# Capacity–Delay–Error Boundaries: A Composable Model of Sources and Systems

Markus Fidler, *Senior Member, IEEE*, Ralf Lübben, *Student Member, IEEE*, and Nico Becker

***Abstract*—This paper develops a notion of capacity–delay–error (CDE) boundaries as a performance model of networked sources and systems. The goal is to provision effective capacities that sustain certain statistical delay guarantees with a small probability of error. We use a stochastic non-equilibrium approach that models the variability of traffic and service to formalize the influence of delay constraints on the effective capacity. Permitting unbounded delays, known ergodic capacity results from information theory are recovered in the limit. We prove that the model has the property of additivity, which enables composing CDE boundaries obtained for sources and systems as if in isolation. A method for construction of CDE boundaries is devised based on moment-generating functions, which includes the large body of results from the theory of effective bandwidths. Solutions for essential sources, channels, and respective coders are derived, including Huffman coding, MPEG video, Rayleigh fading, and hybrid automatic repeat request. Results for tandem channels and for the composition of sources and channels are shown.**

***Index Terms*—Queueing analysis, information theory, channel models, time varying channels, quality of service.**

## I. Introduction

ORIGINATING from the seminal works by Shannon in 1948, the tremendous progress in information and coding theory has fostered ground-breaking applications that range from digital communications to data storage and processing. The fundamental results of information theory are asymptotic limits for the transmission rate of information by a source over a channel. Information theory defines the notion of entropy and channel capacity as the expected information of a source and the expected transinformation of a channel. Coding theory devises practical codes for data compression and reliable transmission that seek to approach the limits established by the entropy and the channel capacity, respectively [3].

In networking, information theory has not become widely accepted, yet. A major challenge for establishing a connection is due to the properties of network data traffic that is highly variable and delay-sensitive [4]. In contrast, information theory mostly neglects the dynamics of information and capacity and focuses on averages, respectively, asymptotic limits. Typically, these limits can be achieved with negligibly small probability of error assuming, however, arbitrarily long codewords and as a consequence arbitrarily large coding delays [5]. In networking though, delay is a key performance parameter that can be traded for capacity or loss as shown, e.g., by queueing theory. Further, the variability of sources is essential in packet data networks, as it potentiates significant resource savings due to statistical multiplexing [4].

The analytical cornerstone of networking is queueing theory that dates back to the works on the dimensioning of circuit-switched networks by Erlang in 1909 and 1917. In 1962 Kleinrock advanced this theory to prove the resource efficiency of packet-switching, which is achieved by bursty sources due to resource sharing. For packet-switched networks queueing theory can provide exact solutions for backlogs and delays. These occur due to the variability of packet inter-arrival and service times. Typically, the inter-arrival and service times obey a certain distribution by assumption, e.g., exponential. Recent approaches like the theory of effective bandwidths [6], [7], the deterministic network calculus [7]–[9] and the stochastic network calculus [7], [10]–[15] compute performance bounds for a wide range of stochastic processes. Despite the need, e.g., for joint coding and scheduling problems or cross-layer optimization in wireless multimedia networks, a tight link between these models and information theory has not been established yet [4], [5], [15].

To bridge the gap towards queuing theory, a non-equilibrium information theory that can model the variability and delay-sensitivity of real sources is required [4], [5]. While [4] envisions "effective bandwidth versus distortion functions," [5] proposes the idea of "throughput-delay-reliability-triplets" to characterize mobile ad-hoc networks. Rate-delay tradeoffs are investigated, e.g., in [16] for networks with multi-path routing and network coding. To provide a connection to queueing networks, [4], [5], [15] mention effective bandwidths, large deviations theory, or stochastic network calculus as potentially promising candidate theories. Further details are, however, omitted and the conclusion is that unifying information and queueing theory remains as one of the most important challenges in this field.

The variability of fading channels is already considered in [17], where a notion of outage capacity is defined. The outage capacity models the probability of errors that occur when the transmission rate is larger than the instantaneous capacity of the channel. A related concept, the delay-limited capacity [18], compensates fluctuations of the fading process using power

Manuscript received April 3, 2014; revised August 5, 2014; accepted October 13, 2014. Date of publication October 30, 2014; date of current version March 6, 2015. The work of R. Lübben was supported through a DFG grant (NetMeter) and the work of N. Becker was supported through an ERC Starting Grant (UnIQue). Portions of this work were presented at IEEE WoWMoM 2012 and the IEEE Sarnoff Symposium 2012. The associate editor coordinating the review of this paper and approving it for publication was L. Song.

The authors are with the Institute of Communications Technology, Department of Electrical Engineering and Computer Science, Leibniz Universität Hannover, Hanover 30167, Germany.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TWC.2014.2365782

control to achieve a constant transmission rate. Subsequent works implement power control subject to additional buffering constraints [19], [20]. Recently, the impact of finite block-length codes on the variability of the capacity is investigated in [21]–[23].

While the definition of outage capacity does not contain any queueing-theoretic aspects, it can be incorporated into a queueing analysis, as shown in [24] using the $\mathrm{M}|\mathrm{G}|1$ model. Markovian queues have been parameterized to model fading channels also in [25] and [26]. While [25] models a block fading process by a variable rate server that is governed by an embedded Markov chain, [26] views fading outages as an impairment process that is described by high priority customers at an $\mathrm{M}|\mathrm{G}|1$ priority queue. The concept of an impairment process was already introduced to the stochastic network calculus in [10] to model outages of a channel, e.g., due to noise, interference, or fading. The impairment model is used in [15] to analyze outages of wireless channels and in [27] to derive a statistical service curve for Markov channels.

Statistical service curves, e.g., [10], [12], [13], are used in the network calculus to specify the input-output-relation of systems, such as links, buffers, and schedulers. A first application of service curves to describe fading channels is presented in [28]. The authors use the MGF, respectively, the Laplace transform of a random process that denotes the service of the channel. The model is named effective capacity as it is dual to the characterization of sources by effective bandwidths [6], [7]. Using arguments from large deviations theory, the effective capacity provides a tail asymptote of the delay $W$ for $W \to \infty$ that arises if constant bit rate traffic is transmitted over a random channel [28]. Effective capacities have subsequently been derived for a variety of systems, e.g., for cognitive radio [29], multi-antenna [30], Rayleigh fading [31], and Nakagami-$m$ fading channels [32].

MGFs are also frequently used in the stochastic network calculus [7], [12], [14]. Compared to the effective capacity, the service curve model of the network calculus provides non-asymptotic results, includes variable bit rate traffic, and extends to multi-hop networks. Related works in the stochastic network calculus use Markov chains to reproduce the state of Gilbert-Elliott fading channels with memory [33], to model Nakagami-$m$ fading channels [34], and to analyze delays of spatial multiplexing MIMO channels [35]. Memoryless Rayleigh fading channels in tandem are investigated in [36] using a two step approach that expresses the MGF as a Mellin transform of an exponential function. The rationale is that exponentiation takes the channel capacity of an additive white Gaussian noise (AWGN) channel from the bit domain to a signal-to-noise ratio (SNR) domain [36].

Regarding traffic sources, networking research typically either assumes certain stochastic processes or employs traffic traces. Methods for construction of an empirical effective bandwidth or an empirical envelope function from a traffic trace, e.g., of MPEG video, are given in [37] and [38], respectively. Recent papers [39], [40] provide an analytical framework for network elements that process and re-scale data, such as a transcoder. While information theoretic concepts are not used in these works, they may facilitate such applications.

An example is [41], where a binary symmetric channel with automatic repeat request is modeled as a network element that re-scales the amount of data to account for retransmissions. Discrete memoryless channels are also investigated in [42], where an error server is defined, that counts the cumulative number of bit errors. Another link between the stochastic network calculus and information theory is established in [43], where a calculus for so-called information-driven networks is introduced. To this end, the authors use the entropy to convert functions of data bits into functions of information bits.

In this paper, we investigate a potentially promising connection of methods from stochastic network calculus with information theory. We develop a notion of capacity-delay-error (CDE)-boundaries for model-based performance evaluation [44] of sources and systems. We use CDE-boundaries as a mathematical model to represent the capacity that is effectively required by a source, respectively, offered by a system under certain maximal delay constraints. Delays are guaranteed with a defined, small probability of error. Compared to information theory, we use a non-equilibrium approach to consider the variability of sources and systems on finite time-scales. In the limit, we recover known ergodic capacity results. We use techniques from convex optimization to prove that our CDE-boundaries are additive. Like the source-channel separation theorem from information theory, the property of additivity enables us to compose results obtained for sources and systems as if in isolation. Tandem systems are expressed as an equivalent single system using properties of the network calculus.

We derive a method for construction of CDE-boundaries from moment generating functions (MGFs). Hence, our model comprises the large body of MGF results provided by the theory of effective bandwidths [6], [7] and the related concept of effective capacity [28]. We derive CDE-boundaries for sources and channels with and without memory including, e.g., Huffman video coding and Rayleigh fading channels. We quantify the increase of delays due to memory and show how error correction codes can be applied in hybrid automatic repeat request (ARQ) systems to combat this effect. We include composition results for fading channels in tandem, and for the transmission of an MPEG source via a fading channel. We expect that our results enable further joint information- and queueing-theoretical investigations that have the potential to provide substantial new insights and applications from a holistic analysis of networked communications systems.

The remainder of this paper is structured as follows. In Section II, we establish the notion of CDE-boundaries and prove important properties of the model. We derive a method for construction of CDE-boundaries and present results for different sources in Section III and systems in Section IV. In Section V, we show results for composition of sources, systems, and tandem systems. In Section VI, we provide brief conclusions.

## II. Capacity-Delay-Error-Boundaries

In this section, we establish capacity-delay-error (CDE)-boundaries as a performance model of sources and systems. We base our definition on a general queueing model, where

data arrivals from sources and the service offered by systems, respectively, are random processes, see Section II-A. In Section II-B, we characterize the arrival and service processes by Legendre transforms. This approach enables us to prove the additivity of CDE-boundaries of sources and systems. This important result facilitates an analysis of networked sources and systems as if in isolation. Finally, in Section II-C, we connect CDE-boundaries to other well-known models: for sources, we make use of a known link between the notion of statistical envelopes and effective bandwidths; for systems, we establish a connection to the effective capacity model.

### A. Queueing Model and Envelope Processes

We employ a queueing model that is defined in the framework of the stochastic network calculus, see [10], [13] and for a broader overview also [15], [45]–[47]. The stochastic network calculus enables the computation of statistical performance bounds of the type $\mathsf{P}[\text{backlog} > y] \leq \varepsilon$ or $\mathsf{P}[\text{delay} > y] \leq \varepsilon$, meaning that the stationary backlog, respectively, delay exceeds a defined threshold $y$ at most with a small probability $\varepsilon$, e.g., $\varepsilon = 10^{-6}$. To this end, it makes use of statistical arrival envelopes and statistical service curves that are statistical bounding functions of the arrivals and the service, respectively.

Denote $A(t)$ the cumulative arrivals of a system, i.e., the cumulative number of bits generated by a source in the time interval $(0,t]$. By convention, there are no arrivals for $t \leq 0$. In the sequel, time $t$ is non-negative and can be either continuous $t \in \mathbb{R}_0^+$ or discrete $t \in \mathbb{N}_0$. By definition $A(t)$ is a non-negative and non-decreasing random process that passes through the origin, i.e., $A(t) \in \mathcal{F}_0$ where $\mathcal{F}_0 = \{f : f(t) \geq f(\tau) \geq 0, \forall\, t \geq \tau \geq 0, f(0) = 0\}$. We use shorthand $A(\tau,t) = A(t) - A(\tau)$ for $t \geq \tau \geq 0$ to describe the cumulative number of bits in $(\tau,t]$. Similarly, the cumulative departures from a system are denoted $D(t)$ where $D(t) \in \mathcal{F}_0$. Besides $\mathcal{F}_0$, we will use $\mathcal{F} = \{f : f(t) \geq f(\tau), \forall\, t \geq \tau \geq 0\}$ that is the set of non-decreasing functions.

We build our work on a basic queueing model that expresses the service offered by a system as a random process $S(\tau,t)$. It defines a lower bound for the departures of the type [7]

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

for all $t \geq 0$. The operator $\otimes$ is referred to as the min-plus convolution.[1] In this work, $S(\tau,t)$ is non-negative, non-increasing in $\tau$, non-decreasing in $t$, and $S(t,t) = 0$ for all $t \geq 0$. Examples of (1) are a work-conserving link with a time varying capacity, where $S(\tau,t)$ is the service that is available in $(\tau,t]$ [7]; the service $S_{\mathrm{lo}}(\tau,t)$ that is left over after scheduling cross traffic at a channel

$$S_{\mathrm{lo}}(\tau,t) = \max\{0, S(\tau,t) - A_{\mathrm{cr}}(\tau,t)\}, \tag{2}$$

where $S(\tau,t)$ is the entire service of the channel and $A_{\mathrm{cr}}(t)$ denotes cross traffic [14]; or the end-to-end service $S_{\mathrm{net}}(\tau,t)$ of a network of $N$ tandem systems[2] where

$$S_{\mathrm{net}}(\tau,t) = S_1 \otimes S_2 \otimes \cdots \otimes S_N(\tau,t) \tag{3}$$

and $S_i(\tau,t)$ are the service processes of the individual systems $i \in [1,2,\ldots,N]$ [7].

