Dimensioning Leaky Buckets in Stochastic Environments*

Peter Buchholz[a], András Mészáros[b,c,*], Miklós Telek[b,c]

[a]TU Dortmund university, Informatik IV, Germany
[b]Budapest University of Technology and Economics, Department of Networked Systems and Services, Hungary
[c]HUNREN-BME Information Systems Research Group, Hungary

Abstract

Leaky buckets are commonly used for access control in networks, where access control stands for traffic regulation at the ingress of the network. In network calculus, which is often applied for performance analysis or dimensioning of networks, leaky buckets are the model behind piecewise linear arrival curves that specify an input bound to a network. In this paper we present the analysis of leaky bucket based access control under stochastic arrivals using fluid queues, when the access control is implemented by possibly more than one leaky buckets. This results in methods to dimension parameters of access control for different stochastic arrival processes including correlated arrivals. The approach is one step to bridge the gap between classical stochastic analysis using queuing networks and deterministic analysis using network calculus. Results are presented for stochastic arrival processes using numerical methods and for measured arrivals using trace driven simulation.

Keywords: stochastic analysis, access control, fluid queues, network calculus

1. Introduction

Single queues and queueing networks (QNs) are widespread models in performance analysis. Two different ways exist to analyze these models. Traditionally, stochastic assumptions are used and mean values for measures of interest are computed [1, 2]. Alternatively, bounds on measures of interest are computed from bounds on arrival and service processes [3, 4]. The latter approach has been further developed in network calculus (NC) [5]. The assumptions of both approaches are rather different. QN analysis (QNA) uses stochastic assumptions to characterize arrival and service processes. Then, either mean values are computed using algorithms like mean values analysis [1] or numerical techniques including matrix geometric methods are applied [2, 6] to obtain detailed probabilistic results for queues or jobs. In contrast, NC assumes strict

*This work is partially supported by the Hungarian Scientific Research Fund OTKA K-138208 project.
*Corresponding author
Email addresses: peter.buchholz@tu-dortmund.de (Peter Buchholz), meszarosa@hit.bme.hu (András Mészáros), telek@hit.bme.hu (Miklós Telek)

bounds on the arrival and service process to compute strict bounds on the delay and the level of buffers.

Thus, QNA is a stochastic approach whereas NC is a deterministic approach. Both approaches have in common that models for the arrival or service process have to be estimated from available data and the resulting models are then used for analysis. In QNA, Markov models are often applied to characterize arrival and service processes. Parameters are estimated from the available data [7]. Markov models allow one to describe correlated arrivals or services, but always have an exponentially decaying tail; therefore, heavy tailed processes can only be approximated [8]. In NC, discrete arrivals are substituted by fluid arrivals and services. Arrival and service processes are bounded by so called arrival and service curves, which are often realized by affine functions or simple piecewise linear functions. By defining an upper bound for the arrivals and a lower bound for the service, upper bounds for buffer level and delay can be determined using min/+ algebra. In real time systems with periodic behavior, bounds for arrivals and services can be determined from the system specification [9]. However, for non-periodic systems the bounds have to be estimated from available data [5, 10].

If NC is applied to non-periodic systems, then it is usually rather conservative because arrival and service bounds have to be valid for all possible sequences of arrivals or services. If bounds result from measured data without any assumptions about the generating process, then strict bounds can only be computed up to some probability using extreme value statistics [11]. If the system does not contain hard real time constraints, then violation of bounds with some small probability can usually be accepted. This consideration led to the development of Stochastic Network Calculus (SNC) [12, 13, 14]. SNC exploits the stochastic nature of traffic and computes bounds that hold up to some small probability. It often uses moment generating functions (MGFs) and often assumes exponentially bounded burstiness which roughly corresponds to exponentially decaying tails, but SNC has also been used for heavy tailed processes [13] which cannot be bounded by an affine function. SNC usually computes simple affine arrival and service curves, often denoted as envelopes, that are then used for an analysis with NC. The resulting bounds on backlog and delay then hold up to some small probability. Often SNC uses a deterministic server and stochastic arrivals [14, 15], which will also be applied in our approach.

In this paper, we go in a similar direction as SNC but use a different approach to generate bounds for the arrival process. The goal is to compute a piecewise linear arrival curve that is exceeded by the arrival process with a small predefined probability at most. In this step, we usually do not consider the capacity of the server which is part of the SNC analysis. Thus, we derive an access control realized by leaky buckets (LBs), which assures that arriving traffic passes immediately with probability $1 - \varepsilon$ for some small ε. Affine arrival curves with only one segment resulting from one LB, are only a specific case which can often be improved by adding additional linear segments or LBs. We assume a constant service but show how uncertainty in the service process can be integrated in the arrival bound. In contrast to SNC, the approach is not based on MGFs. Instead, we show that analysis can be done by fluid queues [16]. For a simple affine arrival curve, a single fluid queue [16, 17] is sufficient for detailed analysis. To analyze piecewise linear curves with more than one segment, a new type of fluid queue with two coupled buffers [18] has to be analyzed. We show how the fluid queue

can be analyzed numerically for stochastic arrival and service processes and by trace driven simulation using real network traces. Consequently, the approach allows the computation of arrival curves from measured data, which is rarely considered in NC or SNC. After the arrival curve is available, NC can be applied to compute backlog and delay bounds for a system with a known service capacity.

The paper is structured as follows. In the next section, basic properties of network calculus and the extension to SNC are summarized. Then, in Section 3, a class of arrival curves is defined and it is shown how they are validated with respect to a trace or from a stochastic description of arrivals by phase type distributions or Markovian arrival processes [7]. In Section 4, we show how to compute an arrival curve for a trace or a stochastic model by considering the fraction of arrivals that exceed the arrival curve. Some statistical measures for a queue or a tandem of queues fed by an arrival stream are introduced afterwards. Results of experiments with real or synthetic data are presented in Section 6. The paper ends with the conclusions.

2. Network Calculus Basics

In the following we summarize some basic results for NC which can be found in the literature [5, 19, 14]. First we show the basic computational steps, then piecewise linear arrival curves are introduced and finally the extension to SNC is outlined.

2.1. Basic Computations for System Analysis

In NC, a system component, which can be a network, a part of network or a simple processing station, is characterized by the load that arrives and the service it provides such that the input stream is delayed and results in an output stream. The input stream or arrival process is denoted by $A(t)$ and describes the accumulated load that arrived until time $t \geq 0$. Consequently, $A(t)$ is non decreasing, left continuous and $A(t) = 0$ for $t \leq 0$. The service process $S(t)$ describes the amount of load that is processed in the interval $(0, t]$ if load to be processed is available for the complete time interval. Again, $S(t)$ is non decreasing, left continuous and $S(t) = 0$ for $t \leq 0$. If arrival process $A(t)$ is fed into a queue with service process $S(t)$, then a departure process $A'(t)$ results, which is a delayed version of $A(t)$ (i.e., $A'(t) \leq A(t)$ for all t). The departure process equals[1]

$$A'(t) = \inf_{0 \leq \tau \leq t} \{A(\tau) + S(t) - S(\tau)\} = A \circledast S(t). \tag{1}$$

The backlog of load in the system is given by

$$b_t = A(t) - A'(t) \text{ and } b_{\max} = \sup_{t \geq 0} \{b_t\}. \tag{2}$$

If we assume that load is processed in First Come First Served (FCFS) order, then the virtual delay of the load arriving at time t equals

$$w_t = \sup_{\tau \geq 0} \{\tau : A(t) \leq A'(t + \tau)\} \text{ and } w_{\max} = \sup_{t \geq 0} \{w_t\}. \tag{3}$$

[1] In NC, the usual notation of min-plus convolution is $\otimes$, which we use for denoting Kronecker products in the sequel.

Since $A(t)$ and $S(t)$ are usually non-deterministic in most systems, either stochastic models or deterministic models for bound computation have to be applied. We consider here deterministic models that are later expanded by some stochastic assumptions. A strict upper arrival curve is defined as a nondecreasing function $\alpha(t)$, with

$$A(t) - A(\tau) \leq \alpha(t - \tau) \text{ for all } 0 \leq \tau \leq t. \tag{4}$$

Similarly, a strict lower service curve is defined as a nondecreasing function $\beta(t)$, with

$$S(t) - S(\tau) \geq \beta(t - \tau) \text{ for all } 0 \leq \tau \leq t. \tag{5}$$

If the arrival process $\alpha(t)$ is fed into a queue with service process $S(t)$, or the arrival process $A(t)$ is fed into a queue with service process $\beta(t)$, then the functions can be plugged into (1)-(3) to compute bounds under FCFS scheduling, e.g. $A'(t) \leq \alpha \circledast S(t)$ and $A'(t) \geq A \circledast \beta(t)$.

2.2. Arrival Curves

From (4) is follows that

$$\alpha(t) \geq \sup_{u \geq 0} \{A(t + u) - A(u)\} \tag{6}$$

where $\sup_{u \geq 0} \{A(t + u) - A(u)\}$ is the smallest upper arrival curve, which is referred to as empirical envelope in [20]. However, if this function is computed for some measured arrival stream, the result is usually a complicated function which makes computations cumbersome. Consequently, simpler function are used as upper arrival curves. The simplest form is an affine function $\alpha(t) = c + r \cdot t$, where c is the initial step and r is the average arrival rate. This function corresponds to arrivals restricted by a LB with capacity c and rate r. More complicated functions result from the combination of n LBs with c_i, r_i $(i = 1, \ldots, n)$ where $c_i < c_{i+1}$ and $r_i > r_{i+1}$. These functions are piecewise linear with decreasing slope.

