Towards the Analysis of Transient Phases with Stochastic Network Calculus

Michael Beck
University of Kaiserslautern, Germany
Distributed Computer Systems Lab (DISCO)
beck@cs.uni-kl.de

***Abstract*—Current analysis methods for queueing systems mostly aim at steady-state results. However, many applications demand results for systems with a transient behavior. In communication networks such transient phases arise from sleep scheduling or the slow-start phase of TCP. In both examples the service offered is separated into a transient and a steady-state phase. In a larger context also the arrivals to a queueing system can show transient behavior. The demand or availability of energy in smart grids, the pool of data generated after the map phase in big data applications, or the arrivals in public transportation networks are a few examples. Analyzing these kind of systems in a steady-state fashion (for example via queueing theory) ignores their behavior in the transient phases. On the other side the analysis of transient phases bears additional challenges. This paper uses stochastic network calculus to describe the non-stationary behavior of queueing systems. With the help of bivariate arrival- and service envelopes time-variant performance bounds are constructed. Two numerical examples exemplify the framework created by these envelopes and compare it to the time-invariant analysis.**

I. Introduction

Many instances of queueing systems contain a transient phase. In such a phase a significant change to the system's steady-state behavior occurs. Both, the arrival or the service process, can be responsible for a transient phase: Transient phases in the arrivals occur for example when a singular event spawns a large number of jobs. Here the timing and the size of the burst can be either deterministic or random. Depleting such job pools [2] results in a transient phase, which ends after all arrivals of the burst are processed. After that the system might renew by accepting the next job pool or converge to a steady-state. In [2] the result of the map phase in big data applications is given as one example for such a job pool. In the broader context of queueing networks also applications in caching networks, traffic engineering, public transport networks, or even evacuation scenarios spring to mind as further examples of job pools.

Transient phases are not limited to the arrival process, but extend to the service process as well. An omnipresent example is the slow-start phase of TCP [15]. Similarly the setup time – for example due to sleep scheduling – of service elements mark another transient phase [1]. Moving more in the direction of $M/M/m/$Setup-systems one might also think of additional service being provided under high loads. This extra service can also be of limited nature. An example could be the reserve energy of a wireless sensor node, harvested in the sleep cycle from renewable sources. Down-times of servers due to repairs form another kind of transient phase. Again all of these phases can, but do not have to, transition into a steady-state.

We see the variety of transient effects on a queueing system is rich. Queueing theory on the other side focuses to analyze the steady-state of a system. Traditionally the queueing system is modeled by recurrent Markov chains and the resulting performance measures are time-invariant. In contrast to that we are interested in time-dependent bounds for transient phases. There exist few results for such an analysis (e.g. [17], [16], [9], [2]), but they are specific to particular problems or queueing systems. A general method, which captures a system's performance in transient phases *and* in its transition to the steady-state is missing.

Stochastic Network Calculus (SNC) [3], [12], [5], [13], [10], [4], [8] offers a alternative approach to analyze queueing systems. This methodology unites concepts from the theory of effective bandwidths (as in [11]) and its predecessor deterministic network calculus (see for example [12], [7]). Typical performance bounds of SNC take the form

$$\mathbb{P}(\mathfrak{b}(t) > x) \leq \varepsilon, \tag{1}$$

where $\mathfrak{b}(t)$ denotes the backlog of the system at time t. We see that these bounds depend on time t at which the system is evaluated. This hints already towards an analysis that might take the different behavior in transient phases into account. Indeed large parts of SNC operate on a bivariate – and hence time-variant – notation. However, the methodology itself eventually retreats to *univariate* – and hence *time-invariant* – bounding functions. As an effect SNC does not resolve transient behavior either. In fact, the above formulated bound, which depends on t, is in literature mostly replaced by a time-invariant version:

$$\lim_{t\to\infty} \mathbb{P}(\mathfrak{b}(t) > x) \leq \varepsilon.$$

The recent paper of Becker and Fidler [1] describes a performance bound which is truly time-variant, though. The therein discussed scenario consists of a service element with sleep scheduling. The wake up time of the service element forms a transient phase. The crucial difference to previous results is, that the bivariate description of the service element is preserved throughout the analysis. This allows to analyze transient phases as well as steady-states of the queueing system with a single method.

This paper extends these bivariate descriptions to arrivals as well. Furthermore the resulting functions are combined into stochastic bounds on the backlog and the virtual delay of a system, as in (1). We present two numerical examples with transient phases. In these the transient phases are reflected directly in the achieved performance bounds. The new bounds lie between two previously available results: 1) a direct application of SNC to the transient phase and 2) an analysis of the system, which neglects the transient phase completely. Eventually these examples showcase the wide variety of transient phases, which can be analyzed by this framework.

