]$ $\mathsf{E}[T]$ increases with decreasing p. Clearly, the higher the transient latency, the longer the relaxation time, i.e., the time it takes to reach stationarity.

Now, we take a deeper look into the transient behavior. We choose $p = 0.1$ and vary only the service rate q, where the mean stationary service rate is $\mathsf{E}[Z] = q$. In Fig.5.4, we present the results for the backlog and delay. We observe that for both bounds, the performance becomes worse for a decreasing service q, while p is constant. Or vice versa if we look at the delay in Fig. 5.4b, we notice that as q increases, the transient overshoot becomes smaller and smaller, with the maximum tending to the origin at $t = 0$. For the backlog in Fig. 5.4a an increasing value of q leads to a reduction of the transient overshoot and, therefore, to faster relaxation times.

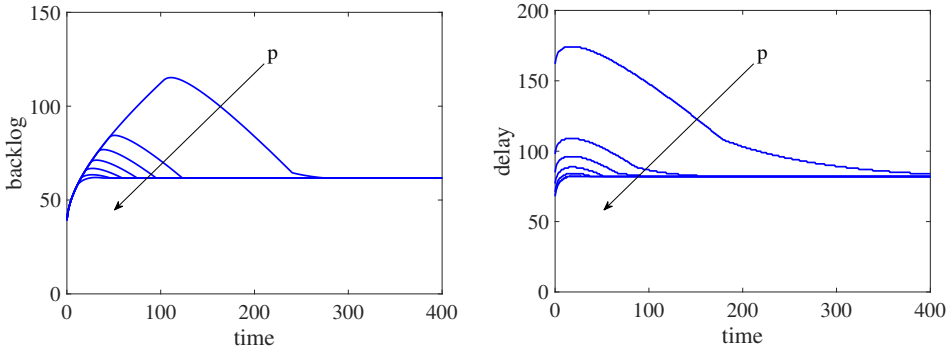
The effect of different parameters of p for fixed $q = 0.9$ is shown in Fig. 5.5. Here, a higher value of p tends to the stationary progress as in Fig. 5.3.

As a consequence, both parameters have a high impact on the performance bounds.



(a) backlog (b) delay

Figure 5.4: Transient backlog and delay of random sleep scheduling with parameter $p = 0.1$ and $q \in [0.4, 0.5, \ldots, 0.9]$. For increasing service rates q the transient overshoot and relaxation times reduces significantly.



(a) backlog (b) delay

Figure 5.5: Transient backlog and delay of random sleep scheduling with parameter $q = 0.9$ and $p \in [0.3, 0.4, \ldots, 0.9]$. For decreasing wake-up times the transient overshoot and relaxation times reduces significantly.

6

MEASUREMENT-METHODS

In order to analyze the transient behavior of systems with sleep scheduling, we derived a model-based approach in Sec. 5.2. There, we have full knowledge of the system's arrivals and services. In real networks, however, this might not be valid. For example, in cellular networks, we cannot guarantee to know anything about the system's internals. Thus, we expect a black-box and only assume linearity. We already discussed measurement methods in black-box and grey-box systems in Sec. 3.2. To identify the system's service curve, we use the measurement methods from Sec. 3.2, i.e., the rate scanning and burst probing method where we adapt the first one such that we can estimate non-stationary service curves for transient phases, too. We show further limitations of the two methods, which are due to the non-convex shape and super-additivity of the services. To overcome these limitations, we design in Sec. 6.4 a new probing method, consisting out of two steps. In the first step, it determines the shape of a suitable probe. We prove that this probe is minimal under certain conditions and lead secondly to a conservative service curve estimate. Please note that we present among others results from [19, 20].

6.1 RATE SCANNING

We start with the rate scanning method from Sec. 3.2.1 and [101, 105]. We modify it such that the definition of a non-stationary service curve Eq. (5.5) is fulfilled. As before, we send constant rate probes $A(t) = rt$ for a set of rates $r \in \mathbb{R}$ and calculate the backlog $B(t)$ from $B(t) = \sup_{\tau\in[0,t]}\{r(t-\tau) - S(\tau,t)\}$. Thus, we also have the lower bound

$$S(\tau,t) \geqslant r(t-\tau) - B(t) \quad \forall\tau \in [0,t]. \tag{6.1}$$

In comparison to the related work we do not use a quantile (3.7) of the stationary backlog $B^{\xi}(r)$, but instead the transient version $B^{\xi}(r,t)$ at time t, where $B^{\xi}(r)$ and $B^{\xi}(r,t)$ indicates the dependence to the rate r. Inserting $B^{\xi}(r,t)$ in Eq.(6.1) yields the form from Eq. (5.3), such that it holds

$$P[S(\tau,t) \geqslant r(t-\tau) - B^{\xi}(r,t), \forall\tau \in [0,t]] \geqslant 1-\xi.$$

It follows with the union bound $\forall t \geqslant \tau \geqslant 0$ that

$$\mathcal{S}^{\varepsilon}_{rs}(\tau, t) = \max_{r \in \mathbb{R}} \{r(t-\tau) - B^{\xi}(r, t)\} \tag{6.2}$$

is a non-stationary service curve as in Eq. (5.5) with $\varepsilon = \sum_{r \in \mathbb{R}} \xi$. The subscript rs shows the relation of the service curve to the rate scanning method.

Next, we evaluate this method. We use the same example of random sleep scheduling from Sec. 5.2.1 with identical parameters $p = 0.1$ and $q = 0.5$. We select ten uniform distributed rates from 0.05 to 0.5, i.e., $r \in \{0.05, 0.1, \ldots, 0.5\}$. We choose 0.5 as maximum rate since the stationary service rate is 0.5, due to $q = 0.5$. Hence, a larger rate does not lead to more information. Because of the union bound in the derivation of the method ε increases with every rate and thus the error probability. For each rate we perform 10^5 repeated experiments and obtain the same amount of backlog samples. Then, for each rate r we compute a backlog quantile $B^{\xi}(r, t)$ for $\xi = 10^{-4}$. Thus, the overflow probability is $\varepsilon = 10^{-3}$ since $\varepsilon = \sum_{r \in \mathbb{R}} \xi$.

For $t = 200$ Fig. 6.1 shows the ten linear segments obtained by each of the probing rates $r \in \mathbb{R}$ (dashed lines). Taking the maximum results in the estimate $\mathcal{S}^{\varepsilon}_{rs}(\tau, t)$. Also, we added as a reference an analytical upper bound and an analytical service curve to the figure. If a function overshoots the upper bound for any $\tau \in [0, t]$ the definition of the service envelope (5.3) is violated. We now compute the bound from a Bernoulli service increment process with parameter q. As before, it starts after a geometrically distributed time T. For the interval $[\tau, t]$ an upper bound of the service is derived as the $(1-\varepsilon)$-quantile of a binomial distribution with parameters $t - \max\{\tau, T\}$ and q, where the number of trials $t - \max\{\tau, T\}$ is a random variable, too. For comparison, we added a second analytical service curve. Again, it follows from the binomial distribution but makes a sample-path argument using the union bound in addition. It can be seen as the lowest expected progression but without being a provable lower bound.

From Fig. 6.1 it is unambiguous that the estimate of the service curve from the rate scanning method cannot follow the pattern of the analytical results. The reason lies in the way the method is constructed. By taking the maximum overall rates $r \in \mathbb{R}$, it cannot recover the non-convex parts as it is fundamentally limited to a convex hull by construction [101]. In other words, taking the point-wise maximum out of a set of linear functions results in a convex function. In summary, we can say that the method is not suitable for analyzing transient phases.

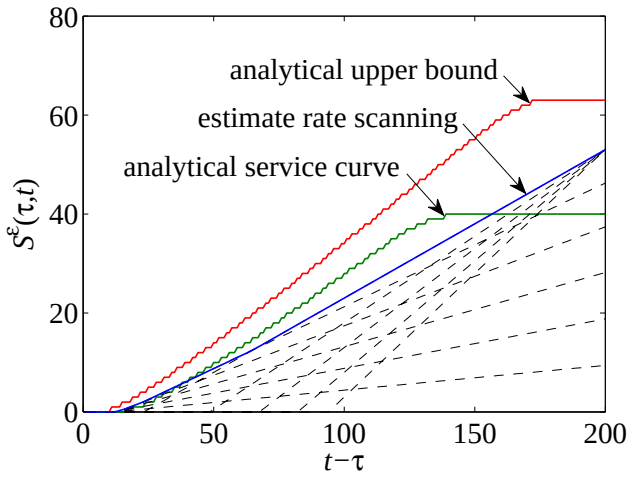


Figure 6.1: Service curve estimates compared to analytical results. The estimate from rate scanning is the maximum of linear rate segments (dashed lines). By construction it can only recover a convex hull.

6.2 BURST RESPONSE

The second method we want to investigate is the burst response method. We already introduced it in Sec. 3.2.2. There, we described that sending the canonical probe for the min-plus theory [96], which is the burst function $\delta(t)$ from Eq. (2.9) yields the service $S(0,t)$. As a reminder, in min-plus algebra, the burst function takes the role of the Dirac delta function and is the neutral element of min-plus convolution. Thus, for all $t \geqslant 0$ the service $S(0,t)$ is obtained by sending a burst probe $A(\tau) = \delta(\tau)$, i.e.,

$$D(t) = \inf_{\tau \in [0,t]} \{\delta(\tau) + S(\tau,t)\} = S(0,t). \tag{6.3}$$

Then, for additive service processes we can compute the service in the interval $(\tau,t]$ immediately from the departures as $S(\tau,t) = D(t) - D(\tau)$. Hence, with Eq. (6.3) we have

$$S(\tau,t) = S(0,t) - S(0,\tau), \tag{6.4}$$

for all $t \geqslant \tau \geqslant 0$, see [86, p. 6].

In order to obtain a stochastic service curve we refer Ω as the set of all feasible sample-paths $D_\omega(t)$ for all $\omega \in \Omega$. Using the assumption of

additivity (6.4) we obtain for each $\omega \in \Omega$ the service process from the burst response (6.3) as

$$S_\omega(\tau, t) = D_\omega(t) - D_\omega(\tau)$$

for all $\tau \in [0, t]$. Next, we select a subset of the sample-paths $\Psi_t \subseteq \Omega$ where we fixed $t > 0$. We discard the worst-case sample-paths as in Eq. (6.6) so it holds with probability $P[\Psi_t] \geqslant 1 - \varepsilon$. Then, we define

$$\mathcal{S}^{\varepsilon}_{br}(\tau, t) = \inf_{\psi \in \Psi_t} \{S_\psi(\tau, t)\}, \tag{6.5}$$

such that for all $\tau \in [0, t]$ and all $\psi \in \Psi_t$ it is true that $S_\psi(\tau, t) \geqslant \mathcal{S}^{\varepsilon}_{br}(\tau, t)$. Since $P[\Psi_t] \geqslant 1 - \varepsilon$ the service curve $\mathcal{S}^{\varepsilon}_{br}(\tau, t)$ satisfies Eq. (5.3) and so is a non-stationary service curve that conforms to Eq. (5.5).

For sure, in practice, the number of repeated measurements is finite, and so the number of feasible sample-paths Ω with the corresponding departures $D_\omega(t)$ at time $t \geqslant 0$. For a fixed $t \geqslant 0$, we select the minimal sample-path ϕ as the one that achieves the minimum

$$S_{\min}(\tau, t) = \min_{\omega \in \Omega} \{S_\omega(\tau, t)\} \tag{6.6}$$

for $\tau \in [0, t]$ most frequently. So, let $\phi = \arg\max_{\omega \in \Omega}\{X_\omega\}$, where

$$X_\omega = \sum_{\tau=0}^{t-1} 1_{S_\omega(\tau,t) = S_{\min}(\tau,t)},$$

for all $\omega \in \Omega$, with indicator function $1_{(\cdot)}$, which is one if the arguments is true and zero otherwise. We remove this sample-path ϕ and obtain the remaining set $\Psi_t = \Omega \backslash \phi$. Then, we repeat the steps described above as long as $P[\Psi_t] \geqslant 1 - \varepsilon$. In other words, for every $\tau \in [0, t]$ we choose the sample-path $\omega \in \Omega$ which has the lowest service in the interval $(\tau, t]$. The sample-path which attains the minimum the most we remove from the set Ω. We repeat this procedure depending on how many sample-paths we want to remove. At the end the non-stationary service curve estimate is obtained from Eq. (6.5) for all $\tau \in [0, t]$.

For the example from the previous section and Fig. 6.1 we present a service curve estimate for burst probing with timestamps $t = 200$, 300 and 400 in Fig. 6.2. We use the same parameters, i.e. the number of repeated experiments is 10^5 with $p = 0.1$, $q = 0.5$ and $\varepsilon = 10^{-3}$. Again, with analytical upper bounds and analytical reference service curves. In comparison to

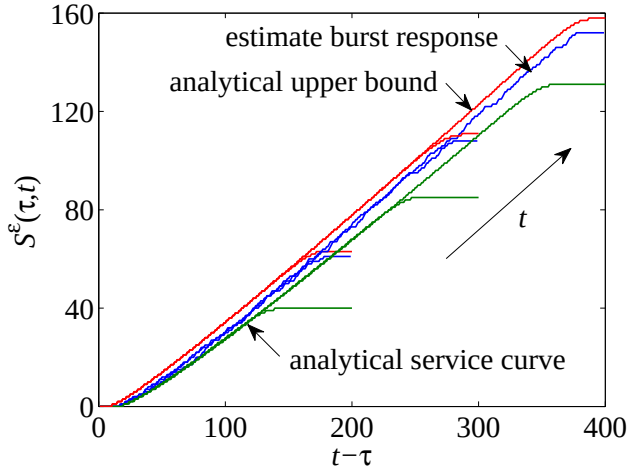


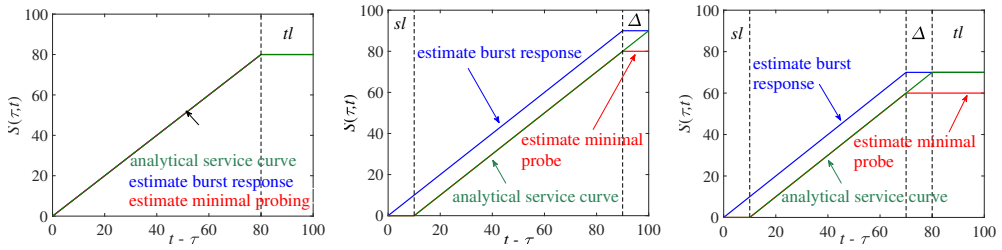
Figure 6.2: Service curve estimates compared to analytical results. Burst probing can estimate non-convex service curves and performs close to the analytical upper bound.

the rate scanning, the burst probing performs close to the analytical upper bound and has the same progression. Hence, it estimates the non-convex part of the service and is therefore suitable to analyze transient effects.

However, as mentioned from the related work section 3.2.2, the burst probing has some major drawbacks. In particular, one basic and intuitive assumption for the construction of the method, namely additive service processes [86], is not guaranteed in general. Basically, Eq. (6.4) would give us an exact service estimate for additive services. For sub-additive service processes we would have a conservative service estimate $S(\tau, t)$, where a function $f(s,t)$ is called sub-additive if $f(s,u) \leqslant f(s,t) + f(t,u)$ for all $u \geqslant t \geqslant s \geqslant 0$. Whereas sub-additivity is not an issue we observed, e.g., from the previous example, that the services are super-additive, i.e., $f(s,u) \geqslant f(s,t) + f(t,u)$ for all $u \geqslant t \geqslant s \geqslant 0$. Hence, using Eq. (6.4) may result in an overestimation of the service $S(\tau, t)$ since $S(\tau, t) \leqslant S(0,t) - S(0,\tau)$. Note that an additive function is super-additive, as well, but not the other way around. In order to quantify the difference from additivity for an super-additive function $f(s,t)$, we define

$$\Delta(s,u) := f(s,u) - \inf_{t\in[s,u]} \{f(s,t) + f(t,u)\}. \tag{6.7}$$

as the maximal deviation from additivity.



(a) Transient latency (b) Stationary latency (c) Transient and stationary latencies

Figure 6.3: Service curve estimates of deterministic sleep scheduling. Latency-rate service curves with a transient latency, with a stationary latency, and with both are compared. The transient latency equals 20 and the stationary latency 10. In the case of the transient latency solely, the latency-rate service curve is additive, and burst probing recovers the exact result. In contrast, in the case of a stationary latency, the service curve is super-additive, and burst probing overestimates the service curve. Minimal probing (see Sec. 6.4) provides a corresponding lower estimate that matches the service curve exactly in case of additivity.

6.3 SUPER-ADDITIVE SERVICE PROCESSES

In the following, we want to explain the impact of super-additive functions on a deterministic case-study, which has the advantage that additional effects such as the outages in the Bernoulli increment process as in Fig. 5.1b do not influence the observations.

For the analysis we choose the deterministic, stationary latency rate function $S^{slr}(\tau, t)$ from Eq. (2.16) and the transient latency rate function $S^{tlr}(\tau, t)$ from Eq. (3.2), respectively. Note that we used transient latency services for the stochastic case in Fig. 6.2. For this scenario we have shown that burst probing provides a good estimate and verifies that these functions are additive. However, if a stationary latency is added to the service as in $S^{slr}(\tau, t)$ we can guarantee super-additivity processes but not a strict additivity, anymore. To see this, let's compute $S^{slr}(0, \tau) + S^{slr}(\tau, t) = R(t - 2T)$, while $S^{slr}(0, t) = R(t - T)$ for $t \geqslant \tau + T$ and $\tau \geqslant T$. Taking the difference as in Eq. (6.7) yields $\Delta(\tau, t) = RT$ for $t \geqslant \tau + 2T$.

To visualize our observations, we plot the transient, the stationary latency, and a combination of both services in Fig. 6.3, which are labeled as analytical service curve. There, we select a transient latency of $T = 20$, a stationary one of $T = 10$ and a rate $R = 1$. For all cases, we present the service curves, as in Fig. 5.1b. Hence, we fix $t = 100$ and increase the interval $t - \tau$, i.e., reducing τ. In Fig. 6.3a, we show the effects of deterministic sleep scheduling with

a transient latency. For large intervals where $t - \tau > 80$, i.e., $\tau < 20$, we observe a flat end of the curve, which corresponds to no additional service in this region. This means that the transient latency which occurs for the first 20 time-slots influences the service. We mark this part with *tl* in Fig. 6.3a.

The stationary latency, which is appropriate to model propagation delays, has an impact on all time-intervals. Thus, for all τ. It has the effect of a right shift in Fig. 6.3b. We marked that region with *sl*. Fig. 6.3c shows both latencies and their corresponding effects, together.

Next, we want to compare the analytical service curves with the estimates obtained from the burst response method. It can be seen, that in case of a transient latency as in Fig. 6.3a the two curves fit perfectly together. That means the burst response yields an exact estimate of the service. Thus, the service is additive. If, nevertheless, stationary latency occurs, the burst response overestimates the service, see Fig. 6.3b and similarly in Fig. 6.3c. This is due to the super-additivity of services with stationary latencies. We marked that region with Δ. Here, the assumption of additivity from Eq. (6.4) erroneously allocates the stationary latency only to large intervals $t - \tau$.

As already pre-announced, we will introduce a method that corrects this overestimation. The method is also included in the figures and labeled as minimal probing, see Sec. 6.4.

Note that, e.g., in cellular networks, we have, in general, both latencies. On the one hand, the transient latency can be seen as the wake-up time and time it takes for a UE to establish a connection to the base station. On the other hand, the stationary latency is the propagation delay. Due to the fact that the propagation delay causes the deviation to additivity and is always greater than zero in production networks, it follows that the burst response method consistently overestimates the service in practice.

Super-additivity of min *and* $\otimes$

In the following, we formalize two lemmas. They show that the $\otimes$ and the min operator are super-additive and confirm additivity only in special cases.

Lemma 1 (Super-additivity of min)**.**

Given two super-additive bivariate functions $f(s,t)$ *and* $g(s,t)$ *for* $t \geqslant s \geqslant 0$*. The minimum* $h(s,t) = \min\{f(s,t), g(s,t)\}$ *is super-additive.*

Note that as a special case of lem. 1 is that the minimum of two additive bivariate functions $f(s,t)$ and $g(s,t)$ for $t \geqslant s \geqslant 0$ is super-additive, but in general not additive. To see this consider the following counterexample where $f(s,t) = t - s$ and $g(s,t) = 2(\lfloor t/2 \rfloor - \lfloor s/2 \rfloor)$. Apparently f and g are additive, however, $h = \min\{f, g\}$ is not.

