

A Non-stationary Service Curve Model for Performance Analysis of Transient Phases

Nico Becker and Markus Fidler

Institute of Communications Technology, Leibniz Universität Hannover

Abstract—Steady-state solutions for a variety of relevant queueing systems are known today, e.g., from queueing theory, effective bandwidths, and network calculus. The behavior during transient phases, on the other hand, is understood to a much lesser extent as its analysis poses significant challenges. Considering the majority of short-lived flows, transient effects that have diverse causes, such as TCP slow start, sleep scheduling in wireless networks, or signalling in cellular networks, are, however, predominant. This paper contributes a general model of regenerative service processes to characterize the transient behavior of systems. The model leads to a notion of non-stationary service curves that can be conveniently integrated into the framework of the stochastic network calculus. We derive respective models of sleep scheduling and show the significant impact of transient phases on backlogs and delays. We also consider measurement methods that estimate the service of an unknown system from observations of selected probe traffic. We find that the prevailing rate scanning method does not recover the service during transient phases well. This limitation is fundamental as it is explained by the non-convexity of non-stationary service curves. A second key difficulty is proven to be due to the super-additivity of network service processes. We devise a novel two-phase probing technique that first determines a minimal pattern of probe traffic. This probe is used to obtain an accurate estimate of the unknown transient service.

I. INTRODUCTION

The majority of flows in today’s computer networks are short-lived [1] and hence dominated by various transient effects that can have a significant impact on their performance. Relevant examples include TCP slow start [1], where the size of the initial congestion window has been repeatedly under debate [2], the convergence of routing protocols, power saving in wireless networks using polling, or signalling and discontinuous reception in cellular networks [3].

The performance analysis of transient phases causes, however, fundamental difficulties and queueing theory offers solutions mainly for the steady-state. As an example consider the basic M|M|1 queue, where the stationary state distribution follows readily from a set of linear balance equations, e.g., [4]. The transient behavior, on the other hand, is expressed by a set of differential equations for which mainly approximate or numerical solutions are known [5]. As a consequence, analytical works on transient effects in computer networks are sparse [6]–[8] and often tailored to specific problems like TCP congestion control [1].

A theory that does without an assumption of stationarity is the deterministic network calculus [9], [10]. It employs

envelope functions of possibly non-stationary processes to analyze the worst-case behavior of systems. Hence, it takes transient phases into the consideration. Once the worst-case is achieved, it can, however, not predict how a system progresses. We explain this effect and present an example in Sec. II-A. The stochastic network calculus [10]–[18], on the other hand, typically either assumes stationarity or uses stationary bounds.

In this work, we use the notion of time-variant systems [19], [20] to model non-stationary service characteristics. Time-variant systems are described by bivariate instead of univariate functions to consider changes over time. We show that non-stationarity can be dealt with in the stochastic network calculus using a similar notational extension. We contribute a non-stationary service curve model that characterizes the service of systems during transient phases. While many known results are recovered for the extended model, certain fundamental properties such as commutativity [19] differ. We derive solutions for systems with sleep scheduling, provide insights into the transient behavior, and quantify the considerable transient overshoot. Secondly, we examine methods for estimation of a system’s service curve from measurements of probe traffic. We refine known measurement methods for estimation of non-stationary service curves, where we encounter additional difficulties that are attributed to the non-convexity and super-additivity of the service. We devise a novel minimal probing method that estimates a non-stationary service curve and provides a measure of accuracy.

The remainder of this work is structured as follows. In Sec. II, we define non-stationary service curves, show a method for construction, and derive models of systems with sleep scheduling. In Sec. III, we investigate the measurement-based estimation of non-stationary service curves. We reveal difficulties that arise and devise a new minimal probing method. We discuss further related works in the respective sections. Sec. IV presents brief conclusions. While we restrict the exposition to non-stationary systems, we note that non-stationary traffic can be dealt with in the same way.

II. SYSTEM MODEL

In this section, we derive a model of non-stationary service curves (Sec. II-B) that enable analyzing the performance of systems during transient phases. The basis of this model is a definition of regenerative service processes (Sec. II-A), where regeneration points mark the start of new transient phases. We include solutions for sleep scheduling and show its impact on the performance.

A. Regenerative Service Processes

We consider a system with cumulative arrivals $A(t)$, where $A(t)$ denotes the number of bits that arrive in the time interval $(0, t]$. By convention, there are no arrivals for $t \leq 0$ so that we generally consider $t \geq 0$. Clearly, $A(t)$ is a non-negative, non-decreasing function, and $A(0) = 0$. Shorthand notation $A(\tau, t) = A(t) - A(\tau)$ is used to denote the arrivals in $(\tau, t]$ where $t \geq \tau \geq 0$. Trivially, $A(t, t) = 0$ for all $t \geq 0$. Similarly, $D(t)$ denotes the cumulative departures from the system.

The service that is provided by the system is characterized by a time-variant service process $S(\tau, t)$ that establishes the departure guarantee [10], [19]–[21]

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

By convention, $S(\tau, t)$ is non-negative and $S(t, t) = 0$ for all $t \geq 0$. The operator \otimes that is defined by (1) is known as convolution under a min-plus algebra [9], [10]. We note that \otimes is associative but not commutative in general. Examples of (1) include a work-conserving server with a time-variant capacity [10] where $S(\tau, t)$ denotes the service that is available in the interval $(\tau, t]$; scheduling with cross-traffic [14]; and networks of systems where the network service process $S^{net}(\tau, t)$ is computed from the service processes $S^i(\tau, t)$ of the individual systems $i = 1 \dots n$ by recursive insertion of (1) as $S^{net}(\tau, t) = S^1 \otimes S^2 \otimes \dots \otimes S^n(\tau, t)$ [10]. The service guarantee (1) enables the derivation of performance bounds. An upper bound of the backlog $B(t) = A(t) - D(t)$ follows by insertion of (1) as

$$B(t) \leq \sup_{\tau \in [0, t]} \{A(\tau, t) - S(\tau, t)\}. \quad (2)$$

Similarly, the delay defined as $W(t) = \inf\{w \geq 0 : A(t) \leq D(t + w)\}$ can be considered.

Throughout this work, we assume that the service is a regenerative process [4] with regeneration points $\mathbb{P} = \{P_0, P_1, P_2, \dots\}$ where $P_0 = 0$ and $P_i < P_{i+1}$ for all $i \geq 0$. We divide $S(\tau, t)$ into segments

$$S_i(\tau, t) = S(\tau + P_i, t + P_i) \quad (3)$$

for all $0 \leq \tau \leq t \leq P_{i+1} - P_i$ and $i \geq 0$, where $S_i(\tau, t)$ is the service process between the i th and the $(i+1)$ th regeneration point. The defining characteristic of a regenerative process is that the $S_i(\tau, t)$ are statistical replicas, i.e.,

$$P[S_i(\tau, t) \leq x] = P[S_j(\tau, t) \leq x] \quad (4)$$

for all $i, j, x \geq 0$, and $0 \leq \tau \leq t \leq \min\{P_{i+1} - P_i, P_{j+1} - P_j\}$. Owing to (4), we omit the index i in the sequel. Also, we will not explicitly mention the constraint $t \leq P_{i+1}$ and assume that the next regeneration point P_{i+1} is spaced sufficiently apart.