II. Stochastic Network Calculus

We introduce the notations and results needed from network calculus. For a detailed introduction see for example [3], [10], [8]. For the ease of presentation we consider time to be slotted and arrivals – such as a stream of data – use the fluid model. Such an arrival flow is defined via the accumulated functions $A(t) = \sum_{s=1}^{t} a(s)$. Here $A(t)$ denotes the number of arrivals up to time t and $a(s)$ is the increment in time-slot $(s-1, s]$. The bivariate extension of A, defined by $A(s,t) := A(t) - A(s)$, simplifies later notations. The function A describes a dimensionless quantity, which abstracts the arrivals to a system (e.g. bits, jobs, or packets).

Now let U be a bivariate function, such that $U(s,t) \leq U(s,t')$ for all $t \leq t'$. The system processing the arrivals is defined as a dynamic U-server, if for all A and time-pairs $s \leq t$ holds

$$D(t) \geq \min_{0 \leq s \leq t} \{A(s) + U(s,t)\}. \tag{2}$$

Here D denotes the cumulative departures of the system. The right-hand side of Equation (2) is known as the min-plus convolution of A and U at $(0,t)$. This operator is defined and denoted by

$$A \otimes U(s,t) = \min_{s \leq r \leq t} \{A(s,r) + U(r,t)\}.$$

The definition of a dynamic U-server is motivated by Lindley's equation. This equation describes the backlog $\mathfrak{b} := A(t) - D(t)$ at time t for a constant rate server with rate u. Its implicit form reads

$$\mathfrak{b}(t) = \max\{0, \mathfrak{b}(t-1) + a(t) - u\}.$$

Equation (2) generalizes above by keeping the service rate variable and replacing the equality by an inequality.

A dynamic U-server bounds the backlog of the system by

$$\mathfrak{b}(t) \overset{(2)}{\leq} A(t) - A \otimes D(0,t) = \max_{0 \leq s \leq t} \{A(s,t) - U(s,t)\}.$$

When we analyze a queueing system the arrival and service process are usually random. Hence, to achieve a quantifiable statement about $\mathfrak{b}(t)$ with the above, more information on A and U is needed.

In the following A denotes a stochastic process indexed by $\mathbb{N}$ and U denotes a stochastic process indexed by the pairs $s \leq t$ with $s, t \in \mathbb{N}$. The information we need to quantify $\mathfrak{b}(t)$ is given in the moment generating function (MGF) of A and U. Denote the MGF of a random variable X by $\phi_X(\theta) := \mathbb{E}(e^{\theta X})$.

Definition 1. *The flow of arrivals A (the service U) has a* univariate *bound $f(\theta, t-s)$ ($g(-\theta, t-s)$) for some $\theta > 0$ and pair $s \leq t$, if its MGF fulfills*

$$\phi_{A(s,t)}(\theta) \leq f(\theta, t-s) \quad \big(\phi_{U(s,t)}(-\theta) \leq g(-\theta, t-s)\big).$$

The flow of arrivals A (the service U) has a bivariate *bound $f(\theta, s, t)$ ($g(-\theta, s, t)$) for some $\theta > 0$ and pair $s \leq t$, if its MGF fulfills*

$$\phi_{A(s,t)}(\theta) \leq f(\theta, s, t) \quad \big(\phi_{U(s,t)}(-\theta) \leq g(-\theta, s, t)\big).$$

Note that the bound on the service is indeed a lower bound, as the MGF is evaluated for negative values. Clearly the second pair of definitions generalizes the first one.

The following example is used later on in Section IV.

Example 2. *We construct a Markov-modulated On-Off (MMOO) flow by defining a discrete-time Markov chain X_t on the states $\{0,1\}$. Further define the increment process I_t of i.i.d. variables. The arrivals increments are now defined by $a(t) = X_t I_t$. The MGF of A is $f(\theta, t-s)$-bounded (e.g. [3]) with*

$$f(\theta, t-s) := \max_{k \in \{0,1\}} E_k \cdot \frac{\max_{i \in \{0,1\}} v_i}{\min_{i \in \{0,1\}} v_i} \cdot \pi(E \cdot T)^{t-s-1}.$$

Here $E_i := \mathbb{E}(e^{\theta a(t)} \mid X_t = i)$ and E is the diagonal matrix with entries E_i. Further T is the transition matrix of the Markov chain, v_i is a positive eigenvector of ET, and π is the spectral radius of a matrix. Similarly one can define more general Markov-modulated arrivals.