Statistical arrival envelopes [10], [13] are used to specify an upper bound of the arrival process, that may be exceeded at most with a defined probability $\varepsilon_A \in [0,1]$. The arrival process has an upper envelope $E_A(t) \in \mathcal{F}$ with overflow profile[3] $\sigma_A(\varepsilon_A) \geq 0$ if for all $t \geq 0$ it holds that [10], [13]

$$\mathsf{P}[\exists \tau \in [0,t] : A(\tau,t) > E_A(t-\tau) + \sigma_A(\varepsilon_A)] \leq \varepsilon_A. \tag{4}$$

Note that (4) specifies a guarantee for entire sample paths of $A(\tau,t)$, i.e., all $\tau \in [0,t]$ are tested.

Similarly, we define statistical service envelopes to specify a lower bound of the service process with underflow probability $\varepsilon_S \in [0,1]$. The service process has lower envelope $E_S(t) \in \mathcal{F}$ with deficit profile $\sigma_S(\varepsilon_S) \geq 0$ if for all $t \geq 0$ it holds that[4]

$$\mathsf{P}[\exists \tau \in [0,t] : S(\tau,t) < E_S(t-\tau) - \sigma_S(\varepsilon_S)] \leq \varepsilon_S. \tag{5}$$

Using the definition of arrival and service envelopes, statistical performance bounds follow readily. We show that backlog and delay bounds can be computed as the maximal vertical and horizontal deviation of $E_A(t)$ and $E_S(t)$, respectively. The backlog at time $t$ is defined as $B(t) = A(t) - D(t)$ and by insertion of (1) $B(t) \leq \sup_{\tau \in [0,t]}\{A(\tau,t) - S(\tau,t)\}$. By substitution of the envelopes $E_A(t-\tau) + \sigma_A(\varepsilon_A)$ from (4) as an upper bound for $A(\tau,t)$ and $E_S(t-\tau) - \sigma_S(\varepsilon_S)$ from (5) as a lower bound for $S(\tau,t)$ it follows that

$$b = \sup_{t \geq 0}\{E_A(t) - E_S(t)\} + \sigma_A(\varepsilon_A) + \sigma_S(\varepsilon_S) \tag{6}$$

is a statistical backlog bound. Since the envelopes (4) and(5) may be violated with probability $\varepsilon_A$ and $\varepsilon_S$, respectively, it holds that the backlog bound is exceeded at most with probability $\mathsf{P}[B(t) > b] \leq \varepsilon_A + \varepsilon_S$ for all $t \geq 0$. Similarly, under the assumption of first-come first-serve (fcfs) order, the delay at time $t$ is defined as $W(t) = \inf\{\tau \geq 0 : A(t) - D(t+\tau) \leq 0\}$. Using the same basic steps it follows that

$$d = \inf\left\{\tau \geq 0 : \sup_{t \geq 0}\{E_A(t) - E_S(t+\tau)\} + \sigma_A(\varepsilon_A) + \sigma_S(\varepsilon_S) \leq 0\right\} \tag{7}$$

is a statistical delay bound with probability $\mathsf{P}[W(t) > d] \leq \varepsilon_A + \varepsilon_S$ for all $t \geq 0$. We refer to $\varepsilon = \varepsilon_A + \varepsilon_S$ as the probability of error. Since $\varepsilon_A$ and $\varepsilon_S$ are free parameters, we can

[1] More commonly known is the min-plus convolution of univariate functions $f(t), g(t)$ that is defined as $f \otimes g(t) := \inf_{\tau \in [0,t]}\{f(\tau) + g(t-\tau)\}$. We use an accordingly extended definition for bivariate functions [7]. Note that min-plus convolution is commutative in case of univariate functions but not in case of bivariate functions.

[2] The composition of tandem systems follows immediately by recursive insertion of (1) and by the associativity of min-plus convolution.

[3] Compared to [10], [13] that define $\varepsilon_A(\sigma_A)$ as a function of $\sigma_A$, we use the inverse $\sigma_A(\varepsilon_A)$. We make this minor change to use $\varepsilon_A$ as a free parameter.

[4] We note that since $S(\tau,t)$ is non-negative, a stronger envelope follows as $\mathsf{P}[\exists \tau \in [0,t] : S(\tau,t) < \max\{0, E_S(t-\tau) - \sigma_S(\varepsilon_S)\}] \leq \varepsilon_S$, that can be used at the expense of additional notation.

fix $\varepsilon$ and choose $\varepsilon_A$ and $\varepsilon_S$ that minimize $\sigma_A(\varepsilon_A) + \sigma_S(\varepsilon_S)$ to derive a minimal backlog and delay bound from (6) and (7), respectively.

The definitions of arrival envelope (4) and service envelope (5) serve as the basis of CDE-boundaries that will be derived in the next section. We conclude this section with a comparison to the related work. Instead of the definition of a random service process $S(\tau, t)$ in (1), the literature on the stochastic network calculus frequently employs a queueing model, where the service is expressed by a statistical service curve. Statistical service curves $\mathcal{S}(t)$ are defined as deterministic functions [48] that give a guarantee of the type [10], [13]

$$\mathsf{P}\left[D(t) < A \otimes \mathcal{S}(t) - \sigma_S(\varepsilon_S)\right] \leq \varepsilon_S \tag{8}$$

for all $t \geq 0$, i.e., statistical service curves provide a lower bound for the departures that may be violated at most with probability $\varepsilon_S \in [0, 1]$. Using the definition of statistical service curve (8) and arrival envelope (4), backlog and delay bounds follow as their maximal vertical and horizontal deviation, respectively. A connection of statistical service curves to random service processes is made in [15], [49], where it is observed that the envelope (5) of a service process (1) also satisfies the definition of statistical service curve (8). In the sequel, we will assume the concept of random service processes as a starting point and derive envelopes thereof. The reason is, that we can readily specify service processes that satisfy (1) for a number of basic, work-conserving systems.

### *B. Definition and Properties of CDE-Boundaries*

In this section, we use Legendre transforms to derive a notion of CDE-boundaries from envelope functions. The model specifies operating points in the capacity-delay-error space that are achievable by a source or a system, respectively. We prove that the model is additive. The additivity is highly useful as it allows composing results that are obtained for sources and systems independently.

*Definition 1 (CDE-Boundaries):* Given an arrival process $A(t)$. We define its $(c, d_A, \varepsilon_A)$-boundary as delay bound $d_A$ with error probability $\varepsilon_A$ if the arrivals are transmitted by a constant rate server with capacity $c$.

Given a service process $S(\tau, t)$. We define its $(c, d_S, \varepsilon_S)$-boundary as delay bound $d_S$ with error probability $\varepsilon_S$ if constant rate arrivals with rate $c$ are transmitted by the server.

To derive relevant properties of the model, we will benefit from the qualities of the Legendre transform of envelope functions. The concave and convex Legendre transforms of $E_A(t)$ and $E_S(t)$, respectively, are defined for $c \geq 0$ as[5]

$$\overline{\mathcal{L}}_A(c) := \sup_{t \geq 0} \left\{E_A(t) - ct\right\}, \tag{9}$$

$$\underline{\mathcal{L}}_S(c) := \sup_{t \geq 0} \left\{ct - E_S(t)\right\}. \tag{10}$$

[5]The Legendre transform is also referred to as Fenchel conjugate [50]. Strictly speaking, the concave conjugate is defined as $\inf_{t \geq 0}\{ct - f(t)\} = -\sup_{t \geq 0}\{f(t) - ct\}$. We slightly adapt the definition for ease of exposition.

The role of Legendre transforms in the min-plus systems theory of the deterministic network calculus has been elaborated, e.g., in [51]. Here, we use Legendre transforms to formulate performance bounds for stochastic arrival processes, service processes, and their composition, respectively.

First, we consider the arrival process $A(t)$ at a work-conserving constant rate server with rate $c > 0$ to derive the $(c, d_A, \varepsilon_A)$-boundary of the arrivals. The constant rate server has service process $S(\tau, t) = c(t - \tau)$ with service envelope $E_S(t) = ct$ and parameter $\sigma_S(\varepsilon_S) = 0$ for all $\varepsilon_S \geq 0$. By insertion into (6), a backlog bound follows as $b_A = \sup_{t \geq 0}\{E_A(t) - ct\} + \sigma_A(\varepsilon_A)$ with error probability $\varepsilon_A$. Since $\sup_{t \geq 0}\{E_A(t) - ct\} = \overline{\mathcal{L}}_A(c)$ it becomes evident that the Legendre transform $\overline{\mathcal{L}}_A(c)$ has the interpretation of a backlog bound. Assuming fcfs order, a delay bound follows in the same way from (7), so that

$$b_A = \overline{\mathcal{L}}_A(c) + \sigma_A(\varepsilon_A), \quad d_A = \frac{\overline{\mathcal{L}}_A(c) + \sigma_A(\varepsilon_A)}{c} \tag{11}$$

defines a $(c, d_A, \varepsilon_A)$-boundary of the source. Conversely, considering the service process $S(\tau, t)$ with constant rate arrivals with rate $c > 0$, a backlog bound and a delay bound with error probability $\varepsilon_S$, respectively, are

$$b_S = \underline{\mathcal{L}}_S(c) + \sigma_S(\varepsilon_S), \quad d_S = \frac{\underline{\mathcal{L}}_S(c) + \sigma_S(\varepsilon_S)}{c}. \tag{12}$$

The bounds (11) and (12) are a dual model of arrival envelopes (4) and service envelopes (5), respectively. The duality is due to the fact that for concave, respectively, convex functions the Legendre transform is its own inverse such that (11) and (12) uniquely determine concave arrival envelopes and convex service envelopes, respectively. While (11) and (12) consider stochastic sources and systems in isolation, we provide a fundamental result for their composition in the following theorem.

*Theorem 1 (Additivity):* Given an arrival process $A(t)$ at a system with service process $S(\tau, t)$. A backlog bound and, assuming fcfs order, a delay bound are

$$b = b_A + b_S, \quad d = d_A + d_S,$$

where $b_A, d_A$ and $b_S, d_S$ are given by (11) and (12), respectively. The error probability of $b$ and $d$ is $\varepsilon = \varepsilon_A + \varepsilon_S$ and $c > 0$ is a free parameter.

If $E_A(t)$ is concave and $E_S(t)$ convex, the minimal backlog bound $b = \inf_{c>0}\{b_A + b_S\}$ and the minimal delay bound $d = \inf_{c>0}\{d_A + d_S\}$ are identical to (6) and (7), respectively.

The proof is provided in Appendix A. We highlight that the performance bounds from Thm. 1 consist of two independent terms that are due to the variability of the source and of the system, respectively. As an important consequence, sources and systems can be analyzed as if in isolation and can be composed afterwards by addition of their individual backlog and delay bounds. Hence, with Thm. 1 we obtain a composable model of sources and systems that are each characterized by their CDE-boundaries $(c, d_A, \varepsilon_A)$ and $(c, d_S, \varepsilon_S)$. Generally, the bounds from Thm. 1 are conservative. Further, for the prevalent case of concave and convex envelope functions, the second part of

Thm. 1 proves that the bounds are as tight as corresponding bounds from the stochastic network calculus, i.e., (6) and (7).

While in this subsection, the overflow profile $\sigma_A$ is a general function of $\varepsilon_A$, we will next provide a method for construction of CDE-boundaries for the class of exponentially bounded arrival processes, where $\varepsilon_A = e^{-\theta\sigma_A}$, respectively, $\sigma_A = -\ln\varepsilon_A/\theta$ for any $\theta > 0$ and $\varepsilon_A \in (0,1]$. For service processes a corresponding deficit profile is derived.

### C. Construction of CDE-Boundaries

In this section, we show how to construct CDE-boundaries from the well-known effective bandwidth model of traffic sources [6], [7] and the effective capacity model of transmission systems [28], respectively. Other methods for construction of CDE-boundaries may be developed, e.g., using service envelopes (5) that are estimated from measurements [49].

The effective bandwidth and the effective capacity are basically moment generating functions (MGFs), respectively, Laplace transforms. Despite their similarity, the two models are essentially researched independently, i.e., a significant body of literature focuses either on deriving the effective bandwidth of sources, e.g., Markovian, On-Off, and fractional Brownian motion, or the effective capacity of wireless systems, e.g., fading channels. The composition of statistically independent sources and systems that are characterized by MGFs has mainly been considered in the stochastic network calculus, see e.g., [7], [14], [33], [52]. Compared to these works, Thm.1 contributes a much simpler rule for composition that creates a strong connection between the effective bandwidth traffic model and the effective capacity channel model without requiring statistical independence.

We use a discrete time[6] model $t \in \mathbb{N}_0$ and assume stationary arrivals, i.e., $\mathsf{P}[A(\tau,\tau+t) > y] = \mathsf{P}[A(t) > y]$ for any $y$ and all $\tau, t \geq 0$. The MGF of the arrivals $A(t)$ is defined as

$$\mathsf{M}_A(\theta,t) = \mathsf{E}\left[e^{\theta A(t)}\right], \tag{13}$$

where $\theta$ is a free parameter. For $\theta < 0$, (13) is also referred to as Laplace transform. MGFs are a convenient model of traffic multiplexing since the MGF of the sum of independent random processes $A(t) = A_1(t) + A_2(t)$ follows immediately as $\mathsf{M}_A(\theta,t) = \mathsf{M}_{A_1}(\theta,t)\mathsf{M}_{A_2}(\theta,t)$.