Deterministic Sleep Scheduling: We present a first application to sleep scheduling, where we consider a transmitter and a receiver that if idle go to sleep state according to a defined protocol. Wake up is scheduled deterministically, T units of time after entering the sleep state. The transmission rate in sleep state is zero and otherwise it is R . Clearly, each

transition to sleep state is a regeneration point and the time-variant service process follows for $t \geq \tau \geq 0$ as

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

so that

$$S(\tau, t) = [R(t - \max\{\tau, T\})]_+ \quad (5)$$

where $[x]_+ = \max\{0, x\}$ is the non-negative part of x .

For numerical evaluation we use a discrete time equivalent of a stationary Poisson arrival process: the number of packet arrivals $N(t)$ in an interval of length t is binomial with parameter $\alpha \in [0, 1]$. The individual packet sizes $Y(i)$ with index $i = 1, 2, \dots$ are independent and identically distributed (iid) geometric random variables with parameter $\beta \in (0, 1]$. Parameter α has the interpretation of an average arrival rate and $1/\beta$ is the average size of packets. For the special case of a system with constant service rate $R = 1$, α/β is the utilization and $\alpha < \beta$ is required for stability.

We characterize the process using an upper envelope function that is derived from its moment generating function (MGF). The MGF of a random variable X is defined as $M_X(\theta) = E[e^{\theta X}]$ for any θ . The respective MGFs of the above processes are $M_N(\vartheta, t) = (\alpha e^\vartheta + 1 - \alpha)^t$ and $M_Y(\theta) = \beta e^\theta / (1 - (1 - \beta)e^\theta)$ for $\theta \in [0, -\ln(1 - \beta))$ [22]. The cumulative arrival process $A(t)$ is the doubly stochastic process $A(t) = \sum_{i=1}^{N(t)} Y(i)$. It has MGF $M_A(\theta, t) = M_N(\ln M_Y(\theta), t)$ [18], [22] so that by insertion

$$M_A(\theta, t) = \left(\frac{\alpha \beta e^\theta}{1 - (1 - \beta)e^\theta} + 1 - \alpha \right)^t. \quad (6)$$

Using Chernoff's theorem $P[X \geq x] \leq e^{-\theta x} M_X(\theta)$ for $\theta \geq 0$ and established methods of the stochastic network calculus [12], [13] it can be shown¹ that the function

$$\mathcal{A}^\varepsilon(t) = \frac{1}{\theta(t)} (\ln M_A(\theta, t) + \rho t - \ln(\rho \varepsilon)) \quad (7)$$

is a statistical envelope function of $A(t)$ that provides the sample path guarantee

$$P[A(\tau, t) \leq \mathcal{A}^\varepsilon(t - \tau), \forall \tau \in [0, t]] \geq 1 - \varepsilon \quad (8)$$

for all $t \geq 0$. Above, $\varepsilon \in (0, 1]$ is a probability of overflow, and $\theta(t) > 0$ and $\rho \in (0, 1/\varepsilon]$ are free parameters². With (8), statistical performance bounds follow readily by substitution of $\mathcal{A}^\varepsilon(t - \tau)$ for $A(\tau, t)$, e.g., the backlog bound (2) yields

$$P\left[B(t) \leq \sup_{\tau \in [0, t]} \{\mathcal{A}^\varepsilon(t - \tau) - S(\tau, t)\}\right] \geq 1 - \varepsilon. \quad (9)$$

In Fig. 1, we illustrate the progression of the backlog bound (9) over time. The parameters of the service process (5)

¹We omit the proof as it is dual to the derivation of (13) and considers only the special case of a stationary process.

²Compared to related works, we use a time-variant parameter $\theta(t)$ instead of a constant θ . This allows optimizing $\theta(t)$ to minimize $\mathcal{A}^\varepsilon(t)$ individually for each t , which facilitates a computationally efficient implementation.

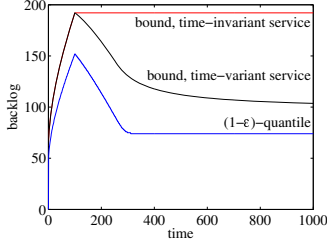


Fig. 1. Progression of the transient backlog over time. The time-variant service model correctly estimates the shape of the $(1 - \varepsilon)$ -quantile.

are $T = 100$ and $R = 1$. For the arrival process (6) we use $\alpha = 0.09$ and $\beta = 0.3$ corresponding to a utilization of 0.3. We choose $\varepsilon = 10^{-9}$ and optimize the free parameters $\theta > 0$ and $\rho \in (0, 1/\varepsilon]$ of (7) numerically. Also, we include a backlog bound from the deterministic network calculus, where the service $S_{inv}(\tau, t) = R[t - \tau - T]_+ = S_{inv}(t - \tau)$ is a univariate time-invariant function that depends only on the width of the interval $(\tau, t]$. Otherwise, it considers the worst-case, that is attained for $\tau = 0$ where $S_{inv}(t) = S(0, t)$ from (5). Fig. 1 shows that both, the time-variant and the time-invariant service model, reveal the same growth of the backlog bound until service starts at $T = 100$. How the transient backlog is cleared after T and eventually converges to a stationary backlog bound is, however, only explained by the time-variant model. To see why the time-invariant model fails note that the sup in (9) is non-decreasing in t if S is a univariate function.

Further, Fig. 1 includes the exact backlog quantile for comparison. For the discrete time Poisson model, the solution can be readily obtained from a discrete time Markov chain. The state of the Markov chain $K(t) \geq 0$ represents the number of arrivals that are in the system at t . For $t \leq T$ the transition matrix $\mathbf{Q}(t)$ is composed of the probabilities $q_{i,i} = 1 - \alpha$, $q_{i,i+1} = \alpha$, and all other $q_{i,j} = 0$. For $t > T$ the service starts so that $q_{i,i-1} = (1 - \alpha)\beta$, $q_{i,i} = (1 - \alpha)(1 - \beta) + \alpha\beta$, $q_{i,i+1} = \alpha(1 - \beta)$, and all other $q_{i,j} = 0$. The Markov chain starts in state $K(0) = 0$, i.e., the initial state distribution $\mathbf{P}(0)$ is the column vector $(1, 0, 0, \dots)$. The state distribution for $t > 0$ follows by repeated insertion of $\mathbf{P}(t) = \mathbf{Q}(t)\mathbf{P}(t - 1)$.

Clearly, for $t \leq T$ the state distribution is binomial, whereas for $t > T$ the distribution makes a transition and for $t \rightarrow \infty$ attains the geometric stationary state distribution [23]

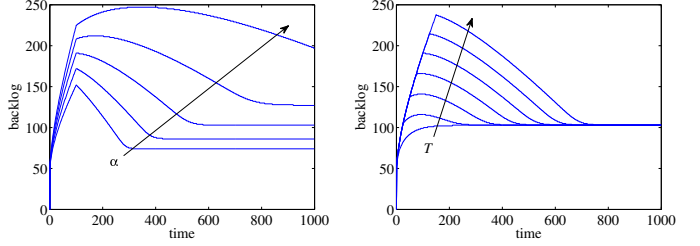
$$\mathbb{P}[K(\infty) = k] = \frac{\beta - \alpha}{\beta(1 - \alpha)} \left(\frac{\alpha(1 - \beta)}{(1 - \alpha)\beta} \right)^k.$$

The backlog distribution can be computed as

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

for $b > 0$ and $\mathbb{P}[B(t) = 0] = \mathbb{P}[K(t) = 0]$ for $b = 0$. The conditional backlog in state k is the sum of k geometric random variables that is negative binomial, i.e., for $k, b > 0$

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



(a) $\alpha = \{0.09, 0.12, \dots, 0.21\}$

(b) $T = \{0, 25, \dots, 150\}$

Fig. 2. Impact of arrival rate α and sleep cycle T on the backlog quantile.