Proof of Lemma 1

By definition of h, we have

$$\begin{aligned}
&h(s,t) + h(t,u) \\
&= \min\{f(s,t), g(s,t)\} + \min\{f(t,u), g(t,u)\} \\
&\leqslant \min\{f(s,u), g(s,u), f(s,t) + g(t,u), g(s,t) + f(t,u)\} \\
&\leqslant h(s,u).
\end{aligned}$$

The super-additivity of f and g is used in the second line, whereas in the third line we have that $\min\{f(s,u), g(s,u)\} = h(s,u)$ and $\min\{h(s,u), x\} \leqslant h(s,u)$ for any x. □

Lemma 2 (Super-additivity of $\otimes$)**.** *Given two bivariate functions* $f(s,t)$ *and* $g(s,t)$ *for* $t \geqslant s \geqslant 0$ *where* $f(t,t), g(t,t) = 0$ *for all* $t \geqslant 0$*. Define* $h(s,t) = f \otimes g(s,t)$*.*

i. If f *and* g *are super-additive, then* h *is super-additive.*

ii. If f *and* g *are additive and univariate, then* h *is additive.*

Similarly as for the first lemma, it follows from Lemma 2 that the convolution of two additive bivariate functions $f(s,t)$ and $g(s,t)$ for $t \geqslant s \geqslant 0$ is super-additive, and additivity cannot be assumed in general.

Again, we consider as a counterexample the additive functions $f(s,t) = t - s$ and $g(s,t) = 2(\lfloor t/2 \rfloor - \lfloor s/2 \rfloor)$, where $h(s,t) = f \otimes g(s,t) = 2\lfloor t/2 \rfloor - s$ is not additive.

Note that for univariate functions Lemma 2 ii) extends a known result for min-plus convolution of sub-additive univariate functions in [96, p. 142].

Proof of Lemma 2

By definition of h, we have

$$\begin{aligned} &h(s,t) + h(t,u) = f \otimes g(s,t) + f \otimes g(t,u) \\ &= \inf_{\tau\in[s,t]} \inf_{v\in[t,u]} \{f(s,\tau) + f(t,v) + g(\tau,t) + g(v,u)\}. \end{aligned} \tag{6.8}$$

i) Given f and g are super-additive. From Eq. (6.8) we have

$$\begin{aligned} &h(s,t) + h(t,u) \\ &\leqslant \inf_{\tau\in[s,t]} \inf_{v\in[t,u]} \{f(s,v) - f(\tau,t) + g(\tau,t) + g(v,u)\} \\ &= \inf_{v\in[t,u]} \{f(s,v) + g(v,u)\} + \inf_{\tau\in[s,t]} \{g(\tau,t) - f(\tau,t)\} \\ &\leqslant \inf_{v\in[t,u]} \{f(s,v) + g(v,u)\}. \end{aligned} \tag{6.9}$$

In the first line, we estimated $f(s,\tau) + f(\tau,t) + f(t,v) \leqslant f(s,v)$ due to the super-additivity of f. In the second line, we rearranged the infima, and in the third line, we estimated $\inf_{\tau\in[s,t]}\{g(\tau,t) - f(\tau,t)\} \leqslant g(t,t) - f(t,t) = 0$ since $f(t,t), g(t,t) = 0$ for all $t \geqslant 0$. Similarly, using the super-additivity of g we derive from Eq. (6.8) that

$$h(s,t) + h(t,u) \leqslant \inf_{\tau\in[s,t]} \{f(s,\tau) + g(\tau,u)\}. \tag{6.10}$$

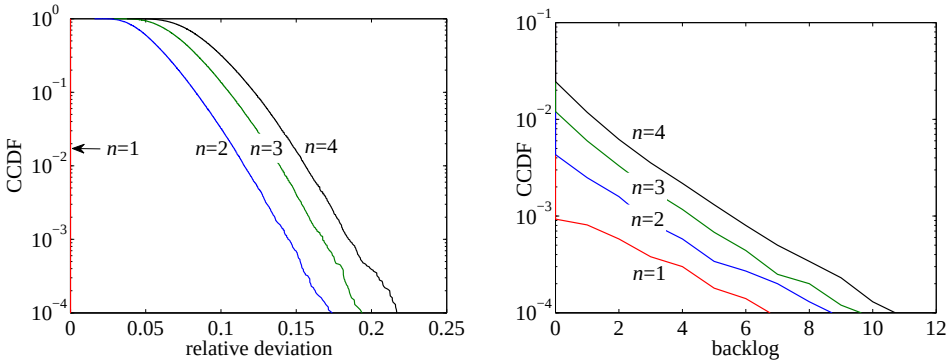
Combining Eq. (6.9) and Eq. (6.10) we obtain

$$h(s,t) + h(t,u) \leqslant \inf_{\tau\in[s,u]} \{f(s,\tau) + g(\tau,u)\} = h(s,u),$$

which proves the super-additivity of h.

ii) For the special case of additive univariate functions $f(s,t) = f(t-s)$ and $g(s,t) = g(t-s)$, that depend only on the difference $t-s$ and not on the absolute values of s and t, it follows that $h(s,t) = f \otimes g(t-s) = h(t-s)$ is also univariate. Using the additivity of f and g, Eq. (6.8) yields that

$$\begin{aligned} &h(t-s) + h(u-t) \\ &= \inf_{\tau\in[s,t]} \inf_{v\in[t,u]} \{f(\tau-s+v-t) + g(t-\tau+u-v)\} \\ &= \inf_{\sigma\in[s+t,t+u]} \{f(\sigma-s-t) + g(t+u-\sigma)\} \\ &= \inf_{\sigma\in[0,u-s]} \{f(\sigma) + g(u-s-\sigma)\} = h(u-s) \end{aligned}$$



(a) Deviation from additivity (b) Accuracy of minimal probing

Figure 6.4: Network of n systems with random sleep scheduling in series. (a) The network service process deviates from additivity. (b) Minimal probing achieves small backlogs, corresponding to a high accuracy of the estimate.

which proves the additivity of h. □

Concerning the property of super-additivity these results lead to a second exception which is noteworthy. It is about the computation of the end-to-end service process $S^{net}(\tau, t)$ for networks of systems. Here, $S^{net}(\tau, t)$ is obtained by the min-plus convolution of the individual systems and their services $S^i(\tau, t)$ for $i = 1, 2, \ldots n$, see Eq. (2.11). Even though all S^i are additive, we have that S^{net} is only super-additive as in lemma 2 i). Strict additivity holds only in special cases. Therefore, we cannot assume that the service processes for tandem systems are additive. As before, the burst probing method is too optimistic, since the estimate is computed by $S^{net}(\tau, t) = S^{net}(0, t) - S^{net}(0, \tau)$ where we only can guarantee $S^{net}(\tau, t) \leqslant S^{net}(0, t) - S^{net}(0, \tau)$.

Let's connect the theory with an example. We choose the same random service processes with random sleep scheduling and the same parameters as, e.g., in Sec. 6.1 for 10^5 sample-paths. We consider a tandem of $n = 1, 2, 3, 4$ networks and compute for additive $S^i(\tau, t)$ the network service process as $S^{net}(\tau, t) = S^1 \otimes \ldots \otimes S^n(\tau, t)$. In Fig. 6.4a we show the distribution of the relative deviation of $S^{net}(0, 400)$ from additivity, i.e., $\Delta(\tau, t)/S^{net}(\tau, t)$. We observe an additive service for $n = 1$, whereas we confirm a significant deviation from additivity for $n > 1$.

6.4 MINIMAL PROBING

As seen so far, it is not trivial to analyze transient phases for additive, non-additive and non-convex service processes and find a valid estimate out of

measurements. To overcome the limitations we have already discovered, we are developing a new probing method. We will be able to analyze transient service processes where we do not assume additivity or convexity. The method consists out of two steps:

1. using the burst response as in Sec. 6.2 we find an upper bound of the service, which leads us to a minimal probe.
2. the minimal probe is used to estimate a conservative non-stationary service curve with a defined accuracy.

In addition to that, we show that the minimal probe is the perfect probe for the system, i.e., any smaller or larger probe estimates only a lower bound of the service. For an intensified discussion regarding the importance of probe traffic, see in [101]. Moreover, minimal backlogging techniques for stationary systems can be found in [140, 111]. Here, the transmission of each probe packet triggers the generation of a new one.

Estimation using Arbitrary Probes

In order to accurately estimate the system's service, the question of suitable probes arises. In general we start with the assumption of arbitrary probes $A(\tau)$ and consider a min-plus linear systems, such that Eq. (2.2) holds with equality, i.e.,

$$D(t) = \inf_{\tau \in [0,t]} \{A(\tau) + S(\tau, t)\}. \tag{6.11}$$

Then, we have for all $\tau \in [0, t]$ that $D(t) \leqslant A(\tau) + S(\tau, t)$. Rearranging yields for all $\tau \in [0, t]$

$$S(\tau, t) \geqslant D(t) - A(\tau). \tag{6.12}$$

By taking standard steps in network calculus, Eq. (6.12) can be reformulated using the backlog $B(t)$ instead of the departures $D(t)$, such that for all $\tau \in [0, t]$ it holds that $S(\tau, t) \geqslant A(\tau, t) - B(t)$. Then, similarly as for the derivation of Eq. (6.2) we use the backlog quantile. Thus,

$$\mathcal{S}^{\varepsilon}(\tau, t) = A(\tau, t) - B^{\varepsilon}(t) \tag{6.13}$$

is a non-stationary service curve as defined by Eq. (5.5), where $\mathcal{S}^{\varepsilon}(\tau, t)$ has the same form as in Eq. (5.3). The backlog quantile is obtained from repeated measurements.

So far, we have not constrained the arrivals. The choice is important. As an example see Sec. 6.1. Here, we used constant rate traffic, i.e., $A(\tau, t) = r(t - \tau)$ that produces a non-stationary service curve but cannot recover transient changes due to the convex form of the service curve. Hence, different probe traffic is needed.

Definition of Minimal Probe

To find suitable probe traffic to analyze transient changes over time, we did not constrain the arrivals so far. The choice is highly non-trivial but imminent since it affects the shape of the service curve, as seen for the rate scanning method 6.1. Moreover, if we choose a probe that is too small, only little information about the system is provided since the arrivals primarily restrict the departures. The other way around, i.e., we choose a probe that is too large, the service estimate will be deteriorated, with the same restriction as in the extreme case of burst probing, see Sec. 6.2.

Now, to obtain the probe traffic that induces the true service of the system from Eq. (6.12), we state the following lemma and define a necessary and sufficient condition for a minimal probe.

Lemma 3 (Minimal Probe)**.** *Fix* $t > 0$ *and define the minimal probe*

$$A_{mp}(\tau) = S(0, t) - S(\tau, t), \tag{6.14}$$

for $\tau \in [0, t]$. *Eq.* (6.12) *holds with equality if and only if* $A(\tau) = A_{mp}(\tau)$ *for all* $\tau \in [0, t]$.

Proof. In Eq. (6.11) we choose the minimal probe from Eq. (6.14) as arrivals. Thus, we substitute Eq. (6.14) into Eq. (6.11) and obtain $D(t) = S(0, t)$, immediately. Then, we show that Eq. (6.14) is sufficient by inserting the minimal probe into Eq. (6.12). Since $D(t) = S(0, t)$, we get $D(t) - A_{mp}(\tau) = S(\tau, t)$ for all $\tau \in [0, t]$.

In order to see that the probe from Eq. (6.14) is necessary, we change our view on the probe and consider any other arrivals then the minimal probe by adding or subtracting something from it. This means, we define the probe as $A(\tau) = A_{mp}(\tau) \pm f(\tau)$ where $f(0) = 0$ and $\exists \tau \in (0, t] : f(\tau) \neq 0$. It follows by the same steps that $D(t) - A(\tau) = S(\tau, t) + \inf_{\upsilon \in [0,t]}\{\pm f(\upsilon)\} \mp f(\tau)$ implying that $\exists \tau \in [0, t] : D(t) - A(\tau) < S(\tau, t)$. □

The significance of Lemma 3 is tremendous. It answers the question of which arrivals we have to send into the network to utilize the system's

service optimally. However, since the minimal probe $A_{mp}(\tau)$ depends on the unknown departures and service, respectively, it is not possible to generate it a priori. Therefore, we have to add a step to prise information about $A_{mp}(\tau)$. From scratch, we cannot assume anything about the system and, in particular, about its service. Thus, sending an arbitrary probe traffic is somehow gambling, e.g., due to the argumentation above about a too low and large probe traffic, respectively.

Anyhow, ignoring the limitations of the burst response method, we obtain information about an unknown system by sending a burst, such that $D(\tau) = \delta \otimes S(\tau) = S(0,\tau)$ and estimate Eq. (6.14) by $\widetilde{A}_{mp}(\tau) = S(0,\tau)$ for $\tau \in [0,t]$. Note that tilde is used to make clear that $\widetilde{A}_{mp}(\tau)$ is an approximation of the minimal probe $A_{mp}(\tau)$, where the approximation is exact in case of an additive service $S(\tau,t)$. Using the burst response method and its stochastic estimate of a non-stationary service curve from Eq. (6.5), an estimate for the minimal probe follows as

$$\widetilde{A}_{mp}(\tau) = \mathcal{S}^{\varepsilon}_{br}(0,t) - \mathcal{S}^{\varepsilon}_{br}(\tau,t), \tag{6.15}$$

for $\tau \in [0,t]$. As a remarkable indication, by construction of $\mathcal{S}^{\varepsilon}_{br}(\tau,t)$ from Eq. (6.5) additivity cannot be assumed, see Lemma 1. Hence, the question arises about the accuracy of the estimate, which will be answered hereafter.

Accuracy of the Estimates

So far, we send a burst and get an estimate of $\widetilde{A}_{mp}(\tau)$ for the minimal probe where we do not know anything about the accuracy.

We start investigations regarding the accuracy with a time-variant and deterministic system. Remembering Eq. (6.12), we obtain a lower estimate of the service out of the arrivals $A(\tau)$ and departures $D(t)$. Further, we get an estimate for the minimal probe $\widetilde{A}_{mp}(\tau) = S(0,\tau)$ from the burst response, see Eq. (6.3). In Eq. (6.12) we replace the departures $D(t)$ with the min-plus convolution from Eq. (6.11). With arrivals $\widetilde{A}_{mp}(\tau)$ we have

$$S(\tau,t) \geqslant \inf_{\upsilon\in[0,t]}\{S(0,\upsilon) + S(\upsilon,t)\} - S(0,\tau),$$

which is a lower bound of the service.

Remembering the upper bound of the service for super-additive function $S(\tau, t)$, we have $S(\tau, t) \leqslant S(0, t) - S(0, \tau)$. Taking the difference of both estimates yields that the service is bounded by an interval of width

$$\Delta(0, t) = S(0, t) - \inf_{v \in [0,t]} \{S(0, v) + S(v, t)\}$$

Now, with the definition from Eq. (6.7) and substituting the service $S(\tau, t)$ into it we get the maximum deviation from additivity of the service for $S(0, t)$. Next, we replace $S(0, t) = \widetilde{A}_{mp}(t)$ and obtain

$$\Delta(0, t) = \widetilde{A}_{mp}(t) - D(t) = B(t). \tag{6.16}$$

Hence, by sending $\widetilde{A}_{mp}(t)$ as a minimal probe, the backlog $B(t)$ at the end of the probe is a measure of accuracy, which is the deviation of a super-additive service process from additivity. Conversely, if the service $S(\tau, t)$ is additive, then the lower and upper bound for $S(\tau, t)$ are equal and the deviation is $\Delta(0, t) = 0$. Ignoring the trivial case of a system with no service, this would mean that sending $\widetilde{A}_{mp}(\tau)$ for additive services $S(\tau, t)$ ends up with a backlog $B(\tau)$ that is zero during the entire probe. Due to the fact that in practice the backlog includes all packets in-flight, this would imply that the propagation delay is zero, which is impossible in real networks. Since the propagation delay can be interpreted as the stationary latency in our scenarios, it follows that the stationary latency is never zero. Remembering that this latency causes the deviation of the service to additivity, we end up that, apart from the trivial case, the service can never be additive in real networks.

Next, we extend the investigations to the stochastic case. From Eq. (6.13) we get a non-stationary service curve out of the arrivals $A(\tau, t)$ and a backlog quantile $B^\varepsilon(t)$. Substituting the minimal probe from Eq. (6.15) we have $\mathcal{S}^\varepsilon_{mp}(\tau, t) = \widetilde{A}_{mp}(\tau, t) - B^\varepsilon(t)$. Since

$$\begin{aligned}
\widetilde{A}_{mp}(\tau, t) &= \widetilde{A}_{mp}(t) - \widetilde{A}_{mp}(\tau) \\
&= \mathcal{S}^\varepsilon_{br}(0, t) - \mathcal{S}^\varepsilon_{br}(t, t) - (\mathcal{S}^\varepsilon_{br}(0, t) - \mathcal{S}^\varepsilon_{br}(\tau, t)) \\
&= \mathcal{S}^\varepsilon_{br}(\tau, t)
\end{aligned} \tag{6.17}$$

where $\mathcal{S}^\varepsilon_{br}(t, t) = 0$, it follows for all $\tau \in [0, t]$ that

$$\mathcal{S}^\varepsilon_{mp}(\tau, t) = \mathcal{S}^\varepsilon_{br}(\tau, t) - B^\varepsilon(t). \tag{6.18}$$

Note, that the superscript ε in Eq. (6.18) is for $\mathcal{S}^{\varepsilon}_{mp}(\tau,t), \mathcal{S}^{\varepsilon}_{br}(\tau,t)$ and $B^{\varepsilon}(t)$ is identical and depends on the backlog quantile as in Eq. (6.13). The subscript*mp* indicates that it is the non-stationary service curve, which results from the procedure of minimal probing.

Thus, we find a conservative service estimate by correcting the possibly too optimistic service estimate from burst probing by the backlog $B^{\varepsilon}(t)$ at the end of the probe from minimal probing, which has been proven to be a measure of accuracy. Therefore, we have a lower and an upper bound for the service, i.e.,

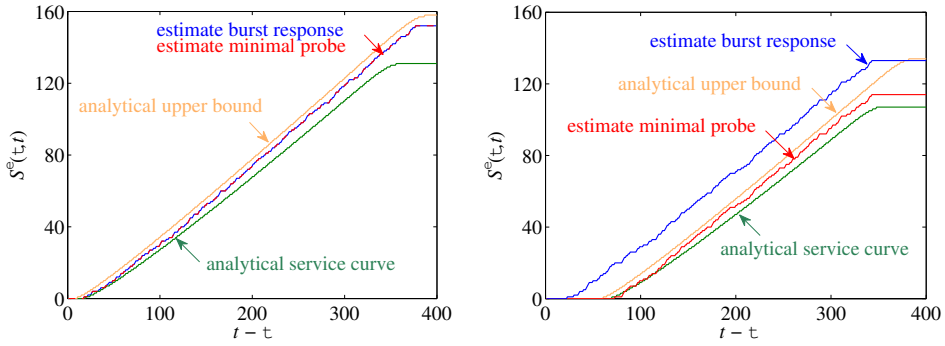
$$\mathcal{S}^{\varepsilon}_{mp}(\tau,t) \leqslant S(\tau,t) \leqslant \mathcal{S}^{\varepsilon}_{br}(\tau,t). \tag{6.19}$$

Now, we further inspect the backlog $B^{\varepsilon}(t)$ and start with the expression $B(t) = \sup_{\tau\in[0,t]}\{A(\tau,t) - S(\tau,t)\}$. By insertion of the estimate for the minimal probe from Eq. (6.15) and Eq. (6.17), respectively, it follows $B(t) = \sup_{\tau\in[0,t]}\{\mathcal{S}^{\varepsilon}_{br}(\tau,t) - S(\tau,t)\}$. From Eq. (6.5) we know that $\mathcal{S}^{\varepsilon}_{br}(\tau,t) = \inf_{\psi\in\Psi_t}\{S_\psi(0,t) - S_\psi(0,\tau)\}$. Thus, for any sample-path $\varphi \in \Psi_t$ we get

$$\begin{aligned} B_\varphi(t) &= \sup_{\tau\in[0,t]}\left\{\inf_{\psi\in\Psi_t}\{S_\psi(0,t) - S_\psi(0,\tau)\} - S_\varphi(\tau,t)\right\} \\ &\leqslant \sup_{\tau\in[0,t]}\{S_\varphi(0,t) - S_\varphi(0,\tau) - S_\varphi(\tau,t)\} \\ &= S_\varphi(0,t) - \inf_{\tau\in[0,t]}\{S_\varphi(0,\tau) + S_\varphi(\tau,t)\} = \Delta_\varphi(0,t). \end{aligned}$$

Due to the definition of the deviation from additivity in Eq. (6.7) we conclude that $B_\varphi(t)$ is bounded by the maximal deviation of $S_\varphi(0,t)$ from additivity. It follows with the same arguments as seen above that $B_\varphi(t) = 0$ for all $\varphi \in \Psi_t$ if $S(\tau,t)$ is additive. Since $P[\Psi_t] \geqslant 1-\varepsilon$, it holds that $B^{\varepsilon}(t) = 0$. Thus, $\mathcal{S}^{\varepsilon}_{mp}(\tau,t)$ is equal to $\mathcal{S}^{\varepsilon}_{br}(\tau,t)$, which means that in this case the service estimate from burst probing $\mathcal{S}^{\varepsilon}_{br}(\tau,t)$ is minimal. The same applies if $S(\tau,t)$ is sub-additive.