The following performance bound is found for example in [3], [5]. It uses univariate bounds on $\phi_{A(s,t)}$ and $\phi_{U(s,t)}$ to bound the backlog or the virtual delay of a system. For an input-output-pair A and D the virtual delay is defined by

$$\mathfrak{d}(t) = \min\{t' \in \mathbb{N}_0 \mid A(t) \leq D(t+t')\}.$$

Theorem 3. *Fix some $\theta > 0$ and assume A and U to be stochastically independent. If A and U are bounded for all $s \leq t + T$ by $f(\theta, t-s)$ and $g(-\theta, t-s)$, respectively, it holds*

$$\mathbb{P}(\mathfrak{b}(t) > x) \leq e^{-\theta x} \sum_{s=0}^{t} f(\theta, s) \cdot g(-\theta, s) \tag{3}$$

$$\mathbb{P}(\mathfrak{d}(t) > T) \leq \sum_{s=0}^{t+T} f(\theta, t-s) \cdot g(-\theta, t+T-s) \tag{4}$$

for all $x > 0$ and all $t, T \in \mathbb{N}$.

The proof is a notational variation of the one given for example in [5]. We use this bound as a representative for the time-invariant view on transient phases.

At this point the question rises if all one needs to do to generalize Theorem 3 is to replace the functions f and g by their bivariate versions. This is however not sufficient. The structure of (3) and (4) enforces the violation probabilities to increase strictly in t, no matter the actual behavior of the underlying system. Hence a substituting f and g in Theorem 3 by their bivariate counterparts is not sufficient to capture the transient phases of the queue.

III. Non-Stationary Arrival and Service Curves

To achieve time-variant performance bounds we need to construct bivariate envelope functions from the MGF-bounds. Becker and Fidler present this method for the service element in [1]. The following theorem extends this to arrivals as well.

Theorem 4. *Fix some $t \in \mathbb{N}$ and assume A is $f(\theta, s, t)$-bounded for all $\theta \in \Theta$ (U is $g(-\theta, s, t)$-bounded) and $s \leq t$. Then A (U) has the following bivariate envelope:*

$$\mathbb{P}\Big(\bigcap_{s=0}^{t} A(s,t) \leq \mathcal{A}^{\varepsilon}(s,t)\Big) \geq 1 - \varepsilon.$$

$$\left(\mathbb{P}\Big(\bigcap_{s=0}^{t} U(s,t) \geq \mathcal{U}^{\varepsilon}(s,t)\Big) \geq 1 - \varepsilon\right)$$

with

$$\mathcal{A}^{\varepsilon}(s,t) = \inf_{\substack{\theta \in \Theta \\ 0 < \delta < \varepsilon^{-1}}} \tfrac{1}{\theta}\left(\log f(\theta,s,t) + \delta(t-s) - \log(\delta\varepsilon)\right). \tag{5}$$

$$\left(\mathcal{U}^{\varepsilon}(s,t) = \sup_{\theta,\delta > 0}\left\{\tfrac{1}{\theta}\left(\log(\delta\varepsilon) - \delta(t-s) - \log g(-\theta,s,t)\right)\right\}\right)$$

Proof. We only construct the envelope for A, as the construction of $\mathcal{U}^{\varepsilon}$ is analogue (see also [1]). Fix some $\theta \in \Theta$ and $\delta, \varepsilon > 0$ such that $\delta\varepsilon < 1$ and define $\mathcal{A}^{\varepsilon}_{\theta,\delta}(s,t)$ as in (5). Then

$$\mathbb{P}\Big(\bigcup_{s=0}^{t} A(s,t) > \mathcal{A}^{\varepsilon}_{\theta,\delta}(s,t)\Big) \leq \sum_{s=0}^{t-1} \mathbb{P}(A(s,t) > \mathcal{A}^{\varepsilon}_{\theta,\delta}(s,t))$$
$$\leq \sum_{s=0}^{t-1} \phi_{A(s,t)}(\theta) e^{-\theta\mathcal{A}^{\varepsilon}_{\theta,\delta}(s,t)} \leq \sum_{s=0}^{t-1} \frac{\phi_{A(s,t)}(\theta)}{f(\theta,s,t)} e^{-\delta(t-s)}\delta\varepsilon$$
$$\leq \delta\varepsilon \sum_{s=0}^{t-1} e^{-\delta(t-s)} \leq \varepsilon.$$

The inequalities are achieved by successively applying the union-bound, Chernoff's inequality, and the definition of $\mathcal{A}^{\varepsilon}_{\theta,\delta}$. The last step bounds the sum from above by $\int_0^{\infty} e^{-\delta x}\mathrm{d}x$. The range of the union in the first line can be limited to the values $\{0, \ldots, t-1\}$ as $A(t,t) = 0 < \mathcal{A}^{\varepsilon}_{\theta,\delta}(t,t)$. Minimizing over all choices for θ and δ completes the proof. □

We now constructs stochastic performance bounds on the backlog and the virtual delay from the bivariate envelopes.

Theorem 5. *Fix some t. Let A be $f(\theta, s, t)$-bounded for all $\theta \in \Theta$ and $s \leq t$ and U be $g(-\theta, s, t)$-bounded for all $\theta > 0$ and $s \leq t$. Then holds*

$$\mathbb{P}(\mathfrak{b}(t) \leq \max_{0 \leq s \leq t}\{\mathcal{A}^{\varepsilon}(s,t) + \mathcal{U}^{\varepsilon}(s,t)\}) \leq 1 - 2\varepsilon \tag{6}$$

for all $\varepsilon > 0$.