The $(1 - \varepsilon)$ -quantile of $B(t)$ is depicted in Fig. 1. We observe that the bound from the time-variant service model provides a good estimate that recovers the shape of the quantile. We note that the deviation is due to bounds that are invoked in the derivation of the Poisson envelope. While the network calculus literature includes envelopes for non-trivial arrival processes including self-similar, long-range dependent [24], [25], and heavy-tailed processes [25], tighter martingale bounds are available, e.g., for Poisson [26], Markov, and autoregressive processes [10].

Fig. 2 presents the impact of the arrival rate α for $T = 100$ and the impact of the sleep cycle T for $\alpha = 0.15$ on the backlog quantile. The remaining parameters are as in Fig. 1. The measures of interest [6] are the maximum overshoot compared to the steady-state and the relaxation time, i.e., the time that is required to reach within a defined range the steady-state backlog. Fig. 2(a) shows that α has a significant impact on both quantities. Interestingly, if α is large, the maximum overshoot occurs after T , i.e., during the transition from binomial to geometric state distribution. The relaxation time reaches values that are larger than T by an order of magnitude.

B. Non-stationary Service Curves

Next, we consider $S(\tau, t)$ as a non-stationary random service process. We define a bivariate envelope function $\mathcal{S}^\varepsilon(\tau, t)$ that conforms to

$$\mathbb{P}[S(\tau, t) \geq \mathcal{S}^\varepsilon(\tau, t), \forall \tau \in [0, t]] \geq 1 - \varepsilon \quad (10)$$

for all $t \geq 0$ where $\varepsilon \in (0, 1]$ is a probability of underflow. Adding $A(\tau)$ to both sides we have

$$\mathbb{P}[A(\tau) + S(\tau, t) \geq A(\tau) + \mathcal{S}^\varepsilon(\tau, t), \forall \tau \in [0, t]] \geq 1 - \varepsilon. \quad (11)$$

Since (11) makes a sample path argument for all $\tau \in [0, t]$, it follows that

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

Using (1) we conclude that

$$\mathbb{P}[D(t) \geq A \otimes \mathcal{S}^\varepsilon(t)] \geq 1 - \varepsilon. \quad (12)$$

We refer to $\mathcal{S}^\varepsilon(\tau, t)$ as non-stationary service curve. It extends the notion of effective service curve [11] that defines (12) for univariate functions $\mathcal{S}^\varepsilon(t - \tau)$. Compared to [11], the definition

of a bivariate function $\mathcal{S}^\varepsilon(\tau, t)$ provides a service guarantee that has the capability to consider transient changes over time.

Next, we use the negative MGF, respectively, Laplace transform of $S(\tau, t)$ denoted $M_S(-\theta, \tau, t) = \mathbb{E}[e^{-\theta S(\tau, t)}]$ to derive a non-stationary service curve. We show that the function

$$\mathcal{S}^\varepsilon(\tau, t) = -\frac{1}{\theta(\tau, t)} (\ln M_S(-\theta, \tau, t) + \rho(t-\tau) - \ln(\rho\varepsilon)), \quad (13)$$

where $\theta(\tau, t) > 0$ and $\rho \in (0, 1/\varepsilon]$ are free parameters, satisfies the service curve guarantee (12).

For completeness, we include the derivation of (13) that extends [27] to non-stationary processes. The derivation employs basic steps from the stochastic network calculus [12], [13]. We use the complementary formulation of (10)

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

to prove that $\mathcal{S}^\varepsilon(\tau, t)$ defined in (13) satisfies (10), or equivalently that $\xi \leq \varepsilon$. Using the union bound and Chernoff's lower bound $\mathbb{P}[X \leq x] \leq e^{\theta x} M_X(-\theta)$ for $\theta \geq 0$ it holds that

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

where $\theta(\tau, t) \geq 0$ is a set of free parameters. The case where $\tau = t$ is omitted since $S(t, t) = 0$ and $\mathcal{S}^\varepsilon(t, t) \leq 0$ by definition. By insertion of $\mathcal{S}^\varepsilon(\tau, t)$ from (13) it follows that

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

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

Random Sleep Scheduling: We extend the deterministic sleep scheduling model from Sec. II-A and consider a non-stationary work-conserving system with random service increments $Z(t)$ for $t \geq 0$. When entering sleep state, the system regenerates and wakes up after a random time $T \geq 0$, i.e., $Z(t) = 0$ for $t \in [0, T]$. The service process is computed as $S(\tau, t) = \sum_{v=\tau+1}^t Z(v)$ for all $t > \tau \geq 0$ and $S(t, t) = 0$ for all $t \geq 0$ [10]. To derive the MGF of $S(\tau, t)$, we first consider the number of usable time-slots in $(\tau, t]$, i.e., after time T

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

The MGF of $U(\tau, t)$ is composed of three terms

$$M_U(\theta, \tau, t) = e^{\theta(t-\tau)} \mathbb{P}[T \leq \tau] + \sum_{v=\tau+1}^t e^{\theta(t-v)} \mathbb{P}[T = v] + \mathbb{P}[T > t], \quad (14)$$

that correspond to the cases where the start of the service T occurs before and including τ , within $(\tau, t]$, and after t , respectively. Given the service increments $Z(t)$ for $t > T$ are iid with MGF $M_Z(\theta)$, the MGF of the service process is [22]

$$M_S(\theta, \tau, t) = \mathbb{E}[(M_Z(\theta))^{U(\tau, t)}] = M_U(\ln M_Z(\theta), \tau, t). \quad (15)$$

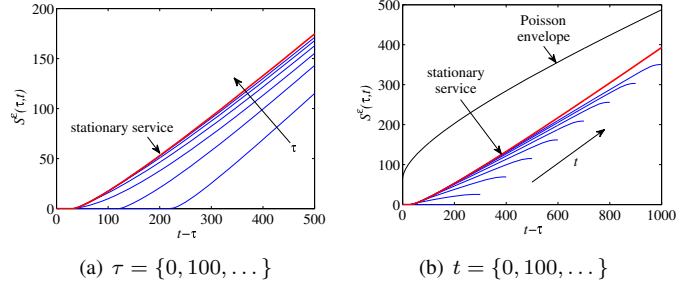


Fig. 3. Non-stationary service curves of random sleep scheduling.

For a concrete example, we model T as a geometric random variable with parameter p , where $\mathbb{P}[T = v] = p(1-p)^v$, and $Z(t)$ for $t > T$ as iid Bernoulli trials with parameter q . Due to the memorylessness of the processes, solutions for this specific example may also be derived, e.g., from a Markov model. We use this example as it enables us to compute certain reference results in Sec. III. Note that the service curve (13) in general is not limited to memoryless processes.

Regarding (14) we have $\mathbb{P}[T \leq \tau] = 1 - (1-p)^{\tau+1}$ and

$$\sum_{v=\tau+1}^t e^{\theta(t-v)} \mathbb{P}[T = v] = e^{\theta t} p \sum_{v=\tau+1}^t (e^{-\theta} (1-p))^v,$$

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

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

Having obtained a solution of (14), the MGF of the service process follows from (15) by insertion of $M_Z(\theta) = qe^\theta + 1 - q$ for a Bernoulli service increment process. Finally, the non-stationary service curve is computed from (13).