As already mentioned, we added the service estimate for minimal probing in Fig. 6.3. Because additional effects, such as the outages from Sec. 6.3, are excluded, we choose this deterministic case as our first example for this method. Again, in Fig. 6.3a, we consider a system that has a service with transient latency only. Here, the estimate of the minimal probe matches the estimate of burst probing and an analytical reference, exactly. Thus, the backlog for the minimal probe is zero, which corresponds to the case that the service is additive. By adding a stationary latency to the service as in



(a) Transient latency

(b) Transient and stationary latency

Figure 6.5: Service curve estimates of random sleep scheduling plus a stationary latency. The estimate of minimal probing stays between the analytical curves, whereas burst probing exceeds the upper bound in case of a stationary latency.

Fig. 6.3b and 6.3c the estimate from minimal probing varies from the burst response method. As in Sec. 6.3, the burst response is above the analytical reference. Hence, burst probing is too optimistic and has to be corrected by the minimal probing procedure. The correction can be done by sending the estimate of the minimal probe $\widetilde{A}_{mp}(\tau)$ and measuring the corresponding backlog at the end of the probe. Here, the backlog at the end of the minimal probe is $B(t) = 10$ and is the deviation of the two methods, see Eq. (6.16). Further, we conclude that the service is super-additive where the maximal deviation from Eq. (6.7) is $\Delta(0,t) = RT = B(t)$.

Generally, the burst response shifts the stationary latency from the beginning of the service curve. We marked this area with *sl* in Figs. 6.3b and 6.3c, to the end of the curve. This way, it overestimates the service till the end. In contrast, minimal probing correctly detects the stationary latency at the region *sl* and matches the analytical curve until a region labeled with Δ in Figs. 6.3b and 6.3c. In this region, the minimal probe is flat and achieves a lower bound with an accuracy of $\Delta(0,t)$, i.e., the backlog $B(t)$ at the end of the probe. In case a transient latency is present, the minimal probe correctly identifies it by emulating the flat area at the end marked with *tl*.

After having a better understanding of the minimal probing procedure, we extend our investigation to the stochastic case with random sleep scheduling, as in Fig. 6.2. We perform the same measurements as in Sec. 6.2 and add the minimal probing, see Fig. 6.5a. For comparison, we added an analytical upper bound and a lower service guarantee specified by Eq. (5.5). In this scenario, we consider a transient latency, such that the minimal probe

matches the burst response, with both methods providing an estimate that lies between the two references.

Next, we include a stationary latency of 50. The result of the measurements can be seen in Fig. 6.5b. As in the deterministic case, the burst response has a higher estimate then the minimal probe method. Similarly, the stationary latency induces a deviation to additivity of the service. The resulting super-additivity leads for the burst response to an overestimation of the service, i.e., the analytical upper bound is exceeded. When the minimal probing is performed, the overly optimistic estimate is corrected by the backlog, so that a valid non-stationary service curve is obtained, which remains between the two analytical references.

Both measurement methods identify the transient latency correctly, represented by the flat area in the upper right area. Furthermore, all curves have the same slope. Hence, the underlying methods pinpoint to the correct rate of the service. However, the burst response falsely assigns the stationary latency to the upper right part of the curve. A precise representation is only observed by the minimal probing method, which allocates it to the lower left of the service curve. Note that besides the stationary latency, which belongs to the left part, we observe an additional delay. It corresponds to some outages caused by the Bernoulli increment process with a service of zero.

Since we investigated that the backlog at the end of minimal probing is a measure of accuracy for the deviation from additivity for super-additive services, we look again at the example of super-additive tandem systems from Sec. 6.3. There, we observed for a system with random sleep scheduling as in Fig. 6.2 that for $n = 1$, the service is additive, whereas for $n > 1$ the relative deviation from additivity is shown in Fig. 6.4a. Howbeit, we compute the backlog of minimal probing for $t = 400$ and for $n = 1 \dots 4$ networks in series with random sleep scheduling and present its distribution of $B(t)$ for 10^5 sample-paths. Again, for $n = 1$ the backlog quantile for $\varepsilon = 10^{-3}$ is $B^{\varepsilon}(t) = 0$. Thus, the estimate from minimal probing is the same as for burst probing, and we conclude an additive service. But for $n > 1$ $B^{\varepsilon}(t)$ is greater than zero, and corresponds to a non-additive service with the deviation shown in Fig. 6.4b. Because we do not know the number of networks in advance and, therefore, whether the service is additive or not, estimates from burst probing are not trustworthy, but, we obtain a conservative estimate with the minimal probe. The estimate has a defined

accuracy given by $B^\varepsilon(t)$. Further, we know that the estimate is accurate because of the small backlog quantile $B^\varepsilon(t)$.

7

CELLULAR NETWORKS

The evolution of mobile networking, to 4G in recent decades and 5G in the near future, has been driven by an ever-increasing growth of mobile data traffic over the years. Today, the use of mobile devices such as smartphones and tablets is part of the daily routine for most of us. In 2017 almost 80% of the Internet users were online through UEs [117] and used applications such as HTTP web browsing traffic, telephony over VoIP, watching, sending, and generating video traffic, file transfers such as music downloads and cloud services, as well as periodic refresh messages, e.g., for news and messenger applications. A more detailed discussion of different types of mobile applications can be found in [117].

However, the evolution started in the late 1970s, and at the beginning of the 1980s, where the first generation (1G) technology, the AMPS standard was introduced. This standard only could send voice traffic as a service.

By the invention of 2G at the end of the 1980s, it was possible to send digital voice, short messages, and packetized data. It was the first time that the public was able to use mobile data traffic in addition to mobile telephony. The data rates for the different types in the evolution of 2G, e.g., GSM, GPRS, and EDGE, were low from today's perspective. The highest possible rates are achieved for EDGE with nominal up- and downlink rates of 220 kbps and 384 kbps, respectively [8]. So, real-time applications and video-streaming were not possible with 2G.

In the next-generation 3G, the data rates increased to 5.76 Mbps in the uplink and 42 Mbps in downlink direction for HSDP+. Although the high latency in 3G systems of the multiple of 100 ms is not tolerable for voice communications [50], the data rates of 3G and the invention of smartphones in the 2000s enabled appropriate use of the devices for HTTP web browsing and other applications. Nevertheless, mobile operators need two core networks in 3G. One circuit-switched network for voice and one packet-switched for data.

This and an ever-increasing growth of user demands for higher data rates initiated a study of the 3GPP [3] on the requirements of the long term evolution (LTE) of 3G. The main objectives were higher data rates and lower latencies of the packets. This led to the 4G LTE standard, which reduced

costs for mobile operators as only one core network is required. Thereby, LTE was designed to reach data rates of 50 Mbps in up- and 100 Mbps in downlink direction [3]. Later releases for LTE Advanced have the potential to reach, e.g., downlink rates of 1 Gbps. An overview of the evolution from cellular technologies can be found in [76, 131].

The need for more efficient technology is evidenced by the fact that the number of smartphones sold per year has increased from about 122 million in 2007 to a factor of ten, i.e., 1,244 million in 2014 [64]. Today, about 1,500 million smartphones are sold per year. This leads to more than 5 billion smartphone users in 2019, which corresponds to approximately 65% of the world population [155]. Considering the fact that in 2018, nearly 45% of the world population lived in rural areas [1], LTE is also a promising candidate to overcome the digital divide with the discrepancy in Internet access between rural and urban areas.

Due to cost efficiency, the providers install fever base stations (eNodeB) in the countryside. With the increasing distance to the eNodeB, the quality of the signals from eNodeB to UE and vice versa reduces dramatically. To analyze the effect of the distance in LTE, a substantial field trial experiment concerning latency, capacity, and throughput performances was done in [145]. The authors present results for a loaded and unloaded network and for near, middle, and far locations, which differ mainly in there corresponding signal-to-interference-plus-noise ratio (SINR).

Further LTE throughput studies have been made in [9, 82, 154]. The authors in [154] consider indoor and outdoor evaluations and the corresponding delay distributions, which are influenced by packet sizes and inter-packet gaps. Besides the performance leap between the cellular generations 3G and 4G, it is shown that the emulations in [154] and the field trial in [145] achieve the theoretical throughput in LTE with a negligible margin. More comparisons of throughput and packet delays in 3G and LTE systems are presented in [43, 81, 82, 107].

The factors behind the performance improvement in LTE include a number of things. A higher bit rate is achieved, e.g., due to OFDM transmission technology, higher coding and modulation schemes, and multiple antenna configuration (MIMO) [50]. The lower latency is often associated with the flat all-IP backhaul network architecture in LTE networks [82, 93, 124, 145].

Another important research topic is the investigation and optimization of the UEs power management in LTE [24, 142]. Similar power consumption models for 3G systems based on MAC-layer power-saving techniques are

given in [130]. To extend the battery life of mobile devices, the 3GPP LTE standard specifies the discontinuous reception scheme (DRX) to reduce the power consumption of a UE [2, 4, 5, 7].

If no data is available to send or receive, power-intensive parts such as the display and radio interface are disabled. In doing so, the UE enters various types of sleep mode and wakes up by a defined schedule to receive information in the downlink or in case an uplink transmission is requested.

Recalling Fig. 1.1 illustrates such a DRX sleep schedule process. Here, the UE is at the beginning in `RRC_IDLE`. Due to a channel request at P_0, the mobile is awaking and goes to `RRC_CONNECTED`. The activation or transition takes T_0 units of time. Subsequently, data can be transmitted. After the data transfer has taken place, the UE returns to sleep mode at P_1, i.e., `RRC_IDLE` and wakes up T_1 units later. The process is now repeated as often as required.

We modeled the sleep duration T_i in the previous chapters where we assumed either deterministic or random wake-up times. Similarly, the service can be deterministic or stochastic during data transmission, e.g., due to outages in wireless transmission. While the mobile is asleep, data is buffered for later transmission. This results in transient backlog and delays which are non-negligible, see Sec. 7.4.1. A more detailed view of DRX is presented in Sec. 7.2.1.

Moreover, the impact of these transient phases has a significant impact on the performance, since the majority of the flows are short-lived [109]. As an example, we have that nearly 50% of all TCP flows are less than five seconds [82].

In the following, we describe our settings in Sec. 7.1, with which we perform all measurements. We present results for the cellular sleep scheduling, i.e., the DRX implementation and other occurring additional latencies, such as HARQ retransmissions on the MAC layer in LTE. Due to real measurements, the question arises after a decent probe selection, e.g., for the selection of a burst in the minimal probing method. We consider that in Sec. 7.3. In Sec.7.4, transient effects in LTE networks are shown and analyzed. Further, we present the first results for the minimal probing method in cellular networks. Besides service curve estimates, we explain characteristics of cellular services, such as transient delays caused by DRX, the stationary latency that is comparable to OWDs, service outages by the radio channel, capacity limits, and the deviation of minimal probing from additivity. A comparison with other cellular technologies, like HSPA+ and

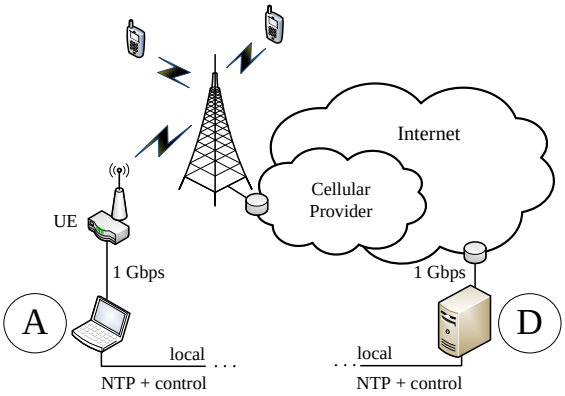


Figure 7.1: The measurement setup comprises a cellular data connection from client (A) to server (D) for estimation, and a separated local control network.

EDGE, is done in Sec. 7.5. Finally, we analyze diurnal characteristics. Note among others we present results from [19, 21].

7.1 MEASUREMENT SETUP

We perform our measurements for two major German commercial internet service provider (ISP). We call them ISP1 and ISP2. Thereby, an extensive set of measurement campaigns is done for ISP1. We also substantiate selected results by confirming them in the other network.

Fig. 7.1 shows the main components of our measurement setup. The cellular Internet connections to the ISPs are made via the cellular client (A), from where we perform all measurements to a wired server (D), which is connected to the national German research and education network at 1 Gbps.

For both providers, we choose as user equipment a stationary Category 3 Teldat RS232j-4G modem for EDGE and LTE and for HSPA, a Teltonika HSPA+ RUT500 modem.

As stated by the ISPs, the nominal uplink rates are 220 kbps for EDGE, 5.76 Mbps for HSPA, and 50 Mbps for LTE.

We send UDP packets and use *rude&crude*[1] as traffic generator. For HSPA and LTE, we choose packets with a size of 1400 bytes, for EDGE, we reduce the packets to 500 bytes to account for the low uplink data rate. To measure the probe arrivals $A(t)$ and departures $D(t)$, packet traces are captured by *libpcap* at the client and the server, respectively. To compute precise

1 http://rude.sourceforge.net/

performance bounds and service estimates, we need accurate timestamps. Therefore, we have a separate local control network for the client and the server that permits time synchronization in a range of a few milliseconds using the Network Time Protocol (NTP). In addition, it enables us to operate at the client and the server remotely without any impact on the cellular network. Finally, all measurements are performed automatically using the tool *sshlauncher*[2], that facilitates repeated execution of distributed network experiments.

7.2 MAC - LAYER

In the following, we investigate specific MAC-layer algorithms that cause an additional delay on packets. On the one hand, we focus on the DRX mode, which we consider the equivalent of the previously introduced random sleep/wake-up times. On the other hand, we introduce the **H**ybrid **A**utomatic **R**epeat re**Q**uest (HARQ) procedure. It is a packet retransmission on the MAC-layer, which leads to an additional delay of up to 8 ms per packet retransmission.

7.2.1 *DRX*

In order to send or receive any data, a mobile device has to be connected to a base station where it has to listen for and send signaling messages. Surely, doing this continuously leads to a battery drain. Therefore, the DRX mode is introduced for energy saving, e.g., for 4G [4, 5, 7]. There, the UE listens on the Physical Downlink Control Channel (PDCCH) for incoming paging messages from the eNodeB. A continuous monitoring of the PDCCH reduces battery power. The DRX mode defines a procedure where the UE only listens at certain time instances on that channel and switches off power-intensive parts, such as the radio interface [4, 7]. Clearly, in between, i.e., during the sleep mode, the UE cannot be paged, potentially incoming packets have to wait until the UE wakes up, which leads to additional latencies.

For a better understanding, we describe the procedure of the DRX mode in LTE, where Fig. 7.2 shows an example of the activity states for a UE. All the parameters, e.g., when the UE is listening or sleeping, are provided by

2 https://github.com/bozakov/sshlauncher

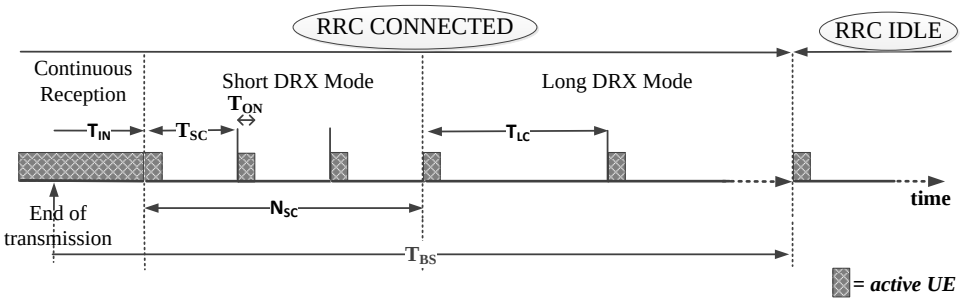


Figure 7.2: A sample path of LTE DRX with RRC_CONNECTED and RRC_IDLE states.

the eNodeB. In general, a mobile can be in two different radio resource control states, i.e., RRC_IDLE or RRC_CONNECTED [7].

If the UE is in RRC_CONNECTED state, the UE receives an inactivity timer T_{BS} from the eNodeB. As long as the UE sends or receives data, it is in *continuous reception* state and the timer T_{BS} is reset. When the data transmission is finished, an inactivity timer T_{IN} starts. The UE remains in the *continuous reception* state according to T_{IN}. The timer is reset in case of new packet transmissions. Otherwise, the UE monitors the PDCCH until T_{IN} expires. The timer can be in the range between 1 ms and 2.56 sec [50]. Afterward, the UE enters the *short DRX* cycle T_{SC} with a length of 2 ms - 640 ms. Thereby, in an interval of length T_{SC}, the UE is listening only for a duration of the time T_{ON} and sleeps otherwise. This can be repeated 1 to 16 times according to a predefined number of *short DRX* cycles N_{sc} if no activity is there. Then, the UE goes to the *long DRX* cycle T_{LC} of length 10 ms - 2.56 sec, where $T_{SC} \leqslant T_{LC}$. The UE remains in the long cycle as long as no data is available for transmission or until the timer T_{BS} expires at the eNodeB. Subsequently, to save resources, the eNodeB initiates an RRC connection release. The eNodeB tears down its data connection to the UE and moves to the RRC_IDLE state.

In RRC_IDLE the UE is no longer actively connected to the eNodeB [7]. Through the paging mechanism, the eNodeB is still able to keep track of the UE. In this state, the mobile reduces activities to the eNodeB to a very low level to save battery power and monitors the PDCCH for paging information according to a DRX cycle. Again, the default DRX cycle length in RRC_IDLE is specified by the eNodeB. For the event of data transmission in up- or downlink direction, the UE must establish a connection to the eNodeB. It changes to the RRC_CONNECTED state in which 16 to 19 signaling messages must be exchanged in the mobile network core in uplink or downlink direction. Data can then be transferred [50].

The time needed to send or receive data by establishing a connection to the base station is what we call cellular sleep scheduling.

Similar mechanisms exist in 2G and 3G. Power consumption models for 3G systems based on MAC layer power-saving techniques are given in [130]. The occurring effects on battery lifetime and power consumption are evaluated, e.g., for 2G and 3G in [116], and for 4G in [81]. Due to the high number of Internet connections we make daily with our mobile devices, battery drain is a performance indicator that is receiving increasing attention due to its impact on the user experience [60, 134, 138].

By comparing the power consumption characteristics of 3G and 4G systems, the authors of [81] conclude that 4G is significantly less power efficient. As stated by the authors, the reason is the OFDM technology in LTE. Although it enables high bit rates, it suffers from low energy efficiency due to a high peak-to-average power ratio.

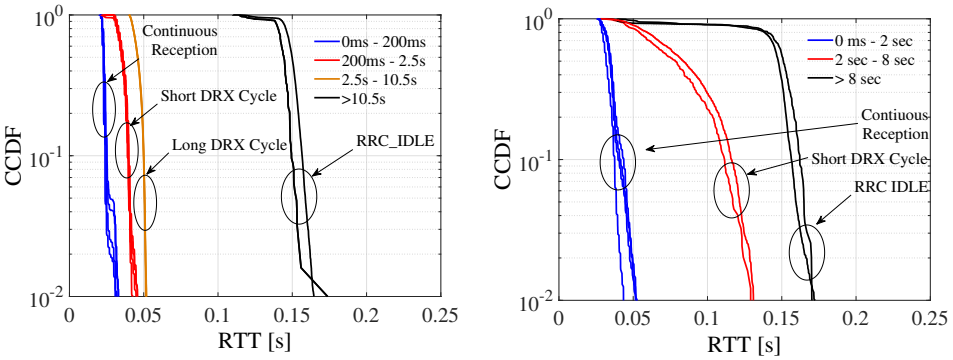
The trade-off between power consumption and delay is obvious [119] and motivated many researcher to optimize the DRX parameters in 3G [147, 148] and 4G [23, 144, 156, 157]. The mean wake-up times for the DRX states can be modeled by semi-Markov chains [23, 144, 148, 156, 157].