If U is $g(-\theta, s, t)$-bounded for all $t \in \mathbb{N}$, then also holds

$$\mathbb{P}(\mathfrak{d}(t) \leq \min\{T' \in \mathbb{N}_0 \mid \mathcal{A}^{\varepsilon} \oslash \mathcal{U}^{\varepsilon}(t+T', t) \leq 0\}) \leq 1 - 2\varepsilon$$

for all $\varepsilon > 0$, where

$$\mathcal{A}^{\varepsilon} \oslash \mathcal{U}^{\varepsilon}(t+T', t) := \max_{0 \leq s \leq t+T'}\{\mathcal{A}^{\varepsilon}(s,t) - \mathcal{U}^{\varepsilon}(s, t+T')\}.$$

Proof. For brevity we give here the proof for the delay-bound only. Equation (6) is a generalization of the bound in [1].

First fix some $t \in \mathbb{N}$ and $\varepsilon > 0$ and define

$$T := \min\{T' \in \mathbb{N}_0 \mid \mathcal{A}^{\varepsilon} \oslash \mathcal{U}^{\varepsilon}(t+T', t) \leq 0\}.$$

Assume for a while

$$A(s,t) \leq \mathcal{A}^{\varepsilon}(s,t) \quad \text{for all } s \leq t$$
$$U(s,t+T) \geq \mathcal{U}^{\varepsilon}(s,t+T) \quad \text{for all } s \leq t+T.$$

Then holds

$$\begin{aligned} 0 &\geq \max_{0 \leq s \leq t+T}\{\mathcal{A}^{\varepsilon}(s,t) - \mathcal{U}^{\varepsilon}(s,t+T)\} \\ &\geq \max_{0 \leq s \leq t+T}\{A(s,t) - U(s,t+T)\} \\ &= A(t) - A \otimes U(0, t+T) \geq A(t) - D(t, t+T), \end{aligned}$$

from which follows $\mathfrak{d}(t) \leq T$. Rewriting this implication as probabilities we get

$$\begin{aligned} \mathbb{P}(\mathfrak{d}(t) \leq T) \geq &\,\mathbb{P}\Big(\bigcap_{s=0}^{t} A(s,t) \leq \mathcal{A}^{\varepsilon}(s,t) \\ &\quad \text{and} \bigcap_{s=0}^{t+T} U(s,t+T) \geq \mathcal{U}^{\varepsilon}(s,t+T)\Big) \\ \geq &\,1 - 2\varepsilon \end{aligned}$$

□

Note that the bivariate $\mathcal{A}^{\varepsilon}$ must be extended to pairs s, t with $s > t$ to determine T. This is achieved by extending the MGF-bounds for such values as well. Indeed the resulting $\mathcal{A}^{\varepsilon}$ generally becomes negative for such values. This ensures the existence of a finite T for each $t \in \mathbb{N}$.

IV. Numerical Evaluation

We now describe two scenarios with transient phases and analyze them with the help of Theorems 3 and 5.

A. Scenario 1

The steady-state of this scenario is a queueing system, which processes MMOO arrivals as presented in Example 2. To model variations in the available service we assume $U(t, t{+}1)$ $(t = 0, \ldots,)$ to be an i.i.d. sequence of exponentially distributed random variables with parameter λ. For more realistic – and elaborate – models, which could be used here instead, see for example [6], [10], [14].

We now add two transient phases to this system: processing a large burst of arrivals, which arrives at time $t = 1$, forms the first transient phase. Furthermore the system reacts by activating additional, yet limited, resources. The depletion of this extra-service forms the second transient phase. In the

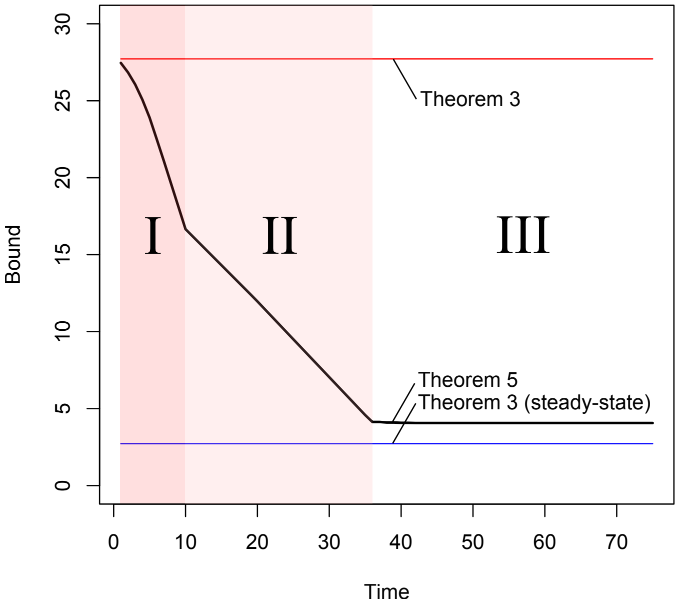

Fig. 1. Numerical Evaluation of Scenario 1 with deterministic $C = 10u_C$. The burst B is set to 25. The graph shows a bound on the backlog of the system, which is broken with a probability of at most $\varepsilon = 10^{-3}$. In the first transient phase (denoted by I) the reserves are consumed. In the second transient phase (denoted by II) the burst is processed, while the reserves are emptied. Eventually the system enters the steady-state in III.