Fig. 3 illustrates $\mathcal{S}^\varepsilon(\tau, t)$ for $p = 0.1$ and $q = 0.5$. The remaining parameters of the service curve are $\varepsilon = 10^{-6}$, $\rho = 10^{-4}$, and $\theta(\tau, t)$ is optimized numerically. In Fig. 3(a), we show how the service in an interval of width $t-\tau$ increases with increasing distance τ from the last regeneration point. For small τ a significant impact of the initial transient phase is noticed. For large τ we observe that $\mathcal{S}^\varepsilon(\tau, t)$ converges towards a stationary service curve that is computed for a Bernoulli service increment process without sleep scheduling. Fig. 3(b) displays service curves for fixed t and variable τ . Here, the transient phase is reflected in the non-convex shape of the curves, rightwards where τ approaches zero. In contrast, the initial delay at the origin, that also applies to the stationary service curve, is caused by the Bernoulli service increment process. For small intervals $t-\tau$ the process results in a service of zero with non-negligible probability. For $t = 100$ and for $t = 200$ the effects bring about a service curve of zero.

The presentation of $\mathcal{S}^\varepsilon(\tau, t)$ in Fig. 3(b) conforms with the formulation of statistical performance bounds, where t is fixed and all $\tau \in [0, t]$ are evaluated. A statistical backlog bound follows from (12) with (8) as

$$\mathbb{P}\left[B(t) \leq \sup_{\tau \in [0, t]} \{\mathcal{A}^\varepsilon(t-\tau) - \mathcal{S}^\varepsilon(\tau, t)\}\right] \geq 1 - 2\varepsilon \quad (16)$$

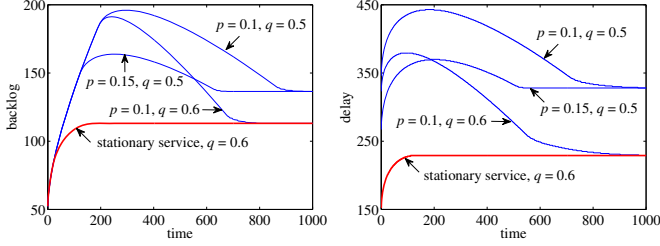


Fig. 4. Transient backlog and delay of random sleep scheduling.

and a first-come first-served delay bound as

$$\mathbb{P}\left[W(t) \leq \inf\left\{w \geq 0 : \sup_{\tau \in [0, t]} \{\mathcal{A}^\varepsilon(t - \tau) - \mathcal{S}^\varepsilon(\tau, t + w)\} \leq 0\right\}\right] \geq 1 - 2\varepsilon. \quad (17)$$

Intuitively, the backlog and delay bound are the maximal vertical and horizontal deviation of $\mathcal{A}^\varepsilon(t - \tau)$ and $\mathcal{S}^\varepsilon(\tau, t)$, respectively. For illustration, Fig. 3(b) includes the Poisson arrival envelope from Sec. II-A for $\alpha = 0.06$, $\beta = 0.3$, and $\varepsilon = 10^{-6}$. Corresponding backlog and delay bounds are shown in Fig. 4 for different parameters of the service p and q where the long-term utilization is $\alpha/(\beta q)$. The remaining parameters θ and ρ are optimized numerically. Compared to the stationary case, the transient overshoot of sleep scheduling is considerable.

III. MEASUREMENT-BASED ESTIMATION

Complementary to the model-based approach, methods for estimation of service curves from measurements of systems have received increasing attention in the past years [28]–[36]. A closely related research area [33] is available bandwidth estimation that seeks to estimate the long-term average unused capacity of a system or a network. In passive measurements [28], [33] the departures of a system are observed given the production traffic of the network, whereas in active probing [29]–[36] artificial test traffic is injected into the system. The different approaches can be further classified to be based either on the assumption of a work-conserving system [28]–[31] or more general min-plus systems [32]–[36] and either use a deterministic model [30]–[34] or a stochastic model [28], [29], [35], [36]. For a more detailed comparison see also [16]. While a method known as rate scanning is available for estimation of stationary service curves for the general class of min-plus linear systems with random service [35], [36], methods for estimation of transient service curves are not elaborated in the current literature.

Like [32]–[36], we consider min-plus linear systems, where (1) holds with equality, i.e., for all $t \geq 0$

$$D(t) = \inf_{\tau \in [0, t]} \{A(\tau) + S(\tau, t)\}. \quad (18)$$

The unknown $S(\tau, t)$ is a time-variant, random service process as in [35], [36]. Compared to [35], [36] we do, however, not assume stationarity of $S(\tau, t)$. Based on (18), the goal of service curve estimation can be phrased as an inversion problem, i.e., solving (18) for $S(\tau, t)$. Due to the infimum, the

min-plus convolution has no inverse operation in general. A solution can, however, be obtained for certain arrival functions $A(t)$ [33]. Using measurements of $S(\tau, t)$, we finally seek to estimate a maximal non-stationary service curve $\mathcal{S}^\varepsilon(\tau, t)$, i.e., find a maximal³ function $\mathcal{S}^\varepsilon(\tau, t)$ that satisfies (12). The measurements are performed assuming a regenerative service process as defined in Sec. II-A, where repeated network probes can observe samples of the service process $S_i(\tau, t)$ at regeneration point P_i as defined by (3).

The remainder of this section is structured as follows: We adapt two known methods, rate scanning (Sec. III-A) and burst response (Sec. III-B), for estimation of non-stationary service curves. Fundamental limitations of these methods are identified that are explained by the non-convexity and the super-additivity of the service (Sec. III-C), respectively. We devise a new two-phase probing method (Sec. III-D) that first determines the shape of a suitable probe, that is proven to be minimal under certain conditions. The probe is used to obtain a service curve estimate with a defined accuracy.

A. Rate Scanning

First, we consider the rate scanning method from [33], [35] that uses constant rate probes $A(t) = rt$ for a set of rates $r \in \mathbb{R}$. While [33] and [35] consider deterministic and stationary service curves, respectively, we adapt the method to provide estimates of non-stationary service curves for transient phases.

For constant rate probes $A(t) = rt$ the backlog of a system follows from (18) as $B(t) = \sup_{\tau \in [0, t]} \{r(t - \tau) - S(\tau, t)\}$. Consequently, it holds that $B(t) \geq r(t - \tau) - S(\tau, t)$ for all $\tau \in [0, t]$. Solving for $S(\tau, t)$ provides the lower bound $S(\tau, t) \geq r(t - \tau) - B(t)$ for all $\tau \in [0, t]$.

Next, we use the definition of $(1 - \xi)$ -quantile

$$X^\xi = \inf\{x \geq 0 : \mathbb{P}\{X \leq x\} \geq 1 - \xi\} \quad (19)$$

and denote $B^\xi(r, t)$ the backlog quantile at time t as a function of r . A non-stationary service curve (12) follows as

$$\mathcal{S}_{rs}^\varepsilon(\tau, t) = \max_{r \in \mathbb{R}} \{r(t - \tau) - B^\xi(r, t)\} \quad (20)$$

for $t \geq \tau \geq 0$. By application of the union bound, $\mathcal{S}_{rs}^\varepsilon(\tau, t)$ satisfies (10) with probability $\varepsilon = \sum_{r \in \mathbb{R}} \xi$ [35]. Compared to [35], (20) uses the transient instead of the stationary backlog to estimate a non-stationary service curve. The service curve that is defined by (20) has the form of a Legendre-Fenchel transform of the backlog [33]. The Legendre-Fenchel transform has a number of useful properties in the network calculus [32], [37]. It is also known as convex conjugate as the result is generally a convex function [38]. For convex functions, the Legendre-Fenchel transform is its own inverse, whereas the bi-conjugate of a non-convex function can only recover the convex hull of the original function [38].