The effective bandwidth of a source is defined for $\theta > 0$ and $t > 0$ as the normalized log MGF [6], [7]

$$\alpha_A(\theta,t) = \frac{1}{\theta t}\ln \mathsf{M}_A(\theta,t). \tag{14}$$

The effective bandwidth increases in $\theta$ from the mean rate of the arrivals in an interval of length $t$ to their peak rate. The parameter $\theta$ is also referred to as the quality of service exponent since the effective bandwidth provides an estimate of the capacity requirements of the source with respect to quality of service constraints. For notational convenience, we define $\alpha_A(\theta,t)$ for $t = 0$ to be a finite constant, so that $t\alpha_A(\theta,t) = \ln \mathsf{M}_A(\theta,t)/\theta = 0$ for $t = 0$.

[6]In continuous time an additional discretization step is required, see [13].

The notion of effective capacity is conceived in [28] as a dual model of the effective bandwidth. Assuming a stationary service process $S(\tau,t)$ with Laplace transform $\mathsf{M}_S(-\theta,t-\tau)$, the effective capacity is defined as $\alpha_S(-\theta,t) = 1/(-\theta t)\ln \mathsf{M}_S(-\theta,t)$ for $\theta > 0$ and $t > 0$. The effective capacity decreases with increasing $\theta$ from the mean service rate to the minimum service rate. It can be interpreted as the maximal data rate a system can support with respect to the quality of service exponent. Again, we define $\alpha_S(-\theta,t)$ for $t = 0$ to be a finite constant.

While MGFs of arrival processes are a convenient model for statistical multiplexing, Laplace transforms of service processes have been effectively used in the network calculus to analyze tandem systems with cross-traffic. The end-to-end service process $S_{\mathrm{net}}(\tau,t)$ of a tandem of $N$ systems each with service process $S_i(\tau,t)$ for $i \in [1,2,\ldots,N]$ is given by min-plus convolution (3). Assuming statistically independent service processes, a bound[7] for the Laplace transform of $S_{\mathrm{net}}(\tau,t)$ for $\theta \geq 0$ and $t \geq 0$ is [14]

$$\mathsf{M}_{S_{\mathrm{net}}}(-\theta,t) \leq \mathsf{M}_{S_1} * \mathsf{M}_{S_2} * \cdots * \mathsf{M}_{S_N}(-\theta,t), \tag{15}$$

where $*$ denotes the convolution in the conventional algebra.[8] The individual systems can, e.g., be described by left over service processes $S_{i,\mathrm{lo}}$ according to (2) where $\mathsf{M}_{S_{i,\mathrm{lo}}}(-\theta,t) \leq \min\{1, \mathsf{M}_{S_i}(-\theta,t)\mathsf{M}_{A_{i,\mathrm{cr}}}(\theta,t)\}$ [14].

Statistical bounds or envelopes can be derived from the effective bandwidth, respectively, the MGF of the arrivals by use of Chernoff's bound. Examples are the exponentially bounded burstiness model [53] as well as effective envelopes [12]. Dual to the exponentially bounded burstiness traffic model, a service model referred to as exponentially bounded fluctuation is developed in [54]. The tightness of these envelopes depends on Chernoff's bound and on the union bound, that is used to evaluate sample paths as defined, e.g., in (4). For a comparison with simulation results see [33] and [55].

Using the connection of CDE-boundaries with envelope functions established in Section II-B, the following theorem derives CDE-boundaries for sources and systems from the effective bandwidth and the effective capacity, respectively.

*Theorem 2 (CDE-Boundaries of Sources and Systems):* Given a stationary arrival process $A(\tau,t)$ with effective bandwidth $\alpha_A(\theta,t)$. A delay bound $d_A$ from (11) is

$$d_A = \frac{\sup_{t\geq 0}\{(\alpha_A(\theta,t) + \rho - c)\,t\}}{c} - \frac{\ln(\theta\rho\varepsilon_A)}{\theta c},$$

where $\varepsilon_A \in (0,1]$ is the error probability, $c > 0$ is the service rate, and $\theta > 0$ and $\rho \in (0, 1/(\theta\varepsilon_A)]$ are free parameters.

Given a system that satisfies (1) with stationary service process $S(\tau,t)$ and effective capacity $\alpha_S(-\theta,t)$. A delay bound $d_S$ from (12) is

$$d_S = \frac{\sup_{t\geq 0}\{(c + \rho - \alpha_S(-\theta,t))\,t\}}{c} - \frac{\ln(\theta\rho\varepsilon_S)}{\theta c},$$

[7]We note that the Laplace transform reverses the order of inequalities such that the upper bound in (15) corresponds to a lower bound of the end-to-end service process. By definition of service processes (1) a lower bound of a service process is also a valid service process.

[8]The conventional convolution is defined for discrete time processes as $f * g(t) = \sum_{\tau=0}^{t} f(\tau)g(t-\tau)$.

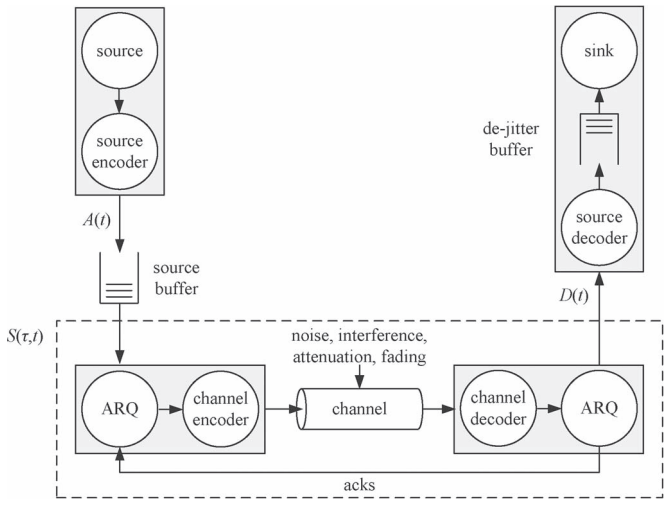


Fig. 1. Joint information- and queueing-theoretic model. A source generates symbols that are encoded and stored in a source buffer, denoted arrivals $A(t)$. A system comprising of an ARQ protocol and a channel coder transmits the arrivals over an error-prone channel. The transmission system including the channel is characterized by a service process $S(\tau, t)$. The departures of the system $D(t)$ are decoded, de-jittered in a buffer, and delivered to the sink.

where $\varepsilon_S \in (0, 1]$ is the error probability, $c > 0$ is the arrival rate, and $\theta > 0$ and $\rho \in (0, 1/(\theta\varepsilon_S)]$ are free parameters.

The proof is provided in Appendix B. Corresponding backlog bounds follow directly as $b_A = cd_A$ and $b_S = cd_S$. Thm. 2 specifies families of functions $d_A$ and $d_S$ each with free parameters $\theta$ and $\rho$. Since any of these functions provides an upper delay bound, a minimal delay bound can be derived by optimization of $\theta$ and $\rho$. Finally, Thm. 1 states that $d = d_A + d_S$ is a delay bound for the composition of the source and the system for all $c > 0$. We note that due to the properties of the $(\min, +)$-algebra, the minimization of $d$ does not depend on the order of optimizing the free parameters.

## III. Sources and Source Coding

In this section, we employ CDE-boundaries to investigate the performance of networked sources. An example is shown in Fig. 1, where a source generates symbols that are encoded and transmitted by a system that offers a random service, e.g., due to a fading channel. Our aim is to combine information- and queueing-theoretic aspects to identify achievable operating points within the capacity-delay-error-space. For example, given the arrival process of the source and the service process of the system, we seek to answer the question can source and/or channel coding achieve a target delay bound $d$ with probability of error $\varepsilon$? The result enables, e.g., the choice of suitable codes or the dimensioning of the playout delay and the de-jitter buffer at the receiver. Due to the additivity established by Thm. 1, we can analyze sources (Section III) and systems (Section IV) as if in isolation and compose the results afterwards (Section V). This separation enables us to consider the gains of source coding and channel coding independently.

After specifying the necessary notation, we investigate CDE-boundaries of different types of sources and source coders. Consider a random variable $X$ that can take any of the values $x_i$, $i \in \mathbb{N}_0$ with probability $p_i$. We also refer to $x_i$ as the symbols and $\mathbb{X} = \{x_0, x_1, \ldots\}$ as the alphabet of the source and denote $|\mathbb{X}|$ its cardinality. Information theory defines that if the event $X = x_i$ occurs, it provides information $\mathsf{I}(x_i) = -\mathrm{ld}\, p_i$ bit, where ld is the logarithm dualis (base 2). The expected information of a source becomes $\mathsf{H}_X := -\sum_i p_i \mathrm{ld}\, p_i$ that is defined as the entropy of $X$.

Shannon established the entropy of a source as a fundamental limit for lossless data compression. To this end, a code expressed as a function $l$ maps symbols $x_i$ to unique codewords of length $l_i$. The compression gain is due to assigning short codewords to frequent symbols. If no codeword is a prefix of any other codeword, the code is referred to as a prefix-free code, where each codeword can be decoded on its own [3]. For an optimal code the expected codeword length $\bar{l} = \sum_i p_i l_i$ is bounded in an interval of one bit width by the entropy as $\mathsf{H}_X \leq \bar{l} < \mathsf{H}_X + 1$ [3]. Without loss of generality, we restrict our investigation to binary codes.

We label successive symbols generated by a discrete source by $n \in \mathbb{N}$ to obtain the random process $X(n)$. By encoding $X(n)$ it follows that $L(n) = l(X(n))$ is a random process of codeword lengths. As $L(n)$ specifies increments, we obtain the cumulative arrival process as $A(n) = \sum_{\nu=1}^{n} L(\nu)$ for $n \geq 1$. We let $A(0) = 0$ by definition.

In the following, we explore the non-equilibrium behavior of encoded sources. We provide analytical solutions for memoryless sources and Huffman coding. We also consider Markov sources, which we parameterize using empirical data from video compression. Further results for a variety of other coders can also be found in the technical report [27].

### A. Memoryless Sources

We start our investigation, with the basic memoryless source, where the $X(n)$ for $n \in \mathbb{N}$ are independent and identically distributed (iid). We use function $l$ to assign a codeword $l_i$ to each symbol $x_i$. By definition, the increment process $L(n) = l(X(n))$ has categorical distribution with MGF

$$\mathsf{M}_L(\theta) = \sum_i p_i e^{\theta l_i}. \tag{16}$$

For the cumulative arrival process $A(n) = \sum_{\nu=1}^{n} L(\nu)$ it follows that $\mathsf{M}_A(\theta, n) = (\mathsf{M}_L(\theta))^n$ is multinomial. Assuming a source that emits symbols at a constant rate of one symbol per time-slot we substitute $n = t$. We will show how to relax this assumption later. From (14) we obtain the effective bandwidth

$$\alpha_A(\theta) = \frac{1}{\theta} \ln \left( \sum_i p_i e^{\theta l_i} \right) \tag{17}$$

for $\theta > 0$. Due to the memorylessness of the source, the effective bandwidth does not depend on $t$. By insertion of (17) into Thm. 2, we derive a delay bound $d_A$. Since $\alpha_A(\theta)$ is independent of $t$, the condition $c > \alpha_A(\theta)$ is sufficient to compute a finite bound. We choose the free parameter $\rho \in (0, 1/(\theta\varepsilon_A)]$ as $\rho = c - \alpha_A(\theta)$, where $c > \alpha_A(\theta)$ for stability. It follows that

$$d_A = \frac{-\ln\left(\theta\left(c - \alpha_A(\theta)\right)\varepsilon_A\right)}{\theta c} \tag{18}$$

for any $\theta > 0$ that satisfies $\theta \leq 1/((c - \alpha_A(\theta))\varepsilon_A)$. In the limit $\theta \to 0$, the effective bandwidth approaches the average codeword length, i.e., $\alpha_A(\theta) \to \bar{l}$. Hence, if arbitrarily large delays are permitted, (18) recovers that a capacity of $\bar{l}$ bits per time-slot is sufficient.

*1) Entropy Coding:* Source coding can achieve a compression gain by using codewords of nonuniform lengths, where short codewords are assigned to frequent symbols. Here, we use Huffman coding to assign codewords to symbols. Results for other codes such as Shannon or Lempel-Ziv coding are provided in the technical report [27]. For completeness, we briefly recapitulate the Huffman algorithm. To construct a Huffman code [3], execute the following steps repeatedly until all symbols have been processed:

- sort the symbols of the alphabet in decreasing order of probability,
- substitute the two least probable symbols by a new compound symbol; assign the sum of the two probabilities; and add one bit to the respective codewords to distinguish the two individual symbols.

The Huffman prefix code achieves the minimal expected codeword length, hence $\mathsf{H}_X \leq \bar{l} < \mathsf{H}_X + 1$. Regarding the individual codeword lengths $l_i$, however, no such simple upper bound exists. In fact, it is shown in [56] that individual codewords of a Huffman code can become as large as approximately 1.45 times the information of the corresponding symbol, i.e., $l_i < -1.45 \mathrm{ld}\, p_i$. As a consequence, the length of actual codewords generated by a Huffman coder for a sequence of symbols may be significantly larger than the entropy.