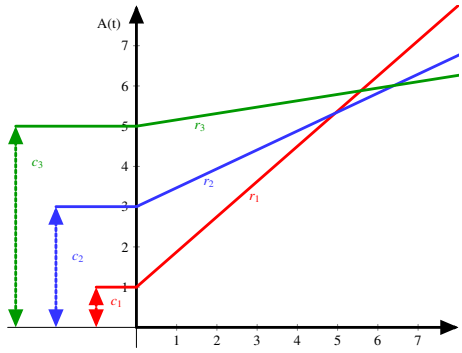

Figure 1: Piecewise linear arrival curve with three segments.

Figure 1 shows an example of an arrival curve with three segments, affine curves are described by each of the three lines. Piecewise linear curves have a tight connection

to combinations of LBs. They can be described as a sequence of LBs as shown below. Alternatively, it can be assumed that the arrival stream is fed to each LB in parallel and if it is delayed in one LB, load cannot pass. We come back to this aspect when stochastic input processes are analyzed.

Service curves can, in principle, be handled similarly. However, we consider here only simple linear servers with service rate s such that $\beta(t) = s \cdot t$ which is then combined with stochastic arrivals.

2.3. Stochastic Arrival Envelopes

Arrival curves in SNC are often denoted as arrival envelopes which are exceeded with some small probability. Usually SNC is applied to a discrete time scale and affine envelope functions $\alpha(t) = \delta(t > 0)c + r \cdot t$ are used [12, 13], where $\delta(e)$ is the indicator function of event e, i.e., $\delta(e) = 1$ if e is true and 0 otherwise. In this situation $A(t) - A(s)$ $(s \leq t)$ is a random variable that describes the arrivals in the interval $(s, t]$. We assume that the underlying process is stationary, i.e., $\mathbb{P}(A(t) - A(s) < x) = \mathbb{P}(A(t - s) < x)$, such that $A(t)$ is a random variable describing the arrival stream in an arbitrary interval of length t. If $A(t)$ is distributed according to a distribution with an infinite support, then $A(t) \leq c + r \cdot t$ cannot hold for finite c and r. Consequently, SNC defines an arrival envelope $\alpha_\varepsilon(t) = \delta(t > 0)c_\varepsilon + r_\varepsilon \cdot t$ that holds up to some probability $\varepsilon > 0$.

$$\forall t \in \mathbb{N}_0 : \ \mathbb{P}(A(t) > \delta(t > 0)c_\varepsilon + r_\varepsilon \cdot t) \leq \varepsilon, \tag{7}$$

which is equivalent with the S^2BB model in [21]. Different approaches exist to derive such envelope functions. The most prominent are based on MGFs and exponentially bounded burstiness (EBB) which are related via Chernoff bounds as shown in [13]. In the following we briefly describe the computation of arrival envelopes using MGFs which can be applied to compute the parameters of leaky buckets assuring a predefined ε with which the load exceeds the bucket content at most. For more sophisticated bounds we refer to the literature [13, 14, 22].

According to [12], a stochastic arrival stream $A(t)$ is upper $(c(\theta), r(\theta))$ constrained for $\theta > 0$ if

$$\frac{1}{\theta} \log\left(\mathbb{E}\left(e^{\theta(A(t)-A(s))}\right)\right) \leq c(\theta) + r(\theta)(t - s) \tag{8}$$

for all $0 \leq s \leq t$ where $c(\theta), r(\theta)$ are non-negative functions of parameter θ. For stationary arrival streams we define

$$r^*(\theta) = \limsup_{t\to\infty} \left\{ \frac{1}{\theta t} \log\left(\mathbb{E}\left(e^{\theta A(t)}\right)\right) \right\}. \tag{9}$$

$r^*(\theta)$ can be interpreted as the infimum of the required rate for leaky buckets with finite capacity $\hat{c}$, because the following relation holds [12, Lemma 7.4.1(ii)]

$$\mathbb{P}(X(t) > \hat{c}) \leq \frac{e^{\theta(c(\theta)-\hat{c})}}{1 - e^{\theta(r^*(\theta)-\hat{r})}} \tag{10}$$

for all $\theta > 0$ and $\hat{r} > r^*(\theta)$, where $X(t)$ is the backlog at the LB formally defined below. The equation defines an upper bound for the probability that the bucket becomes empty

which holds for all $\theta > 0$. Optimization with respect to θ, $\hat{c}$ and $\hat{r}$ allows one to find appropriate parameters for the leaky bucket to assure $\mathbb{P}(X(t) > \hat{c}) \leq \varepsilon$ for any threshold $\varepsilon > 0$.

There are a couple of limitations for this approach. First, the arrival process is often given by a trace measured in a real system. In this case, parameters of the arrival envelope have to be estimated. In [23] statistical network calculus is developed which allows the estimation of the parameters of an affine envelope function from measured arrivals. An alternative approach will be presented below. Second, although SNC allows different arrival envelopes in principle, in practice, affine curves are used in computational approaches. This often has the disadvantage that either r is much larger than the average rate or the size of an accepted burst c becomes very large for small values of ε. However, other possibilities of defining arrival envelopes exist [24] as well. We will consider piecewise linear functions as natural extensions of affine curves and show that even with only two segments the mentioned problem is reduced to a large extent. Third, moment generating functions are useful for independent random variables. The integration of dependencies in the approach is challenging [25] and, if possible at all, results in pessimistic bounds. We present an approach that naturally incorporates various dependencies between inter-arrival times and sizes of arriving load units. Fourth, SNC often requires a discrete time scale which often implies discretization of the continuous time scale with which the system evolves. We consider continuous time analysis.

3. Checking if Traffic Sources Comply with Arrival Curves

We distinguish between the cases when the traffic source is given by a trace, that is available from some measurement or monitoring of the system, and by the stochastic description of the arrival process. In the latter case, we consider only arrivals described by Markov models [7].

3.1. Measured Arrivals

Arrivals in computer networks or related systems occur in discrete portions that arrive at some points in time. A trace $\mathcal{T}$ describes a finite sequence of arrivals that have been measured in a real system or result from a simulation. Let m be the length of the trace. Traces are described by a sequence (t_i, w_i) $(i = 1, \ldots, m)$, where t_i is the inter-arrival time between the $i-1$th and the ith arrival and w_i is the size of the the ith arrival. We assume without loss of generality that $t_0 = 0$.

Now assume that $\mathcal{T}$ should be validated against an upper arrival curve $\alpha = (c_i, r_i)_{i=1,\ldots,n}$. Fig. 2 shows the sequence of LBs to validate the trace. Each LB consists of two buffers: a token bucket, and a buffer. The level of the token bucket is denoted by C_i, which indicates the available capacity which can pass the ith LB immediately. The level of the buffer is denoted by L_i. This buffer contains the delayed load units because of an empty token bucket. The ith token bucket is continuously filled with rate r_i until capacity c_i is reached. Load arrives in discrete pieces but might pass a bucket partially. Thus, if a load unit of size w arrives at time t at a LB with level x and rate r, then the load can pass immediately if $x \geq w$. In this case, the new level of the

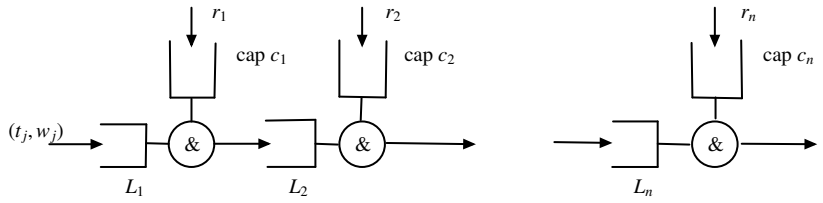

Figure 2: Sequence of LBs for arrival curve validation.

bucket is $x - w$. Otherwise load x passes immediately, the remaining load passes in the interval $(t, t + \frac{w-x}{r}]$ with rate r and the token bucket remains empty in the interval.