Further, Yang [147] used an M|G|1 model with vacations to derive stationary delays for DRX. Another work [142] adapts the DRX cycle parameters to optimize power consumption for certain scenarios while comparing throughput versus power consumption.

Transient power measurements, on the other hand, have significant challenges. For example, consider the basic M|M|1 queue. Here, stationary state distributions are derived from a set of linear balance equations, e.g., in [126]. For transient analysis, we have to solve a set of differential equations which are mostly solved numerically [153]. Thus, the number of transient solutions of queuing systems is rare [59, 80, 141].

However, in the following, we present results for the DRX mode for both ISPs. We send ping messages from client (A) to our server (D) and measure the round-trip-times (RTT) of each ping. We send the next one according to a predefined interpacket-gap and repeat it at least $1,000$ times for each gap. We vary the time between the ping messages with the aim to find characteristics in the network regarding DRX states.

Fig. 7.3 illustrates for both ISPs the RTTs for a set of interpacket-gaps in LTE. We choose the complementary cumulative distribution function (CCDF) for representing since it highlights rare events, i.e., especially high values.



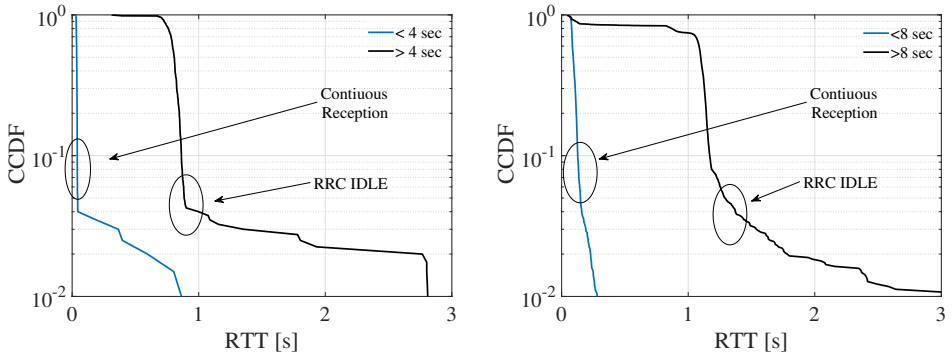
(a) ISP1 (b) ISP2

Figure 7.3: CCDF of `ping` RTT for different inter-packet gaps.

In Fig. 7.3a we show the RTTs for ISP1. We conclude that for interpacket-gaps less than 200 ms, the UE is in continuous reception mode. For higher gaps, i.e., between 200 ms and 2.5 s, we observe higher RTTs and deduce that the UE switched to the short DRX mode. A second jump is observed for time-intervals in a range of 2.5 sec and 10.5 sec, which indicates that the UE is now in the long DRX mode. After 10.5 sec, the RTTs increased to a multiple and is almost surely higher than 100 ms, probably because of the RRC connection establishment, to switch from `RRC_IDLE` to `RRC_CONNECTED`.

For ISP2, we observe a different behavior, as presented in Fig.7.3b. Here, the UE is in continuous reception mode for interpacket-gaps up to 2 sec. In comparison to ISP1, it is ten times longer in this state and close to the maximum value of 2.56 sec, see [50]. For sure, this results in a higher battery drain in the first two seconds for a UE in the network of ISP2. Interestingly, if we further increase the interpacket-gaps for ISP2, we only can identify two more states. The first one is in a range from 2 sec to 8 sec and has RTTs larger than 100 ms in more than 20% of the cases. This might lead to the assumption to infer this to be the `RRC_IDLE` state. Since for gaps greater then 8 sec, the RTTs increase further, we deduce these to `RRC_IDLE` and the first case to the short DRX mode. Thus, we do not have any long DRX cycles available. Apart from the fact that a UE for ISP2 goes to sleep state around 2.5 sec earlier than for ISP1 and reduces so the power consumption, it also saves the battery in short DRX mode in comparison to ISP1. This is due to the longer cycles and times for monitoring the PDCCH.

An analysis of which of these two DRX implementations is more efficient regarding, e.g., battery power management, is out of the scope of this thesis. The various implementations already illustrate that for applications



(a) ISP1 (b) ISP2

Figure 7.4: CCDF of `ping` RTT for different inter-packet gaps for 3G.

such as messaging, the choice of the DRX cycle can have a significant impact on battery performance, and perceived delay. A bad choice of the refreshing period, i.e., just above the inactivity timer of the eNodeB, can lead to an excessive network congestion, as the transition from `RRC_IDLE` to `RRC_CONNECTED` must be made, which harms the network performance [134].

It becomes even more obvious in 3G. The procedure is simpler than in 4G, i.e., there is one inactivity timer threshold and the DRX cycle. Thus, we have two states, which are the `RRC_IDLE` and `RRC_CONNECTED` state[147]. There, the transition from the idle to the connect state takes more than 1 sec, after an inactivity for at least 4 sec in case of ISP1 and 8 sec of ISP2, see Fig.7.4a and 7.4b, respectively. The trade-off between signaling load to power consumption is even more critical in 3G networks[129]. We discuss the effects on throughput and performance bounds such as backlog, e.g., in Sec.7.4.

7.2.2 *HARQ*

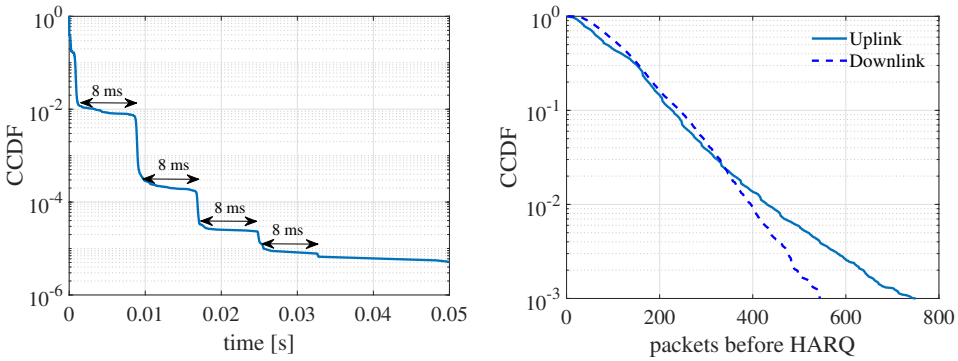
Next, we investigate the HARQ procedure that leads to packet retransmission on the MAC-layer [5, 6]. In a nutshell, it combines (Hybrid) the **A**utomatic **R**epeat re**Q**uest (ARQ) protocol with a **F**orward **E**rror **C**orrection (FEC) code. More precisely, every transmitted data block contains a checksum. If a checksum test fails at the receiver, the data block is considered to be incorrect. A new version of this data block is transmitted to the receiver. The receiver then combines the newly sent packet with the erroneous one to increases the probability of correct decoding. In the meantime, cor-

rectly received out-of-order blocks have to wait in the receiver buffer until predecessor blocks are transmitted.

In LTE, a HARQ retransmission takes 8 ms [5], where the maximum number of HARQ- retransmissions is limited to five (*maxHARQ-Tx*) [90]. An explanation of HARQ and the number of retransmitted packets per retransmission is evaluated, e.g., in [154].

In [21], we already measured the time it takes to perform an HTTP handshake to a common web server and plotted the empirical probability mass function (pmf) where one handshake is the time for sending a SYN packet and receiving the SYN/ACK reply. We added the figure to the appendix, see Fig. A.1. We find that most handshakes are done in around 20 ms. Additionally, we find another significant mode 8 ms later, which we contribute to a HARQ retransmission.

Next, we substantiate this result and perform measurements from our client (A) to our server (D) in up- and downlink direction. We send constant bit rate traffic and analyze the interpacket-gaps at the receiver for HARQ retransmissions in Fig.7.5a. The CCDF shows four regions with an 8 ms increase, which coincides with the period for HARQ retransmissions. There is the possibility of a fifth transmission, which is not detectable due to the low probability. Further, we notice that for less than 1% of the packets, a HARQ retransmission occurs. Thus, we conclude that we have very good channel conditions, because LTE has the goal to have around 10% of HARQ retransmissions. In Fig.7.5b, we performed uplink and downlink measurements and analyzed how many packets are received before any HARQ retransmission. In more than 99% of the cases, there are not many differences in the number of successfully transmitted packets in up- and downlink, where, e.g., at the 1% quantile, we have nearly 400 successful transmitted packets before any HARQ retransmission happens. Then, with lower probability, the uplink performs better, e.g., with a probability of 0.1% we have around 750 packets, which is about 35% higher than the downlink case. Whereas the almost linear slope in Fig. 7.5b, i.e. exponential decrease on the log-scale indicates that the probability of a HARQ retransmission of a packet is independent of other packets, the CCDF in Fig. 7.5a shows that this is not true. Otherwise, the square of the probability for one HARQ-retranmission equals the probability that a packet sees two retransmissions which is obviously not the case.



(a) Interpacket gaps at the receiver (b) Train length before HARQ retransmission

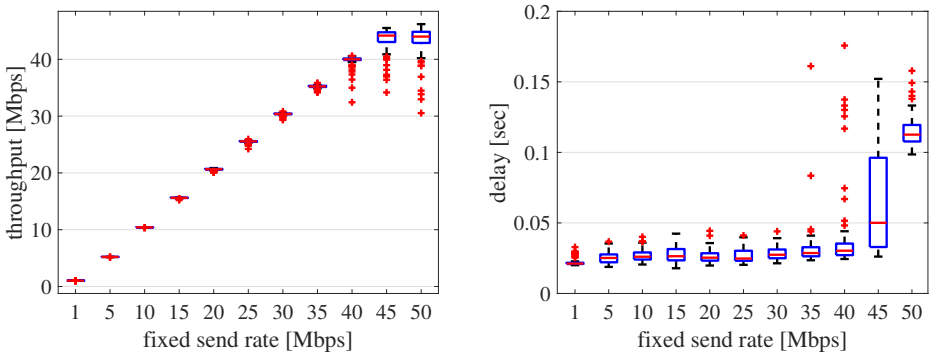
Figure 7.5: HARQ retransmissions

7.3 BURSTS IN PRACTICE

In order to estimate transient service curves in cellular networks we follow the description from Sec. 6.4, i.e., the minimal probing procedure. Thereby, we send a burst in the first step to find an upper bound of the service. Out of the departures, we are able to compute the minimal probe, which, unlike all other traffic, is optimal under certain conditions. Then, this minimal probe is sent as arrival traffic. Finally, we obtain from the backlog at time t a lower bound of the service.

However, to get a better understanding of which effects bursts have, we perform some baseline measurements and present the results in Fig.7.6. Here, we send in the LTE network of ISP1 CBR traffic for two seconds in the uplink direction from the client (A) to server (D). We repeat the measurements for each rate 50 times with a fixed packet size of 1400 bytes. The nominal uplink rate is 50 Mbps. Fig. 7.6a shows boxplot for the throughput measurements. The center of the box represents the median, whereas the borders are the 0.25 and 0.75 percentiles, respectively, while the lower and upper whiskers correspond to the 0.01 and 0.99 percentiles.

Computing the average over two seconds, we find a limiting rate around 44 Mbps. Below this rate, the fluctuations in the data rates are very low. This is due to the fact that the link capacity is not fully utilized, i.e., the probe traffic is too small. Generally, if we choose too small probe traffic, the departures are mostly limited by the arrivals. For rates larger then 44 Mbps, we are exhausting the link, such that the variations in the data rate increases. Apart from that, it has further effects, as shown in Fig. 7.6b.



(a) Throughput (b) Delay

Figure 7.6: Greedy throughput and delay distributions for different UDP traffic rates between client (A) and server (D) in uplink direction

There we observe that the OWD for the last packet of each measurement is almost stable around 26 ms for low rates and increases slightly to 30 ms at 40 Mbps. Higher rates, e.g., 50 Mbps lead to a significant jump to 110 ms. For more details about packet delays, loss and network buffers, see [21], where we also included measurements for the downlink.

In practice, there are many instruments for bandwidth estimation that seeks to find the sending rate at which the delay increases. For example, pathload [84], which sends packet trains through the network at a certain rate and measures the delay. The rate is gradually increased until the delay increases. A higher delay is then interpreted as network overload, i.e., the rate exceeds the available bandwidth. The idea of pathchirp [120] is similar. Here, it is sent only one train, where the rate is varied within this train. An overview of bandwidth estimation tools can be found in [41].

Likewise, as the pathload and pathchirp methods, we assume for the moment that sending with a rate above the limiting rate exhausts the network and is therefore like a burst. In consequence, during connection establishment packets are buffered to the maximum buffer size, which leads then to packet loss and higher delays. Of course, large buffers can handle bursty traffic or bad channel conditions better before a loss occurs. A drawback is a performance deterioration, e.g., for users who generate a traffic mix of large file transfers and delay-sensitive live video streaming. In this scenario, the user will likely experience a lagged video call.

As we will see, the minimum probe sends the traffic so that the rate is as high as possible before the delay and backlog increase. Further, we know from network calculus that sending a burst immediately yields the

service, see Sec. 6.2 where we replaced the arrivals with the burst function in Eq. (6.3).

Practically, arrivals can not be infinite. Additionally, we do not know how the service changes over time. Thus, we cannot adjust the arrivals accordingly. Therefore we do without a variable bitrate as burst and choose a finite CBR arrival traffic with the rate r, which has the same properties for all t as $\delta(t)$.

An initial guess is to take the accumulated service $S(0,t)$ from Eq. (3.10) and divide it by the amount of time t, such that we could set $r = S(0,t)/t$. The next example clarifies that this gives us only in certain cases a valid burst rate and might underestimate it otherwise.

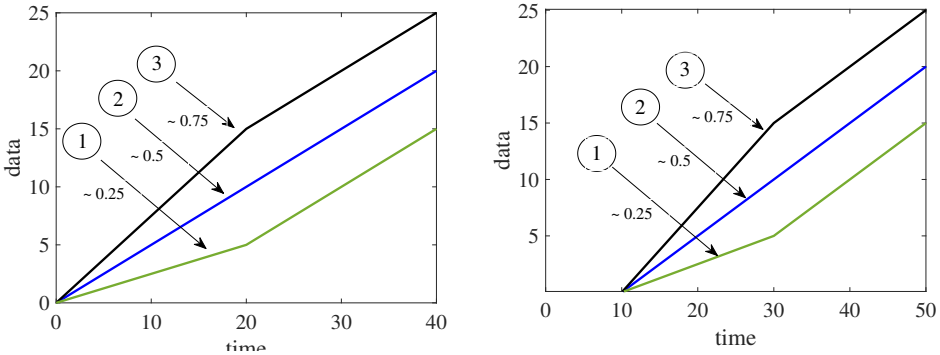
We assume to have two cases, each with three service functions $S_i(0,t)$ for $i = 1,2,3$. In the first cases we have no stationary latency, i.e., $T = 0$ and individual rates of $0.25 \cdot i$ for service curve $i = 1,2,3$ and $t \in [0,20]$ where we fix the rate to 0.5 afterwards. In the second scenario the stationary latency is $T = 10$ with the same individual rates up to $t = 30$ and 0.5 subsequently. A visual representation is given in Fig. 7.7. We start with $T = 0$ and compute for all $S_i(0,t)$ with $i = 1,2,3$ the corresponding potential burst rate as $r_i = S_i(0,t)/t$ for $t = 40$. Hence, $r_1 = 3/8$, $r_2 = 1/2$ and $r_3 = 5/8$. Clearly, $r_1 \cdot t$ and $r_2 \cdot t$ are always greater or equal than $S_1(t)$ and $S_2(t)$ for all t. For $S_3(t)$, this does not apply which is due to the fact that for the first 20 time-slots the maximum rate of $S_3(t)$ is 0.75 whereas $r_3 = 5/8$. As a consequence, in this part r_3 is not a burst rate such that a possible service estimate as in Eq. (6.3) is limited by the arrivals from $r_3 \cdot t$.

It becomes even more apparent if we add a stationary latency as in the second case with $T = 10$. Here, $r_1 = 3/10$, $r_2 = 2/5$ and $r_3 = 1/2$ for $t = 50$. Consequently non of the potential burst rates r_i for $i = 1,2,3$ are above the maximum rate of the corresponding $S_i(t)$. Remember that a stationary latency is comparable to packets OWD. This means for our throughput measurements from Fig. 7.6a that taking the limiting rate as burst rate could lead to a lower service estimate.

Next, we try to find a valid burst rate r by computing the maximum service rate over all $\tau \in [0,t]$ for a fixed t where $S(\tau,t) = S(t) - S(\tau)$. In this sense, we have

$$r = \max_{\tau \in [0,t]} \{S(\tau,t)/(t-\tau)\}. \tag{7.1}$$

Using this approach we have for the case with $T = 10$ the following rates $r_1 = 1/2$, $r_2 = 1/2$ and $r_3 = 5/8$. Hence, we find an acceptable rate r for



(a) $T = 0$ (b) $T = 10$

Figure 7.7: Latency rate service curves $S_i(0,t)$ for $i = 1,2,3$ with latency $T = 0$ and $T = 10$ and individual rates of $0.25, 0.5$ and 0.75 for 20 timeslots after T and 0.5 afterwards

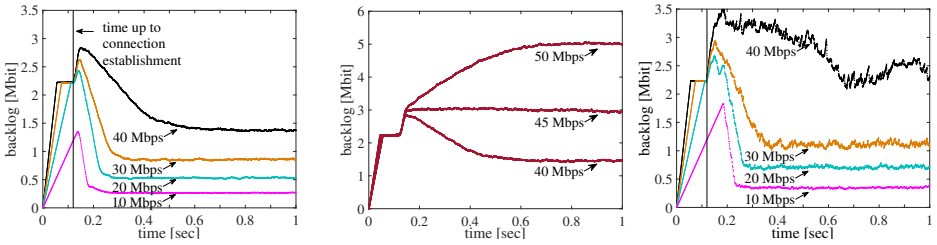
$S_1(t)$ and $S_2(t)$ whereas for $S_3(t)$ the estimate is still too low, which is because of the higher rate of 0.75 between $t \in [10,30]$. As a consequence, in Eq. (7.1) we have to take the maximum not just over all $\tau \in [0,t]$, but over all t as well.

For the practical measurements in 2G, 3G, and 4G networks, we do not know the rates in advance. Moreover, to find the minimal possible burst rate r for all networks is complex and takes a lot of effort. Nevertheless, we know the nominal uplink rates, as stated by the providers, are 220 kbps for EDGE, 5.76 Mbps for HSPA, and 50 Mbps for LTE. Thus, every rate above is a valid burst rate.

7.4 TRANSIENT SERVICE OF LTE

Next, we perform measurements in LTE networks with sleep scheduling, i.e., with DRX mode. Results are shown for carrier ISP1 and ISP2. In particular, we are interested in transient effects on performance bounds such as backlog and delay, as well as in the estimation of service curves resulting from the use of the minimum probing method of Sec. 6.4. Thus, we present non-stationary service curves that provide time-varying performance measures, such as the transient overshoot and the relaxation time, as introduced in Sec.3.1 and [141].

The results presented show the influence of DRX on the uplink, whereby it is guaranteed that the UE is dormant, i.e., that the radio resource control protocol is in `RRC_IDLE` state as in Fig. 7.2, which corresponds to the idle



(a) Mean (b) Mean around limiting rate (c) 0.95-Quantile

Figure 7.8: Transient backlog of LTE for CBR traffic.