For an analytical investigation, assume a source with an infinite alphabet with geometrically distributed symbols $p_i = p(1-p)^i$ for $i \in \mathbb{N}_0$. The entropy can be derived by use of the geometric sum as

$$\mathsf{H}_X = -\frac{p \,\mathrm{ld}\, p + (1-p)\mathrm{ld}(1-p)}{p}.$$

We let $p = 1/2$ to obtain a dyadic source where $\mathsf{I}(x_i) = -\mathrm{ld}\, p_i = i + 1$ is integer. The entropy becomes $\mathsf{H}_X = 2$. The corresponding Huffman code uses codewords of lengths $l_i = i + 1$. Consequently, the average codeword length is $\bar{l} = 2$ and the code achieves the entropy. The effective bandwidth follows from (17) after some algebra where $\theta \in (0, \ln 2)$ as

$$\alpha_A(\theta) = -\frac{1}{\theta} \ln(2e^{-\theta} - 1). \tag{19}$$

In Fig. 2, we show the $(c, d_A, \varepsilon_A)$-boundary obtained for the Huffman coded source by insertion of (19) into (18). We optimized the parameter $\theta$ numerically to obtain the smallest delay bound. The surface plot depicts the capacity that achieves a delay bound subject to a defined probability of error. The delays are due to the randomness that is introduced by variable codeword lengths. For $c > \mathsf{H}_X$ finite delay bounds can be computed, whereas the delay grows unbounded for $c \to \mathsf{H}_X$. Also, Fig. 2 shows the logarithmic growth of $d_A$ for decaying $\varepsilon_A$ that is characteristic of the approach.

*2) Variable Symbol Rate:* So far, we assumed that sources generate symbols at a constant rate. Next, we show how sources

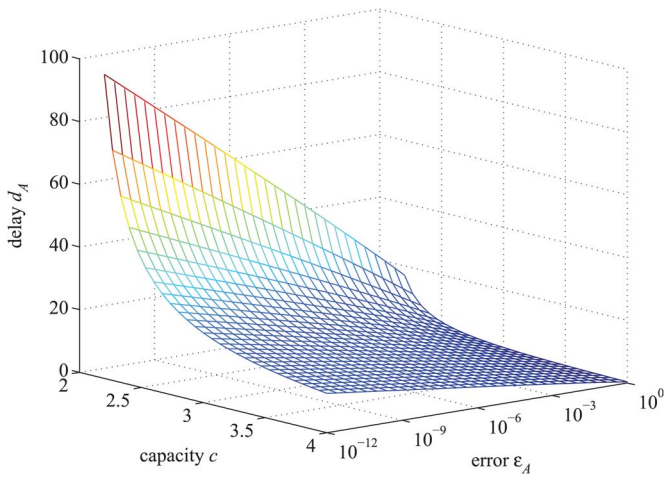


Fig. 2. $(c, d_A, \varepsilon_A)$-boundary of a Huffman coded dyadic source with entropy $\mathsf{H}_X = 2$. The delay grows unbounded if the capacity approaches the expected codeword length $\bar{l} = 2$.

with a variable symbol rate can be modeled using conditional MGFs. We provide a solution for the Poisson process. Given a memoryless source and denote $\mathsf{M}_L(\theta)$ the MGF of the increments (16). The conditional MGF of $n$ increments becomes $\mathsf{M}_A(\theta, n) = (\mathsf{M}_L(\theta))^n$. Here, the count of increments $N(t) \in \mathbb{N}_0$ in the interval $(0, t]$ is a random process with probability mass function $p_N(n, t)$. The effective bandwidth of the arrival process $A(t)$ follows by unconditioning as

$$\alpha_A(\theta, t) = \frac{1}{\theta t} \ln\left(\sum_n (\mathsf{M}_L(\theta))^n \, p_N(n, t)\right) \tag{20}$$

for $\theta > 0$ and $t > 0$. A Poisson process with mean rate $\lambda$ has probability mass function $p_N(n, t) = e^{-\lambda t}(\lambda t)^n / n!$. By insertion into (20) we derive

$$\alpha_A(\theta) = \frac{\lambda}{\theta} (\mathsf{M}_L(\theta) - 1)$$

for $\theta > 0$, where we used that $\sum_{n=0}^{\infty} a^n / n! = e^a$.

Fig. 3(a) shows the $(c, d_A, \varepsilon_A)$-boundary from (18) for a Huffman coded Poisson source. The remaining parameters are as in Fig. 2, i.e., the generation of symbols is memoryless and obeys the dyadic distribution. For illustration, we fix $\varepsilon_A = 10^{-6}$. For comparison, we show results for a Huffman coded source with constant symbol rate, as in Fig. 2, and a Poisson source with constant word size of 2 bit. The average symbol rate of all sources is $\lambda = 1$. Consequently, for all cases finite delay bounds exist if $c > \mathsf{H}_X = 2$. The double randomness of the Huffman coded Poisson source causes, however, a considerable increase of the delays.

*3) Length Limited Codewords:* While source coding can achieve a significant compression gain, unfavorable sequences of symbols that are mapped to large codewords can cause noticeable transmission delays, as shown in Fig. 2. The Huffman algorithm can be modified to comply with a given maximum codeword length at the expense of an increase of the average codeword length. We use the corresponding encoding algorithm from [57].

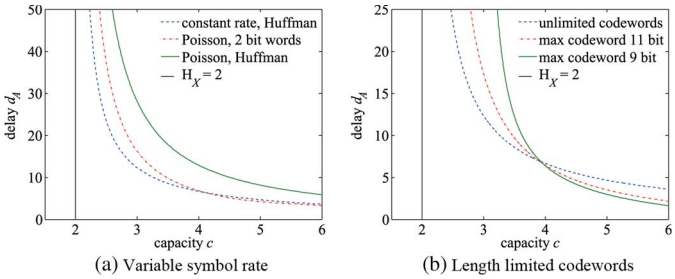


Fig. 3. (a) Huffman coded Poisson source compared to a Huffman coded constant rate source and a Poisson source with constant word size of 2 bit. (b) Length-limited Huffman codes compared to unlimited codeword lengths.

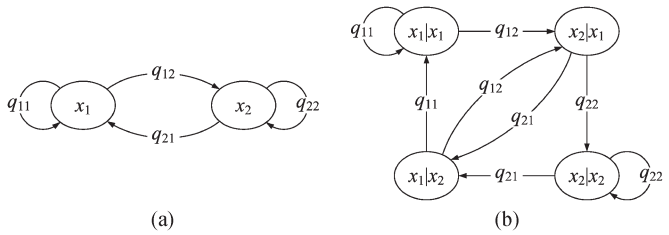


Fig. 4. (a) Example two-state Markov chain. (b) Extended Markov model where the information generated by symbol $x_j$ given the previous symbol was $x_i$ is uniquely determined by the state $x_j|x_i$ itself.

To show the effects of length limited codewords, we constrain the dyadic source from Fig. 2 to 256 symbols. We truncate and re-normalize the geometric symbol distribution accordingly. The entropy of the source remains virtually unchanged, i.e., 2 bit. In Fig. 3(b) we compare the unlimited Huffman code that uses codewords of up to 255 bit to the limited code with a length limit of 11 bit and 9 bit, respectively. The average codeword lengths are 2 bit, 2.24 bit, and 2.96 bit, respectively. The $(c, d_A, \varepsilon_A)$-boundaries make a fundamental tradeoff evident: due to its higher compression gain, the unlimited code outperforms the length limited codes if the capacity is close to the entropy; otherwise, the more balanced codewords of the length limited codes achieve smaller delay bounds.

### *B. Markov Sources*

In this section, we consider the encoding of discrete, stationary Markov sources, i.e., random processes $X(n)$ with first order dependence where the symbol $x_j$ that occurs in step $n$ depends only on the previous symbol $x_i$ in step $n-1$. The symbol $x_i$ determines the state of the Markov chain that can take any of the values $i = 1, 2, \ldots, |\mathbb{X}|$. An example of a two state Markov chain is shown in Fig. 4(a). We denote $p_i$ the stationary state distribution of the chain and $q_{ij}$ the transition probabilities from state $i$ to state $j$. Define $\mathbf{P}$ to be the row vector $(p_1, p_2, \ldots)$ and $\mathbf{Q}$ to be the state transition matrix. The stationary state distribution is the solution of $\mathbf{P} = \mathbf{PQ}$ under the normalization condition $\mathbf{P1} = 1$ where $\mathbf{1}$ is a column vector of ones. We let $n = t$ assuming a source that emits symbols at a constant rate of one symbol per time-slot.

*1) Direct Coding:* First, we consider a coder that maps individual symbols $x_i$ to codewords of length $l_i$. Let $\mathbf{L}(\theta)$ be the diagonal matrix $\mathbf{L}(\theta) = \mathrm{diag}(e^{\theta l_1}, e^{\theta l_2}, \ldots)$. The effective bandwidth of the coded Markov source results from [7] for $\theta > 0$ and $t > 0$ as

$$\alpha_A(\theta, t) = \frac{1}{\theta t} \ln \left( \mathbf{P} \left( \mathbf{L}(\theta)\mathbf{Q} \right)^{t-1} \mathbf{L}(\theta)\mathbf{1} \right). \qquad (21)$$

To compute the delay bound from Thm. 2 we choose the free parameter $\rho \in (0, 1/(\theta\varepsilon_A)]$ as $\rho = c - \sup_{t>0}\{\alpha_A(\theta, t)\}$. It follows that

$$d_A = \frac{-\ln\left(\theta\left(c - \sup_{t>0}\left\{\alpha_A(\theta, t)\right\}\right)\varepsilon_A\right)}{\theta c} \qquad (22)$$

for any $\theta > 0$ that satisfies $\theta \leq 1/((c - \sup_{t>0}\{\alpha_A(\theta, t)\})\varepsilon_A)$.

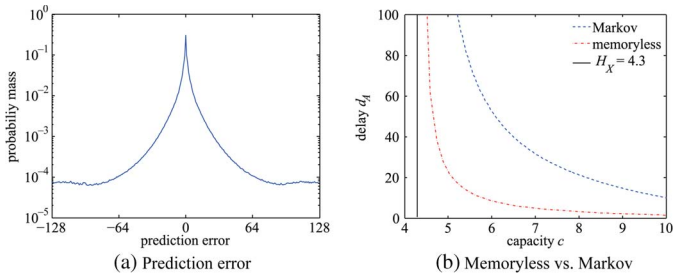


Fig. 5. (a) The HuffYUV prediction error approximates a two-sided geometric distribution. (b) HuffYUV encoded source with memoryless compared to Markovian prediction error. The memoryless model underestimates delays.

We show numerical results for lossless video coding. A common technique in lossless image and video compression is to exploit the correlation of neighboring pixels by prediction. The prediction error, i.e., the difference between adjacent pixels, typically follows a two-sided geometric distribution. Hence, it can be efficiently compressed using entropy coding. The method is implemented, e.g., in the HuffYUV ffmpeg coder, which we use for compression of the Big Buck Bunny cartoon. The video comprises 14315 frames in PNG format with a resolution of 640 × 360 pixels each. We consider the prediction error of the 8 bit luminance component. Fig. 5(a) depicts the probability mass function of the prediction error. It has an entropy of $\mathsf{H}_X = 4.29$ bit. The average codeword length of the Huffman code is $\bar{l} = 4.32$ bit.

Frequently, statistical independence of the prediction error is assumed. Using the memoryless model from (17), Fig. 5(b) depicts the corresponding $(c, d_A, \varepsilon_A)$-boundary after compression of the video by Huffman coding. The empirical data for the given video sequence shows, however, considerable first-order dependencies. Measuring the empirical dependence of the prediction error, we parameterize a Markov model and compute the $(c, d_A, \varepsilon_A)$-boundary from (22). Fig. 5(b) shows that both models have the same limiting behavior, i.e., the delay bounds grow unboundedly if the capacity approaches the average codeword length. Compared to small delays that apply for the memoryless model, significantly larger delays can be noticed if the dependencies are considered.

*2) State-Dependent Coding:* As the memory of a Markov source helps predicting the next symbol, its conditional entropy $\mathsf{H}(X(n)|X(n-1)) = -\sum_i \sum_j p_i q_{ij} \mathrm{ld}\, q_{ij}$ [3] can be significantly smaller than its entropy $\mathsf{H}_X = -\sum_i p_i \mathrm{ld}\, p_i$. Consequently, a code that encodes individual symbols is limited by

$\mathsf{H}_X$, whereas codes that take advantage of the memory of the source can achieve a compression down to $\mathsf{H}(X(n)|X(n-1))$. For a first order dependent source, it is sufficient to consider only the previous symbol, i.e., the state of the Markov chain. One option are individual codes for each of the states, i.e., the last symbol determines the code that is used to encode the next symbol. Alternatively, a coder can encode groups of symbols instead of single symbols to exploit the memory, see [1] and the technical report [27].

We propose a model for state-dependent codes, where we extend the state space of the Markov chain from $|\mathbb{X}|$ to $|\mathbb{X}|^2$. We label the states $x_j|x_i$ meaning that symbol $x_j$ occurs in the current time-slot after symbol $x_i$ occurred in the previous time-slot. We denote $l_{j|i}$ the length of the codeword that applies for state $x_j|x_i$, i.e., the codeword that is used to encode symbol $x_j$ given the previous symbol was $x_i$. The transition probabilities from state $x_j|x_i$ to state $x_k|x_j$ are $q_{jk}$ for any $i, j, k$ and zero otherwise. Fig. 4(b) shows the accordingly extended Markov model for the example from Fig. 4(a). As before, the effective bandwidth can be computed from (21) and the $(c, d_A, \varepsilon_A)$-boundary from (22). For the Big Buck Bunny cartoon as used in Fig. 5(b), the conditional entropy of the prediction error is $\mathsf{H}(X(n)|X(n-1)) = 3.80$ bit, compared to its entropy of $\mathsf{H}_X = 4.29$ bit. Hence, the use of state-dependent codes can achieve a higher compression gain that causes a shift of the limit of the CDE-boundary depicted in Fig. 5(b) to the left.