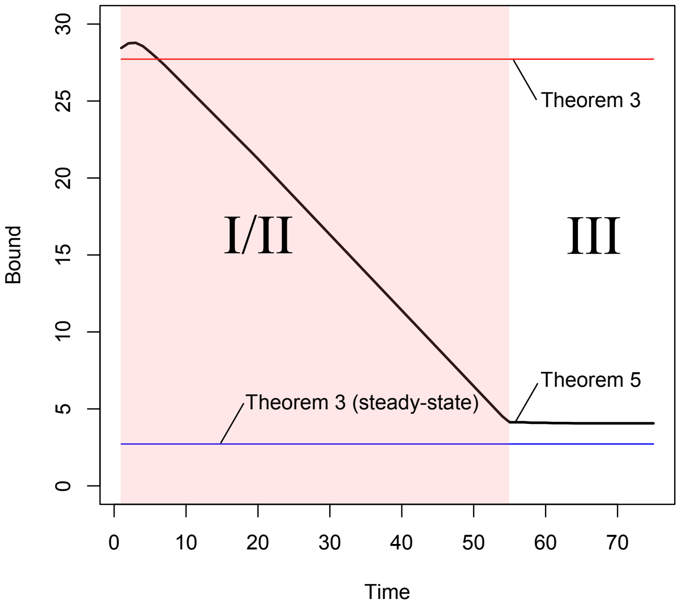

Fig. 2. Numerical Evaluation of Scenario 1 with random C and burst $B = 25$. The graph shows a bound on the backlog of the system, which is broken with a probability of at most $\varepsilon = 10^{-3}$. The transient phases I and II are smoothed out, due to C being a random variable. The system enters the steady-state III after processing the burst B.

beginning both phases overlap. The question which of the transient phases ends first depends on the size of the burst and the reserves, respectively.

To apply Theorems 3 and 5 we must bound the total arrival and service process.

1) Bounding the Total Arrivals: In addition to the Markov-modulated arrivals we have a burst of size B arriving at time $t = 1$. Hence the increments of the total arrivals A_{tot} are $a_{tot}(1) = a(1) + B, a_{tot}(2) = a(2), a_{tot}(3) = a(3), \ldots$, where B is a random variable independent of A. We can model and bound A_{tot} in different ways. First we could construct a Markov chain, which equals the Markov chain of A, but with a transient burst state added. The initial distribution of X would be set, such that the chain starts in the burst-state with probability one. After one time step the chain would move to the recurrent subchain of A. As a result we would see a burst in time-step 1 and the usual behavior of A afterwards. However, finding corresponding MGF-bounds is harder in this scenario compared to Example 2, as the involved Markov chain is reducible.

The total arrivals A_{tot} are easier bounded by exploiting $\phi_{A_{tot}(s,t)}(\theta) = \phi_{A(s,t)}(\theta)\phi_{B(s,t)}(\theta)$, where $B(s,t) = B\mathbf{1}_{\{s=0\}}$. Now A can be bounded by $f_A(\theta, t-s)$ as in Example 2 and all missing is a bound on $\phi_{B(s,t)}(\theta)$ to achieve $\phi_{A_{tot}(s,t)}(\theta) \leq f_B(\theta, s, t) f_A(\theta, s, t)$. An easy – and univariate – MGF-Bound on B is $\phi_{B(s,t)}(\theta) = \phi_B(\theta)$, resulting in

$$\phi_{A_{tot}(s,t)}(\theta) \leq \phi_B(\theta) \max_{k \in \{0,1\}} E_k \frac{\max_{i\in\{0,1\}} v_i}{\min_{i\in\{0,1\}} v_i} \pi (E \cdot T)^{t-s-1},$$

where the notations are the same as in Example 2.

In contrast a bivariate bound on the MGF of A_{tot} replaces the factor $\phi_B(\theta)$ in the above by $\phi_{B(s,t)}(\theta)$.