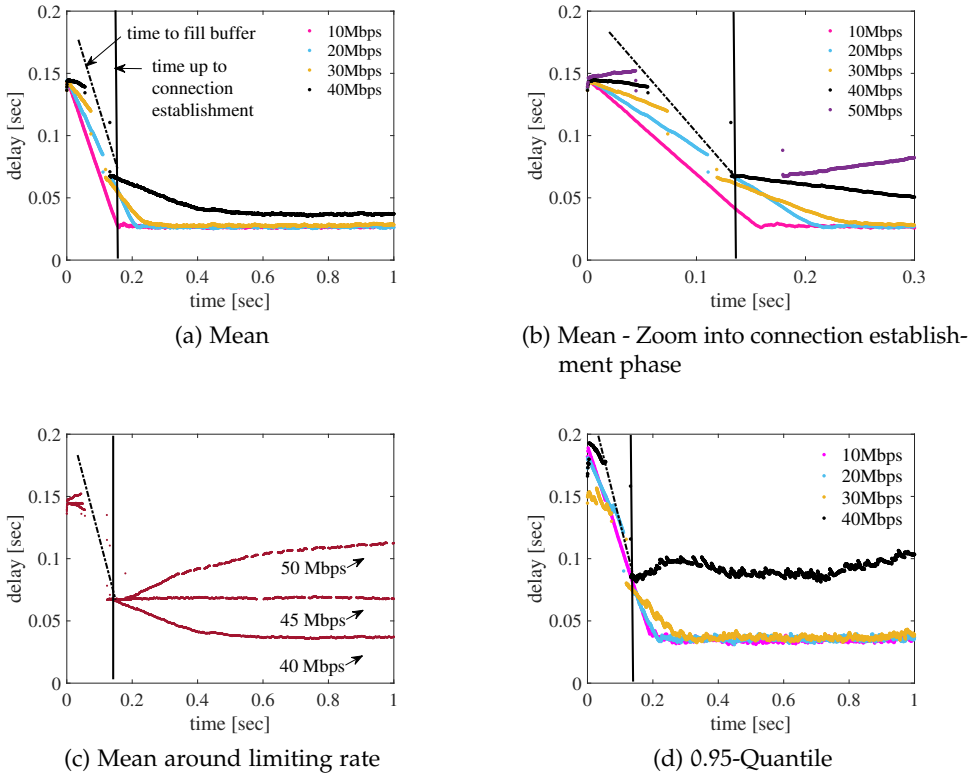
state in Fig. 1.1. To ensure that the UE enters RRC_IDLE state we have to wait at least 10.5 sec for ISP1 and 8 sec for ISP2 before each measurement, see Fig.7.3. We expect that studies for the downlink will provide similar results. Apart from that, we mention that sleep scheduling is not only done at the UE but also at the base station, where we speak of green cellular networks with base station sleeping [137].

7.4.1 *Transient Overshoot and Relaxation Time*

In Fig. 7.6b, we analyzed the delay of the last packet for different CBR traffic, where we considered this delay as stationary. To investigate the behavior of LTE regarding transient overshoot and relaxation times, we plot for ISP1 the mean and 95 percentiles of the backlog progression over time in Fig. 7.8. Similarly, Fig. 7.9 shows the result for the delay.

For the backlog, we observe that all curves show the transient overshoot and relaxation times as for the model in Sec. 3.1. In detail, in the beginning, a connection establishment is performed with the eNodeB. The duration is marked with the vertical line after 120 ms in Fig. 7.8a. Up to this time, all data is queued in the buffer at the UE. The buffer fills linearly according to the sending rate. For 10 Mbps, we have a backlog of 1.2 Mbps. While no packet loss occurs for 10 Mbps, we observe a deterministical loss for rates of 20 Mbps and higher as soon as the backlog reaches 2.2 Mb, respectively, 200 packets. Note that we also measured the buffer limit in the UE for different packet sizes. The experiments substantiate the size of 200 packets, regardless of which packet size we choose.

An interesting fact is that the backlog continues to increase after the connection is established, which is due to the way we calculate it, i.e., $B(t) = A(t) - D(t)$. More precisely, after the successful connection establishment, the packets from the client's buffer (A) are sent to the receiver (D) and,



(a) Mean

(b) Mean - Zoom into connection establishment phase

(c) Mean around limiting rate

(d) 0.95-Quantile

Figure 7.9: Transient delay of LTE for CBR traffic.

therefore, still in transmission. During this time, new packets are generated with a specific rate and fill the buffer again. For the mean backlog, this effect takes as long as the mean stationary delays for low rates in Fig. 7.6b, i.e., around 25 ms, whereas the 0.95 quantile is around 50 ms. Afterward, the backlog depletes, according to a rate that is the difference in the service and arrival rate. Consequently, for rates close to the limiting rate of 44 Mbps, it takes significantly longer to relax the transient overshoot for the backlog. And for an arrival rate that is close or even equal to the limiting rate, the backlog remains and even increases for higher rates, as in Fig.7.8b. This is due to the fact that the probability that the buffer is full increases with the rate. Thus, packets wait longer to be served in the buffer. In comparison to the mean, the volatility of the backlog for the 0.95-quantile strongly grows for higher rates. If it reaches stationarity, the value of the backlog is mainly caused due to the packets in flight and can be approximated by the OWD times the arrival rate. As an example, we choose the mean OWD of 25 ms and an arrival rate of 10 Mpbs that leads to 0.25 Mb.

Next, we take a more in-depth look into the behavior of the delay and present results for the mean and 0.95 quantiles in Fig.7.9. We start with

the mean delay for CBR arrival traffic and focus our investigations to the beginning of each curve in Fig.7.9a, where we marked similarly to the backlog the time up to connection establishment at $T = 120$ ms with a vertical line. At the very beginning we notice a behavior of the delay comparable to the model in Fig.5.4b and Fig. 5.5b. To see this we zoom in Fig. 7.9b into the early phase of the delay distribution from Fig.7.9a. There we observe that no or only little transient overshoot is present at low rates. Thus the delay for the next packets immediately decrease until stationarity is reached, because the time to the connection establishment also decreases. For larger rates, we notice a transient overshoot, which is, however, limited to the buffer in the UE, i.e., after 200 packets, new arrivals are discarded until the connection establishment is done. The declining dashed line shows the time needed to fill the buffer for different rates. Theoretically, there is no packet loss at rates up to about 18.6 Mbps, where 18.6 Mbps is the rate at which 200 packets are sent in the first 120 ms at a packet size of 1400 Bytes. Interestingly, after connection establishment, all rates with losses start with a delay of around 67 ms. If we assume a corresponding declining delay progression for a sending rate of 18.6 Mpbs as for, e.g., 10 Mpbs, then we find that it coincides with the delay value we would have sent 18.6 Mpbs, i.e., the largest possible rate without a deterministic loss.

Back to Fig.7.9a where we sent CBR traffic of $10, 20, 30$ and 40 Mbps, i.e., below the limiting rate of 44 Mbps. We observe that the delay decreases according to the difference of service and arrival rate until we obtain the stationary delay. The stationary delay increases from 27 ms for lower rates to 36 ms for 40 Mbps. Then, in Fig.7.9c we analyze the delay behavior for CBR traffic around the limiting rate and choose arrival rates of $40, 45$ and 50 Mbps. By sending 45 Mbps, which is almost the limiting rate of 44 Mbps, the delay remains nearly unchanged and increases slightly by 2 ms to 69 ms in Fig. 7.9c, since there is scarcely any difference of the service and arrival rate, wheres for larger rates the delay increases up to 110 ms, see also Fig.7.6 for, e.g., the median of the stationary delays and other quantiles. Last but not least, Fig. 7.9d shows the 0.95 quantile of the delay. We observe that the delay is around 80 ms after connection establishment. Moreover, the delay for lower rates increases a bit in comparison to the mean, where the stationary delay is now around 35 ms. For 40 Mbps, the delay fluctuates around 80 ms.

Generally, we observe that the buffer size and time up to connection establishment influence the backlog and delay performances a lot. Especially

the clearance of transient overshoot takes several hundreds of milliseconds and is therefore non-negligible. Even more, knowing the service and adjusting the arrival rate accordingly can make the difference in delay-sensitive applications such as streaming between a fluent and a lagging stream.

7.4.2 *Non-stationary Service Curves*

Next, we identify the transient services by non-stationary service curves for ISP1 and ISP2 and start with their LTE networks. In comparison to the analysis of the transient backlog and delay where we sent different CBR traffic to investigate transient behavior, the service curve is a single characteristic function of the system, providing transient performance measures for each type of traffic arrival.

We follow the procedure from section 6.4. The method consists out of two steps. In a first step, we send burst traffic to find an estimate of the minimal probe $\widetilde{A}_{mp}(\tau)$. Then the second step is to use this $\widetilde{A}_{mp}(\tau)$ as new arrival traffic to estimate a service curve $S^{\varepsilon}_{mp}(\tau, t)$. Thereby, we already know that the backlog quantile at time t $B^{\varepsilon}(t)$ is a measure of accuracy for $S^{\varepsilon}_{mp}(\tau, t)$.

Further, $B^{\varepsilon}(t)$ can be used to obtain an upper service curve estimate $S^{\varepsilon}_{br}(\tau, t)$ from the burst response where $S^{\varepsilon}_{br}(\tau, t) = S^{\varepsilon}_{mp}(\tau, t) + B^{\varepsilon}(t)$ from Eq. (6.18).

The choice of a practical burst rate r we already discussed in Sec. 7.3. In order to avoid that the service is limited by the arrivals, we concluded that the rate r has to be greater or equal then the service at any time. As stated by the service providers ISP1 and ISP2, the nominal uplink capacity is 50 Mbps. Therefore, any rate r that exceeds 50 Mbps emulates a burst. We particularly mention that the burst does not have to be CBR. However, it is the most practical choice.

For both providers we choose $t = 1$ sec. We send 100 bursts and obtain for each sample path the individual burst response. For $\varepsilon = 0.05$ we compute $S^{\varepsilon}_{br}(\tau, t)$ from Eq. (6.5).

An estimate of the minimal probe $\widetilde{A}_{mp}(\tau)$ immediately follows from $S^{\varepsilon}_{br}(\tau, t)$ by Eq. (6.15). Then, we send the minimal probe 100 times. For each measurement we take the backlog $B(t)$ at $t = 1$ sec and compute the 0.95-quantile $B^{\varepsilon}(t)$. We get the non-stationary service curve $S^{\varepsilon}_{mp}(\tau, t)$ by inserting the minimal probe $\widetilde{A}_{mp}(\tau)$ and the backlog quantile $B^{\varepsilon}(t)$ for $\varepsilon = 0.05$ into Eq. (6.13).

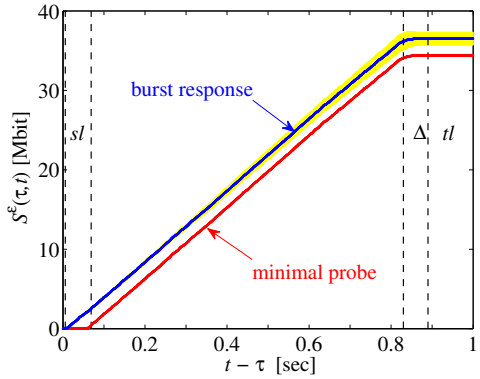
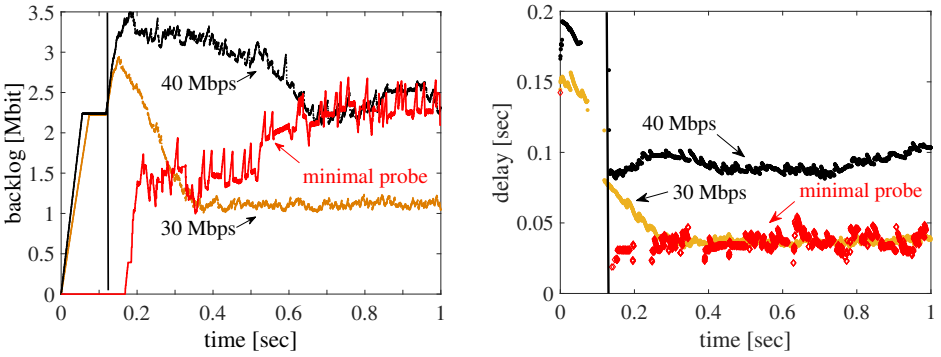


Figure 7.10: LTE service curve estimate at night for ISP1

We present first results for ISP1 in Fig. 7.10 for measurements during the night. Here, we show the mean of ten estimates of $\mathcal{S}^\varepsilon_{mp}(\tau,t)$ and $\mathcal{S}^\varepsilon_{br}(\tau,t)$ obtained by minimal probing and the burst response, respectively. Moreover, we add 0.95-confidence interval of $\mathcal{S}^\varepsilon_{br}(\tau,t)$, depicted as a yellow area. It confirms stable estimates. Note, that a confidence interval for $\mathcal{S}^\varepsilon_{mp}(\tau,t)$ provides only little more information. Consequently, we omit it in this and all other figures. We conclude a good accuracy of our estimate $\mathcal{S}^\varepsilon_{mp}(\tau,t)$, since the backlog $B^\varepsilon(t)$ of the minimal probe is small, i.e., around 2.1 MB. It separates the too optimistic bound of burst probing $\mathcal{S}^\varepsilon_{br}(\tau,t)$ from the lower service estimate $\mathcal{S}^\varepsilon_{mp}(\tau,t)$, where the backlog is the deviation resulting out of the super-additivity of the service, see Sec. 6.4. Comparing the service curve from Fig. 7.10 with the Figs. 6.3c and 6.5b it is noteworthy that the estimates show the same characteristics. In particular, we have the following service characteristics **S1** - **S5**:

s1 - SERVICE OUTAGES: For intervals $t-\tau \leqslant 8$ ms, both service curve estimates $\mathcal{S}^\varepsilon_{mp}(\tau,t)$ and $\mathcal{S}^\varepsilon_{br}(\tau,t)$ are equal to zero, indicating service outages on short time-scales, e.g., due to the characteristics of the radio channel.

s2 - STATIONARY LATENCY: The region marked with *sl* expresses a stationary latency of about 50 ms. As in Fig. 6.3c, $\mathcal{S}^\varepsilon_{mp}(\tau,t)$ identifies this region correctly, whereas $\mathcal{S}^\varepsilon_{br}(\tau,t)$ overestimates the service and attributes the stationary latency to the region marked with Δ. The effect is due to the super-additivity of the service that is caused by the stationary latency, see Sec. 6.3.



(a) 0.95-Quantile of the backlog

(b) 0.95-Quantile of the delay

Figure 7.11: 0.95-Quantile of backlog and delay for CBR LTE traffic of ISP1 including minimal probing

s3 - TRANSIENT LATENCY: The region marked with *tl* shows a transient latency of about 120 ms that is due to sleep scheduling. Comparing to Fig.7.3a validates this value.

s4 - CAPACITY LIMIT: The upward segment at the center has a slope of 44 Mbps. It denotes the effective capacity limit with respect to ε. The almost constant slope evidences a stable transmission rate for intervals of $t - \tau \geqslant 58$ ms.

s5 - $B^{\varepsilon}(t)$: The maximal vertical difference of the burst response $\mathcal{S}^{\varepsilon}_{br}(\tau, t)$ and minimal probe $\mathcal{S}^{\varepsilon}_{mp}(\tau, t)$ is defined by the accuracy of the method which is determined by the backlog for the minimal probe $\widetilde{A}_{mp}(\tau)$ at $t = 1$, i.e., 2.1 Mb.

In the case of Fig. 7.10, the minimal probe is adapted to the system's transient service characteristics. Thereby, the minimal probe sends a single packet at the beginning of each probe, which acts like a trigger to initiate the wake-up procedure of the UE. During connection establishment, the minimal probe does not send any further packets. Then, after the initial waiting time, the minimal probe has an average rate of 44 Mbps and remains very stable over the time, as we will see in the following, this cannot be expected, generally. However, it results in a 0.95-backlog quantile of about 2.1 Mb at the end of the probe.

The effects of the minimal probe for the backlog and delay over the time are shown in Fig. 7.11a and Fig. 7.11b. In both figures, we compare the 0.95-backlog and delay quantile of 30 Mbps and 40 Mbps CBR traffic, respectively, with a representative sample path from the minimal probe. In comparison to the CBR traffic, the transient overshoot and the following relaxation

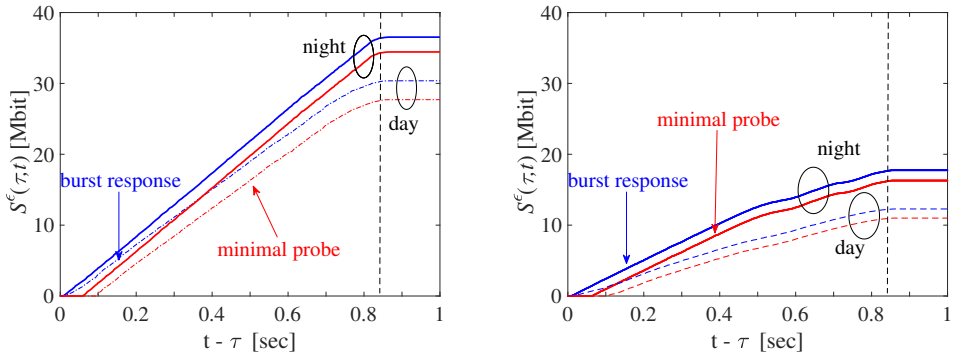


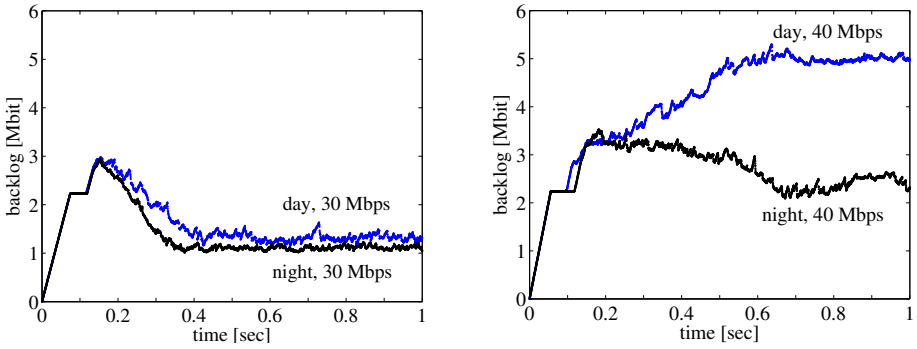
Figure 7.12: Service curve estimates of LTE, for ISP1 and ISP2. Solid lines show estimates obtained during the night, and dashed lines during the day, respectively.

time are eliminated. Furthermore, by adapting the minimal probe to the service characteristics of the network, we have neither a buffer overflow nor extensive delays over 100 ms for high arrival rates. This is because we perfectly utilize the network without overloading it.

7.4.3 *Diurnal Characteristics of LTE*

The results shown in Fig. 7.10 are made for carrier ISP1, where all measurements are done during night time. However, we know, e.g., from our work in [14], that the throughput is affected by many parameters, such as the location, SINR, and the day time. We will not investigate all of the different parameters, refer to [14], and concentrate the investigations on the day time.

In Fig.7.12a, we include the service curve estimates for carrier ISP1 during the day, plotted as dashed lines. We added a vertical dashed line which marks the transient latency (tl) and the Δ-region from Fig. 7.10 for the night. Note, that we will integrate this line for all non-stationary service curve measurements. In order to maintain clarity, we do not include additional lines, e.g. those for the day. In comparison to the night, we observe a reduction of the service. The effect confirms the results from [14] and may be attributed to the fact that other users follow a diurnal pattern, which leads to a higher utilization of the network at day times. As a consequence, the effective capacity limit is reduced from 44 Mbps at night to 38 Mbps during the day. Further, the backlog quantile $B^{\epsilon}(t)$ increases from 2.1 Mb to 2.7 Mb. We substantiate this pattern by sending CBR traffic at day and night.



(a) Backlog 30 Mbps (b) Backlog 40 Mbps

Figure 7.13: Transient 0.95-backlog quantiles of LTE during day and night for ISP1.

The corresponding backlog progressions for chosen rates, i.e., 30 Mbps and 40 Mbps are presented in Fig.7.13. There, we observe only a little difference between day and night for 30 Mbps and a significant increase in the backlog for 40 Mbps. Apart from that, the curves in Fig. 7.12a have the same general shape. In particular, we note that the transient plus stationary latency is only marginally affected by the time of measurement, see at the flat area in the upper right of the curves and Table 7.1.