In a practical implementation of rate scanning, the set of probing rates \mathbb{R} has to be selected. Options are, e.g., linearly or geometrically spaced rates combined with suitable tests

³We note that in general there is no unique maximal solution but only Pareto efficient solutions.

that determine the maximum probing rate [33], [35]. Then, estimates of the backlog quantile $B^\xi(r, t)$ for $t \geq 0$ are obtained from repeated measurements for each $r \in \mathbb{R}$. A service curve estimate $S_{r_s}^\varepsilon(\tau, t)$ follows from (20).

For evaluation of the method, we consider a simulation of the random sleep scheduling model from Sec. (II-B) with identical parameters $p = 0.1$ and $q = 0.5$, i.e., the stationary service rate is 0.5. We perform rate scanning with ten rates $r \in \{0.05, 0.1, \dots, 0.5\}$. For each rate we obtain 10^5 backlog samples from repeated experiments from which we extract an estimate of the backlog quantile $B^\xi(r, t)$ for $\xi = 10^{-4}$ so that $\varepsilon = \sum_{r \in \mathbb{R}} \xi = 10^{-3}$. Fig. 5(a) shows the linear segments obtained by each of the ten probing rates $r \in \mathbb{R}$ and the resulting estimate $S_{r_s}^\varepsilon(\tau, t)$ for $t = 200$.

As a reference, we include an analytical upper bound⁴ in Fig. 5(a). Any function that exceeds the upper bound for some $\tau \in [0, t]$ violates the definition of service envelope (10). Also, we show an analytical service curve⁵ for comparison. Clearly, the service curve estimate from rate scanning cannot recover the non-convex parts of the analytical results as it is fundamentally limited to a convex hull by construction.

B. Burst Response

Motivated by the min-plus systems theory of the network calculus [9], a canonical probe for system identification is the burst function $\delta(\tau)$ that is defined as

$$\delta(\tau) = \begin{cases} 0 & \text{for } \tau = 0, \\ \infty & \text{for } \tau > 0. \end{cases} \quad (21)$$

In min-plus algebra, the burst function takes the role of the Dirac delta function, i.e., it is the neutral element of min-plus convolution. Consequently, sending a burst probe $A(\tau) = \delta(\tau)$ reveals the service $S(0, t)$ for all $t \geq 0$ as the burst response of the system

$$D(t) = \inf_{\tau \in [0, t]} \{\delta(\tau) + S(\tau, t)\} = S(0, t). \quad (22)$$

For additive service processes as defined in [15, p. 6], $S(\tau, t)$ can be obtained for all $t \geq \tau \geq 0$ as

$$S(\tau, t) = S(0, t) - S(0, \tau). \quad (23)$$

For a stochastic analysis, we denote Ω the set of all feasible sample paths $D_\omega(t)$. For each $\omega \in \Omega$ we use the additivity (23) to obtain the service process from the burst response (22) as

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

for all $\tau \in [0, t]$. We fix $t > 0$ and select a subset of the sample paths $\Psi_t \subseteq \Omega$ with probability $\mathbb{P}[\Psi_t] \geq 1 - \varepsilon$. By definition of

$$S_{br}^\varepsilon(\tau, t) = \inf_{\psi \in \Psi_t} \{S_\psi(\tau, t)\}, \quad (24)$$

⁴For a Bernoulli service increment process with parameter q that starts after a geometrically distributed time T , an upper bound of the service provided in $[\tau, t]$ is derived as the $(1 - \varepsilon)$ -quantile of a binomial distribution with parameters $t - \max\{\tau, T\}$ and q . We note that the first parameter, i.e., the number of trials $t - \max\{\tau, T\}$, is a random variable.

⁵The analytical service curve again uses the binomial distribution but makes a sample path argument using the union bound in addition.

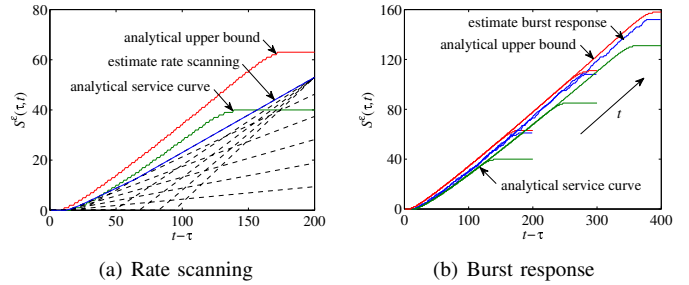


Fig. 5. Service curve estimates compared to analytical results. The estimate from rate scanning is the maximum of linear rate segments (dashed lines). By construction it can only recover a convex hull. Burst probing can estimate non-convex service curves and performs close to the analytical upper bound.

it holds that $S_\psi(\tau, t) \geq S_{br}^\varepsilon(\tau, t)$ for all $\tau \in [0, t]$ and all $\psi \in \Psi_t$. Further, we have $\mathbb{P}[\Psi_t] \geq 1 - \varepsilon$ so that $S_{br}^\varepsilon(\tau, t)$ satisfies (10). Hence, $S_{br}^\varepsilon(\tau, t)$ is a non-stationary service curve that conforms to (12).

In a practical probing scheme, we can only observe a finite set of sample paths Ω from repeated measurements of the departures $D_\omega(t)$ for $t \geq 0$. We fix $t > 0$ and estimate the service $S_\omega(\tau, t) = D_\omega(t) - D_\omega(\tau)$ for all $\tau \in [0, t]$ and $\omega \in \Omega$. To construct Ψ_t we remove a set of minimal sample paths from Ω . We define the minimal sample path ϕ to be the sample path that attains the minimum

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

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

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

The indicator function $1_{(\cdot)}$ is one if the argument is true and zero otherwise. We remove ϕ to obtain $\Psi_t = \Omega \setminus \phi$ and repeat the above steps for Ψ_t as long as $\mathbb{P}[\Psi_t] \geq 1 - \varepsilon$. Finally, we obtain the service curve estimate from (24) for all $\tau \in [0, t]$.

Fig. 5(b) shows service curve estimates from burst probing for $t = 200, 300$, and 400 together with analytical upper bounds and analytical reference service curves for the same system as used for Fig. 5(a). We observe that burst probing performs very well as it provides service curve estimates that closely follow the analytical upper bound.