To validate a trace against an arrival curve, a discrete event based approach can be applied, even though the model contains continuous parts. Let $(C_i(t), L_i(t))$ be the state of token bucket i and buffer i at time t. This implies $0 \leq C_i(t) \leq c_i$, $0 \leq L_i(t)$ and $C_i(t) \cdot L_i(t) = 0$. Let $T_j = \sum_{i=1}^{j} t_j$ be the time of the jth arrival. Now define for $i = 1, \ldots, n$

$$\Delta_0(t) = \min_{T_j \geq t} \left\{ T_j - t \right\}, \quad \Delta_i(t) = \begin{cases} \frac{L_i(t)}{r_i} & \text{if } L_i(t) > 0, \\ \frac{c_i - C_i(t)}{r_i} & \text{if } L_i(t) = 0 \wedge C_i(t) < c_i, \\ \infty & \text{otherwise,} \end{cases} \tag{11}$$

where $\Delta_0(t)$ is the time of the next arrival after t, and $\Delta_i(t)$ (i=1,...,n) is the time of the next event associated with LB i. Then $\Delta(t) = \min_{0 \leq i \leq n} \{\Delta_i(t)\}$ is the next event after time t. Let $L_0(t) > 0$ and $r_0 = 0$ by definition and $k_i(t)$ be the last busy buffer before node i at time t, i.e. $k_i(t) = \text{argmax}_{0 \leq k < i} L_k(t) > 0$. If $\Delta(t) \neq \Delta_0(t)$, then at $t + \Delta(t)$ the new state $(C_i(t + \Delta(t)), L_i(t + \Delta(t)))$ becomes

$$\begin{aligned} C_i(t + \Delta(t)) &= \begin{cases} 0 & \text{if } L_i(t) > 0, \\ C_i(t) + \Delta(t)(r_i - r_{k_i(t)}) & \text{if } 0 \leq C_i(t) < c_i \wedge L_i(t) = 0, \\ c_i & \text{if } C_i(t) = c_i, \end{cases} \\ L_i(t + \Delta(t)) &= \begin{cases} L_i(t) - \Delta(t)(r_i - r_{k_i(t)}) & \text{if } L_i(t) > 0, \\ 0 & \text{otherwise.} \end{cases} \end{aligned} \tag{12}$$

If an arrival (t_j, w_j) takes place at time $t + \Delta(t)$ (i.e., $\Delta_0(t) = \Delta(t)$), define values

$$v_{j,0} = w_j, \quad C_0(t) = \infty \quad \text{and} \quad v_{j,i} = \min\{v_{j,i-1}, C_{i-1}(t)\} \tag{13}$$

for $i = 1, \ldots, n$. That is, $v_{j,i}$ is the amount of load arriving to LB i from LB $i - 1$ at the time of the jth arrival to the system. With these values the new buffer levels are given by $L_i(t + \Delta(t)) = L_i(t) + v_i - v_{i+1}$ and $C_i(t + \Delta(t)) = \max\{C_i(t) - v_i, 0\}$. Let T_{m+1} be the time when the mth arrival leaves the last LB completely. Then,

$$P_\alpha^a = \frac{\sum_{j=1}^{m} \delta\left(\max_{i=1,\ldots,n} \left\{ L_i(T_j^+) \right\} > 0\right)}{m} \quad \text{and} \quad P_\alpha^s = \frac{\int_0^{T_{m+1}} \delta\left(\max_{i=1,\ldots,n} \{L_i(t)\} > 0\right) dt}{T_{m+1}} \tag{14}$$

are estimates for the probability that an arrival exceeds arrival curve α and the probability that a random observer finds backlog in the system (which approximates the fraction of time that the arrival process exceeds the arrival curve), respectively. The earlier is more commonly used in practice.

3.2. Stochastic Arrival and Service Times

For stochastic analysis we assume that inter-arrival times and the amount of arriving load are distributed according to a phase type distribution (PHD) with representation $(\boldsymbol{p}, \boldsymbol{D})$, denoted as PHD$(\boldsymbol{p}, \boldsymbol{D})$, or a Markovian arrival process (MAP) with representation $(\boldsymbol{D}, \boldsymbol{C})$ [7], denoted as MAP$(\boldsymbol{D}, \boldsymbol{C})$, where $\boldsymbol{p}$ is a probability distribution vector, $\boldsymbol{D}$ is a sub-generator with only transient states (i.e., $-\boldsymbol{D}^{-1}$ exists and is non-negative) and $\boldsymbol{C}$ is a non-negative matrix such that $\boldsymbol{D} + \boldsymbol{C}$ is an irreducible generator. For a MAP the embedded initial vector is defined as the unique solution of $-\boldsymbol{p}\boldsymbol{D}^{-1}\boldsymbol{C} = \boldsymbol{p}$ and $\boldsymbol{p}\mathbb{1} = 1$, where $\mathbb{1}$ is the column vector of ones. In this case the marginal distribution of the MAP is PHD$(\boldsymbol{p}, \boldsymbol{D})$. Furthermore, we define $\boldsymbol{d} = -\boldsymbol{D}\mathbb{1}$. Each arrival adds an amount of work to the system which is distributed according to the corresponding PHD$(\boldsymbol{p}^s, \boldsymbol{D}^s)$ or MAP$(\boldsymbol{D}^s, \boldsymbol{C}^s)$, and the time between two arrivals is defined according to PHD$(\boldsymbol{p}^a, \boldsymbol{D}^a)$ or MAP$(\boldsymbol{D}^a, \boldsymbol{C}^a)$. Since the amount of work that arrives corresponds to the time required to serve the load divided by the speed of the server, we can denote the load size as service time. If this load is fed to a single LB (c, r), then the resulting model can be interpreted as a fluid queue with service rate r, which will be described first, before we introduce the input to a sequence of LBs.

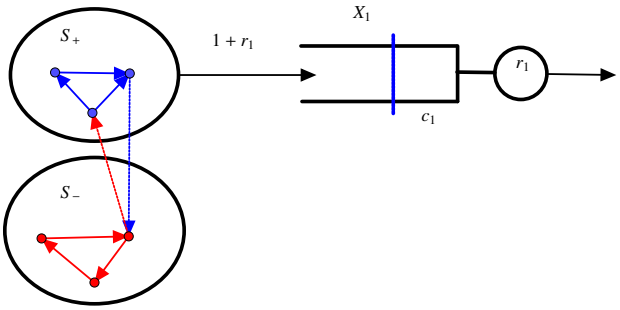

Figure 3: Fluid queue modeling the stochastic input driven by an affine arrival curve.

For an affine arrival curve the resulting fluid model is shown in Fig. 3. The fluid level in the fluid queue is driven by a Markov chain [17]. The state space of the driving Markov chain can be decomposed into two subsets: $\mathcal{S}_+$ and $\mathcal{S}_-$. In states from $\mathcal{S}_+$ fluid arrives to the queue with rate 1, in states from $\mathcal{S}_-$ the queue is emptied with rate r. The sojourn time in $\mathcal{S}_+$ corresponds to the service time and the sojourn time in $\mathcal{S}_-$ corresponds to the inter-arrival time. Fig. 4 shows an example trajectory of the fluid level in the queue. In the left graph, we see the behavior of the Markov fluid queue. In states from $\mathcal{S}_+$, the queue grows linearly, in states from $\mathcal{S}_-$ the queue shrinks linearly or remains zero. In the original system, load arrives in batches and the behavior is as in

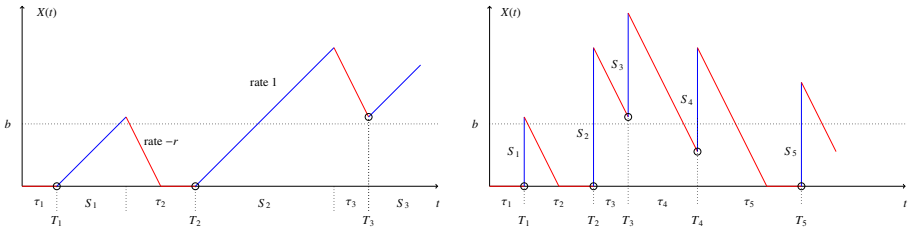

Figure 4: Example trajectory of the fluid model (left) and the same trajectory restricted to $\mathcal{S}_-$ (right).

the right graph in Fig. 4, which presents the same behavior of the queue that is shown in the left graph restricted to the states in $\mathcal{S}_-$. If we know the results for the Markov fluid queue model (left graph), then it is easy to derive results for the modified model (right graph) by considering conditional probabilities, which is well established in the analysis of fluid queues [26].

Let $X_1(t) = c_1 - C_1(t) + L_1(t)$ denote the continuous fluid level which represents the state of the LB at time t. Since $0 \le C_1(t) \le c_1$, $0 \le L_1(t)$ and $C_1(t) \cdot L_1(t) = 0$ hold, $X_1(t)$ completely characterizes the $C_1(t)$ token bucket and the $L_1(t)$ buffer contents. That is, $C_1(t) = c_1 - X_1(t)$ and $L_1(t) = 0$ if $X_1(t) \le c_1$, while $C_1(t) = 0$ and $L_1(t) = X_1(t) - c$ if $X_1(t) > c_1$. Consequently, the knowledge of $X_1(t)$ allows us to derive all relevant quantities from the model as shown below. The fluid queue in Fig. 3 is driven by an irreducible Markov chain with generator matrix

$$\boldsymbol{Q} = \begin{pmatrix} \boldsymbol{Q}_{++} & \boldsymbol{Q}_{+-} \\ \boldsymbol{Q}_{-+} & \boldsymbol{Q}_{--} \end{pmatrix},$$

where $\boldsymbol{Q}_{++}$ contains the transitions between the states in $\mathcal{S}_+$, $\boldsymbol{Q}_{+-}$ from the states in $\mathcal{S}_+$ to the states in $\mathcal{S}_-$, etc. The steady state results for the queue can be analyzed using matrix geometric methods [16] and the fluid level distribution in steady state has a matrix exponential representation. The $\boldsymbol{Q}$ matrix, in our case, can have one of the following structures