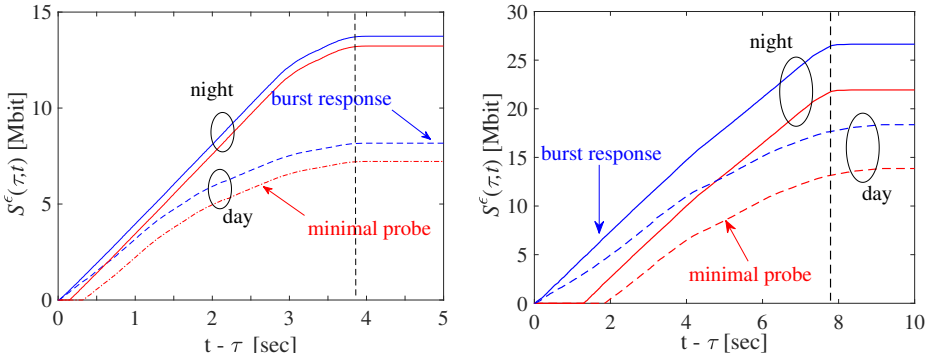
7.4.4 *Comparison of ISP1 and ISP2 in LTE*

Next, we compare our findings for ISP1 with ISP2. The corresponding non-stationary service curves for day and night are shown in Fig.7.12b. As for ISP1, we are able to extract the service characteristics **S1** - **S5** out of the non-stationary service curve, see Table 7.1. In detail, the service outages (**S1**) are 9 ms at night and increase to 12 ms at day time. Thus, slightly higher than for ISP1. The stationary latency (**S2**) is during the night very similar as for ISP1, but increases during the day to 76 ms or by 43%, which is 10 ms more than for ISP1. The transient latency is, however, better and even decreases from 100 ms to 90 ms for the day. Due to the higher stationary latency, the total time for the connection establishment is nevertheless comparable to the night. A noticeable difference, though, can be seen in the capacity limit. During the night, it is on top at 26 Mbps and decreases at day to 17 Mbps. In comparison to ISP1, it is around 20 Mbps lower, which also influences the mean backlog, i.e., it is reduced to 1.44 Mbps and 1.3 Mbps.

Over all we conclude that ISP1 outperforms ISP2 at least in terms of the service characteristics **S1** - **S5**.

	ISP1		ISP2	
	LTE		LTE	
	Day	Night	Day	Night
S1 - Service Outages	9.5 ms	8 ms	12 ms	9 ms
S2 - Stationary Latency	65 ms	50 ms	76 ms	53 ms
S3 - Transient Latency	115 ms	120 ms	90 ms	100 ms
S4 - Capacity Limit	38 Mbps	44 Mbps	17 Mbps	26 Mbps
S5 - Backlog $B^{\varepsilon}(t)$	2.7 Mb	2.1 Mb	1.3 Mb	1.45 Mb

Table 7.1: Service characteristics of LTE for ISP1 and ISP2



(a) ISP1 (b) ISP2

Figure 7.14: Service curve estimates of HSUPA, for ISP1 and ISP2. Solid lines show estimates obtained during the night, and dashed lines during the day, respectively.

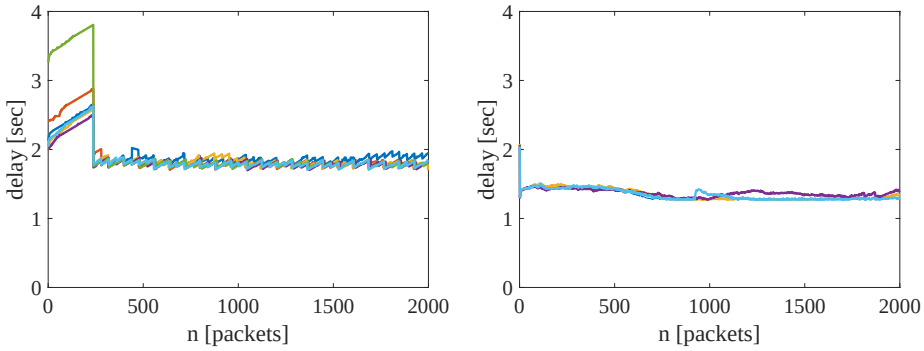
7.5 COMPARISON WITH HSPA AND EDGE

In the following, we perform the minimal probing method at day and night to estimate the non-stationary service curve for the preceding technologies EDGE (2G) and HSUPA (3G). Thereby, we compare the corresponding time-dependent services for ISP1 and ISP2. Additionally, we show results of the transient overshoot and relaxation time for ISP1 sending CBR arrival traffic.

To compensate the lower capacity and higher latencies, we reduce the burst rate and increase the measurement time accordingly.

Due to the same explanations as before, we are able to observe the service characteristics **S1** - **S5** for EDGE and HSUPA. An overview for all technologies and carriers is given in Table 7.2.

We start with measurements for HSUPA in Fig.7.14 where the solid lines are results for the night and dashed lines for the day, respectively. Fig.7.14a represents the results for ISP1 and Fig.7.14b for ISP2. For the sake of clarity,



(a) delay burst response (b) delay minimal probe

Figure 7.15: Example for the exhausted delay for ISP2 in a 3G network

we start with the nightly measurements. We choose a measurement duration of 5 sec for ISP1 and 10 sec for ISP2.

For short time-scales we observe service outages (**S1**), that occur for ISP1 in intervals $t - \tau \leqslant 15$ms, see Fig.7.14a. Similarly in Fig.7.14b the outages for ISP2, but with an increase of 50%, i.e. for times-scales $t - \tau \leqslant 21$ms. Next, we have a transient latency (**S3**) of 1 sec for both carriers. This can be compared with the DRX-cycles in Fig. 7.4. We identify 4 Mbps as capacity limit (**S4**) for ISP1. In comparison to LTE, the maximum capacity is reached more slowly. It is represented by the bent segment for time instances between $3 \leqslant t - \tau \leqslant 3.9$ sec. In this case, we do not have this tendency for ISP2. Nevertheless, the capacity limit is a bit lower, namely 3.4 Mbps.

However, the biggest difference between the two carriers we have for the stationary latency (**S2**). Whereas ISP1 has a moderate OWD of 130 ms, we identify a stationary latency of 1.2 sec for ISP2. Which is the reason for the doubled measurement period since the non-convex area at the upper right in Fig.7.14b is more than 2 sec long, resulting from the high transient and the stationary latency of 1 sec each. Surely, the larger stationary delays effects **S5** - the backlog $B^{\varepsilon}(t)$, since more packets are in flight such that we have 0.7 Mb for ISP1 and 4.7 Mb for ISP2.

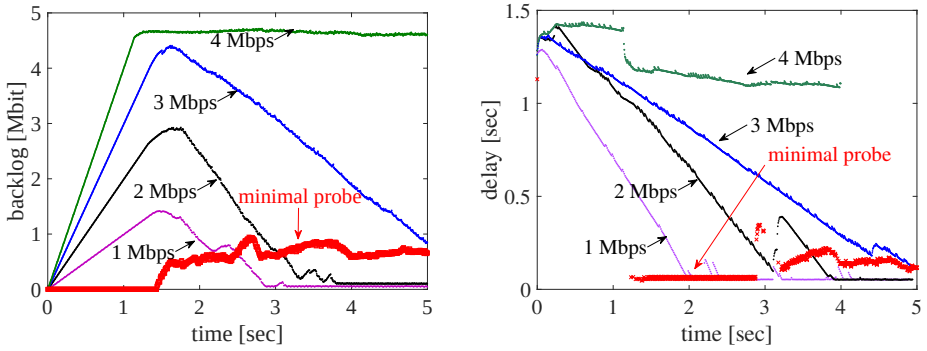
To investigate this effect, we take a closer look at the OWD delays that we get by performing the minimal probe method for ISP2. In Fig.7.15, we show the delay we obtain with the burst response method and minimal probe. For the burst response in Fig.7.15a, we have the typical behavior. For the first packet, we have the delay for the connection establishment in addition to the OWD. Then we send further packets although the connection is not open. Thus, the delay increases until the buffer is full, which is the case for

238 packets. As a reminder, for HSPA, we use a Teltonika HSPA+ RUT500 modem. After connection establishment, we observe a constant delay of around 1.75 sec for every burst. In comparison, for the minimal probe, we only identify the first packet with a delay larger than 2 sec, which is due to the fact that we need a trigger packet to open the connection to the base station. Then, the traffic of the minimal probe waits for a second until the connection between the base station and UE is there to send the carrier dependent traffic for ISP2. Interestingly, every packet has a delay of around 1.3 sec. So, it is less than for the burst response but still very large. The reason for this, we find in the current literature. In [42], the authors stated that the RTT could vary widely according to the distance to the base station, geographical location, and coverage of the base station. Fabini [65] finds that the payload of ICMP packets has a high impact on the RTT that increases up to several hundreds of ms for an ICMP packet. Further, [146] found out that high-throughput flows have a high impact on the OWD, such that delays of several seconds are possible. Hence, these results support our findings.

The results illustrate the advantages of the minimal probe, which provides estimates of non-stationary - service curve $\mathcal{S}^{\varepsilon}_{mp}(\tau, t)$ according to the different service characteristics of each carrier, such as the reduced rate at the beginning of a measurement or the carrier-dependent high latencies for the corresponding transmission technology.

It becomes even more apparent when we look at the transient effects that occur when sending CBR traffic, see Fig. 7.16. Here, we present the 0.95-quantiles of the backlog and delay progression for CBR traffic up to the limiting rate of 4 Mbps for ISP1. For the backlog in Fig.7.16a we observe the typical transient overshoot after the connection is established.

In contrast to LTE, the backlog for HSPA has a less sharp peak, as can be seen from the rounded region at 1.5 seconds. This effect is due to the slow increase in the transmission rate after the connection is established. The relaxation time then takes, e.g., for 3 Mbps clearly more than 5 seconds to reach a stationary state and probably does not relax for 4 Mbps at all. In contrast to that, we also added a backlog sample path for the minimal probe. Due to the fact that the minimal probe takes the system's characteristic into account, it is obvious and remarkable that the backlog reaches the stationary state without any relaxation time and transient overshoot. Similarly, we identify large delays at the beginning for all CBR arrivals in Fig.7.16b. Significant transient overshoot is observed at rates like 2 Mbps and higher.



(a) Backlog + MP for ISP1

(b) Delay + MP for ISP1

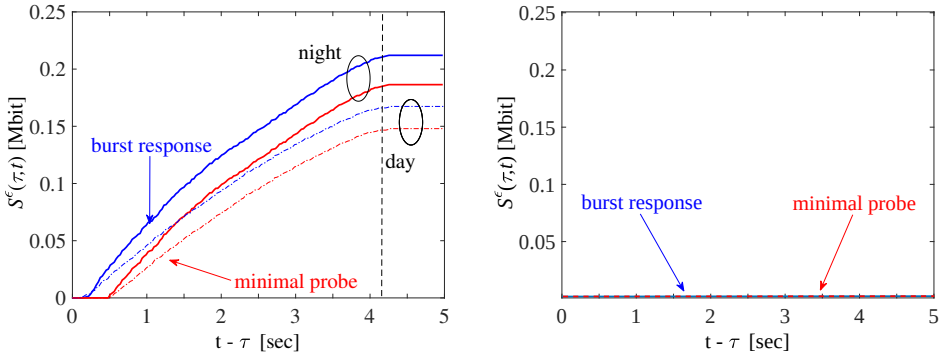
Figure 7.16: Transient 0.95-backlog and delay quantiles of HSUPA for CBR traffic and ISP1 including the results for the minimal probe

Furthermore, the relaxation time at all rates is very long and lasts several seconds, starting with delays of more than one second, which is not practical for delay-sensitive applications. Sending the traffic according to the minimal probe eliminates the transient delays. Hence, an application such as high-quality up-streaming is more likely.

Note that we also included the non-stationary service curves at day time for ISP1 and ISP2 in Fig.7.14 in the form of the dashed lines. As before, we find the service characteristics **S1** - **S5**, see Table 7.2. Generally, we find that all measurements during the day have the same trend. In comparison to the night, we observe a service reduction, e.g., a decrease of the capacity limit (**S4**) to 3.2 Mbps for ISP1, while the stationary latency (**S2**) increases to 340 ms. Again, these effects may be linked to higher activities of users during the day and lead to a reduced accuracy as can be seen from the larger deviation $B^{\varepsilon}(t)$ of the lower estimate $\mathcal{S}^{\varepsilon}_{mp}(\tau, t)$ from the upper estimate $\mathcal{S}^{\varepsilon}_{br}(\tau, t)$.

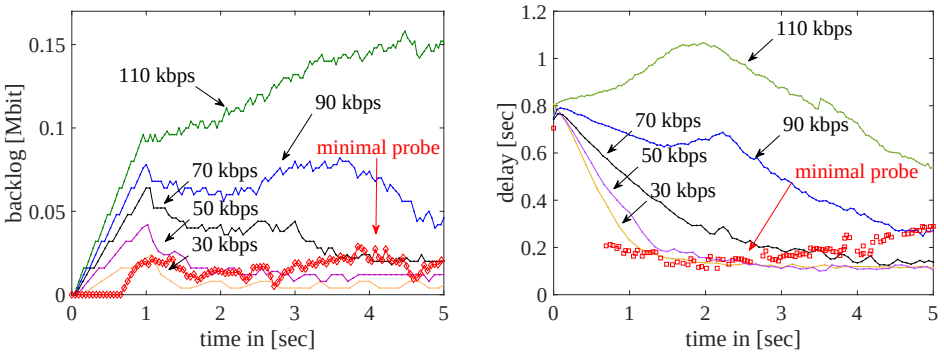
Finally, we estimate the non-stationary service curve $\mathcal{S}^{\varepsilon}_{mp}(\tau, t)$ and service characteristics **S1** - **S5** for EDGE with a measurement duration of 5 sec and a burst rate of 1 Mbps. Due to the low capacity we reduced the packet size to 500 Bytes. The service curve estimates for ISP1 and ISP2 are presented in Fig. 7.17.

We start with ISP1 from Fig.7.17a and the night measurements, i.e., the solid lines. Similarly to the corresponding results for HSPA the estimate of the service curve has a bend which is even stronger than for HSPA after connection establishment. It indicates that the parameters **S1** - **S5** are less explicit. In particular, we have service outages **S1** of $t - \tau \leqslant 200$ ms, a



(a) ISP1 (b) ISP2

Figure 7.17: Service curve estimates of EDGE, for ISP1 and ISP2. Solid lines show estimates obtained during the night, and dashed lines during the day, respectively.



(a) EDGE backlog (b) EDGE delay

Figure 7.18: Transient 0.95-backlog and delay quantiles of EDGE for CBR traffic and ISP1 including the results of the minimal probe.

stationary latency **S2** of 300 ms, a transient latency **S3** of 500 ms, a capacity limit **S4** of 70 kbps and a backlog **S5** of 256 kb. For daytimes we notice lower service outages **S1** of $t - \tau \leqslant 120$ ms, but also higher stationary **S2** and transient **S3** latencies of 380 ms and 600 ms, respectively. Additionally, we observe 51 kbps as capacity limit **S4**, which is only 73% to the nightly capacity. It further reduces **S5** to 200 kb and increases in this case the accuracy $B^\epsilon(t)$. A possible explanation is because the lower capacity limit leads to a lower number of packets sent by the minimal probe such that the number of packets in flight reduces as well.

Note that we also performed measurements for ISP2. For LTE and HSPA we already investigated lower service characteristics **S1** - **S5** in comparison to ISP1. For EDGE, we even observed the extreme case, i.e., we only

measured a service in really rare cases with a service of around 10 kbps. Therefore, the 0.95-quantile lead to a service of 0 kbps. As a result, we end up with the special case where the service estimate for burst response $\mathcal{S}^{\varepsilon}_{br}(\tau, t)$ and minimal probe $\mathcal{S}^{\varepsilon}_{mp}(\tau, t)$ are equal to zero for all $t \geqslant 0$.

Hence, the minimal probing method does not only work for good channel qualities; it also provides estimates in cases where only low or no service at all is available. Furthermore, we are able to send arrival traffic in such a way that it eliminates transient overshoots and the following relaxation time by taking the system's specific service into account. This holds for EDGE, as well, as we can see from the 0.95-quantiles of the backlog and delay for CBR traffic in an EDGE network of ISP1. As before, we have a noticeable transient overshoot for the backlog, which is relaxed after several seconds depending on the rate, whereas the minimal probe in Fig.7.18 reaches the stationary backlog immediately. This is also valid for the delay progression of CBR and minimal probe traffic in Fig.7.18b.

Therefore, we conclude that regardless of the network type, the minimum probing method results in an accurate, non-stationary service curve estimate that is capable of determining system characteristics while eliminating transient overshoots and relaxation times because minimum probing takes into account time-dependent network characteristics such as DRX mode.

	ISP1						ISP2					
	EDGE		HSUPA		LTE		EDGE		HSUPA		LTE	
	Day	Night	Day	Night	Day	Night	Day	Night	Day	Night	Day	Night
S1 - Service Outages	120 ms	200 ms	19 ms	15 ms	9.5 ms	8 ms	-	-	23 ms	21 ms	12 ms	9 ms
S2 - Stat. Latency	380 ms	300 ms	340 ms	130 ms	65 ms	50 ms	-	-	1.7 sec	1.2 sec	76 ms	53 ms
S3 - Trans. Latency	600 ms	500 ms	800 ms	1 sec	115 ms	120 ms	-	-	0.8 sec	1 sec	90 ms	100 ms
S4 - Capacity Limit	51 kbps	70 kbps	3.2 Mbps	4 Mbps	38 Mbps	44 Mbps	-	-	2.8 Mbps	3.4 Mbps	17 Mbps	26 Mbps
S5 - Backlog $B^{\varepsilon}(t)$	20 kb	25 kb	1.05 Mb	0.7 Mb	2.7 Mb	2.1 Mb	-	-	4.5 Mb	4.7 Mb	1.3 Mb	1.45 Mb

Table 7.2: Comparison of service characteristics for EDGE, HSUPA and LTE for both carriers

8

CONCLUSION AND FUTURE WORK

In this thesis, we contributed a notion of non-stationary service curve, which allowed us to analyze transient phases in computer networks such as in cellular networks. The established theory is integrated into the framework of stochastic network calculus, where time-variant systems are handled typically by using stationary random processes or stationary bounds. In order to model time-variance and therefore changes over time, bivariate instead of univariate functions were used.

We modeled systems with deterministic sleep scheduling and derived time-dependent performance bounds for backlog and delay, where the measures of interests were the transient overshoot and the relaxation time, which is the time it takes to reach steady-state. Then, with the help of an exact solution, we illustrated that, on the one hand, only a time-variant service description could follow the exact progress in the same way. On the other hand, the time-invariant formulation remained in the worst-case because of the univariate functions that consider the length of the time-intervals and not the time instances itself. The importance and benefits of the extension are illustrated by studying the influence of arrival rate α and length of sleep cycles T. Depending on the parameters, the maximum overshoot occurred after T, and the relaxation time was able to reach values that were a magnitude larger than T. Furthermore, our results showed that time-invariant performance bounds could easily be many times larger than the equivalent steady-state time-variant bounds.

Based on regenerative processes and our findings from the time-variant concepts, we extended the theory from the deterministic to the stochastic network calculus and derived non-stationary service curves. In this way, we can model systems with sleep scheduling, which had random wake-up times and services. Thus, we laid the theoretical foundation to study the transient and stationary behavior of systems like DRX mode in cellular networks such that we obtained new insights into the specific implementation of sleep scheduling in these networks.

Since we could not expect to know the service in cellular networks in advance, we generally assumed to have a black-box. Therefore, we investigated measurement-based methods to identify the characteristics

and services of an unknown system. More precisely, our goal was to get along without certain assumptions about the internals of a network, while estimating a general service model of a linear system with a time-variant, regenerative service.

In doing so, we refined well-known measurement methods such as the rate scanning and burst response method to estimate a non-stationary service curve. We encountered additional difficulties due to the non-convexity and super-additivity of the service.

To overcome these limitations, we developed a novel measurement method to obtain a non-stationary service curve. This is the minimal probing procedure, which consisted of two steps and also provided a measure for the accuracy. First, the method estimated a probe that is minimal under certain conditions. The second step was to use this minimal probe and estimate a non-stationary service curve with a defined accuracy. In comparison to many tools that estimate, e.g., the available bandwidth as a single value, our new method showed the network's actual service progression. It also provided a wide range of additional information, including transient delays due to sleep scheduling, stationary OWDs, time-dependent service rates, service outages due to wireless transmission characteristics, and the method's accuracy.

After we showed by simulations that the minimal probe method gave valid results, we performed a huge measurement campaign in cellular networks. The evaluation was the first practical validation of our new method in several real production networks of EDGE, HSPA, and LTE. We provided new insights into the corresponding DRX mode and estimated the transient service characteristics for two providers at day and night. Based on our estimation of the service curve, we were able to show characteristic features of the cellular data service that explained the observation of significant transient overshoots and long relaxation times. Our measurements had shown that delays in the range of seconds are common with EDGE and HSPA. For LTE, however, we found an improvement in the order of magnitude. Generally, we believe that the measurement results provided good evidence that our model of non-stationary service curves and the method for estimating the shape of the service curve are suitable for characterizing key aspects of mobile network service, such as stationary and transient delays and rate limitations. Additionally, we observed that sending the minimal probe as arrival traffic lead to an elimination of the transient overshoot

and relaxation times such that steady-state delay and backlog values were reached as soon as possible with the highest possible rate.