2) Bounding the Service: We move now to the service description. For simplicity the setup-time is set to zero (see the second scenario for the modeling of setup-times). Similarly to the arrivals we can split the service into two parts: the time-invariant service U and the additional service taken from the reserves. The capacity of the reserves is given by the random variable C and we assume they are tapped with a deterministic rate u_C. As a result the reserves are depleted at time $\frac{C}{u_C}$. Further $C(s,t) = u_C \max\{0, \min\{t, t_C\} - s\}$ describes an upper bound on the additional service in the interval $(s,t]$, where $t_C := \lfloor \frac{C}{u_C} \rfloor$.

The MGF of the total service splits into $\phi_{U_{tot}(s,t)}(-\theta) = \phi_{U(s,t)}(-\theta)\phi_{C(s,t)}(-\theta)$. However, if we want to find a univariate $g_C(-\theta, t-s)$ with $\phi_{C(s,t)} \leq g_C(-\theta, t-s)$ we see, that only $g_C = 1$ fulfills this inequality for all pairs s, t. This is due to the fact, that the considered interval $(s,t]$ can always be shifted beyond time t_C, such that $C(s,t) = 0$. In expression, the univariate formulation cannot resolve the transient phase of the service.

For a bivariate g_C consider the events $\{\lfloor \frac{C}{u_C} \rfloor = t\}$ and denote their probabilities by p_t. Such an event means there is enough capacity to provide t time-slots (but no more) of

TABLE I
PARAMETERS FOR SCENARIO 1 AND 2

Parameter	Value
Transition Probability T_{00}	0.7
Transition Probability T_{11}	0.7
Arrival Rate in On-State λ_A	4
Service Rate λ_U	0.25
Rate of Additional Service u_C	1
Expected Capacity $1/\lambda_C$ (Scenario 1 only)	10
Expected Setup-Time $1/p_W$ (Scenario 2 only)	20

additional service. Then holds

$$\begin{aligned}
&\phi_{U_C(s,t)}(-\theta)\\
&= \sum_{t_C=0}^{\infty} \mathbb{E}(e^{-\theta U_C(s,t)} \mid \lfloor \tfrac{C}{u_C} \rfloor = t_C)p_{t_C}\\
&\leq \sum_{t_C=0}^{\infty} \mathbb{E}(e^{-\theta \max\{0,\min\{t,t_C\}-s\}u_C} \mid \lfloor \tfrac{C}{u_C} \rfloor = t_C)p_{t_C}\\
&= \sum_{t_C=0}^{s} p_{t_C} + \sum_{t_C=s+1}^{t} e^{\theta(t_C-s)u_C}p_{t_C} + \sum_{t_C=t+1}^{\infty} e^{\theta(t-s)u_C}p_{t_C}
\end{aligned}$$

for all $s \leq t$.

The above bound can be used to construct a bivariate envelope for U_{tot} (Theorem 4). However, this requires knowledge about the distribution of C. In this simple scenario we assume C to be exponentially distributed with parameter λ_C. Inserting the distribution of C and using the formula for geometric sums the above becomes

$$\begin{aligned}
\phi_{U_C(s,t)}(-\theta) \leq{}& (1 - e^{-\lambda_C u_C(s+1)})\\
&+ (1 - e^{(t-s)(\theta u_C - \lambda_C u_C)})e^{u_C(\theta-\lambda_C-\lambda_C s)}\\
&+ e^{\theta(t-s)u_C}e^{-\lambda_C u_C(t+1)}.
\end{aligned}$$

3) Evaluation: With the bivariate envelopes for arrivals and service in place we can now analyze the system.

Figure 1 presents the numerical results for this scenario (for a full list of the used parameters see Table I). It shows a backlog bound in the sense of Equation (1) for varying t and a fixed violation probability of $\varepsilon = 10^{-3}$. For this graph the quantities B and C are chosen non-random. The two transient phases can be easily identified in that case. Figure 2, shows the same scenario, but with C exponentially distributed instead. We see the transient phases are smoothed out by doing so.

For comparison we included two univariate bounds: the red lines show the backlog bound, when applying Theorem 3 directly. We see, that the time-invariant bounds cannot resolve the transient phase: The backlog-bound does not decrease as the system evolves. The blue lines represent a time-invariant analysis but for a system without the transient phases. These bounds exclude the additional burst and reserves. We actually see a slight improvement compared to the results of Theorem 5, even after the steady-state is reached. This effect is due to the parameter δ in Theorem 4, which effectively reduces the long-term service rate slightly and similarly increases the arrivals long-term rate. This gap is the price we pay to capture the transient behavior at the beginning of the system's evolution.

B. Scenario 2

For this scenario we use the same steady-state system as in Scenario 1.

Again we add two transient effects. The first is an initial burst as before. The second is the activation of additional resources, which happens after some setup time W. This is a generalization to the model in [1]. Indeed if the service in the "steady-state" would be zero our model would reduce to the one in [1]. As in [1] we model the setup time by a random variable W, which is geometrically distributed with parameter p_W. After the setup time the service element has access to additional service V.