$$\begin{aligned} \boldsymbol{Q}_0 &= \begin{pmatrix} \boldsymbol{D}^s & \boldsymbol{d}^s\boldsymbol{p}^a \\ \boldsymbol{d}^a\boldsymbol{p}^s & \boldsymbol{D}^a \end{pmatrix}, & \boldsymbol{Q}_1 &= \begin{pmatrix} \boldsymbol{I}\otimes\boldsymbol{D}^s & \boldsymbol{I}\otimes\boldsymbol{d}^s \\ \boldsymbol{C}^a\otimes\boldsymbol{p}^s & \boldsymbol{D}^a \end{pmatrix}, \\ \boldsymbol{Q}_2 &= \begin{pmatrix} \boldsymbol{D}^s & \boldsymbol{p}^a\otimes\boldsymbol{C}^s \\ \boldsymbol{d}^a\otimes\boldsymbol{I} & \boldsymbol{D}^a\otimes\boldsymbol{I} \end{pmatrix}, & \boldsymbol{Q}_3 &= \begin{pmatrix} \boldsymbol{I}\otimes\boldsymbol{D}^s & \boldsymbol{I}\otimes\boldsymbol{C}^s \\ \boldsymbol{C}^a\otimes\boldsymbol{I} & \boldsymbol{D}^a\otimes\boldsymbol{I} \end{pmatrix}, \end{aligned} \tag{15}$$

where $\otimes$ denotes the Kronecker product. $\boldsymbol{Q}_0$ describes PH distributed *iid* arrival and service. For example, if the inter arrival time is phase type distributed with representation $\boldsymbol{p}^s, \boldsymbol{D}^s$, and the packet size is phase type distributed with representation $\boldsymbol{p}^a, \boldsymbol{D}^a$, where $\boldsymbol{p}^s = [1,0]$, $\boldsymbol{D}^s = \begin{pmatrix} -3 & 1 \\ 0 & -1 \end{pmatrix}$ and $\boldsymbol{p}^a = [0.5, 0.5]$, $\boldsymbol{D}^a = \begin{pmatrix} -4 & 1 \\ 0 & -2 \end{pmatrix}$ then $\boldsymbol{d}^s = -\boldsymbol{D}^s\mathbb{1} = \begin{pmatrix} 2 \\ 1 \end{pmatrix}$, $\boldsymbol{d}^a = -\boldsymbol{D}^a\mathbb{1} = \begin{pmatrix} 3 \\ 2 \end{pmatrix}$ and $\boldsymbol{Q}_0 = \begin{pmatrix} \boldsymbol{D}^s & \boldsymbol{d}^s\boldsymbol{p}^a \\ \boldsymbol{d}^a\boldsymbol{p}^s & \boldsymbol{D}^a \end{pmatrix} = \begin{pmatrix} -3 & 1 & 1 & 1 \\ 0 & -1 & 0.5 & 0.5 \\ 3 & 0 & -4 & 1 \\ 2 & 0 & 0 & -2 \end{pmatrix}$.

$\boldsymbol{Q}_1$ describes correlated (MAP) inter-arrival times and *iid* (PH) service times, $\boldsymbol{Q}_2$ *iid* (PH) inter-arrival times and correlated (MAP) service times, and $\boldsymbol{Q}_3$ correlated (MAP) inter-arrival times and correlated (MAP) service times. If n^a and n^s are the dimensions of the Markov chains describing arrivals and service, then matrices $\boldsymbol{Q}_0$ through $\boldsymbol{Q}_3$ have dimensions $n^a + n^s$, $n^a n^s + n^a$, $n^a + n^a n^s$ and $2n^a n^s$, respectively. Furthermore, the following matrix

$$\boldsymbol{Q}_4 = \begin{pmatrix} \boldsymbol{D}^s & \boldsymbol{B}^s \\ \boldsymbol{B}^a & \boldsymbol{D}^a \end{pmatrix} \tag{16}$$

with $\boldsymbol{B}^s, \boldsymbol{B}^a \geq \boldsymbol{0}$, $\boldsymbol{D}^s_{ij}, \boldsymbol{D}^a_{ij} \geq 0$ for $i \neq j$, $\boldsymbol{D}^s \mathbb{1} = -\boldsymbol{B}^s \mathbb{1}$ and $\boldsymbol{D}^a \mathbb{1} = -\boldsymbol{B}^a \mathbb{1}$ allows the representation of correlation between inter-arrival and service times. For the analysis of the correlation and the computation of parameters to obtain specific coefficients of correlations we refer to [27].

The introduced model allows stochastic inter-arrival times and a stochastic distribution of the arriving load. If the service of the fluid queue has some stochastic fluctuation, this can be encoded in the fluid model. E.g., if the server fails from time to time and requires some recovery time, then set $\mathcal{S}_-$ can be split into a set $\mathcal{S}_-$, where the fluid level in the buffer decreases and $\mathcal{S}_0$, where the fluid level remains constant. In a similar way, varying rates of the fluid server according to some background Markov process can be modeled and analyzed [16].

Let $f(x)$ be the pdf of the buffer level of the fluid queue restricted to states from $\mathcal{S}_-$ and $f^+(x)$ the pdf of the buffer level immediately after entering $\mathcal{S}_-$. Both quantities can be derived from the matrix exponential stationary solution of the fluid queue (c.f. [28, Theorem 1]). From these quantities, the probability of an arrival and a random observer finding backlog in the system can be computed, respectively, as

$$P^a_{(c,r)} = \int_c^\infty f^+(x)dx \quad \text{and} \quad P^s_{(c,r)} = \int_c^\infty f(x)dx. \tag{17}$$

Now assume that we have an arrival curve according to $(c_1, r_1), (c_2, r_2)$ with $c_1 < c_2$ and $r_1 > r_2$, as shown in Fig. 1 by the red and blue lines. Let $f_1(x)$ and $f_1^+(x)$ be the pdfs for the arrival curve (c_1, r_1). To obtain results for the two LBs together, we need to know the joint density of fluid in the buffer defined by LB 1 with (c_1, r_1) and the buffer defined by LB 2 with (c_2, r_2). This can be done by considering an extended fluid model with two buffers that are filled simultaneously, as shown in Fig. 5.

Let $g(x, y)$ be the joint pdf of both buffer levels of the two buffer fluid model restricted to states from $\mathcal{S}_-$ and $g^+(x, y)$ the pdf of the joint buffer levels immediately after entering a state from $\mathcal{S}_-$ from a state in $\mathcal{S}_+$. Then the required results, considering that $c_1 < c_2$, can be computed as

$$\begin{aligned} P^a_{(c_1,r_1),(c_2,r_2)} &= \int_{x=c_1}^\infty f_1^+(x)dx + \int_{x=0}^{c_1}\int_{y=c_2}^\infty g^+(x,y)dydx, \\ P^s_{(c_1,r_1),(c_2,r_2)} &= \int_{x=c_1}^\infty f_1(x)dx + \int_{x=0}^{c_1}\int_{y=c_2}^\infty g(x,y)dydx, \end{aligned}$$

where the first term contains the probability that LB 1 has backlog ($X_1 > c_1$) and the second term represents the probability that LB 1 has no backlog but LB 2 has ($X_1 <$

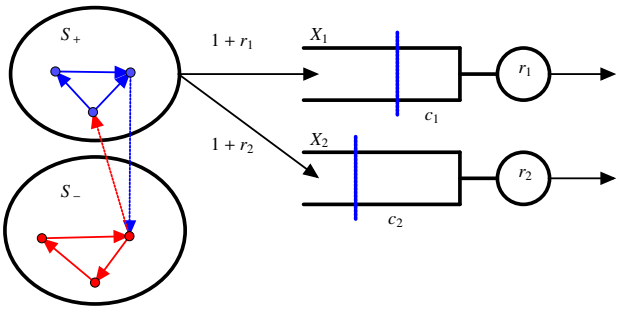

Figure 5: Fluid queue modeling the stochastic input driven by a piecewise linear curve.

$c_1, X_2 > c_2$). Analysis of the two buffer model is, unfortunately, more complex than the analysis of the single buffer system. In [18] we developed a numerical approach to analyze the system with sufficient accuracy for $P_\alpha^{s/a}$ not too small.

4. Generation of Arrival Curves

Let $\bar{r} = \lim_{t\to\infty} \frac{A(t)}{t}$ denote the mean arrival rate. For a queue with inter-arrival time distribution $(\boldsymbol{p}^a, \boldsymbol{D}^a)$ and service time distribution $(\boldsymbol{p}^s, \boldsymbol{D}^s)$ (see the description above), the mean arrival rate equals to $\bar{r} = \frac{\boldsymbol{p}^s(\boldsymbol{D}^s)^{-1}\mathbf{1}}{\boldsymbol{p}^a(\boldsymbol{D}^a)^{-1}\mathbf{1}}$. Observe that the mean arrival rate depends only on the marginal distributions and not on the correlation structure. For finite traces $\bar{r}$ has to be estimated using standard methods to compute one-sided confidence intervals from possibly dependent observations [29].