Despite its good performance, burst probing has, however, limitations. Particularly, we discover that the intuitive assumption of additive service processes [15], that is a basis of the method, is not justifiable in general. A notable exception are networks of systems, where the end-to-end network service process $S^{net}(\tau, t) = S^1 \otimes S^2 \otimes \dots \otimes S^n(\tau, t)$ is derived by min-plus convolution of the service processes of the individual systems $S^i(\tau, t)$ for $i = 1, 2, \dots, n$. For additive S^i , the following Lemma 2 proves that S^{net} is super-additive in general and additive only in certain special cases. The result is significant since it refutes the assumption of additive service processes for the large class of tandem systems. It implies that the burst probing method is too optimistic as it estimates $S^{net}(\tau, t) = S^{net}(0, t) - S^{net}(0, \tau)$ from (23) whereas in

general only $S^{net}(\tau, t) \leq S^{net}(0, t) - S^{net}(0, \tau)$ holds due to the super-additivity of S^{net} . Besides, burst probes can be considered intrusive as they cause non-linear behavior of certain systems [33]. An important example are FIFO multiplexers, where a burst probe can preempt other traffic, resulting in a too optimistic service estimate.

C. Super-additivity of min and \otimes

The burst response method in Sec. III-B is applicable to additive service processes. The following lemmas show that the min and \otimes operators in general maintain only super-additivity but not additivity. A function $f(s, t)$ is super-additive if $f(s, u) \geq f(s, t) + f(t, u)$ for all $u \geq t \geq s \geq 0$. Trivially, an additive function $f(s, u) = f(s, t) + f(t, u)$ for all $u \geq t \geq s \geq 0$ is also super-additive.

Lemma 1 (Super-additivity of min). *Given two super-additive bivariate functions $f(s, t)$ and $g(s, t)$ for $t \geq s \geq 0$. The minimum $h(s, t) = \min\{f(s, t), g(s, t)\}$ is super-additive.*

We note that the minimum of two additive bivariate functions $f(s, t)$ and $g(s, t)$ for $t \geq s \geq 0$ is super-additive, as a special case of Lemma 1, but in general not additive. A counterexample is $f(s, t) = t - s$ and $g(s, t) = 2(\lfloor t/2 \rfloor - \lfloor s/2 \rfloor)$. Clearly f and g are additive, however, $h = \min\{f, g\}$ is not.

Proof of Lemma 1. By definition of h we have

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

In the second line we used the super-additivity of f and g and in the third line $\min\{f(s, u), g(s, u)\} = h(s, u)$ and $\min\{h(s, u), x\} \leq h(s, u)$ for any x . \square

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

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

We note that the convolution of two additive bivariate functions $f(s, t)$ and $g(s, t)$ for $t \geq s \geq 0$ is super-additive, but in general not additive. As a counterexample consider the additive functions $f(s, t) = t - s$ and $g(t, u) = 2(\lfloor u/2 \rfloor - \lfloor t/2 \rfloor)$ for which $h(s, u) = f \otimes g(s, u) = 2\lfloor u/2 \rfloor - s$ is not additive. To quantify this effect, we define the maximal deviation of a super-additive function $h(s, u)$ from additivity as

$$\Delta(s, u) := h(s, u) - \inf_{t \in [s, u]} \{h(s, t) + h(t, u)\}. \quad (25)$$

Additivity is, however, achieved in case of univariate functions, where Lemma 2 ii) extends a result for min-plus convolution of sub-additive univariate functions [9, p. 142].

In Fig. 6(a), we evaluate the network service process $S^{net}(\tau, t) = S^1 \otimes \dots \otimes S^n(\tau, t)$ for a tandem of $n = 2, 3, 4$ systems with random sleep scheduling as in Sec. II-B, i.e., the

service processes $S^i(\tau, t)$ are additive. We consider 10^5 sample paths and show the distribution of the relative deviation of $S^{net}(0, 400)$ from additivity, i.e., $\Delta(\tau, t)/S^{net}(\tau, t)$. Fig. 6(a) confirms significant deviations from additivity.

Proof of Lemma 2. By definition of h we have

$$\begin{aligned} h(s, t) + h(t, u) &= f \otimes g(s, t) + f \otimes g(t, u) \\ &= \inf_{\tau \in [s, t]} \inf_{v \in [t, u]} \{f(s, \tau) + f(t, v) + g(\tau, t) + g(v, u)\}. \quad (26) \end{aligned}$$

i) Given f and g are super-additive. From (26) we have

$$\begin{aligned} h(s, t) + h(t, u) &\leq \inf_{\tau \in [s, t]} \inf_{v \in [t, u]} \{f(s, v) - f(\tau, t) + g(\tau, t) + g(v, u)\} \\ &= \inf_{v \in [t, u]} \{f(s, v) + g(v, u)\} + \inf_{\tau \in [s, t]} \{g(\tau, t) - f(\tau, t)\} \\ &\leq \inf_{v \in [t, u]} \{f(s, v) + g(v, u)\}. \quad (27) \end{aligned}$$

In the first line, we estimated $f(s, \tau) + f(\tau, t) + f(t, v) \leq f(s, v)$ due to the super-additivity of f . In the second line, we rearranged the infima, and in the third line, we estimated $\inf_{\tau \in [s, t]} \{g(\tau, t) - f(\tau, t)\} \leq g(t, t) - f(t, t) = 0$ since $f(t, t), g(t, t) = 0$ for all $t \geq 0$. Similarly, using the super-additivity of g we derive from (26) that

$$h(s, t) + h(t, u) \leq \inf_{\tau \in [s, t]} \{f(s, \tau) + g(\tau, u)\}. \quad (28)$$

Combining (27) and (28) we obtain

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

which proves the super-additivity of h .

ii) For the special case of univariate functions $f(s, t) = f(t - s)$ and $g(s, t) = g(t - s)$, that depend only on the difference $t - s$ and not on the absolute values of s and t , it follows that $h(s, t) = f \otimes g(t - s) = h(t - s)$ is also univariate. Using the additivity of f and g , (26) yields that

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

which proves the additivity of h . \square

D. Minimal Probing

Next, we devise a new probing method to characterize the transient behavior of service curves that are neither convex nor additive. The method comprises two phases. In the first step, the method estimates a minimal probe as well as an upper bound of the service curve from burst probing. In the second step, the minimal probe is used to estimate a non-stationary service curve with a defined accuracy. We show that the minimal probe can reveal the service of the system, ideally without causing backlogs. The importance of the probe traffic intensity is also discussed in [33] and estimation methods for stationary systems that use minimal backlogging are [29], [39].

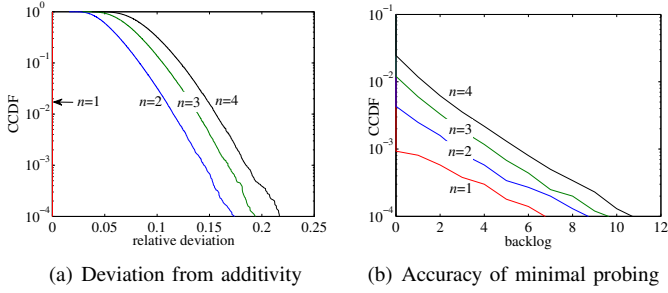


Fig. 6. Network of n systems with random sleep scheduling in series. (a) The network service process deviates from additivity. (b) Minimal probing achieves small backlogs, corresponding to a high accuracy of the estimate.

1) *Estimation using Arbitrary Probes:* First, we consider how to obtain a service curve estimate from an arbitrary probe. The estimate will then be used to derive conditions for the shape of a minimal probe. From (18) it follows that $D(t) \leq A(\tau) + S(\tau, t)$ for all $\tau \in [0, t]$ so that

$$S(\tau, t) \geq D(t) - A(\tau), \quad (29)$$

for all $\tau \in [0, t]$. An equivalent expression using the backlog is $S(\tau, t) \geq A(\tau, t) - B(t)$ for all $\tau \in [0, t]$. By insertion of the backlog quantile it follows that

$$S^\varepsilon(\tau, t) = A(\tau, t) - B^\varepsilon(t) \quad (30)$$

satisfies (10), i.e., $S^\varepsilon(\tau, t)$ is a non-stationary service curve (12). For the special case of $A(\tau, t) = r(t - \tau)$ the rate scanning method is recovered.