In this way, the minimal probing method provided knowledge about the system so that video streaming applications could benefit by keeping latencies to a minimum with the best possible video quality. As a future work, integrating the minimal probing method into video coding systems can lead to an improvement of the user experience, regarding the observed latencies versus quality. Apart from this, we are convinced that the non-stationary service curve model and minimal probing method can be extended to new cellular technologies such as the upcoming 5G standard and the analysis of other transient effects such as TCP slow start. For the latter case, the theory must be adapted by using the max-Plus instead of the min-Plus algebra.

Part II

APPENDIX

A

APPENDIX

PROPERTIES IN SYSTEM THEORY

A right-continuous function is defined as follows:

Definition 2. *(Right-continuity) Let* f *be a function in* $\mathbb{R}$*. Then is* f *right-continuous at* x_0 *iff*

$$f(x_0^+) = f(x_0)$$

where $f(x_0^+) = \lim_{x \downarrow x_0} f(x)$

In the classical system theory it is common to consider linear and time-invariant systems. Therefore, we define linearity and time-invariance in system theory.

Definition 3. *(Linearity) Let* V *and* W *be two vector spaces over the same field* K*. A function* $f : V \to V$ *is called to be a linear map if for any two vectors* $u,v \in V$ *and any scalar* $c \in K$ *the following two conditions are satisfied:*

1. *additivity:* $f(u+v) = f(u) + f(v)$
2. *homogeneity:* $f(cu) = cf(u)$

In our case the field K are the real numbers $\mathbb{R}$. In a time-invariant system we have the effect that a time-shift in the input signal produces also a time-shift in the output signal, i.e.,

Definition 4. *(Time-Invariance) A system is time-invariant if a time-delay of the input signal* $x(t+\delta)$ *that yields to time-delay of the output* $y(t+\delta) = \Pi(x(t+\delta))$ *is equal to the signal* $y(t) = \Pi(x(t))$ *for all* $\delta \geqslant 0$*.*

PROPERTIES IN MIN-PLUS NETWORK CALCULUS

In comparison to the classical system theory we changed the sum (or integral in the continuous case) with the minimum (infimum) and the multiplication with the addition. Hence, we have to adapt the definition of linearity 3 and time-invariance 4.

Definition 5. *(Min-Plus-Linearity) Let* $A_1(t)$ *and* $A_2(t)$ *be two input signals at time* t *and* $D_1(t)$ *and* $D_2(t)$ *the corresponding output signals. Then, the* $\otimes$*-operator is linear, iff*

1. *additivity:* $\inf(D_1(t), D_2(t)) = \Pi(\inf(A_1(t), A_2(t)))$
2. *homogeneity:* $D_1(t) + a = \Pi(A_1(t) + a)$

The definition of the time-invariance changes to the following:

Definition 6. *(Time-Invariance) Let* $A(t)$ *be the input at time* t *that results into the output* $D(t) = \Pi(A(t))$*, where* Π *is the underlying operator. If a time-shift of* $\delta \geqslant 0$ *yields the same but time-shifted output* $D(\delta, t + \delta) = \Pi(A(\delta, t + \delta))$ $\forall \delta \geqslant 0$ *then the operator* Π *has the property of time-invariance.*

For the min-operator ($\wedge$) and the $+$-operator we can show that the algebraic structures $(\mathcal{R} \cup \infty, \wedge, +)$ is a commutative semifield with the following properties:

1. (Associativity) $\forall a, b, c \in \mathbb{R} \cup \infty$
   $(a + b) + c = a + (b + c)$
   $(a \wedge b) \wedge c = a \wedge (b \wedge c)$
2. (Commutativity) $\forall a, b \in \mathbb{R} \cup \infty$
   $a + b = b + a$
   $a \wedge b = b \wedge a$
3. (Distributivity) $\forall a, b, c \in \mathbb{R} \cup \infty$
   $(a \wedge b) + c = (a + c) \wedge (b + c)$

Let $f(t), g(t) \in \mathcal{F}_0$. Then, with the following operations we have a set of properties:

1. Let $\otimes$ be the min-plus operator from Eq. (2.2), i.e, for f, g we have $(f \otimes g)(t) = \min_{0 \leqslant s \leqslant t}\{f(s) + g(t - s)\}$.
2. Let $\star$ be the pointwise minimum of f and g, i.e., $(f \star g)(t) = \min\{f(t), g(t)\}$

Then, we have the following properties:

1. (Associativity) $\forall f \in \mathcal{F}_0$
   $(f \star g) \star c = f \star (g \star c)$
   $(f \otimes g) \otimes h = f \otimes (g \otimes h)$
2. (Commutativity) $\forall f \in \mathcal{F}_0$
   $f \star g = g \star f$
   $f \otimes g = g \otimes f$

3. (Distributivity) $\forall f, g, h \in \mathcal{F}_0$
   $(f \star b) \otimes h = (f \otimes h) \star (g \otimes h)$

4. (Zero element) $\forall f \in \mathcal{F}_0$
   $f \otimes \epsilon = f$, where ϵ is the sequence with $\epsilon(t) = \infty$

5. (Absorbing Zero element)$\forall f \in \mathcal{F}_0$
   $f \star \epsilon = \epsilon \star f = \epsilon$

6. (Identity element) $\forall f \in \mathcal{F}_0$
   $f \otimes \mathbf{e} = \mathbf{e} \otimes f = f$, where $\mathbf{e}$ is the sequence with $\mathbf{e}(0) = 0$ and $\mathbf{e}(t) = \infty$ for all $t > 0$.

7. (Idempotency of addition) $\forall f \in \mathcal{F}_0$
   $f \star f = f$

8. (Monotonicity) $\forall f, \tilde{f}, g, \tilde{g} \in \mathcal{F}_0$, if $f \leqslant \tilde{f}$ and $g \leqslant \tilde{g}$ then
   $f \star g \leqslant \tilde{f} \star \tilde{g} \leqslant f$
   $f \otimes g \leqslant \tilde{f} \otimes \tilde{g}$
   If g is also in $\mathcal{F}_0$, then $f \otimes g \leqslant f$. If both f and g are in $\mathcal{F}_0$, then $f \star g \geqslant f \otimes g$.

CHERNOFF'S THEOREM

Theorem 1. *Chernoff Bound: Let X be a random variable and $M_X(\theta)$ be the MGF of X with the free parameter $\theta \geqslant 0$. Then it holds that*

$$P[X > x] \leqslant e^{-\theta x} M_X(\theta) \tag{A.1}$$

MARTINGALE

Definition 7. *Martingale [62]: The sequence $\{x_n, n \geqslant 1\}$ is called a martingale relative to a filtration $\{\mathbf{F}_n; n \geqslant 1\}$ if each x_n has an expectation, and if for $m < n$ the expected value of x_n given the past up to time m is X_m, that is*

$$E[x_n|\mathbf{F}_m] = x_m, \tag{A.2}$$

where a filtration is a finite or infinite increasing sequence of σ-algebras in the sense that $\mathbf{F}_1 \subset \mathbf{F}_2 \subset \ldots$.

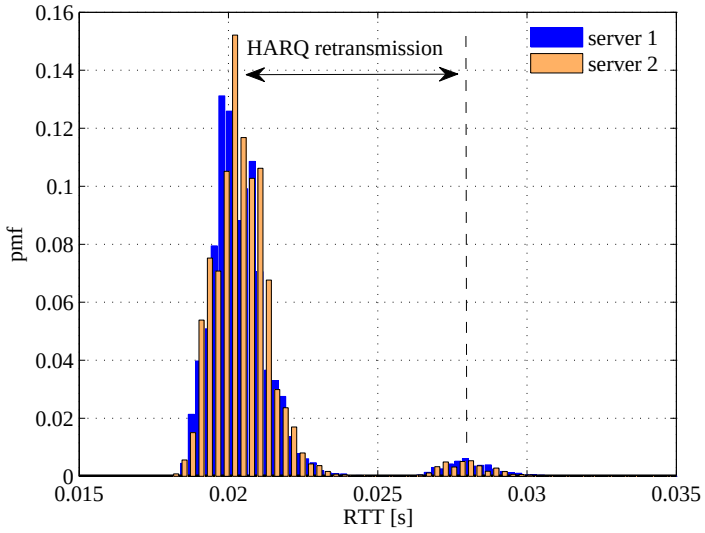


Figure A.1: PMF of the RTTs for TCP connection establishment handshakes

HARQ

We use a result from our previous work [21] to illustrate the occurrence of HARQ retransmissions in LTE networks. To do so, we perform 5×10^4 independent HTTP handshakes, where we send HTTP requests from client (A) to two popular web servers. Here, we are not interested to analyze effects from DRX implementations. Thus, we make sure to be in `RRC_CONNECTED` state and *continuous reception* mode. Then, a single handshake is the time between the sending timestamp of a SYN message and receiving timestamp of the corresponding SYN/ACK reply. We present the empirical pmf in Fig.A.1, where we observe for both webservers that typically a single handshake requires 20 ms. In addition to that, we find a characteristic second mode at about 28 ms. This additional 8 ms coincides with a HARQ retransmission.

BIBLIOGRAPHY

[1] World bank. https://data.worldbank.org/indicator/sp.rur.totl.zs? [accessed 04.04.2020].

[2] 3GPP specification TR 23.770 (2012). 3rd Generation Partnership Project; Technical Specification Group Services and System Aspects; Study on system impacts of extended Discontinuous Reception (DRX) cycle for power consumption optimization . Release 13, version 0.0.

[3] 3GPP specification TS 25.913 (2009). Requirements for Evolved UTRA (E-UTRA) and Evolved UTRAN (E-UTRAN). Release 8, version 8.0.

[4] 3GPP specification TS 36.304 (2011). Evolved Universal Terrestrial Radio Access (E-UTRA); User Equipment (UE) procedures in idle mode. Release 8, version 8.10.

[5] 3GPP specification TS 36.321 (2012). Evolved Universal Terrestrial Radio Access (E-UTRA); Medium Access Control (MAC) protocol specification. Release 8, version 8.12.

[6] 3GPP specification TS 36.322 (2010). Evolved Universal Terrestrial Radio Access (E-UTRA); Radio Link Control (RLC) protocol specification. Release 8, version 8.8.

[7] 3GPP specification TS 36.331 (2013). Evolved Universal Terrestrial Radio Access (E-UTRA); Radio Resource Control (RRC); Protocol specification. Release 8, version 8.20.

[8] 3GPP specification TS 50.059. (2001). Enhanced Data rates for GSM Evolution (EDGE); Project scheduling and open issues for EDGE. Release 4, version 0.1.

[9] Abed, G. A., Ismail, M., and Jumari, K. (2011). Traffic modeling of lte mobile broadband network based on ns-2 simulator. In *2011 Third International Conference on Computational Intelligence, Communication Systems and Networks*, pages 120–125. IEEE.

[10] Agharebparast, F. and Leung, V. C. M. (2006). Slope domain modeling and analysis of data communication networks: A network calculus complement. In *Proc. of IEEE ICC*, pages 591–596.

[11] Agrawal, R., Baccelli, F., and Rajan, R. (2004). An algebra for queueing networks with time-varying service and its application to the analysis of Integrated Service networks. *Mathematics of Operation Research*, 29(3):559–591.

[12] Agrawal, R., Cruz, R. L., Okino, C. M., and Rajan, R. (1999). Performance bounds for flow control protocols. *IEEE/ACM Trans. Networking*, 7(3):310–323.

[13] Agrawal, R. and Rajan, R. (1996). Performance bounds for guaranteed and adaptive services. Technical Report RC 20649, IBM.

[14] Akselrod, M., Becker, N., Fidler, M., and Luebben, R. (2017). 4g lte on the road-what impacts download speeds most? In *2017 IEEE 86th Vehicular Technology Conference (VTC-Fall)*, pages 1–6. IEEE.

[15] Alcuri, L., Barbera, G., and D'Acquisto, G. (2005). Service curve estimation by measurement: An input output analysis of a softswitch model. In *Proc. of QoS-IP*, pages 49–60.

[16] Arita, C. and Yanagisawa, D. (2010). Exclusive queueing process with discrete time. Technical Report arXiv:1008.4651v2, arXiv.

[17] Baccelli, F., Cohen, G., Olsder, G. J., and Quadrat, J.-P. (1992). *Synchronization and Linearity: An Algebra for Discrete Event Systems*. Wiley.

[18] Beck, M. (2016). Towards the analysis of transient phases with stochastic network calculus. In *2016 17th International Telecommunications Network Strategy and Planning Symposium (Networks)*, pages 164–169. IEEE.

[19] Becker, N. and Fidler, M. (2015). A non-stationary service curve model for performance analysis of transient phases. In *Proc. of ITC 27*, pages 116–124.

[20] Becker, N. and Fidler, M. (2018). A non-stationary service curve model for estimation of cellular sleep scheduling. *IEEE Transactions on Mobile Computing*, 18(1):28–41.

[21] Becker, N., Rizk, A., and Fidler, M. (2014). A measurement study on the application-level performance of LTE. In *Proc. of IFIP Networking*.

[22] Bennett, J. C. R., Benson, K., Courtney, W. F., and Le Boudec, J.-Y. (2002). Delay jitter bounds and packet scale rate guarantee for expedited forwarding. *IEEE/ACM Trans. Networking*, 10(4):529–540.

[23] Bhamber, R. S., Fowler, S., Braimiotis, C., and Mellouk, A. (2013). Analytic analysis of lte/lte-advanced power saving and delay with bursty traffic. In *Communications (ICC), 2013 IEEE International Conference on*, pages 2964–2968. IEEE.

[24] Bontu, C. and Illidge, E. (2009). Drx mechanism for power saving in lte. *Communications Magazine, IEEE*, 47(6):48–55.

[25] Boorstyn, R.-R., Burchard, A., Liebeherr, J., and Oottamakorn, C. (2000). Statistical service assurances for traffic scheduling algorithms. *IEEE J. Select. Areas Commun.*, 18(12):2651–2664.

[26] Borys, A., Wasielewska, K., and Rybarczyk, D. (2013). A contribution to the system-theoretic approach to bandwidth estimation. *Int. Journal of Electronics and Telecommunications*, 59(2):141–149.

[27] Bredel, M., Bozakov, Z., and Jiang, Y. (2010). Analyzing router performance using network calculus with external measurements. In *Proc. of IEEE IWQoS*.

[28] Bredel, M. and Fidler, M. (2008). A measurement study of bandwidth estimation in ieee 802.11 g wireless lans using the dcf. In *International Conference on Research in Networking*, pages 314–325. Springer.

[29] Bredel, M. and Fidler, M. (2009). Understanding fairness and its impact on quality of service in IEEE 802.11. In *Proc. of IEEE INFOCOM*, pages 1098–1106.

[30] Burchard, A., Liebeherr, J., and Ciucu, F. (2011). On superlinear scaling of network delays. *IEEE/ACM Trans. Networking*, 19(4):1043–1056.

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

[32] Cetinkaya, C., Kanodia, V., and Knightly, E. W. (2001). Scalable services via egress admission control. *IEEE Trans. Multimedia*, 3(1):69–81.

[33] Champati, J. P., Al-Zubaidy, H., and Gross, J. (2018). Transient delay bounds for multi-hop wireless networks. *arXiv preprint arXiv:1806.03328*.

[34] Chang, C.-S. (1994). Stability, queue length and delay of deterministic and stochastic queueing networks. *IEEE Trans. Automat. Contr.*, 39(5):913–931.

[35] Chang, C.-S. (1996). On the exponentiality of stochastic linear systems under the max-plus algebra. *IEEE Trans. Automat. Contr.*, 41(8):1182–1188.

[36] Chang, C.-S. (1998). On deterministic traffic regulation and service guarantees: A systematic approach by filtering. *IEEE Trans. Inform. Theory*, 44(3):1097–1110.

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

[38] Chang, C.-S. and Cruz, R. L. (1999). A time varying filtering theory for constrained traffic regulation and dynamic service guarantees. In *Proc. of IEEE INFOCOM*, pages 63–70.

[39] Chang, C.-S., Cruz, R. L., Le Boudec, J.-Y., and Thiran, P. (1999). A min-plus system theory for constrained traffic regulation and dynamic service guarantees. Technical Report SSC/1999/024, EPFL.

[40] Chang, C.-S., Cruz, R. L., Le Boudec, J.-Y., and Thiran, P. (2002). A min, + system theory for constrained traffic regulation and dynamic service guarantees. *IEEE/ACM Trans. Networking*, 10(6):805–817.

[41] Chaudhari, S. S. and Biradar, R. C. (2015). Survey of bandwidth estimation techniques in communication networks. *wireless personal communications*, 83(2):1425–1476.

[42] Chen, Y., Duffield, N., Haffner, P., Hsu, W.-L., Jacobson, G., Jin, Y., Sen, S., Venkataraman, S., and Zhang, Z.-L. (2013a). Understanding the complexity of 3g umts network performance. In *2013 IFIP Networking Conference*, pages 1–9. IEEE.

[43] Chen, Y.-C., Lim, Y.-s., Gibbens, R. J., Nahum, E. M., Khalili, R., and Towsley, D. (2013b). A Measurement-based Study of MultiPath TCP Performance over Wireless Networks. In *Proc. of ACM IMC*, pages 455–468.

[44] Choe, J. and Shroff, N. B. (1998). A central-limit-theorem-based approach for analyzing queue behavior in high-speed networks. *IEEE/ACM Trans. Networking*, 6(5):659–671.

[45] Ciucu, F. (2007a). Network calculus delay bounds in queueing networks with exact solutions. In *Proc. of ITC-20*, volume 4516 of *LNCS*, pages 495–506. Springer.

[46] Ciucu, F. (2007b). *Scaling Properties in the Stochastic Network Calculus*. PhD thesis, Univ. of Virginia.

[47] Ciucu, F., Burchard, A., and Liebeherr, J. (2006). Scaling properties of statistical end-to-end bounds in the network calculus. *IEEE/ACM Trans. Networking*, 14(6):2300–2312.

[48] Ciucu, F. and Schmitt, J. (2012). Perspectives on network calculus - no free lunch but still good value. In *Proc. of ACM SIGCOMM*, pages 311–322.

[49] Ciucu, F., Schmitt, J., and Wang, H. (2011). On expressing networks with flow transformations in convolution-form. In *Proc. IEEE INFOCOM*.

[50] Cox, C. (2012). *An Introduction to LTE: LTE, LTE-Advanced, SAE and 4G Mobile Communications*. Wiley.

[51] Cruz, R. and Taneja, M. (1998). An analysis of traffic clipping. In *Proc. of Conference on Information Science and Systems, Princeton University*.

[52] Cruz, R. L. (1991a). A calculus for network delay, part I and II: Network elements in isolation and network analysis. *IEEE Trans. Inform. Theory*, 37(1):114–141.

[53] Cruz, R. L. (1991b). A calculus for network delay, part I: Network elements in isolation. *IEEE Trans. Inform. Theory*, 37(1):114–131.

[54] Cruz, R. L. (1991c). A calculus for network delay, part II: Network analysis. *IEEE Trans. Inform. Theory*, 37(1):132–141.

[55] Cruz, R. L. (1995). Quality of service guarantees in virtual circuit switched networks. *IEEE J. Select. Areas Commun.*, 13(6):1048–1056.

[56] Cruz, R. L. (1996). Quality of service management in Integrated Services networks. In *Proc. of Semi-Annual Research Review, Center of Wireless Communication, UCSD*.

[57] Cruz, R. L. (1998). SCED+: Efficient management of quality of service guarantees. In *Proc. of IEEE INFOCOM*, pages 625–634.

[58] Cruz, R. L. and Okino, C. M. (1996). Service guarantees for window flow control. In *Prof. of Allerton Conference on Communication, Control, and Computing*.

[59] de Souza e Silva, E. and Gail, H. R. (1998). An algorithm to calculate transient distributions of cumulative rate and impulse based reward. *Stochastic models*, 14(3):509–536.

[60] Deng, S. and Balakrishnan, H. (2012). Traffic-aware Techniques to Reduce 3G/LTE Wireless Energy Consumption. In *Proc. of ACM CoNEXT*, pages 181–192.

[61] Doob, J. L. (1953). *Stochastic Processes*. Wiley.