For the ease of presentation we assume again i.i.d. exponentially distributed increments with parameter λ_V.

How to bound A_{tot} was discussed in the previous scenario already. We focus hence on the bounding of

$$\phi_{U_{tot}(s,t)}(-\theta) = \phi_{U(s,t)}(-\theta)\phi_{V(s,t)}(-\theta).$$

For a univariate bound observe first, that $V(s,t) \geq V(0,t-s)$ as the setup-time always starts in $t = 0$. Hence we have the univariate bound

$$\begin{aligned}
&\phi_{V(s,t)}(-\theta) \leq \phi_{V(0,t-s)}(-\theta)\\
&= \sum_{r=0}^{\infty} \mathbb{E}(e^{-\theta V(s,t)} \mid W = r)\mathbb{P}(W = r)\\
&= \sum_{r=0}^{t-s-1} \Big(\frac{\lambda}{\lambda+\theta}\Big)^{t-s-r}(1-p_W)^r p_W + \sum_{r=t-s}^{\infty}(1-p_W)^r p_W\\
&= p_W\Big(\frac{\lambda}{\lambda+\theta}\Big)^{t-s} \cdot \frac{1-(\frac{(\lambda_V+\theta)(1-p_W)}{\lambda_V})^{t-s}}{1-\frac{(\lambda_V+\theta)(1-p_W)}{\lambda_V}} + (1-p_W)^{t-s}\\
&=: g_V(-\theta, t-s).
\end{aligned}$$

To bound $\phi_{V(s,t)}(-\theta)$ with a bivariate function we note first

$$V(s,t) = \max\{0, t - \max\{s, W\}\}.$$

With an analysis very similar to the one in the previous scenario we have

$$\begin{aligned}
&\phi_{V(s,t)}(\theta) =\\
&(1-p_W)^t + \left(\frac{\lambda_V}{\lambda_V+\theta}\right)^{t-s}(1-(1-p_W)^s)\\
&+ \left(\frac{\lambda_V}{\lambda_V+\theta}\right)^{t-s}\left(p_W(1-p_W)^s\frac{1-(\frac{(\lambda_V+\theta)(1-p_W)}{\lambda_V})^{t-s}}{1-\frac{(\lambda_V+\theta)(1-p_W)}{\lambda_V}}\right)\\
&=: g_V(-\theta, s, t)
\end{aligned}$$

for all $s \leq t$.

As before the univariate and bivariate functions g_V are used within Theorems 3 and 4, respectively.

Figures 3 and 4 show the gain of a bivariate analysis for this scenario. Here Figure 3 evaluates the scenario with a fixed setup-time $W = 20$ and Figure 4 with a geometrically

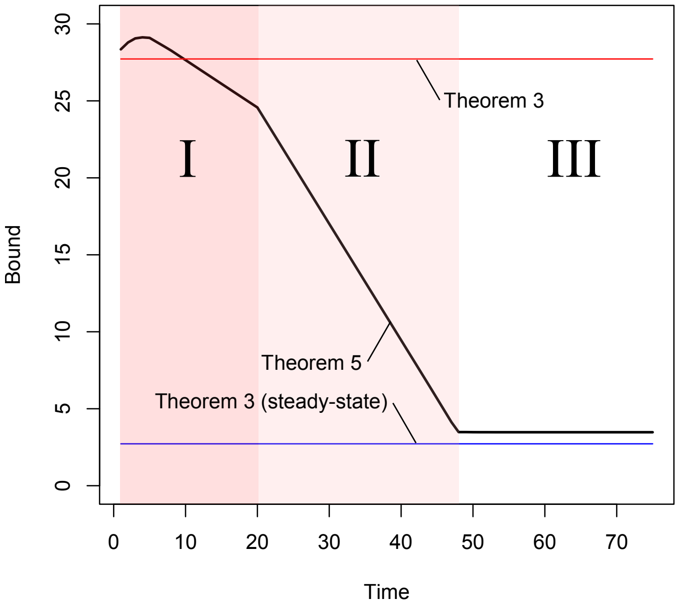

Fig. 3. Numerical Evaluation of Scenario 2 with a setup-time of $W = 20$ and burst $B = 25$. The graph shows a bound on the backlog of the system, which is broken with a probability of at most $\varepsilon = 10^{-3}$. In the first transient phase (denoted by I) the additional service is still setting up. In the second transient phase (denoted by II) the burst is processed by the total service U_{tot}. Eventually the system enters the steady-state in III.