2) *Definition of Minimal Probes:* So far, we did not constrain the shape of the probes and defining a suitable probe is non-trivial. Intuitively, a too small probe will provide little information about the service as the observed departures are mostly limited by the arrivals [33]. A too large probe, on the other hand, will deteriorate the estimate, e.g., in the extreme case of a burst probe $A(\tau) = \delta(\tau)$, the lower bound in (29) will only be useful for $\tau = 0$ but not for $\tau > 0$. The same restriction applied to burst probing earlier. Consequently, we seek to find the minimal probe $A_{mp}(\tau)$ that satisfies $D(t) = \inf_{\tau \in [0, t]} \{A_{mp}(\tau) + S(\tau, t)\} = S(0, t)$ for a fixed $t > 0$, i.e., the minimal probe that allows estimating the service from observations of the departures. The minimal probe is

$$A_{mp}(\tau) = S(0, t) - S(\tau, t) \quad (31)$$

since any function $A(\tau)$ for which $\exists \tau \in [0, t] : A(\tau) < A_{mp}(\tau)$ will only reveal $D(t) < S(0, t)$. We note that $A_{mp}(\tau)$ maximizes the right hand side of (29), where (18) confirms that $D(t) - A_{mp}(\tau) = S(\tau, t)$. For any other larger or smaller probe $A(\tau) = A_{mp}(\tau) + f(\tau)$ it holds that $D(t) - A(\tau) = S(\tau, t) + \inf_{v \in [0, t]} \{f(v)\} - f(\tau) \leq S(\tau, t)$.

Since $A_{mp}(\tau)$ depends on the unknown service, it cannot be constructed a priori. To gather information that facilitates estimating $A_{mp}(\tau)$, we initially consider the system's burst response $D(t) = S(0, t)$ for $A(t) = \delta(t)$. An estimate of (31) is $\tilde{A}_{mp}(\tau) = S(0, \tau)$ that is exact if $S(\tau, t)$ is additive. In the stochastic case, we use $S_{br}^\varepsilon(\tau, t)$ from (24) to estimate

$$\tilde{A}_{mp}(\tau) = S_{br}^\varepsilon(0, t) - S_{br}^\varepsilon(\tau, t). \quad (32)$$

We note that additivity of $S_{br}^\varepsilon(\tau, t)$ cannot be assumed by construction of (24), see Lemma 1.

3) *Accuracy of the Estimates:* We start the investigation for a time-variant but otherwise deterministic system, where (29) provides a lower estimate of the service. Given the probe $\tilde{A}_{mp}(\tau) = S(0, \tau)$, where $S(0, \tau)$ is obtained as the system's burst response. By insertion into (29), a lower estimate is

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

where we used (18) to compute $D(t)$ for $\tilde{A}_{mp}(\tau)$. On the other hand, the burst response provides the upper estimate $S(\tau, t) \leq S(0, t) - S(0, \tau)$ if $S(\tau, t)$ is super-additive. Taking the difference of the upper and the lower estimate, the service process is bounded in an interval of width

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

that is the maximum deviation of $S(0, t)$ from additivity as defined in (25). Further, using $\tilde{A}_{mp}(t) = S(0, t)$ and (18),

$$\Delta(t) = \tilde{A}_{mp}(t) - D(t) = B(t),$$

i.e., the estimation accuracy can be directly seen from the backlog at the end of the probe. Consequently, if $S(\tau, t)$ is additive, $\Delta(t) = 0$, i.e., both the lower and the upper estimate recover $S(\tau, t)$ exactly and the backlog $B(\tau)$ that is caused by $\tilde{A}_{mp}(\tau)$ is zero during the entire probe, i.e., for all $\tau \in [0, t]$.

In the stochastic case, we obtain a non-stationary service curve estimate from (30) as $S_{mp}^\varepsilon(\tau, t) = \tilde{A}_{mp}(\tau, t) - B^\varepsilon(t)$ using the probe defined in (32). By use of (32) we have $\tilde{A}_{mp}(\tau, t) = \tilde{A}_{mp}(t) - \tilde{A}_{mp}(\tau) = S_{br}^\varepsilon(\tau, t)$ since $S_{br}^\varepsilon(t, t) = 0$ by definition. Note that we do not assume additivity of $S(\tau, t)$, so far. By insertion into (30), it holds that

$$S_{mp}^\varepsilon(\tau, t) = S_{br}^\varepsilon(\tau, t) - B^\varepsilon(t). \quad (33)$$

We conclude that $B^\varepsilon(t)$ observed by minimal probing is a measure of accuracy that separates the conservative estimate of minimal probing from the possibly too optimistic estimate of burst probing.

To investigate $B^\varepsilon(t)$, consider the backlog expression $B(t) = \sup_{\tau \in [0, t]} \{A(\tau, t) - S(\tau, t)\}$ that is derived from (18). By insertion of (32) we have $B(t) = \sup_{\tau \in [0, t]} \{S_{br}^\varepsilon(\tau, t) - S(\tau, t)\}$. By definition (24), $S_{br}^\varepsilon(\tau, t) = \inf_{\psi \in \Psi_t} \{S_\psi(0, t) - S_\psi(0, \tau)\}$ so that for any sample path $\varphi \in \Psi_t$

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

i.e., $B_\varphi(t)$ is bounded by the maximal deviation of $S_\varphi(0, t)$ from additivity. Finally, if $S(\tau, t)$ is additive, it follows that $B_\varphi(t) = 0$ for all $\varphi \in \Psi_t$. Since $\mathbb{P}[\Psi_t] \geq 1 - \varepsilon$ it holds that $B^\varepsilon(t) = 0$ and $S_{mp}^\varepsilon(\tau, t)$ recovers $S_{br}^\varepsilon(\tau, t)$ exactly.

Fig. 6(b) quantifies the distribution of $B(t)$ that is observed by minimal probing at $t = 400$. A network of $n = 1 \dots 4$

systems in series, each with random sleep scheduling as in Sec. II-B, is considered. The network service process is additive for $n = 1$, but not for $n > 1$ as confirmed by Fig. 6(a). Since in general, it is not known whether the service process is additive or not, estimates of burst probing are not reliable. Using minimal probing, we can either if $B^\varepsilon(t) = 0$ confirm the estimate of burst probing or otherwise obtain a conservative estimate with a defined accuracy that is given by $B^\varepsilon(t)$. In Fig. 6(b) we observe that $B^\varepsilon(t) > 0$ for $\varepsilon = 10^{-3}$ and $n > 1$, i.e., the estimate from burst probing is not confirmed. $B^\varepsilon(t)$ is, however, small, i.e., minimal probing is accurate. We omit showing further service curve estimates of minimal probing since, except for the additive offset $B^\varepsilon(t)$, they generally match the estimates of burst probing as determined by (33).