We conclude this section with some remarks on further applications. While we focused on delays, backlogs can be analyzed in the same way, where $\varepsilon_A$ can be interpreted as the probability of buffer overflow if buffer space is limited [7]. We note that the envelopes from Lem. 2 can also be used for traffic policing at the source, where excess traffic, that occurs with probability $\varepsilon_A$, can be discarded pro-actively to prevent that backlog and delay bounds are violated in the network.

## IV. Transmission Systems

In this section, we investigate basic transmission systems. We examine Rayleigh fading, hybrid ARQ, and Gilbert-Elliott channels and derive respective CDE-boundaries. Firstly, we are concerned with systems in isolation. The composition of sources and systems is covered in Section V.

### A. Rayleigh Fading

We consider the transmission of data via a fading channel. To begin with, we omit the ARQ protocol that is depicted in Fig. 1. Instead, we assume that the transmitter has perfect channel state information. It adapts the encoding accordingly such that data can generally be decoded at the receiver. We adopt the basic approach taken, e.g., in [36] and estimate the transmission rate from Shannon's capacity $C = \delta_f \mathrm{ld}(1+\gamma)$, where $\delta_f$ is the bandwidth of the channel and $\gamma$ is the signal-to-noise ratio (SNR). The SNR $\gamma(t)$ is governed by a random fading process with discrete time index $t$ and time-slot duration $\delta_t$. With $\beta = \delta_f \delta_t / \ln 2$ the number of bits served in the $t$-th time-slot is

$$Y(t) = \beta \ln\left(1+\gamma(t)\right). \tag{23}$$

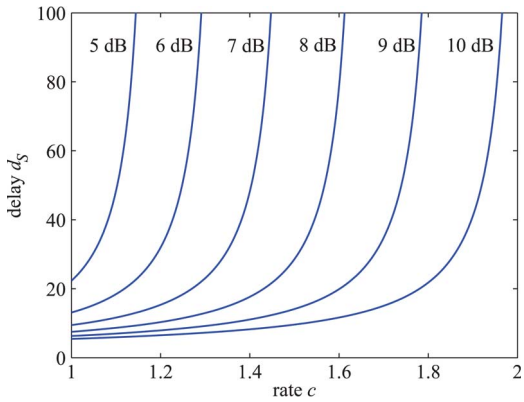


Fig. 6. $(c, d_S, \varepsilon_S)$-boundaries of Rayleigh fading channels with different SNR.

In case of Rayleigh fading $\gamma$ is exponentially distributed with density $f(\gamma) = \lambda e^{-\gamma\lambda}$ and mean $1/\lambda$. After some algebra the Laplace transform of the increment process $Y(t)$ follows as $\mathsf{M}_Y(-\theta) = e^{\lambda}\lambda^{\theta\beta}\Gamma(-\theta\beta+1, \lambda)$, where we used the incomplete Gamma function $\Gamma(a,\lambda) = \int_\lambda^\infty z^{a-1}e^{-z}dz$. The service process is computed as $S(\tau,t) = \sum_{\vartheta=\tau+1}^{t} Y(\vartheta)$ for $t > \tau \geq 0$. Under the assumption that fading samples are iid, that is justified if the time-slot duration is sufficiently large compared to the channel coherence time [36], the effective capacity of the service process is $\alpha_S(-\theta) = -\ln \mathsf{M}_Y(-\theta)/\theta$ and by insertion

$$\alpha_S(-\theta) = -\frac{1}{\theta}\ln\left(e^{\lambda}\lambda^{\theta\beta}\Gamma(-\theta\beta+1,\lambda)\right)$$

for $\theta > 0$. Based on the effective capacity, the $(c, d_S, \varepsilon_S)$-boundary of the fading channel is provided by Thm. 2. We choose the free parameter $\rho \in (0, 1/(\theta\varepsilon_S)]$ as $\rho = \alpha_S(-\theta) - c$, where $\alpha_S(-\theta) > c$ for stability to obtain

$$d_S = \frac{-\ln\left(\theta\left(\alpha_S(-\theta)-c\right)\varepsilon_S\right)}{\theta c} \tag{24}$$

for any $\theta > 0$ that satisfies $\theta \leq 1/((\alpha_S(-\theta)-c)\varepsilon_S)$.

We provide numerical results for parameter $\beta = 1$, i.e., the capacity is normalized and given in nats/Hz/s. The mean SNR is $10\log_{10}(1/\lambda) = \{5, 6, \ldots, 10\}$ dB and the corresponding expected service per time-slot is determined by the Shannon capacity (23) as $\mathsf{E}[Y(t)] = \{1.19, 1.34, 1.50, 1.66, 1.83, 2.01\}$. We show the delay bound $d_S$ from (24) that applies for constant rate arrivals with rate $c$. We let $\varepsilon = 10^{-6}$ and optimize parameter $\theta$ numerically. Fig. 6 reveals the influence of the SNR on the delay. Clearly, if the arrival rate approaches the expected service, delays grow unboundedly. The effect is due to the variability of the channel. The capacity limit is determined by the SNR. For moderate utilizations small delays prevail where a higher SNR generally leads to smaller delays.

### B. Hybrid ARQ

In this section, we avoid the previous assumption of perfect channel state information and consider the hybrid ARQ transmission system depicted in Fig. 1. Blocks of $l$ data symbols are encoded using an error correcting code. The resulting $k > l$ code symbols are transmitted over an error-prone channel,

where the receiver has no a priori knowledge whether a symbol is received correctly or not. The code has a symbol error correcting capability of $s$, i.e., the receiver can correctly decode all blocks that have up to $s$ symbol errors. The ARQ protocol uses a frame check sequence to detect remaining errors. If no error is detected, the block is acknowledged and the sender proceeds with the next block. Otherwise, the sender immediately retransmits the current block. The transmission time of a block is one time-slot. An implementation of this policy is stop-and-wait ARQ as used, e.g., in IEEE 802.11 WiFi networks. For simplicity, we assume that acknowledgements are not delayed nor lost. Models of different ARQ schemes are also provided in [41].

Given symbol errors occur independently with probability $p$. The block error probability after decoding is computed as

$$P = \sum_{i=s+1}^{k} \binom{k}{i} p^i (1-p)^{k-i}, \tag{25}$$

i.e., a block cannot be reliably decoded if it contains more than $s$ symbol errors. The transmission of each block is an iid Bernoulli trial, i.e., during a time-slot no data is delivered with probability $P$ and $ml$ bits are delivered with probability $1-P$, where we assume $m$-bit symbols. We denote $Y(t)$ the amount of service that is available in the $t$-th time-slot. The Laplace transform becomes $\mathsf{M}_Y(-\theta) = P + (1-P)e^{-\theta ml}$. Since the increments $Y(t)$ are independent, the service process $S(\tau, t)$ has effective capacity $\alpha_S(-\theta) = -\ln \mathsf{M}_Y(-\theta)/\theta$ for $\theta > 0$. By insertion of $\mathsf{M}_Y(-\theta)$ we obtain for $\theta > 0$ that

$$\alpha_S(-\theta) = -\frac{1}{\theta} \ln \left( P + (1-P)e^{-\theta ml} \right).$$

Based on the effective capacity, the $(c, d_S, \varepsilon_S)$-boundary of the hybrid ARQ system follows from (24).

For an implementation, we consider conventional Reed-Solomon codes. Reed-Solomon $(k, l)$ codes with $k = 2^m - 1$ code symbols and $l = 2^m - 1 - 2s$ data symbols are available for any $m > 1$ [58]. We use the code family for $m = 6$, i.e., blocks comprise of $k = 63$ coded symbols generated from $l = k - 2s$ data symbols. The code family enables us to examine how the symbol error correcting capability $s$ impacts the delay performance. Parameter $s$ has two counteracting effects on the capacity of the hybrid ARQ system: increasing $s$ helps avoiding retransmissions since more errors can be corrected. On the other hand, it increases the overhead as the number of data symbols per block $l$ is reduced. For evaluation we compute the $(c, d_S, \varepsilon_S)$-boundary of the system from (24) for $\varepsilon_S = 10^{-6}$. We optimize the free parameter $\theta$ numerically. The symbol error probability is $p = 0.01$.

The results are depicted in Fig. 7. With increasing $s$ the capacity of the system first grows, i.e., the curves are shifted to the right. In this region the additional errors that can be corrected by a stronger code outweigh its overhead. Then, starting with $s = 3$ the capacity shrinks, as the increasing overhead prevails. Consequently, there exists an optimal $s$ with respect to the capacity. Regarding delays, increasing $s$ is mostly beneficial, i.e., the curves are shifted downwards. The effect is due to the fact that a stronger error correcting code requires fewer time-consuming retransmissions.

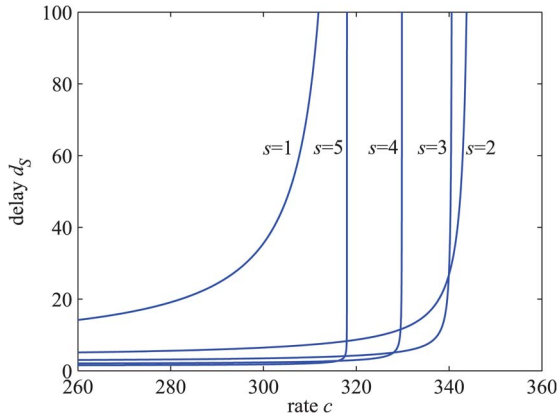


Fig. 7. $(c, d_S, \varepsilon_S)$-boundaries of hybrid ARQ using Reed-Solomon codes. The results show the impact of the error correcting capability $s$.

### C. Gilbert-Elliott Channel

Next, we relax the assumption of a memoryless channel and consider the impact of first-order dependencies on the performance of hybrid ARQ. We use the Gilbert-Elliott model to characterize the channel by a finite-state Markov chain. In each of the states $i$ the transmission rate is given as $r_i$. As in Section III-B denote $\mathbf{P}$ the stationary state distribution, $\mathbf{Q}$ the transition matrix, and $\mathbf{R}(-\theta) = \mathrm{diag}(e^{-\theta r_1}, e^{-\theta r_2}, \ldots)$. The effective capacity for $\theta > 0$ and $t > 0$ is

$$\alpha_S(-\theta, t) = -\frac{1}{\theta t} \ln \left( \mathbf{P} \left( \mathbf{R}(-\theta)\mathbf{Q} \right)^{t-1} \mathbf{R}(-\theta)\mathbf{1} \right).$$

To compute the delay bound from Thm. 2, we choose the free parameter $\rho \in (0, 1/(\theta\varepsilon_S)]$ as $\rho = \inf_{t>0}\{\alpha_S(-\theta, t)\} - c$ and it follows that

$$d_S = \frac{-\ln \left( \theta \left( \inf_{t>0} \{\alpha_S(-\theta, t)\} - c \right) \varepsilon_S \right)}{\theta c} \tag{26}$$

for any $\theta > 0$ that satisfies $\theta \leq 1/((\inf_{t>0}\{\alpha_S(-\theta, t)\} - c)\varepsilon_S)$.

As a model of the hybrid ARQ system from Section IV-B, we obtain a two-state Markov chain with states On (state 1) and Off (state 2). In On state, data can be decoded correctly so that $r_1 = ml$ bits are transmitted per time-slot, whereas in Off state, data cannot be decoded correctly so that $r_2 = 0$. The stationary state probabilities are $p_1 = 1 - P$ and $p_2 = P$, where $P$ is determined by the error correcting capability $s$ of the code (25). The remaining free parameter of the two-state Markov model can be used to adjust the mean state holding times. We fix the mean time to change state twice $1/q_{12} + 1/q_{21} = aT$. We use $T = 1/p_1 + 1/p_2$ so that parameter $a = 1$ corresponds to the memoryless case from Section IV-B, whereas memory increases with increasing $a > 1$.

In Fig. 8, we depict the impact of the memory for Reed-Solomon codes with symbol error correcting capabilities of $s = 1$ and $s = 5$, respectively. The remaining parameters of the code are as in Section IV-B. We optimized parameter $\theta$ numerically. Generally, Fig. 8 shows that the delay bounds become larger in case of memory. The effect is due to longer Off periods. While both codes achieve approximately the same maximum throughput, it is interesting to note that codes with a higher error correction capability achieve smaller delay bounds due to less pronounced Off periods.

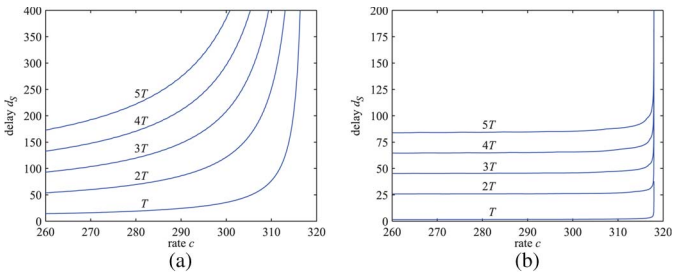


Fig. 8. Gilbert-Elliott channels with different state holding times $aT$. The error correcting capability of the code is $s = 1$ (a) and $s = 5$ (b).