[62] Doob, J. L. (1971). What is a martingale? *The American Mathematical Monthly*, 78(5):451–463.

[63] Dukkipati, N. et al. (2010). An argument for increasing TCP's initial congestion window. *ACM Comp. Com. Rev.*, 40(3):26–33.

[64] Duzgun, F. and Yamamoto, G. (2016). The effect of promoter incentive to the smartphone sales in retail chains: A turkish case. *International Journal of Economics & Management Sciences*, 5.

[65] Fabini, J., Karner, W., Wallentin, L., and Baumgartner, T. (2009). The illusion of being deterministic–application-level considerations on delay in 3g hspa networks. In *International Conference on Research in Networking*, pages 301–312. Springer.

[66] Falk, J., Dürr, F., and Rothermel, K. (2019). Modeling time-triggered service intermittence in network calculus. In *Proceedings of the 27th International Conference on Real-Time Networks and Systems*, pages 90–100.

[67] Ferrari, D. (1990). Client requirements for real-time communication services. *IEEE Communications Magazine*, 28(11):65–72.

[68] Fidler, M. (2006a). An end-to-end probabilistic network calculus with moment generating functions. In *Proc. of IWQoS*, pages 261–270.

[69] Fidler, M. (2006b). A network calculus approach to probabilistic quality of service analysis of fading channels. In *Proc. of IEEE Globecom*.

[70] Fidler, M. (2010). A survey of deterministic and stochastic service curve models in the network calculus. 12(1):59–86.

[71] Fidler, M., Lübben, R., and Becker, N. (2015). Capacity-Delay-Error-Boundaries: A Composable Model of Sources and Systems. *IEEE Trans. Wireless Commun.*, 14(3):1280–1294.

[72] Fidler, M. and Recker, S. (2006). Conjugate network calculus: A dual approach applying the Legendre transform. *Computer Networks*, 50(8):1026–1039.

[73] Fidler, M. and Rizk, A. (2015). A guide to the stochastic network calculus. 17(1):92–105.

[74] Fitzek, F. H. P. and Reisslein, M. (2001). MPEG–4 and H.263 video traces for network performance evaluation. *IEEE Network*, 15(6):40–54.

[75] García Villalba, L. J., Sandoval Orozco, A. L., Trivino Cabrera, A., and Barenco Abbas, C. J. (2009). Routing protocols in wireless sensor networks. *Sensors*, 9(11):8399–8421.

[76] Gawas, A. (2015). An overview on evolution of mobile wireless communication networks: 1g-6g. *International Journal on Recent and Innovation Trends in Computing and Communication*, 3(5):3130–3133.

[77] Goyal, P., Lam, S. S., and Vin, H. M. (1997). Determining end-to-end delay bounds in heterogeneous networks. *Multimedia Systems*, 5(3):157–163.

[78] Grimmett, G. and Stirzaker, D. (2001). *Probability And Random Processes*. Oxford University Press, third edition.

[79] Hisakado, T., Okumura, K., Vukadinovic, V., and Trajkovic, L. (2003). Characterization of a simple communication network using Legendre transform. In *Proc. of ISCAS*, pages 738–741.

[80] Horváth, A., Paolieri, M., Ridi, L., and Vicario, E. (2012). Transient analysis of non-Markovian models using stochastic state classes. *Performance Evaluation*, 69(7):315–335.

[81] Huang, J., Qian, F., Gerber, A., Mao, Z. M., Sen, S., and Spatscheck, O. (2012). A close examination of performance and power characteristics of 4g LTE networks. In *Proc. of ACM MobiSys*, pages 225–238.

[82] Huang, J., Qian, F., Guo, Y., Zhou, Y., Xu, Q., Mao, Z. M., Sen, S., and Spatscheck, O. (2013). An in-depth study of LTE: Effect of network protocol and application behavior on performance. *ACM SIGCOMM Comput. Commun. Rev.*, 43(4):363–374.

[83] Jain, M. and Dovrolis, C. (2002a). End-to-end available bandwidth: Measurement methodology, dynamics, and relation with TCP throughput. In *Proc. of ACM SIGCOMM*, pages 295–308.

[84] Jain, M. and Dovrolis, C. (2002b). Pathload: A measurement tool for end-to-end available bandwidth. In *Proc. of PAM*.

[85] Jiang, Y. (2010). A note on applying stochastic network calculus. Technical report.

[86] Jiang, Y. and Liu, Y. (2008). *Stochastic Network Calculus*. Springer-Verlag.

[87] Jiang, Y., Yin, Q., Liu, Y., and Jiang, S. (2009). Fundamental calculus on generalized stochastically bounded bursty traffic for communication networks. *Computer Networks*, 53(12):2011–2021.

[88] Karumanchi, A., Talabattula, S., Rao, K., and Varadarajan, S. (2004). A shift varying filtering theory for dynamic service guarantees. In *International Workshop on Quality of Service in Multiservice IP Networks*, pages 613–625. Springer.

[89] Knightly, E. W. (1997). Second moment resource allocation in multi-service networks. In *Proc. of ACM SIGMETRICS*, pages 181–191.

[90] Kreher, R. and Gaenger, K. (2010). *LTE signaling: troubleshooting and optimization*. John Wiley & Sons.

[91] Kumar, A., Manjunath, D., and Kuri, J. (2004). *Communication Networking: An Analytical Approach*. Elsevier.

[92] Kurose, J. (1992). On computing per-session performance bounds in high-speed multi-hop computer networks. In *Proc. of ACM SIGMETRICS*, pages 128–139.

[93] Laner, M., Svoboda, P., Romirer-Maierhofer, P., Nikaein, N., Ricciato, F., and Rupp, M. (2012). A comparison between one-way delays in operating HSPA and LTE networks. In *Proc. of IEEE WiOpt*, pages 286–292.

[94] Le Boudec, J.-Y. (1996). Network calculus made easy. Technical Report EPFL-DI 96/218, EPFL.

[95] Le Boudec, J.-Y. (1998). Application of network calculus to guaranteed service networks. *IEEE Trans. Inform. Theory*, 44(3):1087–1096.

[96] Le Boudec, J.-Y. and Thiran, P. (2001). *Network Calculus A Theory of Deterministic Queuing Systems for the Internet*. Springer-Verlag.

[97] Le Boudec, J.-Y. and Thiran, P. (2004). Network calculus a theory of deterministic queuing systems for the Internet. `http://ica1www.epfl.ch/PS_files/netCalBookv4.pdf`. extended and revised online version of [96].

[98] Li, C., Burchard, A., and Liebeherr, J. (2007). A network calculus with effective bandwidth. *IEEE/ACM Trans. Networking*, 15(6):1442–1453.

[99] Li, M., Claypool, M., and Kinicki, R. (2008). Wbest: A bandwidth estimation tool for ieee 802.11 wireless networks. In *2008 33rd IEEE Conference on Local Computer Networks (LCN)*, pages 374–381. IEEE.

[100] Liebeherr, J., Burchard, A., and Ciucu, F. (2012). Delay bounds in communication networks with heavy-tailed and self-similar traffic. *IEEE Trans. Inform. Theory*, 58(2):1010–1024.

[101] Liebeherr, J., Fidler, M., and Valaee, S. (2010). A system theoretic approach to bandwidth estimation. *IEEE/ACM Trans. Networking*, 18(4):1040–1053.

[102] Liebeherr, J., Wrege, D. E., and Ferrari, D. (1996). Exact admission control for networks with a bounded delay service. *IEEE/ACM Trans. Networking*, 4(6):885–901.

[103] Lübben, R. (2013). *System Identification of Computer Networks with Random Service*. PhD thesis, Leibniz Universität Hannover.

[104] Lübben, R., Fidler, M., and Liebeherr, J. (2011). A foundation for stochastic bandwidth estimation of networks with random service. In *IEEE INFOCOM*, pages 1817–1825.

[105] Lübben, R., Fidler, M., and Liebeherr, J. (2014). Stochastic bandwidth estimation in networks with random service. *IEEE/ACM Trans. Networking*, 22(2):484–497.

[106] Mao, S. and Panwar, S. S. (2006). A survey of envelope processes and their applications in quality of service provisioning. *IEEE Communications Surveys and Tutorials*, 8(3):2–20.

[107] McWilliams, B., Le Pezennec, Y., and Collins, G. (2012). HSPA+ (2100 MHz) vs LTE (2600 MHz) spectral efficiency and latency comparison. In *Proc. of IEEE Telecommunications Network Strategy and Planning Symposium*, pages 1–6.

[108] Melander, B., Björkman, M., and Gunningberg, P. (2000). A new end-to-end probing and analysis method for estimating bandwidth bottlenecks. In *Proc. of IEEE Globecom*, pages 415–420.

[109] Mellia, M., Stoica, I., and Zhang, H. (2002). TCP model for short lived flows. *IEEE Commun. Lett.*, 6(2):85–87.

[110] Michelinakis, F., Bui, N., Fioravantti, G., Widmer, J., Kaup, F., and Hausheer, D. (2016). Lightweight capacity measurements for mobile networks. *Computer Communications*, 84:73–83.

[111] Nam, S. Y., Kim, S., and Park, W. (2002). Analysis of minimal backlogging-based available bandwidth estimation mechanism. *Computer Communications*, 35(4):431–443.

[112] Nikolaus, P. and Schmitt, J. (2020). On the stochastic end-to-end delay analysis in sink trees under independent and dependent arrivals. In *International Conference on Measurement, Modelling and Evaluation of Computing Systems*, pages 136–154. Springer.

[113] Nikolaus, P., Schmitt, J., and Ciucu, F. (2019). Dealing with dependence in stochastic network calculus–using independence as a bound. In *International Conference on Analytical and Stochastic Modeling Techniques and Applications*, pages 71–84. Springer.

[114] Oshiba, T. and Nakajima, K. (2010). Quick end-to-end available bandwidth estimation for qos of real-time multimedia communication. In *The IEEE symposium on Computers and Communications*, pages 162–167. IEEE.

[115] Papoulis, A. and Pillai, S. U. (2002). *Probability, Random Variables and Stochastic Processes*. McGraw-Hill, forth edition.

[116] Perrucci, G. P., Fitzek, F. H., Sasso, G., Kellerer, W., and Widmer, J. (2009). On the impact of 2g and 3g network usage for mobile phones' battery life. In *Wireless Conference, 2009. EW 2009. European*, pages 255–259. IEEE.

[117] Phongtraychack, A. and Dolgaya, D. (2018). Evolution of mobile applications. In *MATEC Web of Conferences*, volume 155, page 01027. EDP Sciences.

[118] Poloczek, F. and Ciucu, F. (2014). Scheduling analysis with martingales. In *Performance Evaluation Special Issue IFIP Performance*, volume 79, pages 56–72.

[119] Ra, M.-R., Paek, J., Sharma, A. B., Govindan, R., Krieger, M. H., and Neely, M. J. (2010). Energy-delay tradeoffs in smartphone applications. In *Proceedings of the 8th International Conference on Mobile Systems, Applications, and Services*, MobiSys '10, pages 255–270, New York, NY, USA. ACM.

[120] Ribeiro, V., Riedi, R., Baraniuk, R., Navratil, J., and Cottrell, L. (2003). PathChirp: Efficient available bandwidth estimation for network paths. In *Proc. of PAM*.

[121] Rizk, A. and Fidler, M. (2008). On the identifiability of link service curves from end-host measurements. In *Proc. of Euro-NF NET-COOP*, volume 5425 of *LNCS*, pages 53–61.

[122] Rizk, A. and Fidler, M. (2012). Non-asymptotic end-to-end performance bounds for networks with long range dependent FBM cross traffic. *Computer Networks*, 56(1):127–141.

[123] Rockafellar, R. T. (1972). *Convex Analysis*. Princeton University Press.

[124] Romirer-Maierhofer, P., Ricciato, F., and Coluccia, A. (2008). Explorative analysis of one-way delays in a mobile 3g network. In *2008 16th IEEE Workshop on Local and Metropolitan Area Networks*, pages 73–78. IEEE.

[125] Ross, S. (2002). *A First Course in Probability*. Prentice Hall, 6 edition.

[126] Ross, S. M. (2007). *Introduction to Probability Models*. Academic Press, 9 edition.

[127] Sariowan, H., Cruz, R. L., and Polyzos, G. C. (1995). Scheduling for quality of service guarantees via service curves. In *Proc. of IEEE ICCCN*, pages 512–520.

[128] Schmitt, J. B., Zdarsky, F. A., and Fidler, M. (2008). Delay bounds under arbitrary multiplexing: When network calculus leaves you in the lurch ... In *Proc. of IEEE INFOCOM*, pages 1669–1677.

[129] Schwartz, C., Hossfeld, T., Lehrieder, F., and Tran-Gia, P. (2013a). Performance analysis of the trade-off between signalling load and power consumption for popular smartphone apps in 3g networks. *University of Würzburg, Tech. Rep*, 485.

[130] Schwartz, C., Lehrieder, F., Wamser, F., Hoßfeld, T., and Tran-Gia, P. (2013b). Smart-Phone Energy Consumption Vs. 3G Signaling Load: The Influence of Application Traffic Patterns. In *Proc. of IEEE TIWDC on Green ICT*.

[131] Sharma, P. (2013). Evolution of mobile wireless communication networks-1g to 5g as well as future prospective of next generation communication network. *International Journal of Computer Science and Mobile Computing*, 2(8):47–53.

[132] Shiobara, S. and Okamawari, T. (2017). A novel available bandwidth estimation method for mobile networks using a train of packet groups. In *Proceedings of the 11th International Conference on Ubiquitous Information Management and Communication*, pages 1–7.

[133] Shriram, A. and Kaur, J. (2007). Empirical evaluation of techniques for measuring available bandwidth. In *Proc. of IEEE INFOCOM*, pages 2162–2170.

[134] Siekkinen, M., Hoque, M. A., Nurminen, J. K., and Aalto, M. (2013). Streaming over 3G and LTE: How to Save Smartphone Energy in Radio Access Network-friendly Way. In *Proc. of ACM MoVid*, pages 13–18.

[135] Starobinski, D. and Sidi, M. (2000). Stochastically bounded burstiness for communication networks. *IEEE Trans. Inform. Theory*, 46(1):206–212.

[136] Strauss, J., Katabi, D., and Kaashoek, F. (2003). A measurement study of available bandwidth estimation tools. In *Proc. of ACM IMC*, pages 39–44.

[137] Tabassum, H., Siddique, U., Hossain, E., and Hossain, M. J. (2014). Downlink performance of cellular systems with base station sleeping, user association, and scheduling. *IEEE Trans. Wireless Commun.*, 13(10):5752–5767.

[138] Tirronen, T., Larmo, A., Sachs, J., Lindoff, B., and Wiberg, N. (2012). Reducing energy consumption of LTE devices for machine-to-machine communication. In *IEEE GLOBECOM*, pages 1650–1656.

[139] Undheim, A., Jiang, Y., and Emstad, P. J. (2007). Network calculus approach to router modeling with external measurements. In *Proc. of CHINACOM*, pages 276–280.

[140] Valaee, S. and Li, B. (2002). Distributed call admission control for ad hoc networks. In *Proc. of IEEE VTC*, pages 1244–1248.

[141] Wang, C.-Y., Logothetis, D., Trivedi, K. S., and Viniotis, I. (1996). Transient behavior of ATM networks under overloads. In *Proc. of IEEE INFOCOM*, pages 978–985.

[142] Wigard, J., Kolding, T., Dalsgaard, L., and Coletti, C. (2009). On the user performance of LTE UE power savings schemes with discontinuous reception in LTE. In *Proc. of IEEE ICC*, pages 1–5.

[143] Wu, D. O. and Negi, R. (2003). Effective capacity: A wireless link model for support of quality of service. *IEEE Trans. Wireless Commun.*, 2(4):630–643.

[144] Wu, J., Zhang, T., and Zeng, Z. (2013). Performance analysis of discontinuous reception mechanism with web traffic in lte networks. In *Personal Indoor and Mobile Radio Communications (PIMRC), 2013 IEEE 24th International Symposium on*, pages 1676–1681. IEEE.

[145] Wylie-Green, M. P. and Svensson, T. (2010). Throughput, capacity, handover and latency performance in a 3GPP LTE FDD field trial. In *Proc. of IEEE GLOBECOM*, pages 1–6.

[146] Xu, Y., Wang, Z., Leong, W. K., and Leong, B. (2014). An end-to-end measurement study of modern cellular data networks. In *International Conference on Passive and Active Network Measurement*, pages 34–45. Springer.

[147] Yang, S.-R. (2007). Dynamic power saving mechanism for 3g umts system. *Mobile Networks and Applications*, 12(1):5–14.

[148] Yang, S.-R., Yan, S.-Y., and Hung, H.-N. (2007). Modeling umts power saving with bursty packet data traffic. *IEEE Transactions on Mobile Computing*, 6(12).

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

[150] Yaron, O. and Sidi, M. (1994). Generalized processor sharing networks with exponentially bounded burstiness arrivals. *Journal of High Speed Networks*, 3:375–387.

[151] Yin, Q., Jiang, Y., Jiang, S., and Kong, P. Y. (2002). Analysis of generalized stochastically bounded bursty traffic for communication networks. In *Proc. of IEEE LCN*, pages 141–149.

[152] Zhang, H. and Knightly, E. (1994). Providing end-to-end statistical performance guarantees with interval dependent stochastic models. In *Proc. of ACM SIGMETRICS*, pages 211–220.

[153] Zhang, J. and Coyle, E. J. (1991). The transient solution of time-dependent M/M/1 queues. *IEEE Trans. Inform. Theory*, 37(6):1690–1696.

[154] Zhang, L., Okamawari, T., and Fujii, T. (2012). Performance Evaluation of End-to-End Communication Quality of LTE. In *Proc. of IEEE VTC*, pages 1–5.

[155] Zhang, S. and Costa, S. (2018). Mobile phone usage patterns, security concerns, and security practices of digital generation. *International Journal of Mobile Human Computer Interaction (IJMHCI)*, 10(1):23–39.

[156] Zhou, K., Nikaein, N., and Spyropoulos, T. (2013). Lte/lte-a discontinuous reception modeling for machine type communications. *IEEE Wireless Communications Letters*, 2(1):102–105.

[157] Zhou, L., Xu, H., Tian, H., Gao, Y., Du, L., and Chen, L. (2008). Performance analysis of power saving mechanism with adjustable drx cycles in 3gpp lte. In *Vehicular Technology Conference, 2008. VTC 2008-Fall. IEEE 68th*, pages 1–5. IEEE.

OWN PUBLICATIONS

[1] Becker, N. and Fidler, M. (2018). A non-stationary service curve model for estimation of cellular sleep scheduling. *IEEE Transactions on Mobile Computing*, 18(1):28–41.

[2] Becker, N. and Fidler, M. (2015). A non-stationary service curve model for performance analysis of transient phases. In *Proc. of ITC 27*, pages 116–124.

[3] Becker, N., Rizk, A., and Fidler, M. (2014). A measurement study on the application-level performance of LTE. In *Proc. of IFIP Networking*.

[4] Akselrod, M., Becker, N., Fidler, M., and Lübben, R. (2017). 4g lte on the road-what impacts download speeds most? In *2017 IEEE 86th Vehicular Technology Conference (VTC-Fall)*, pages 1–6. IEEE.

[5] Fidler, M., Lübben, R., and Becker, N. (2015). Capacity-Delay-Error-*Boundaries: A Composable Model of Sources and Systems*. *IEEE Trans. Wireless Commun.*, 14(3):1280–1294.

SCIENTIFIC CARREER

Curriculum Vitae	
Name	Nico Becker
Day of birth	15.08.1984
Education	
11/12 - 12/17	Research assistant and PhD student at the Institute of Communications Technology: *Leibniz Universität Hannover*
10/04 - 01/12	Studies of Mathematics (Dipl.-Math.): *Leibniz Universität Hannover*
10/07 - 09/08	Studies of Computational Mathematics with Modelling (M.Sc.): *Brunel University London*

COLOPHON

This document was typeset using the typographical look-and-feel `classicthesis` developed by André Miede. The style was inspired by Robert Bringhurst's seminal book on typography "*The Elements of Typographic Style*".

Hermann Zapf's *Palatino* and *Euler* type faces (Type 1 PostScript fonts *URW Palladio L* and *FPL*) are used. The "typewriter" text is typeset in *Bera Mono*, originally developed by Bitstream, Inc. as "Bitstream Vera". (Type 1 PostScript fonts were made available by Malte Rosenau and Ulrich Dirr.)

Final Version as of March 24, 2021 at 21:21.