An affine arrival curve is characterized by $\alpha = (c_1, r_1)$ where $r_1 = (1 + h)\bar{r}$ for some constant $h > 0$. If the allocated bandwidth should be minimized, then a small value $h \ll 1$ should be chosen. However, the value for c is usually determined by physical restrictions of the system which cannot accept arbitrarily large batches. For a given bound on P_α^a or P_α^s, a smaller r implies a larger c, which means that larger batches are allowed to enter the system immediately. Usually some cost function $\Psi(c, r) = \zeta_c c + \zeta_r r$, with positive coefficients ζ_c and ζ_r has to be optimized according to some bound ε for $P_{(c,r)}^a$ or $P_{(c,r)}^s$. This results in the following optimization problem.

$$\min_{c,r} (\zeta_c c + \zeta_r r) \text{ s.t. } P_{(c,r)} \le \varepsilon, \bar{r} < r \le r^{\max}, c \le c^{\max}, \tag{18}$$

where $P_{(c,r)}$ equals $P_{(c,r)}^a$ or $P_{(c,r)}^s$ and $r^{\max}, c^{\max}$ are bounds for r and c which are usually derived from constraints of the system. Since the optimization problem is non-linear, one has to apply some intelligent search strategy. It should be noted that the evaluation of $P_{(c,r)}$ is cheap for fixed r and varying c because $P_{(c,r)}$ is a matrix exponential function of c, whereas the modification of r requires the solution of a new set of non-linear matrix equations. However, usually the optimization is fairly smooth such that the effort for computing the optimum, if matrix $\boldsymbol{Q}$ is not too large, is not much higher than in SNC, which also requires numerical computations.

To generate arrival curves with more than one linear segment one can, in principle, extend the approach by defining appropriate cost functions. However, even for only two segments, an optimization problem with 4 parameters has to be solved, which is more expensive than the optimization problem (18) because the two buffer fluid queue is harder to solve than the one buffer system. It is often realistic to assume that some parameters are fixed. A simple and practical approach is to set r_1 according to the limits of the system, e.g., $r_1 = r_{\max}$, and then determine c_1 from the analysis of a single buffer fluid queue such that $P_{(c_1,r_1)} \leq \varepsilon$. This single LB can then be extended to a sequence of two LBs by solving the following optimization problem.

$$\min_{c_2,r_2} \left(\zeta_{c_2} c_2 + \zeta_{r_2} r_2\right) \text{ s.t. } P_{((c_1,r_1)(c_2,r_2))} \leq \varepsilon, \bar{r} < r_2 \leq r_1 = r_{\max}, c_1 \leq c_2 \leq c_{\max} \tag{19}$$

Often, c_2 can also be fixed and only r_2 computed. This problem can then be solved by repeated solution of the two buffer fluid system. Again, modifications of r_2 require new computations whereas results for different values of c_2 are much easier to compute.

5. Queueing Analysis

Arrival curves are then fed into some system which can be analyzed using the standard NC approach. We consider here only single queues and tandems of queues without cross traffic assuming FCFS scheduling. More complicated networks can be analyzed as described in the literature [19, 5, 4, 30].

We begin with a single queue with upper arrival curve α and lower service curve β. Then it follows from (1-3) that a lower bound for the departure process is given by

$$\alpha'(t) \geq \inf_{0 \leq \tau \leq t} \{\alpha(\tau) + S(t) - S(\tau)\} = \alpha \circledast S(t) \tag{20}$$

and upper bounds for the backlog and delay under FCFS are

$$b_{max} \leq \sup_{s \geq 0} \{\alpha(s) - \alpha'(s)\} \text{ and } w_{max} \leq \sup_{\tau \geq 0} \{\tau : \alpha(t) \leq \alpha'(t+\tau)\}. \tag{21}$$

The equations describe the maximal horizontal and vertical difference between the arrival and the service curves. If the arriving load passes a sequence of q servers with lower service curves β_p $(p = 1, \ldots, q)$, then the service curves can first be combined by computing $\beta = \beta_1 \circledast \beta_2 \circledast \ldots \circledast \beta_q$. The resulting model can be analyzed like a system with a single server.

6. Examples

We consider two types of examples. First, stochastic queues and afterwards a traffic trace that is fed into a server.

6.1. Queueing Models

We begin with a simple queueing model, where exact results are available and extend this model afterwards by introducing correlation such that numerical analysis of the fluid queue is required to obtain the results.

6.1.1. M/M/1 queue

The first model we consider is the classical M/M/1 queue. Load arrives with exponentially distributed inter-arrival times with rate λ and the size of a load unit is exponentially distributed with rate μ. The load is fed into a LB which is represented by a fluid queue (see Fig. 3). Let X be the buffer level in steady state. $X = L + c - C$ describes the connection between the LB and the fluid queue, where $L = \lim_{t\to\infty} L(t)$ and $C = \lim_{t\to\infty} C(t)$ are the stationary backlog and token bucket levels. Furthermore, let X^a be the fluid level immediately after an arrival. The results for this simple example can be computed in closed form using results from queuing theory. It is known that for an M/M/1 queue with arrival rate λ, service rate μ and utilization $\rho = \lambda/\mu < 1$, the steady state backlog of the server (time to empty the queue with no further arrival) (B^s) and the backlog immediately after an arrival (B^a) satisfy $\mathbb{P}(B^s > x) = \rho e^{\mu(1-\rho)x}$ and $\mathbb{P}(B^a > x) = e^{\mu(1-\rho)x}$, respectively (see [31] and [32, Chap. 11.2]). In our case the server has speed r and the size of arriving load units is exponentially distributed with parameter μ. Therefore $\rho = \lambda/(r\mu)$. Then, $\mathbb{P}(X^a > c) = e^{-\frac{r\mu-\lambda}{r}c}$ for $r\mu > \lambda$. Fixing $\varepsilon = \mathbb{P}(X^a > c_a)$ implies $c_a = -\log(\varepsilon)\,\frac{r}{r\mu-\lambda}$. If we set c_a according to this expression, the resulting LB (c_a, r) assures $P^a_{(c_a,r)} \leq \varepsilon$. To compute $P^s_{(c,r)}$ we have $\mathbb{P}(X^s > c) = \rho e^{-\frac{r\mu-\lambda}{r}c}$ for $r\mu > \lambda$. Setting $\varepsilon = \mathbb{P}(X^s > c_s)$, we have $c_s = -\log\left(\frac{\varepsilon}{\rho}\right)\frac{r}{r\mu-\lambda}$, which ensures $P^s_{(c_s,r)} \leq \varepsilon$.

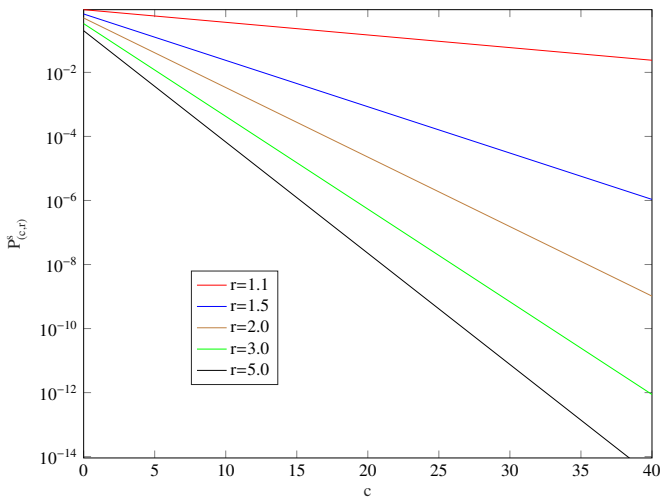

Figure 6: $P^s_{(c,r)}$ as a function of c for different rates r in the M/M/1-system with $\lambda = \mu = 1$ and logarithmic y-axis.

Results for the fluid model are exact and correspond to the analytical results of the M/M/1 queue. Fig. 6 and Fig. 7 show an example with $\lambda = \mu = 1$. For r only slightly larger than 1, a huge c is necessary to obtain small $P^s_{(c,r)}$ and $P^a_{(c,r)}$. Due to the simple structure of the model, the difference between $P^a_{(c,r)}$ and $P^s_{(c,r)}$ is small. Tab. 1 shows

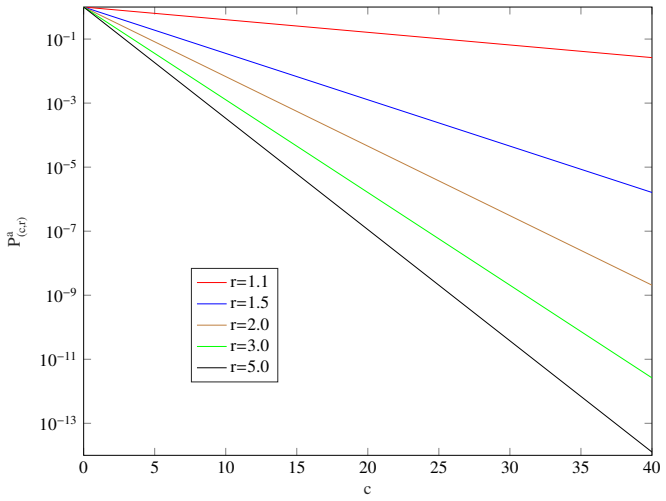

Figure 7: $P^a_{(c,r)}$ as a function of c for different rates r in the M/M/1-system with $\lambda = \mu = 1$ and logarithmic y-axis.