We note that a variety of different systems that have been characterized by Markov chains can be included into the CDE-model in the same way. These include among others correlated fading [33], [35], [59], opportunistic scheduling [33], cognitive radio [29], MIMO [35], and Chase combining [2].

## V. Composition Results

In this section, we consider networks of systems. We present an approach for concatenation of tandem systems, where the service process of each system is characterized by its Laplace transform. Hence, the approach includes the systems from Section IV as well as the set of systems with known effective capacity from the literature. We provide a specific solution for the case of iid service increments as assumed, e.g., for the Rayleigh fading channel in Section IV-A. The result is a CDE-boundary for an entire network. Finally, we take advantage of the additivity of CDE-boundaries to compose a source and a (tandem) system. Using Thm. 2, the method generally applies to sources and systems with known effective bandwidth, respectively, effective capacity, including the sources from Section III and systems from Section IV. We note that CDE-boundaries of, e.g., more complex network topologies, systems with feedback flow control, joint source-channel coding, or scalers [39], [40] have not been investigated, yet.

### A. Networks of Systems

We consider a network of $i = 1 \dots N$ systems in series. The network calculus features a characterization of the network by a single service process that is computed by (3) as the min-plus convolution of the service processes of the individual systems. Given statistically independent systems, a bound for the Laplace transform of the network service process $\mathsf{M}_{S_{\text{net}}}(-\theta, t)$ is provided by (15) as

$$\mathsf{M}_{S_{\text{net}}}(-\theta,t) \leq \mathsf{M}_{S_1} * \mathsf{M}_{S_2} * \cdots * \mathsf{M}_{S_N}(-\theta,t)$$
$$= \sum_{\tau_i \geq 0 : \sum_{i=1}^{N} \tau_i = t} \mathsf{M}_{S_1}(-\theta,\tau_1)\mathsf{M}_{S_2}(-\theta,\tau_2)\cdots\mathsf{M}_{S_N}(-\theta,\tau_N). \tag{27}$$

A lower bound of the effective capacity of the network follows from $\alpha_{S_{\text{net}}}(-\theta, t) = -\ln \mathsf{M}_{S_{\text{net}}}(-\theta, t)/(\theta t)$ for $\theta > 0$ and $t > 0$ by insertion of (27).

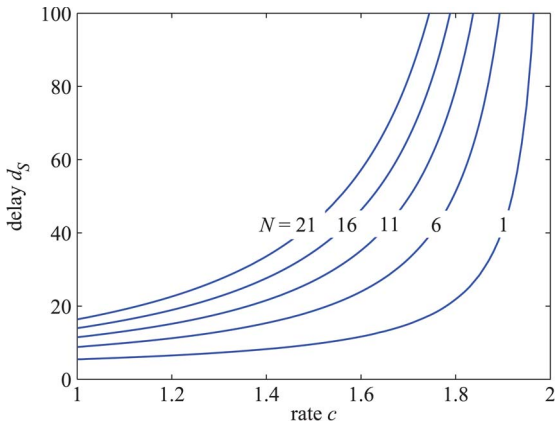


Fig. 9. $(c, d_S, \varepsilon_S)$-boundary of a tandem of $N$ Rayleigh fading channels.

In the following, we include a solution for homogeneous channels with iid service increments[9] $Y_i(t)$ as analyzed in [36], i.e., $\mathsf{M}_{S_i}(-\theta, t) = (\mathsf{M}_{Y_i}(-\theta))^t$. Consequently, (27) comprises $\binom{t+N-1}{N-1}$ identical summands $(\mathsf{M}_{Y_i}(-\theta))^t$ [36], [60]. A lower bound of the effective capacity of the network follows for $\theta > 0$ and $t > 0$ as

$$\alpha_{S_{\text{net}}}(-\theta,t) \geq \alpha_{S_i}(-\theta) - \frac{1}{\theta t}\ln\binom{t+N-1}{N-1}, \tag{28}$$

where we substituted $\alpha_{S_i}(-\theta) = -\ln(\mathsf{M}_{Y_i}(-\theta))^t/(\theta t)$. By insertion of (28), Thm. 2 gives the $(c, d_S, \varepsilon_S)$-boundary of the network.

In addition to numerical results from Thm. 2, we provide an analytical solution to analyze the growth of delays in networks in Appendix C. The analytical solution recovers the finding from [14] that performance bounds for $N$ independent systems in series are in $\mathcal{O}(N)$ and hence proves that the CDE-boundary of the network grows additively with $N$.

Fig. 9 shows $(c, d_S, \varepsilon_S)$-boundaries for a tandem of $N = 1, 6, 11, 16, 21$ Rayleigh fading channels as in Section IV-A. The mean SNR is 10 dB and $\varepsilon = 10^{-6}$. We optimized parameters $\theta$ and $\rho$ numerically. The numerical results confirm the linear growth of $d_S$ with $N$.

### B. Composition of Sources and Systems

In this section, we show how our model facilitates the convenient composition of sources and systems. Key to the composition is the additivity of CDE-boundaries that is established by Thm. 1.

We consider an MPEG video source that is transmitted via a Rayleigh fading channel. The video is the Big Buck Bunny cartoon from Section III-B that is encoded using the ffmpeg coder. We use different quantization parameters (qp) to obtain a set of encoded streams to enable source rate adaptation. Corresponding PSNR values are given in Table I as a quality measure. For the encoded video streams we apply the method from [38] to compute empirical envelopes. The $(c, d_A, \varepsilon_A)$-boundary follows by insertion of the envelope into (9) and subsequently

[9]Heterogeneous channels and non-iid service increments can be analyzed using the same basic steps and linear envelopes of the type $\mathsf{M}_{S_i}(-\theta, t) \leq e^{-\theta(\varsigma_i + \varrho_i t)}$ [47]. Also, we omit results on scheduling cross-traffic.

TABLE I
MEAN PSNR OF THE VIDEO FOR DIFFERENT QUANTIZATION PARAMETERS

| qp | 19 | 20 | 21 | 22 | 23 |
|---|---|---|---|---|---|
| PSNR | 46.39 | 45.61 | 44.99 | 44.35 | 43.59 |

| qp | 24 | 25 | 26 | 27 | 28 |
|---|---|---|---|---|---|
| PSNR | 42.97 | 42.34 | 41.61 | 40.99 | 40.39 |

(11). Since the model from [38] is deterministic, parameter $\sigma_A(\varepsilon_A)=0$ for all $\varepsilon_A\in[0,1]$. The transmission system is a Rayleigh fading channel as in Section IV-A, where we fix the bandwidth $\delta_f=4$ MHz, the duration of a time-slot $\delta_t=1$ ms, and $\varepsilon_S=10^{-6}$. While we perform the analysis for a single channel, we note that it applies to tandem channels in the same way.

Fig. 10(a) shows the $(c,d_A,\varepsilon_A)$-boundary of the video source and the $(c,d_S,\varepsilon_S)$-boundary of the Rayleigh fading channel for different video qp and channel SNR, respectively. A delay bound for the composition of the source and the system is computed from Thm.1 by addition of the individual CDE-boundaries for any $c>0$. The minimal delay bound that is obtained by choice of the optimal $c$ is depicted in Fig. 10(b).

In Fig. 10(c) we show the delay bound of the composition for different qp as a function of the SNR. Given the qp of the source and the SNR of the channel, Fig. 10(c) provides a method to determine the playout delay such that video frames arrive late for reproduction at the receiver at most with probability $\varepsilon=10^{-6}$. In Fig. 10(d), we show combinations of qp and SNR that ensure a playout delay of 0.1, 0.15, 0.2, 0.25, and 0.3 ms, respectively, with error probability $\varepsilon=10^{-6}$. All operating points in the upper right of the curves are feasible. Given a target playout delay, Fig. 10(d) enables adaptive video applications that adjust the qp of the video to the SNR of the channel such that the playout is ensured with probability $1-\varepsilon$.

## VI. CONCLUSION

We devised a notion of CDE-boundaries to facilitate the performance analysis of networked sources and systems. Intuitively, the model defines achievable operating points in the capacity-delay-error-space. A fundamental insight is the additivity of CDE-boundaries that is established for transmission of a source over a system. It enables the convenient composition of results obtained for sources and systems in isolation and creates a strong link between the individual research areas of effective bandwidth and effective capacity. We derived CDE-boundaries for various essential sources and systems, including the HuffYUV ffmpeg coder, Rayleigh fading channels, and hybrid ARQ systems. We presented composition results for tandem channels and for the transmission of an MPEG source over a fading channel. Extending the CDE-model to more complex systems and topologies is subject of future work, where we expect that solutions may be obtained using, e.g., the connection to MGFs and effective envelopes. Applications of the theory include, e.g., wireless quality of service, cross-layer optimization, and adaptive systems. As an example, we showed how to dimension the playout delay of a video transmission and how an adaptive video source can be adjusted to the SNR of the channel, e.g., by use of the quantizer.

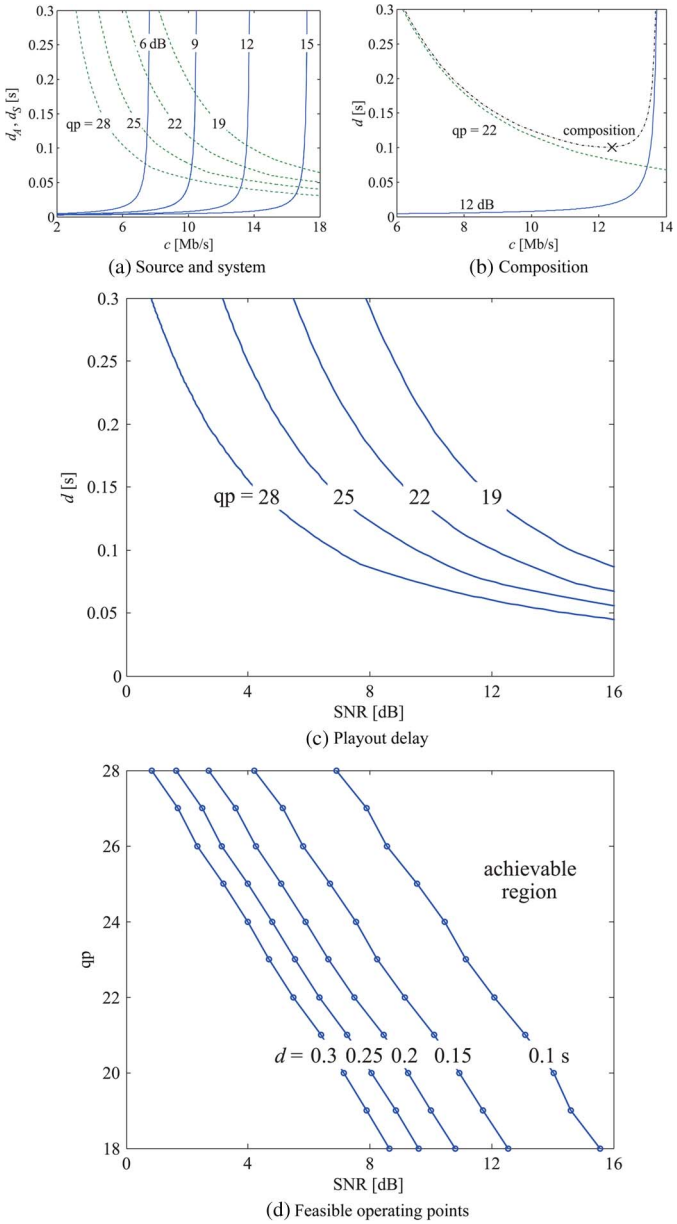


Fig. 10. Transmission of an MPEG video via a Rayleigh fading channel. (a) $(c,d_A,\varepsilon_A)$-boundary of the source for different qp and $(c,d_S,\varepsilon_S)$-boundary of the system for different SNR. (b) The delay bound of the composition follows as the minimum of the sum of the individual CDE-boundaries. (c) Playout delay as a function of the SNR. (d) All combinations of qp and SNR in the upper right of the curves are feasible operating points with respect to the target playout delay.

## APPENDIX

### A. Proof of Thm.1

The proof of Thm.1 uses the following Lem. 1.

*Lemma 1 (Duality):* Given the envelope functions $E_A(t)$ and $E_S(t)\in\mathcal{F}$ with Legendre transform $\overline{\mathcal{L}}_A(c)$ and $\underline{\mathcal{L}}_S(c)$, respectively. It holds that

$$\sup_{t\geq 0}\{E_A(t)-E_S(t)\}\leq\inf_{c>0}\left\{\overline{\mathcal{L}}_A(c)+\underline{\mathcal{L}}_S(c)\right\},$$

$$\inf\left\{\tau\geq 0:\sup_{t\geq 0}\{E_A(t)-E_S(t+\tau)\}\leq 0\right\}\leq\inf_{c>0}\left\{\overline{\mathcal{L}}_A(c)/c+\underline{\mathcal{L}}_S(c)/c\right\}.$$

If $E_A(t)$ is concave and $E_S(t)$ is convex,[10] both statements hold with equality.

*Proof of Lem. 1:* Let $v = \sup_{t\geq 0}\{E_A(t) - E_S(t)\}$ be the maximum vertical deviation of $E_A(t)$ and $E_S(t)$. For any $c > 0$ we rewrite $v = \sup_{t\geq 0}\{E_A(t) - ct + ct - E_S(t)\}$. It follows that

$$v \leq \sup_{t\geq 0}\{E_A(t) - ct\} + \sup_{t\geq 0}\{ct - E_S(t)\} = \overline{\mathcal{L}}_A(c) + \underline{\mathcal{L}}_S(c)$$

for all $c>0$. Taking the infimum for $c>0$ proves the first claim.