IV. CONCLUSIONS

This work contributed a notion of non-stationary service curves to the stochastic network calculus. We analyzed systems with sleep scheduling and provided insights into the transient behavior. Noticeable transient overshoots and large relaxation times were observed that have a significant impact on network performance. Due to its generality, the service curve model can include a variety of further systems. Beyond modelling, we considered measurement-based methods for identification of systems. We discovered that existing probing methods cannot accurately estimate non-stationary service curves. The difficulties were related to the non-convexity and super-additivity of the service. We devised a novel two-phase method that first estimates a minimal probe that is used in a second step to obtain an accurate service curve estimate. Simulation results confirmed the fidelity of the approach.

REFERENCES

- [1] M. Mellia, I. Stoica, and H. Zhang, "Tcp model for short lived flows," *IEEE Commun. Lett.*, vol. 6, no. 2, pp. 85–87, Feb. 2002.
- [2] N. Dukkipati *et al.*, "An argument for increasing TCP's initial congestion window," *ACM Comp. Com. Rev.*, vol. 40, no. 3, pp. 26–33, 2010.
- [3] N. Becker, A. Rizk, and M. Fidler, "A measurement study on the application-level performance of LTE," in *Proc. of IFIP Networking*, Jun. 2014.
- [4] S. M. Ross, *Introduction to Probability Models*, 9th ed. Academic Press, 2007.
- [5] J. Zhang and E. J. Coyle, "The transient solution of time-dependent M/M/1 queues," *IEEE Trans. Inf. Theory*, vol. 37, no. 6, pp. 1690–1696, Nov. 1991.
- [6] C.-Y. Wang *et al.*, "Transient behavior of ATM networks under overloads," in *Proc. of IEEE INFOCOM*, Mar. 1996, pp. 978–985.
- [7] E. de Souza e Silva and H. R. Gail, "An algorithm to calculate transient distributions of cumulative rate and impulse based reward," *Stochastic models*, vol. 14, no. 3, pp. 509–536, 1998.
- [8] A. Horváth, M. Paolieri, L. Ridi, and E. Vicario, "Transient analysis of non-Markovian models using stochastic state classes," *Performance Evaluation*, vol. 69, no. 7, pp. 315–335, 2012.
- [9] J.-Y. Le Boudec and P. Thiran, *Network Calculus A Theory of Deterministic Queuing Systems for the Internet*. Springer-Verlag, 2001.
- [10] C.-S. Chang, *Performance Guarantees in Communication Networks*. Springer-Verlag, 2000.
- [11] A. Burchard, J. Liebeherr, and S. Patek, "A min-plus calculus for end-to-end statistical service guarantees," *IEEE Trans. Inf. Theory*, vol. 52, no. 9, pp. 4105–4114, Aug. 2006.
- [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/ACM Trans. Netw.*, vol. 14, no. 6, pp. 2300–2312, Jun. 2006.
- [14] M. Fidler, "An end-to-end probabilistic network calculus with moment generating functions," in *Proc. of IWQoS*, Jun. 2006, pp. 261–270.
- [15] Y. Jiang and Y. Liu, *Stochastic Network Calculus*. Springer-Verlag, Sep. 2008.
- [16] 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.
- [17] F. Ciucu and J. Schmitt, "Perspectives on network calculus - no free lunch but still good value," in *Proc. of ACM SIGCOMM*, Aug. 2012, pp. 311–322.
- [18] M. Fidler and A. Rizk, "A guide to the stochastic network calculus," *IEEE Commun. Surveys Tuts.*, Jul. 2014, 10.1109/COMST.2014.2337060.
- [19] C.-S. Chang *et al.*, "A min, + system theory for constrained traffic regulation and dynamic service guarantees," *IEEE/ACM Trans. Netw.*, vol. 10, no. 6, pp. 805–817, Dec. 2002.
- [20] R. Agrawal, F. Baccelli, and R. Rajan, "An algebra for queueing networks with time-varying service and its application to the analysis of Integrated Service networks," *Mathematics of Operation Research*, vol. 29, no. 3, pp. 559–591, Aug. 2004.
- [21] C.-S. Chang and R. L. Cruz, "A time varying filtering theory for constrained traffic regulation and dynamic service guarantees," in *Proc. of IEEE INFOCOM*, Mar. 1999, pp. 63–70.
- [22] S. Ross, *A First Course in Probability*, 6th ed. Prentice Hall, 2002.
- [23] C. Arita and D. Yanagisawa, "Exclusive queueing process with discrete time," arXiv, Tech. Rep. arXiv:1008.4651v2, Nov. 2010.
- [24] A. Rizk and M. Fidler, "Non-asymptotic end-to-end performance bounds for networks with long range dependent FBM cross traffic," *Computer Networks*, vol. 56, no. 1, pp. 127–141, Jan. 2012.
- [25] J. Liebeherr, A. Burchard, and F. Ciucu, "Delay bounds in communication networks with heavy-tailed and self-similar traffic," *IEEE Trans. Inf. Theory*, vol. 58, no. 2, pp. 1010–1024, Feb. 2012.
- [26] F. Ciucu, "Network calculus delay bounds in queueing networks with exact solutions," in *Proc. of ITC-20*, ser. LNCS, vol. 4516. Springer, Jun. 2007, pp. 495–506.
- [27] M. Fidler, R. Lübben, and N. Becker, "Capacity-Delay-Error-Boundaries: A Composable Model of Sources and Systems," *IEEE Trans. Wireless Commun.*, vol. 14, no. 3, pp. 1280–1294, Mar. 2015.
- [28] C. Cetinkaya, V. Kanodia, and E. W. Knightly, "Scalable services via egress admission control," *IEEE Trans. Multimedia*, vol. 3, no. 1, pp. 69–81, Mar. 2001.
- [29] S. Valaee and B. Li, "Distributed call admission control for ad hoc networks," in *Proc. of IEEE VTC*, Sep. 2002, pp. 1244–1248.
- [30] A. Undheim, Y. Jiang, and P. J. Emstad, "Network calculus approach to router modeling with external measurements," in *Proc. of CHINACOM*, Aug. 2007, pp. 276–280.
- [31] M. Bredel, Z. Bozakov, and M. Fidler, "Analyzing router performance using network calculus with external measurements," in *Proc. of IEEE IWQoS*, Jun. 2010.
- [32] T. Hisakado *et al.*, "Characterization of a simple communication network using Legendre transform," in *Proc. of ISCAS*, May 2003, pp. 738–741.
- [33] J. Liebeherr, M. Fidler, and S. Valaee, "A system theoretic approach to bandwidth estimation," *IEEE/ACM Trans. Netw.*, vol. 18, no. 4, pp. 1040–1053, Aug. 2010.
- [34] A. Rizk and M. Fidler, "On the identifiability of link service curves from end-host measurements," in *Proc. of Euro-NF NET-COOP*, ser. LNCS, vol. 5425, Sep. 2008, pp. 53–61.
- [35] 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.
- [36] R. Lübben, "System identification of computer networks with random service," Ph.D. dissertation, Leibniz Universität Hannover, 2013.
- [37] M. Fidler and S. Recker, "Conjugate network calculus: A dual approach applying the Legendre transform," *Computer Networks*, vol. 50, no. 8, pp. 1026–1039, Jun. 2006.
- [38] R. T. Rockafellar, *Convex Analysis*. Princeton University Press, 1972.
- [39] S. Y. Nam, S. Kim, and W. Park, "Analysis of minimal backlogging-based available bandwidth estimation mechanism," *Computer Communications*, vol. 35, no. 4, pp. 431–443, Feb. 2002.