	$\zeta_c = \zeta_r = 1$		$\zeta_c = 2, \zeta_r = 1$		$\zeta_c = 1, \zeta_r = 2$	
ε	c	r	c	r	c	r
0.1	3.88	2.46	3.41	3.08	4.54	2.03
0.01	6.81	3.09	6.21	3.87	7.76	2.46
0.001	9.74	3.44	8.79	4.67	10.79	2.78
0.0001	12.42	3.87	11.52	4.99	13.66	3.07
0.00001	15.10	4.21	14.21	5.27	16.31	3.40
0.000001	17.53	4.70	16.70	5.79	19.15	3.59

Table 1: Optimized LBs for different cost functions, different bounds ε.

results for the LB configuration for different cost functions and different ε bounds. Results for $P^s_{(c,r)}$ are similar. It can be seen that by selecting an appropriate cost function the relation between r and c can be adequately chosen.

Often the maximal size of load units that can be submitted to a system is restricted and parameter r has to be adjusted to obtain $P^a_{(c,r)} \leq \varepsilon$. We consider this situation for the example and assume that $c \leq 15$. Choosing $c = 15$ implies $P^a_{(15,r)} \geq P^a_{(15,\infty)} = 3.059 \cdot 10^{-7}$. The second column in Tab. 2 contains the values for r that are necessary to obtain $P^a_{(15,r)} \leq \varepsilon$ using a single LB. In this situation the arrival envelope can be improved by adding a second LB (c_2, r_2) such that $c_2 > c_1, r_2 < r_1$. There are different ways to compute the parameters for the second LB, here we minimize r_2 for $c_2 = 100$. Results are given in column 4 of Tab. 2. It can be noticed that the second LB allows one to reduce the final rate significantly, especially for smaller values of ε.

ε	r_1	c_1	r_2	c_2
0.1	1.19	15.00	1.05	100.00
0.01	1.45	15.00	1.09	100.00
0.001	1.86	15.00	1.11	100.00
0.0001	2.60	15.00	1.19	100.00
0.00001	4.31	15.00	1.23	100.00
0.000001	12.67	15.00	1.45	100.00

Table 2: Optimized sequences of two LBs with $c_1 = 15$ and $c_2 = 100$ for different bounds ε.

	$\zeta_c = \zeta_r = 1$		$\zeta_c = 2, \zeta_r = 1$		$\zeta_c = 1, \zeta_r = 2$	
ε	c	r	c	r	c	r
0.1	4.18	2.78	3.61	3.54	5.02	2.22
0.01	7.65	3.58	6.83	4.66	8.65	2.88
0.001	10.79	4.31	9.70	5.83	11.99	3.42
0.0001	13.95	4.79	12.52	6.76	15.73	3.60
0.00001	16.83	5.40	15.47	7.22	18.85	4.00
0.000001	19.81	5.81	18.09	8.15	22.54	4.07

Table 3: Optimized LBs for different cost functions and different bounds ε in case of correlated exponentially distributed inter-arrival times and load sizes.

6.1.2. Dependent exponentially distributed inter-arrival times and load sizes

We consider the fluid model governed by matrix

$$\boldsymbol{Q} = \left(\begin{array}{ccc|ccc} -1 & 1 & 0 & 0 & 0 & 0 \\ 0 & -2 & 2 & 0 & 0 & 0 \\ 0 & 0 & -3 & 3 & 0 & 0 \\ \hline 1 & 0 & 0 & -3 & 2 & 0 \\ 0 & 1 & 0 & 0 & -2 & 1 \\ 0 & 0 & 1 & 0 & 0 & -1 \end{array}\right).$$

This matrix is of type $\boldsymbol{Q}_4$ as defined in (16). The matrix describes the same exponential marginal distributions as before but the sojourn times in $\mathcal{S}_-$ and $\mathcal{S}_+$ are negatively correlated with correlation coefficient $-13/36$ (see [33, 34] for further details). This implies for our model that after short inter-arrival times the probability of a larger arrival increases. For this model a closed form solution is not available, even for the single buffer case. In Fig. 8 and Fig. 9, we plot $P^s_{(c,r)}$ and $P^a_{(c,r)}$, which are obtained from the numerical analysis of the fluid model. Comparing the result in Fig. 6/7 and Fig. 8/9, it can be seen that the negative correlation increases the tail probabilities by about two orders of magnitude for the same exponential marginal distributions. This, of course, also has an effect on the dimensioning of the LBs as it is exemplified in Tab. 3 and 4. A comparison of these results with the ones in Tab. 1 and 2 indicates that correlation increases the necessary rate and size of the LBs for all ε values.

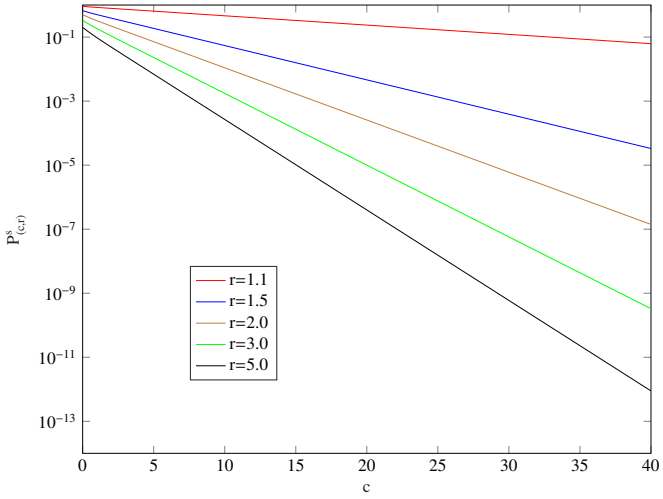

Figure 8: $P^s_{(c,r)}$ as a function of c for different rates r in case of correlated exponentially distributed inter-arrival times and load sizes.

ε	r_1	c_1	r_2	c_2
0.1	1.26	15.00	1.08	100.0
0.01	1.67	15.00	1.15	100.0
0.001	2.42	15.00	1.21	100.0
0.0001	4.00	15.00	2.30	100.0
0.00001	8.20	15.00	4.40	100.0
0.000001	30.39	15.00	9.90	100.0

Table 4: Optimized sequences of two LBs with $c_1 = 15$ and $c2 = 100$ for different bounds ε in case of correlated exponentially distributed inter-arrival times and load sizes.

6.1.3. Dependent complex inter-arrival times and load sizes

In this example we consider the fluid queue governed by matrix

$$\boldsymbol{Q} = \left(\begin{array}{ccc|cc} -2 & 2 & 0 & 0 & 0 \\ 0 & -2 & 0 & 2 & 0 \\ 0 & 0 & -10 & 0 & 10 \\ \hline 1.782 & 0 & 0.018 & -1.8 & 0 \\ 0.018 & 0 & 0.182 & 0 & -0.2 \end{array}\right).$$

This matrix is also of the type $\boldsymbol{Q}_4$ according to (16). We have a MAP as arrival process with a hyper-exponential marginal distribution and positive correlation between the arrivals. The size of the arriving load depends on the subsequent phase of the arrival process. If the next arrival is from the first (fast) phase of the MAP, then the size of load is Erlang 2 distributed with mean 1, otherwise load is exponentially distributed

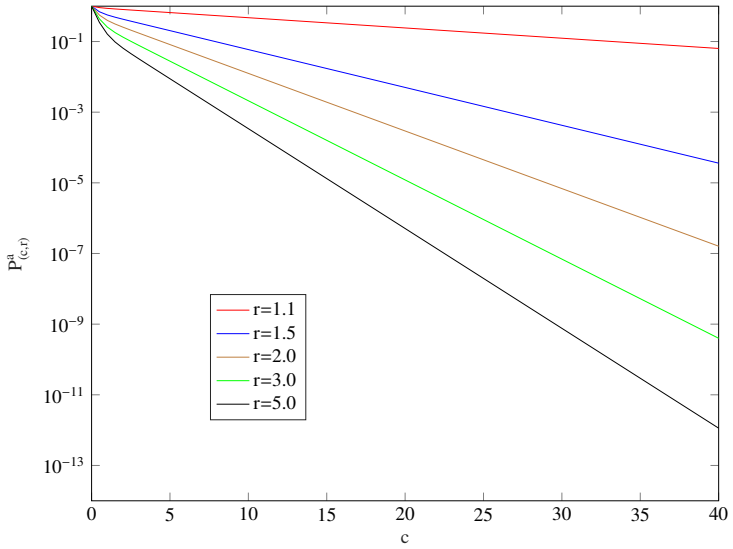

Figure 9: $P^a_{(c,r)}$ as a function of c for different rates r in case of correlated exponentially distributed inter-arrival times and load sizes.

with rate 10. The example indicates that complicated dependencies can be described by matrix $\boldsymbol{Q}$.

The mean inter-arrival time is 1 and the mean size of the load is $0.9 \cdot 1 + 0.1 \cdot 0.1 = 0.91$ which is less than the mean load of the previous examples. Fig. 10 and Fig. 11 plot again $P^s_{(c,r)}$ and $P^a_{(c,r)}$ as a function of c for different rates r. For smaller values of r, the probabilities of large buffer levels remain high because of the correlated arrivals, whereas for larger values of r, the probabilities are smaller than in the previous example because of the Erlang distributed arrival sizes.

The results for the dimensioning of the LBs, in Tab. 5 and 6, also confirm that for smaller values of ε larger values of r and/or c are required than before but for smaller ε this no longer is the case because the Erlang distributed load sizes and the low arrival rate in the second state of the MAP become dominant.