If $E_A(t)$ is concave and $E_S(t)$ is convex, it follows immediately from Fenchel's duality theorem [61] that the first statement holds with equality.

Let $h = \inf\{\tau \geq 0 : \sup_{t\geq 0}\{E_A(t) - E_S(t+\tau)\} \leq 0\}$ be the maximum horizontal deviation of $E_A(t)$ and $E_S(t)$. Note that an equivalent expression is $h = \inf\{\tau \geq 0 : E_A(t) - E_S(t+\tau) \leq 0, \forall\, t \geq 0\}$. For any $c > 0$ and $\vartheta \geq 0$ we rewrite

$$h = \inf\{\tau \geq 0 : E_A(t) - c(t+\vartheta) + c(t+\vartheta) - E_S(t+\tau) \leq 0,\ \forall\, t \geq 0\}.$$

We let $\vartheta = \sup_{t'\geq 0}\{E_A(t') - ct'\}/c$ so that $E_A(t) - c(t+\vartheta) = E_A(t) - ct - \sup_{t'\geq 0}\{E_A(t') - ct'\} \leq 0$ for all $t \geq 0$ and estimate

$$h \leq \inf\{\tau \geq 0 : c(t+\vartheta) - E_S(t+\tau) \leq 0,\ \forall\, t \geq 0\}.$$

After some reordering we arrive at the equivalent form

$$h \leq \inf\left\{\tau \geq 0 : \sup_{t\geq 0}\{c(t+\tau) - E_S(t+\tau)\} \leq c(\tau - \vartheta)\right\}.$$

Since $\sup_{t\geq 0}\{c(t+\tau) - E_S(t+\tau)\} \leq \sup_{t\geq 0}\{ct - E_S(t)\}$ for all $\tau \geq 0$ it follows with some further reordering that

$$h \leq \inf\left\{\tau \geq 0 : \vartheta + \sup_{t\geq 0}\{ct - E_S(t)\}/c \leq \tau\right\}.$$

The solution of the infimum is $\tau = \vartheta + \sup_{t\geq 0}\{ct - E_S(t)\}/c$ so that by insertion of $\vartheta$

$$h \leq \sup_{t'\geq 0}\{E_A(t') - ct'\}/c + \sup_{t\geq 0}\{ct - E_S(t)\}/c$$

for all $c > 0$. An equivalent expression is $h \leq \inf_{c>0}\{\overline{\mathcal{L}}_A(c)/c + \underline{\mathcal{L}}_S(c)/c\}$ which proves the second claim.

Finally, we show that if $E_A(t)$ is concave and $E_S(t)$ is convex, the second statement holds with equality. By definition of $h$ it follows that $E_S(t+h+\epsilon) \geq E_A(t)$ for all $t \geq 0$ and any $\epsilon > 0$. From the separation theorem for convex and concave functions [61] there exists at least one tuple $\vartheta, c$ so that $E_S(t+h+\epsilon) \geq c(t+\vartheta)$ for all $t \geq -(h+\epsilon)$ and $c(t+\vartheta) \geq E_A(t)$ for all $t \geq 0$. Since $E_A(t)$ and $E_S(t)$ are non-decreasing and excluding the trivial case $E_A(t) = E_S(t) = 0$ for all $t \geq 0$, it holds that $c > 0$. From $c(t+\vartheta) \geq E_A(t)$ we estimate $(E_A(t) - ct)/c \leq \vartheta$ for all $t \geq 0$ such that $\sup_{t\geq 0}\{E_A(t) - ct\}/c \leq \vartheta$. Similarly, we obtain from $E_S(t+h+\epsilon) \geq c(t+\vartheta)$ by substitution of $t' = t + h + \epsilon$ that $E_S(t') \geq c(t' - h - \epsilon + \vartheta)$ for all $t' \geq 0$. It follows that $(ct' - E_S(t'))/c \leq h + \epsilon - \vartheta$ for all $t' \geq 0$ such that $\sup_{t'\geq 0}\{ct' - E_S(t')\}/c \leq h + \epsilon - \vartheta$. Adding the expressions for $E_A(t)$ and $E_S(t)$ we conclude that there exists $c > 0$ so that $h \geq \sup_{t\geq 0}\{E_A(t) - ct\}/c + \sup_{t\geq 0}\{ct - E_S(t)\}/c - \epsilon$. Consequently, $h \geq \inf_{c>0}\{\overline{\mathcal{L}}_A(c)/c + \underline{\mathcal{L}}_S(c)/c - \epsilon\}$. Letting $\epsilon \to 0$ this proves that the upper bound from Lem. 1 is also a lower bound so that equality holds. ■

*Proof of Thm.1:* Thm. 1 follows immediately by insertion of $E_A(t) = E'_A(t) + \sigma_A(\varepsilon_A)$ and $E_S(t) = E'_S(t) - \sigma_S(\varepsilon_S)$ into Lem. 1. ■

### B. Proof of Thm. 2

The proof of Thm. 2 uses the following Lem. 2 for construction of arrival and service envelopes, respectively.

*Lemma 2 (Arrival and Service Envelopes):* Given a stationary arrival process $A(\tau, t)$ with effective bandwidth $\alpha_A(\theta, t)$. Define

$$E_A(t) = (\alpha_A(\theta, t) + \rho)\, t - \frac{\ln(\theta\rho)}{\theta}$$

for $t \geq 0$, where $\theta > 0$ and $\rho \in (0, 1/(\theta\varepsilon_A)]$ are free parameters. $E_A(t)$ is a statistical arrival envelope (4) of $A(\tau, t)$ with overflow profile $\sigma_A(\varepsilon_A) = -\ln \varepsilon_A/\theta$ for $\varepsilon_A \in (0, 1]$.

Given a stationary service process $S(\tau, t)$ with effective capacity $\alpha_S(-\theta, t)$. Define

$$E_S(t) = (\alpha_S(-\theta, t) - \rho)\, t + \frac{\ln(\theta\rho)}{\theta}$$

for $t \geq 0$, where $\theta > 0$ and $\rho \in (0, 1/(\theta\varepsilon_S)]$ are free parameters. $E_S(t)$ is a statistical service envelope (5) of $S(\tau, t)$ with deficit profile $\sigma_S(\varepsilon_S) = -\ln \varepsilon_S/\theta$ for $\varepsilon_S \in (0, 1]$.

*Proof of Lem. 2:* We only show the proof of the service envelope and omit the arrival envelope that follows in the same way. The derivation of the envelopes uses basic steps from the stochastic network calculus, see e.g., [13]. We define

$$\xi = \mathsf{P}\left[\exists \tau \in [0, t] : S(\tau, t) < E_S(t - \tau) - \sigma_S(\varepsilon_S)\right]$$

that is the left hand side of (5) and use the union bound to estimate

$$\xi \leq \sum_{\tau=0}^{t-1} \mathsf{P}\left[S(\tau, t) < E_S(t - \tau) - \sigma_S(\varepsilon_S)\right],$$

where we omitted the addend at $\tau = t$ that is zero trivially since $E_S(0) - \sigma_S(\varepsilon_S) \leq 0$ and $S(\tau, t)$ is non-negative.

For each of the addends we apply Chernoff's bound, i.e., $\mathsf{P}[S(\tau, t) \leq y] \leq e^{\theta y}\mathsf{M}_S(-\theta, t - \tau) = e^{\theta(y - (t-\tau)\alpha_S(-\theta, t-\tau))}$ for $\theta > 0$. By substitution of $y = E_S(t - \tau) - \sigma_S(\varepsilon_S)$ we have

$$\xi \leq \sum_{\tau=0}^{t-1} e^{\theta(E_S(t-\tau) - \sigma_S(\varepsilon_S) - (t-\tau)\alpha_S(-\theta, t-\tau))}. \tag{29}$$

We let $E_S(t) = (\alpha_S(-\theta, t) - \rho)t + \ln(\theta\rho)/\theta$ from Lem. 2. The free parameter $\rho > 0$ will be used as a slack rate to achieve

[10] If time is discrete $t \in \mathbb{N}_0$, we require that there exist functions $E'_A(y)$ concave and $E'_S(y)$ convex, $y \in \mathbb{R}_0^+$ such that $E_A(t) = E'_A(t)$ and $E_S(t) = E'_S(t)$ for all $t \in \mathbb{N}_0$ [61].

summability. Further $\rho \leq e^{\theta\sigma_S(\varepsilon_S)}/\theta$ ensures that $E_S(0) - \sigma_S(\varepsilon_S) \leq 0$. By insertion of $E_S(t)$ we get

$$\xi \leq \theta\rho e^{-\theta\sigma_S(\varepsilon_S)} \sum_{\tau=0}^{t-1} e^{-\theta\rho(t-\tau)}.$$

We substitute $\sum_{\tau=0}^{t-1} e^{-\theta\rho(t-\tau)} = \sum_{\delta=1}^{t} e^{-\theta\rho\delta}$ and estimate $\sum_{\delta=1}^{t} e^{-\theta\rho\delta} \leq \sum_{\delta=1}^{\infty} e^{-\theta\rho\delta}$ for all $t \geq 0$. Since $e^{-\theta\rho\delta}$, where $\theta > 0$ and $\rho > 0$, is decreasing in $\delta > 0$, we can bound each summand by $e^{-\theta\rho\delta} \leq \int_{\delta-1}^{\delta} e^{-\theta\rho y} dy$ and compute

$$\xi \leq \theta\rho e^{-\theta\sigma_S(\varepsilon_S)} \int_0^{\infty} e^{-\theta\rho y} dy = e^{-\theta\sigma_S(\varepsilon_S)}. \tag{30}$$

Equating $\varepsilon_S = e^{-\theta\sigma_S(\varepsilon_S)}$ and solving for $\sigma_S(\varepsilon_S) = -\ln\varepsilon_S/\theta$ completes the proof of the service envelope. ■

We note that a closely related statistical envelope function

$$E_S(t) = (\alpha_S(-\theta, t) - \rho)\, t + \frac{\ln(1 - e^{-\theta\rho})}{\theta} \tag{31}$$

for $\theta > 0$ and $\rho > 0$ is derived if we solve the geometric sum $\sum_{\delta=0}^{\infty} e^{-\theta\rho\delta} = 1/(1 - e^{-\theta\rho})$ instead of the integral (30).

*Proof of Thm. 2:* Thm. 2 follows immediately by insertion of $E_A(t)$ from Lem. 2 into (9) and subsequently into (11), respectively, by insertion of $E_S(t)$ from Lem. 2 into (10) and (12). ■

### C. CDE-Boundaries of Networks

The proof of the additive growth of the CDE-boundary of networks uses the following lemma.

*Lemma 3 (Network Service Envelopes):* Given a tandem of $N$ homogeneous systems each with service process $S_i(\tau, t)$, iid service increments, and effective capacity $\alpha_{S_i}(-\theta)$. The tandem has network service process $S_{\text{net}}(\tau, t)$ (3), and effective capacity $\alpha_{S_{\text{net}}}(-\theta, t)$ (28). Define

$$E_{S_{\text{net}}}(t) = (\alpha_{S_i}(-\theta) - \rho)\, t + \frac{N \ln(1 - e^{-\theta\rho})}{\theta}$$

for $t \geq 0$, where $\theta > 0$ and $\rho > 0$ are free parameters. $E_{S_{\text{net}}}(t)$ is a statistical service envelope (5) of $S_{\text{net}}(\tau, t)$ with deficit profile $\sigma_{S_{\text{net}}}(\varepsilon_{S_{\text{net}}}) = -\ln\varepsilon_{S_{\text{net}}}/\theta$ for $\varepsilon_{S_{\text{net}}} \in (0, 1]$.

For $N = 1$, Lem. 3 recovers the previous result for a single system (31) where $\alpha_S(-\theta, t) = \alpha_S(-\theta)$ in case of iid service increments.

*Proof of Lem. 3:* The first steps of the proof are identical to the proof of Lem. 2. By insertion of $E_{S_{\text{net}}}(t)$ from Lem. 2 and $\alpha_{S_{\text{net}}}(-\theta, t)$ from (28) into (29) we obtain for the overflow probability $\xi$ that

$$\xi \leq e^{-\theta\sigma_{S_{\text{net}}}(\varepsilon_{S_{\text{net}}})} \sum_{\tau=0}^{t-1} \binom{t-\tau+N-1}{N-1} \left(e^{-\theta\rho}\right)^{t-\tau} \left(1 - e^{-\theta\rho}\right)^N,$$

where $\rho > 0$ is a free parameter. We substitute $\delta = t - \tau$ and let $t \to \infty$ to estimate

$$\xi \leq e^{-\theta\sigma_{S_{\text{net}}}(\varepsilon_{S_{\text{net}}})} \sum_{\delta=0}^{\infty} \binom{\delta + N - 1}{N - 1} (e^{-\theta\rho})^{\delta} (1 - e^{-\theta\rho})^N.$$

Since $\theta\rho > 0$ we have $0 < e^{-\theta\rho} < 1$ and the argument of the sum is the negative binomial distribution so that the infinite sum equals one and hence

$$\xi \leq e^{-\theta\sigma_{S_{\text{net}}}(\varepsilon_{S_{\text{net}}})}.$$

Letting $\varepsilon_{S_{\text{net}}} = e^{-\theta\sigma_{S_{\text{net}}}(\varepsilon_{S_{\text{net}}})}$ and solving for $\sigma_{S_{\text{net}}}(\varepsilon_{S_{\text{net}}}) = -\ln\varepsilon_{S_{\text{net}}}/\theta$ completes the proof. ■

We insert $E_{S_{\text{net}}}(t)$ from Lem. 3 into (10) and subsequently into (12) to obtain for $\theta > 0$, $\rho > 0$, and $c > 0$ that

$$d_{S_{\text{net}}} = \frac{\sup_{t \geq 0}\{(c + \rho - \alpha_{S_i}(-\theta))t\}}{c} - \frac{N \ln(1 - e^{-\theta\rho}) + \ln\varepsilon_{S_{\text{net}}}}{\theta c}.$$

We choose the free parameter $\rho = \alpha_{S_i}(-\theta) - c$ where $\alpha_{S_i}(-\theta) > c$ for stability so that

$$d_{S_{\text{net}}} = -\frac{N \ln\left(1 - e^{-\theta(\alpha_{S_i}(-\theta) - c)}\right) + \ln\varepsilon_{S_{\text{net}}}}{\theta c},$$

which proves the linear growth of $d_{S_{\text{net}}}$ with $N$.