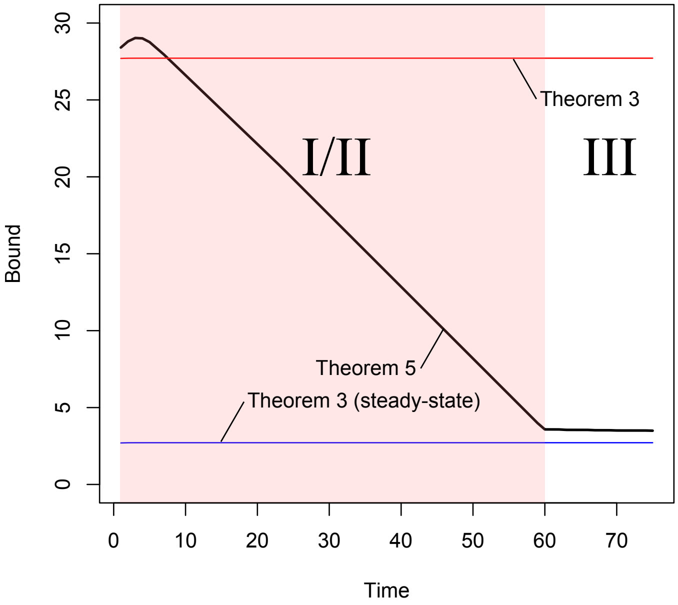

Fig. 4. Numerical Evaluation of Scenario 2 with random setup-time W and burst $B = 25$. The graph shows a bound on the backlog of the system, which is broken with a probability of at most $\varepsilon = 10^{-3}$. The transient phases I and II are smoothed out, due to W being a random variable. The system enters the steady-state III after processing the burst B.

distributed setup-time. Again we can observe how the bivariate formulation captures the system's transient phase and its steady-state. As in the previous scenario the time-dependent bounds lie between the time-invariant bound (red line) and the time-invariant bound for the steady-state system (blue line).

V. Conclusion

In this paper we extended the notion of bivariate envelopes of [1] for service elements to arrival curves. Further we have proven time-variant delay-bounds for this new type of envelopes. Two numerical examples showcase the wide variety of transient systems, which can be analyzed by this approach. In both the backlog-bound is captured by the bivariate envelopes and the transient phases are visibly resolved. Furthermore the analysis captures the transient phases, the steady-state of the system, and the transition from one into the other.

Future work includes the extension of Theorem 4 to a fully fleshed out network calculus as in [3], [12], [10]. In expression fundamental operations like the concatenation property must be verified.

References

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

[2] D. Cerotti, M. Gribaudo, R. Pinciroli, and G. Serazzi. Stochastic analysis of energy consumption in pool depletion systems. In *Measurement, Modelling and Evaluation of Dependable Computer and Communication Systems*, pages 25–39. Springer, 2016.

[3] C.-S. Chang. *Performance Guarantees in Communication Networks*. Telecommunication Networks and Computer Systems. Springer-Verlag, 2000.

[4] F. Ciucu and J. B. Schmitt. Perspectives on network calculus: no free lunch, but still good value. *Proc. of ACM SIGCOMM*, August 2012.

[5] M. Fidler. An end-to-end probabilistic network calculus with moment generating functions. In *Proc. of IEEE IWQoS*, pages 261–270, June 2006.

[6] M. Fidler. A network calculus approach to probabilistic quality of service analysis of fading channels. In *Proc. of IEEE Globecom 2006*, pages 1–6, Nov 2006.

[7] M. Fidler. Survey of deterministic and stochastic service curve models in the network calculus. *IEEE Communications Surveys Tutorials*, 12(1):59–86, 2010.

[8] M. Fidler and A. Rizk. A guide to the stochastic network calculus. *IEEE Communications Surveys Tutorials*, 17(1):92–105, Firstquarter 2015.

[9] A. Horvth, M. Paolieri, L. Ridi, and E. Vicario. Transient analysis of non-markovian models using stochastic state classes. *Performance Evaluation*, 69(78):315 – 335, 2012. Selected papers from {QEST} 2010.

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

[11] F. P. Kelly. Notes on effective bandwidths. In F. P. Kelly, S. Zachary, and I. Ziedins, editors, *Stochastic Networks: Theory and Applications*, number 4 in Royal Statistical Society Lecture Notes, pages 141–168. Oxford University Press, 1996.

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

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

[14] R. Lübben and M. Fidler. On the delay performance of block codes for discrete memoryless channels with feedback. In *Sarnoff Symposium (SARNOFF), 2012 35th IEEE*, pages 1–6, May 2012.

[15] M. Mellia, I. Stoica, and H. Zhang. Tcp model for short lived flows. *IEEE Communications Letters*, 6(2):85–87, 2002.

[16] C.-Y. Wang, D. Logothetis, K. S. Trivedi, and I. Viniotis. Transient behavior of atm networks under overloads. In *INFOCOM '96. Fifteenth Annual Joint Conference of the IEEE Computer Societies. Networking the Next Generation. Proceedings IEEE*, volume 3, pages 978–985 vol.3, Mar 1996.

[17] J. Zhang and E. J. Coyle. The transient solution of time-dependent m/m/1 queues. *IEEE Transactions on Information Theory*, 37(6):1690–1696, Nov 1991.