	$\zeta_c = \zeta_r = 1$		$\zeta_c = 2, \zeta_r = 1$		$\zeta_c = 1, \zeta_r = 2$	
ε	c	r	c	r	c	r
0.1	3.66	3.36	3.10	4.08	4.38	2.90
0.01	5.88	4.09	5.06	5.07	6.88	3.44
0.001	7.88	4.66	6.80	6.12	9.18	3.82
0.0001	9.70	5.21	8.60	6.67	11.29	4.16
0.00001	11.51	5.65	10.28	7.28	13.16	4.53
0.000001	13.25	6.07	11.93	7.82	15.22	4.75

Table 5: Optimized LB with different cost functions and different bounds ε in case of correlated inter-arrival times and load sizes.

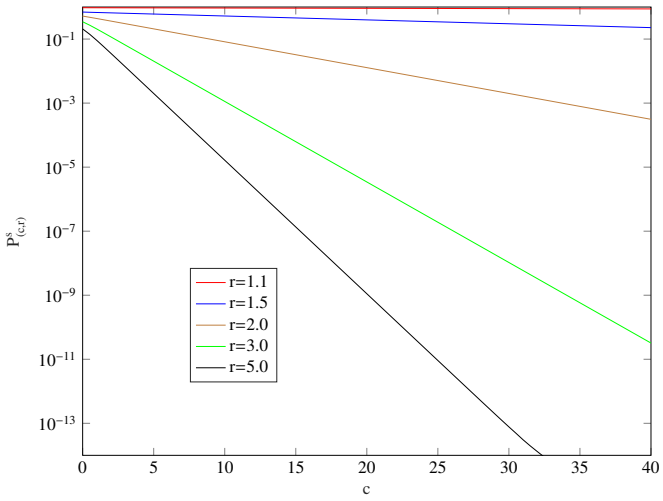

Figure 10: $P^s_{(c,r)}$ as a function of c for different rates r in case of correlated inter-arrival times and load sizes.

ε	r_1	c_1	r_2	c_2
0.1	1.90	15.00	1.58	100.0
0.01	2.25	15.00	1.71	100.0
0.001	2.64	15.00	1.94	100.0
0.0001	3.15	15.00	2.29	100.0
0.00001	3.85	15.00	2.84	100.0
0.000001	4.86	15.00	3.30	100.0

Table 6: Optimized sequences of two LBs with $c_1 = 15$, $c2 = 100$ and different bounds ε in case of correlated inter-arrival times and load sizes.

6.1.4. Tandem Networks

We consider a tandem network consisting of a sequence of stations without cross traffic. In QNA it is assumed that service times at the stations are independently drawn from the service times distribution. In NC, each load unit has the same size in each station and the service times depends only on the speed of the server. The latter view is often more realistic, e.g., in case of the transmission of packets over a sequence of connections in a computer network.

If we consider a tandem network with q stations with lower service curves $\beta_p(t) = s_p \cdot t$ for $p \in \{1, \ldots, q\}$, then $\beta = \beta_1 \circledast \beta_2 \circledast \ldots \circledast \beta_q = s \cdot t$ with $s = \min_{p=1,\ldots,q} s_p$. By defining the arrival curve that corresponds to an LB with rate $r = s$ and capacity c, we can compute the output flow, backlog level and delay for all $t \geq 0$ according to (1)-(3)

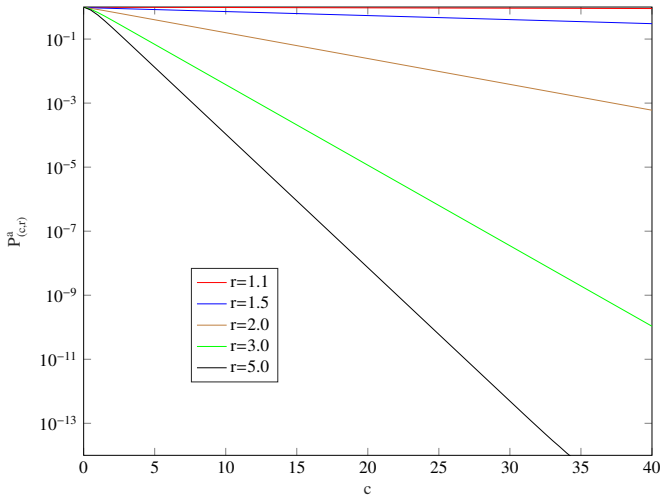

Figure 11: $P^a_{(c,r)}$ as a function of c for different rates r in case of correlated inter-arrival times and load sizes.

as follows

$$\begin{aligned}
\alpha'(t) &\geq \inf_{0\leq\tau\leq t} (c\delta(\tau > 0) + \tau \cdot r + (t-\tau)\cdot s) = s \cdot t,\\
b_t &\leq \alpha(t) - \alpha'(t) = c,\\
w_t &\leq \frac{b_t}{s} = \frac{c}{s}.
\end{aligned}$$

Thus, backlog and delay result from the parameters of the LB in this case. An exact delay bound, which is exceeded with at most probability ε, can be computed for stochastic input processes with the approach proposed in this paper. For input processes defined by a trace, trace driven simulation has to be applied.

For more complex service curves or envelopes, like latency-rate servers, the simple relation between the LB at the input and backlog and delay no longer holds. In this case, results from SNC have to be combined with the results presented here for the determination of the arrival envelope.

6.2. Network Traffic Traces

As an example for a real trace we use a trace from the *MAWI Working Group Traffic Archive* [35, 36]. The archive contains traces for almost every day since 2006 from a 1 Gbps upstream link. We selected the trace from January 1st, 2023 and analyze only the IP packets in it. Timestamps are used to compute the inter-arrival times and the payload to define the packet size. The trace contains 68, 209, 829 IP packets, the statistics of the trace can be found in Tab. 7. It can be seen that the link has an average load of about 47.2 Mbyte/sec.

	mean	variance	min	max
inter-arrival time (msec)	0.0132035	0.000344	0.0	0.8280277
packet size (byte)	623.54328	448434.5	28	1500

Table 7: Statistics of the MAWI trace from January 1st 2023.

ε	n_{opt}	c	r	$P^a_{(c,r)}$
0.1	100	9,000	130,000	0.08683225
	1,000	9,000	130,000	0.08683225
0.01	100	25,500	140,000	0.00669713
	1,000	22,500	140,000	0.00835178
0.001	100	30,000	190,000	0.00037201
	1,000	33,000	170,000	0.00077814
0.0001	100	36,000	180,000	**0.00049550**
	1,000	24,000	220,000	0.00004381
	10,000	18,000	230,000	0.00001755
0.00001	100	22,500	230,000	**0.00003314**
	1,000	22,500	230,000	**0.00003314**
	10,000	21,000	240,000	0.00000320
0.000001	100	25,500	220,000	**0.00004043**
	1,000	27,000	240,000	0.00000000
	10,000	25,500	240,000	0.00000000

Table 8: Results for the LB with trace input (c in bytes, r in bytes/msec).

For dimensioning the LB we use trace driven simulation. From the trace sub-traces of length $n_{sub} = 10,000$, containing consecutive elements, are collected and n_{opt} sub-traces are selected randomly to determine the parameters. For each sub-trace a trace driven simulation of the fluid queue is performed and the probabilities of exceeding given thresholds are estimated after neglecting the first $0.1n_{sub}$ arrivals to avoid the transient phase. For optimization of the rate and size parameters the linear cost function with $\zeta_c = \zeta_r = 1$ is used, where the rate is given in $bytes/msec$ and the size in $bytes$. Furthermore, the parameters are computed with respect to the upper bound of the one sided confidence interval with significance level 0.95. Optimization is done on a grid with grid size 10,000 bytes/msec for the rates and 1,500 bytes for the size. The maximal size of the buffer is restricted to 36,000 bytes. After the optimization has been performed, results are validated by determining $P^a_{(c,r)}$ for a trace driven simulation of the whole trace. We keep the sub-traces for one optimization with a fixed ε but draw new sub-traces whenever ε is modified.

Tab. 8 contains the results for a single LB with different values of ε and n_{opt}. Results that are above the required probability of exceeding the capacity of the LB are printed in boldface. It can be noticed that the for larger values of ε, reliable results are achieved with a small number of replications. However, for $\varepsilon = 10^{-5}$ or 10^{-6}, results are much more sensitive. Usually, these small probabilities of exceeding the available bucket

ε	n_{opt}	c_1	r_1	c_2	r_2	$P^a_{(30000,r_1)(120000,r_2)}$
0.1	100	30,000	130,000	120,000	120,000	0.03323425
	1,000	30,000	130,000	120,000	120,000	0.03323425
0.01	100	30,000	140,000	120,000	130,000	0.00536459
	1,000	30,000	140,000	120,000	130,000	0.00536459
0.001	100	30,000	180,000	120,000	130,000	**0.00163844**
	1,000	30,000	190,000	120,000	160,000	0.00039136
0.0001	100	30,000	230,000	120,000	180,000	0.00005484
	1,000	30,000	230,000	120,000	200,000	0.00002632
0.00001	100	30,000	220,000	120,000	140,000	**0.00042842**
	1,000	30,000	230,000	120,000	180,000	**0.00005484**
0.000001	100	30,000	230,000	120,000	180,000	**0.00005484**
	1,000	30,000	240,000	120,000	230,000	0.00000000

Table 9: Optimized sequences of two leaky buckets with $c_1 = 30,000$ and $c2 = 120,000$ for the trace input and different bounds ε.

size depend on a few large packets which often arrive one after the other with small inter-arrival times. If these packets are not present in a trace driven simulation, or in a real system in the trace used for dimensioning, results can become wrong. In our experiments we took the trace as *ground truth*, which is obviously not the case in a real system where a trace is only one random observation of the real behavior. Our results indicate that for dimensioning LBs large traces are necessary if small probabilities are required. However, even probabilities around 10^{-6} seem to be achievable if sufficiently reliable data is available.