## REFERENCES


[1] R. Lübben and M. Fidler, "Non-equilibrium information envelopes and the capacity-delay-error-tradeoff of source coding," in *Proc. IEEE WoWMoM*, Jun. 2012, pp. 1–9.

[2] R. Lübben and M. Fidler, "On the delay performance of block codes for discrete memoryless channels with feedback," in *Proc. IEEE Sarnoff Symp.*, May 2012, pp. 1–6.

[3] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. Hoboken, NJ, USA: Wiley, 2006.

[4] A. Ephremides and B. Hajek, "Information theory and communications networks: An unconsummated union," *IEEE Trans. Inf. Theory*, vol. 44, no. 6, pp. 2416–2434, Oct. 1998.

[5] J. Andrews *et al.*, "Rethinking information theory for mobile ad hoc networks," *IEEE Commun. Mag.*, vol. 46, no. 12, pp. 94–101, Dec. 2008.

[6] F. P. Kelly, *Notes on Effective Bandwidths*. London, U.K.: Oxford Univ. Press, 1996, ser. Royal Statistical Society Lecture Notes, pp. 141–168, no. 4.

[7] C.-S. Chang, *Performance Guarantees in Communication Networks*. New York, NY, USA: Springer-Verlag, 2000.

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

[9] J.-Y. Le Boudec and P. Thiran, *Network Calculus a Theory of Deterministic Queuing Systems for the Internet*. New York, NY, USA: Springer-Verlag, 2001.

[10] R. L. Cruz, "Quality of service management in Integrated Services networks," in *Proc. Semi-Annual Res. Rev., Center Wireless Commun., UCSD*, Jun. 1996, pp. 1–9.

[11] M. Reisslein, K. W. Ross, and S. Rajagopal, "A framework for guaranteeing statistical QoS," *IEEE/ACM Trans. Netw.*, vol. 10, no. 1, pp. 27–42, Feb. 2002.

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

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

[14] M. Fidler, "An end-to-end probabilistic network calculus with moment generating functions," in *Proc. IEEE IWQoS*, 2006, pp. 261–270.

[15] Y. Jiang and Y. Liu, *Stochastic Network Calculus*. New York, NY, USA: Springer-Verlag, 2008.

[16] J. M. Walsh, S. Weber, and C. W. Maina, "Optimal rate-delay tradeoffs and delay mitigating codes for multipath routed and network coded networks," *IEEE Trans. Inf. Theory*, vol. 55, no. 12, pp. 5491–5510, Dec. 2009.

[17] L. H. Ozarow, S. Shamai, and A. D. Wyner, "Information theoretic considerations for cellular mobile radio," *IEEE Trans. Veh. Technol.*, vol. 43, no. 2, pp. 359–378, May 1994.

[18] S. V. Hanly and D. N. Tse, "Multiaccess fading channels-part ii: Delay-limited capacities," *IEEE Trans. Inf. Theory*, vol. 44, no. 7, pp. 2816–2831, Nov. 1998.

[19] R. A. Berry and R. G. Gallager, "Communication over fading channels with delay constraints," *IEEE Trans. Inf. Theory*, vol. 48, no. 5, pp. 1135–1149, May 2002.

[20] X. Li, X. Dong, and D. Wu, "Queue length aware power control for delay-constrained communication over fading channels," *Wireless Commun. Mobile Comput.*, vol. 12, no. 10, pp. 901–909, Jul. 2012.
[21] D. Baron, M. A. Khojastepour, and R. G. Baraniuk, "How quickly can we approach channel capacity?" in *Conf. Rec. 38th Asilomar Conf. Signals, Syst. Comput.*, Nov. 2004, pp. 1096–1100.
[22] Y. Polyanskiy, H. V. Poor, and S. Verdú, "Dispersion of the Gilbert-Elliott channel," in *Proc. IEEE ISIT*, Jun. 2009, pp. 2209–2213.
[23] Y. Polyanskiy, H. V. Poor, and S. Verdú, "Channel coding rate in the finite blocklength regime," *IEEE Trans. Inf. Theory*, vol. 56, no. 5, pp. 2307–2359, May 2010.
[24] N. Ahmed and R. G. Baraniuk, "Throughput measures for delay-constrained communications in fading channels," in *Proc. Allerton Conf. Commun., Control Comput.*, 2003, pp. 1–11.
[25] I. Bettesh and S. Shamai, "Queuing analysis of the single user fading channel," in *Proc. 21st IEEE Conv. Electr. Electron. Eng. Israel*, 2000, pp. 274–277.
[26] J. Burdin and R. Landry, "Delay analysis of wireless Nakagami fading channels," in *Proc. IEEE GLOBECOM*, Nov. 2008, pp. 4191–4195.
[27] R. Lübben and M. Fidler, "Non-equilibrium information envelopes and the capacity-delay-error-tradeoff of source coding," in *Proc. IEEE Int. Symp. World Wireless, Mobile Multimedia Netw.*, Jul. 2011, pp. 1–9, [Online]. Available: http://arxiv.org/pdf/1107.3087.pdf
[28] D. Wu and R. Negi, "Effective capacity: A wireless link model for support of quality of service," *IEEE Trans. Wireless Commun.*, vol. 2, no. 4, pp. 630–643, Jul. 2003.
[29] S. Akin and M. C. Gursoy, "Effective capacity analysis of cognitive radio channels for quality of service provisioning," *IEEE Trans. Wireless Commun.*, vol. 9, no. 11, pp. 3354–3364, Nov. 2010.
[30] E. A. Jorswieck, R. Mochaourab, and M. Mittelbach, "Effective capacity maximization in multi-antenna channels with covariance feedback," *IEEE Trans. Wireless Commun.*, vol. 9, no. 10, pp. 2988–2993, Oct. 2010.
[31] Q. Wang, D. Wu, and P. Fan, "Effective capacity of a correlated Rayleigh fading channel," *Wireless Commun. Mobile Comput.*, vol. 11, no. 11, pp. 1485–1494, Nov. 2011.
[32] Q. Wang, D. Wu, and P. Fan, "Effective capacity of a correlated Nakagami-m fading channel," *Wireless Commun. Mobile Comput.*, vol. 12, no. 14, pp. 1225–1238, Oct. 2012.
[33] M. Fidler, "WLC15-2:A network calculus approach to probabilistic quality of service analysis of fading channels," in *Proc. IEEE GLOBECOM*, 2006, pp. 1–6.
[34] C. Li, H. Che, and S. Li, "A wireless channel capacity model for quality of service," *IEEE Trans. Wireless Commun.*, vol. 6, no. 1, pp. 356–366, Jan. 2007.
[35] K. Mahmood, A. Rizk, and Y. Jiang, "On the flow-level delay of a spatial multiplexing MIMO wireless channel," in *Proc. IEEE ICC*, Jun. 2011, pp. 1–6.
[36] H. Al-Zubaidy, J. Liebeherr, and A. Burchard, "A (min, x) network calculus for multi-hop fading channels," in *Proc. IEEE INFOCOM*, 2013, pp. 1833–1841.
[37] R. J. Gibbens, *Traffic Characterisation and Effective Bandwidths for Broadband Network Traces*. London, U.K.: Oxford Univ. Press, 1996, ser. Royal Statistical Society Lecture Notes, pp. 169–179, no. 4.
[38] D. Wrege, E. Knightly, H. Zhang, and J. Liebeherr, "Deterministic delay bounds for VBR video in packet-switching networks: Fundamental limits and practical trade-offs," *IEEE/ACM Trans. Netw.*, vol. 4, no. 3, pp. 352–362, Jun. 1996.
[39] M. Fidler and J. B. Schmitt, "On the way to a distributed systems calculus: An end-to-end network calculus with data scaling," in *Proc. ACM SIGMETRICS/Perform.*, 2006, pp. 287–298.
[40] F. Ciucu, J. B. Schmitt, and H. Wang, "On expressing networks with flow transformations in convolution-form," in *Proc. IEEE INFOCOM*, Apr. 2011, pp. 1979–1987.
[41] H. Wang and J. B. Schmitt, "On the way to a wireless network calculus—The single node case with retransmissions," Univ. Kaiserslautern, Kaiserslautern, Germany, Tech. Rep. 375/10, Jan. 2010.
[42] J. Xie and Y. Jiang, "An analysis on error servers for stochastic network calculus," in *Proc. IEEE LCN*, Oct. 2008, pp. 184–191.
[43] K. Wu, Y. Jiang, and G. Hu, "A calculus for information-driven networks," in *Proc. IEEE IWQoS*, Jul. 2009, pp. 1–9.
[44] B. Haverkort, *Performance of Computer Communication Systems: A Model-Based Approach*. Hoboken, NJ, USA: Wiley, 1998.
[45] M. Fidler, "A survey of deterministic and stochastic service curve models in the network calculus," *IEEE Commun. Surveys Tuts.*, vol. 12, no. 1, pp. 59–86, 2010.
[46] F. Ciucu and J. Schmitt, "Perspectives on network calculus: No free lunch, but still good value," in *Proc. ACM SIGCOMM*, Aug. 2012, pp. 311–322.
[47] M. Fidler and A. Rizk, "A guide to the stochastic network calculus," *IEEE Commun. Surveys Tuts.*, DOI: 10.1109/COMST.2014.2337060, to be published.
[48] F. Ciucu, "Scaling properties in the stochastic network calculus," Ph.D. dissertation, School Eng. Appl. Sci., Univ. Virginia, Charlottesville, VA, USA, Aug. 2007.
[49] R. Lübben, M. Fidler, and J. Liebeherr, "Stochastic bandwidth estimation in networks with random service," *IEEE/ACM Trans. Netw.*, vol. 22, no. 2, pp. 484–497, Apr. 2014.
[50] R. T. Rockafellar, *Convex Analysis*. Princeton, NJ, USA: Princeton Univ. Press, 1970.
[51] M. Fidler and S. Recker, "Conjugate network calculus: A dual approach applying the Legendre transform," *Comput. Netw.*, vol. 50, no. 8, pp. 1026–1039, Jun. 2006.
[52] K. Angrishi, "An end-to-end stochastic network calculus with effective bandwidth and effective capacity," *Comput. Netw.*, vol. 57, no. 1, pp. 78–84, Jan. 2013.
[53] O. Yaron and M. Sidi, "Performance and stability of communication networks via robust exponential bounds," *IEEE/ACM Trans. Netw.*, vol. 1, no. 3, pp. 372–385, Jun. 1993.
[54] K. Lee, "Performance bounds in communication networks with variable-rate links," in *Proc. ACM SIGCOMM*, 1995, pp. 126–136.
[55] A. Burchard, J. Liebeherr, and F. Ciucu, "On superlinear scaling of network delays," *IEEE/ACM Trans. Netw.*, vol. 19, no. 4, pp. 1043–1056, Aug. 2011.
[56] G. Katana and O. Nemetz, "Huffman codes and self-information," *IEEE Trans. Inf. Theory*, vol. IT-22, no. 3, pp. 337–340, May 1976.
[57] A. Turpin and A. Moffat, "Practical length-limited coding for large alphabets," *Comput. J.*, vol. 38, no. 5, pp. 339–347, 1995.
[58] B. Sklar, *Digital Communications: Fundamentals and Applications*, 2nd ed. Englewood Cliffs, NJ, USA: Prentice-Hall, 2001.
[59] Q. Zhang and S. A. Kassam, "Finite-state Markov model for Rayleigh fading channels," *IEEE Trans. Commun.*, vol. 47, no. 11, pp. 1688–1692, Nov. 1999.
[60] S. Ross, *A First Course in Probability*, 8th ed. Englewood Cliffs, NJ, USA: Prentice-Hall, 2010.
[61] K. Murota, *Discrete Convex Analysis*. Philadelphia, PA, USA: SIAM, 2003.

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

**Ralf Lübben** (S'11) received the M.Sc. degree in computer science and the Ph.D. degree from Leibniz Universität Hannover, Hanover, Germany, in 2007 and 2014, respectively. He is currently a postdoctoral researcher with the Institute of Communications Technology, Leibniz Universität Hannover.

**Nico Becker** received the Diploma in mathematics from Leibniz Universität Hannover, Hanover, Germany, in 2012. Since 2012, he has been a Ph.D. candidate with the Institute of Communications Technology, Leibniz Universität Hannover.