We now consider two LBs in series and assume that the first LB has a size of 30,000 bytes (i.e., 20 Ethernet packets of maximal size). For this LB we determine the rate r such that threshold ε is not exceeded. The value is determined from the same set of randomly selected sub-traces that are used for optimization of the parameters of the second LB. The values are then used as (c_1, r_1) for two LBs in series. Afterwards, r_2 is computed for a LB with $c_2 = 120,000$ bytes.

Results for the two LBs can be found in Tab. 9. It can be noticed that the use of a second LB has not as drastic effects as for the stochastic input models, which indicates that the trace exhibits more correlation and long range dependencies, but the rate for the second bucket is always smaller than for the first bucket.

7. Conclusions

In this paper we develop an approach to compute the parameters of affine and piecewise linear arrival envelopes, corresponding to sequences of LBs, for stochastic input processes. It has been shown how these models are related to fluid queues and how the parameters can be derived from the stationary analysis of appropriate fluid models. With the presented method it is possible to compute or estimate a combination of LBs that assures for a given arrival process that $1 - \varepsilon$ of the arriving load is immediately submitted to the system or network.

There are several topics for extension of the presented approach. First, an integration into stochastic network calculus seems to be interesting and important. In the presented analysis it is assumed that arriving load that exceeds the capacity of the LB is delayed but not dropped. Since fluid queues with finite capacity can also be analyzed, it should be possible to consider also the case when load is dropped. The case where load units can be partially dropped if they are too a large for the available capacity seems to be relatively simple, whereas the case where the whole packet is dropped if it exceeds the capacity is more challenging.

References

[1] E. D. Lazowska, J. Zahorjan, G. S. Graham, K. C. Sevcik, Quantitative System Performance - Computer Systems Analysis Using Queueing Networks, Prentice-Hall, 1984, available oline: http://www.cs.washington.edu/home/lazowska/qsp/.

[2] M. F. Neuts, Matrix-geometric solutions in stochastic models: an algorithmic approach, Courier Corporation, 1994.

[3] R. L. Cruz, A calculus for network delay, part I: network elements in isolation, IEEE Transactions on Information Theory 37 (1) (1991) 114–131.

[4] R. L. Cruz, A calculus for network delay, part II: network analysis, IEEE Transactions on Information Theory 37 (1) (1991) 132–141.

[5] J. L. Boudec, P. Thiran, Network Calculus: A Theory of Deterministic Queuing Systems for the Internet, Vol. 2050 of Lecture Notes in Computer Science, Springer, 2001.

[6] W. J. Stewart, Introduction to the numerical solution of Markov chains, Princeton University Press, 1995.

[7] P. Buchholz, J. Kriege, I. Felko, Input Modeling with Phase-Type Distributions and Markov Models - Theory and Applications, Springer Briefs in Mathematics, Springer, 2014. doi:10.1007/978-3-319-06674-5.

[8] A. Horvath, M. Telek, Approximating heavy tailed behaviour with phase type distributions, Advances in Algorithmic Methods for Stochastic Models (2000) 191–214.

[9] M. Moy, K. Altisen, Arrival curves for real-time calculus: The causality problem and its solutions, in: Tools and Algorithms for the Construction and Analysis of Systems TACAS, 2010, pp. 358–372.

[10] S. Vastag, Arrival and delay curve estimation for SLA calculus, in: Winter Simulation Conference, WSC '12, Berlin, Germany, December 9-12, 2012, 2012.

[11] D. L. Iglehart, Extreme values in the GI/G/1 queue, The Annals of Mathematical Statistics (1972) 627–635.

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

[13] M. Fidler, A. Rizk, A guide to the stochastic network calculus, IEEE Communications Surveys & Tutorials 17 (1) (2014) 92–105. doi:10.1109/COMST.2014.2337060.

[14] Y. Jiang, Y. Liu, Stochastic Network Calculus, Springer, 2008.

[15] C. Li, A. Burchard, J. Liebeherr, A network calculus with effective bandwidth, IEEE/ACM Trans. Netw. 15 (6) (2007) 1442–1453. doi:10.1145/1373476.1373494.

[16] S. Ahn, V. Ramaswami, Fluid flow models and queues – a connection by stochastic coupling, Stochastic Models 19 (3) (2003) 325–348.

[17] V. Ramaswami, Matrix analytic methods for stochastic fluid flows, Teletraffic science and engineering (1999) 1019–1030.

[18] P. Buchholz, A. Meszaros, M. Telek, Stationary analysis of a constrained markov fluid model with two buffers, Stochastic Models (to appear) (2024). doi:10.1080/15326349.2024.2339247.

[19] M. Boyer, E. Le Corronc, A. Bouillard, Deterministic network calculus: From theory to practical implementation, John Wiley & Sons, 2018.

[20] D. Wrege, E. Knightly, H. Zhang, J. Liebeherr, Deterministic delay bounds for vbr video in packet-switching networks: fundamental limits and practical trade-offs, IEEE/ACM Transactions on Networking 4 (3) (1996) 352–362. doi:10.1109/90.502234.

[21] F. Ciucu, J. Schmitt, Perspectives on network calculus: no free lunch, but still good value, SIGCOMM Comput. Commun. Rev. 42 (4) (2012) 311–322. doi:10.1145/2377677.2377747.

[22] F. Poloczek, F. Ciucu, Scheduling analysis with martingales, Perform. Evaluation 79 (2014) 56–72. doi:10.1016/j.peva.2014.07.004.

[23] M. A. Beck, S. A. Henningsen, S. B. Birnbach, J. B. Schmitt, Towards a statistical network calculus - dealing with uncertainty in arrivals, in: 2014 IEEE Conference on Computer Communications, INFOCOM 2014, Toronto, Canada, April 27 - May 2, 2014, IEEE, 2014, pp. 2382–2390. doi:10.1109/INFOCOM.2014.6848183.

[24] S. Mao, S. S. Panwar, A survey of envelope processes and their applications in quality of service provisioning, IEEE Commun. Surv. Tutorials 8 (1-4) (2006) 2–20. doi:10.1109/COMST.2006.253272.

[25] P. Nikolaus, J. B. Schmitt, F. Ciucu, Dealing with dependence in stochastic network calculus - using independence as a bound, in: M. Gribaudo, E. S. Sopin, I. A. Kochetkova (Eds.), Analytical and Stochastic Modelling Techniques and Applications - 25th International Conference, ASMTA 2019, Moscow, Russia, October 21-25, 2019, Proceedings, Vol. 12023 of Lecture Notes in Computer Science, Springer, 2019, pp. 71–84. doi:10.1007/978-3-030-62885-7_6.

[26] A. Badescu, L. Breuer, A. Da Silva Soares, G. Latouche, M.-A. Remiche, D. Stanford, Risk processes analyzed as fluid queues, Scandinavian Actuarial Journal 2005 (2) (2005) 127–141.

[27] P. Buchholz, J. Kriege, Fitting correlated arrival and service times and related queueing performance, Queueing Syst. Theory Appl. 85 (3-4) (2017) 337–359. doi:10.1007/s11134-017-9514-5.

[28] N. Akar, O. Gursoy, G. Horvath, M. Telek, Transient and first passage time distributions of first-and second-order multi-regime markov fluid queues via mefication, Methodology and Computing in Applied Probability 23 (2021) 1257–1283.

[29] A. M. Law, Simulation modeling and analysis, 5th Edition, Mcgraw-hill New York, 2015.

[30] A. Scheffler, S. Bondorf, Network calculus for bounding delays in feedforward networks of FIFO queueing systems, in: A. Abate, A. Marin (Eds.), Quantitative Evaluation of Systems - 18th International Conference, QEST 2021, Paris, France, August 23-27, 2021, Proceedings, Vol. 12846 of Lecture Notes in Computer Science, Springer, 2021, pp. 149–167. doi:10.1007/978-3-030-85172-9_8.

[31] P. G. Harrison, Response time distributions in queueing network models, in: IFIP International Symposium on Computer Performance Modeling, Measurement and Evaluation, Springer, 1993, pp. 147–164.

[32] W. J. Stewart, Probability, Markov chains, queues, and simulation: the mathematical basis of performance modeling, Princeton university press, 2009.

[33] M. Bladt, B. F. Nielsen, On the construction of bivariate exponential distributions with an arbitrary correlation coefficient, Stochastic Models 26 (2) (2010) 295–308.

[34] P. Buchholz, On the representation of correlated exponential distributions by phase type distributions, ACM SIGMETRICS Performance Evaluation Review 49 (3) (2022) 73–78.

[35] Mawi working group traffic archive, `https://mawi.wide.ad.jp/mawi/`.

[36] K. Cho, K. Mitsuya, A. Kato, Traffic data repository at the {WIDE} project, in: 2000 USENIX Annual Technical Conference (USENIX ATC 00), 2000.