A First Course in Stochastic Network Calculus

Dipl.-Math. Michael Alexander Beck

January 9, 2014

Contents

Foreword

This *First Course in Stochastic Network Calculus* is a “First” course in two different perspectives. One can see it as a introductory course to Stochastic Network Calculus. On the other side it was my first course I have given about SNC to a few students in June 2012. It builds on the lecture *Performance Modelling in Distributed Systems [6]* at the University of Kaiserslautern. The concepts of stochastic network calclulus parallels those of deterministic network calculus. This is why I reference on the lecture of 2011 at several points to stress these connections. This document however is thought of a stand-alone course and hence a deep study of [6] is not necessary (but recommended).

This course contains a rather large probability primer to ensure the students can really grasp the expressions, which appear in stochastic network calculus. A student familiar with probability theory might skip this first chapter and delve directly into the stochastic network calculus. For each topic exercises are given, which can (and should) be used to strengthen the understanding of the presented definitions and theory.

This document is still in process and hopefully will evolve at some day into a fully grown course about Stochastic Network Calculus, providing a good overview over this exciting theory. Hence, please provide feedback to `beck@cs.uni-kl.de`.

- Michael Beck

1 Basics of Probability

2 Stochastic Arrivals and Service

In this chapter we carefully motivate the need for stochastic arrivals and service and see two approaches how such stochastic processes can be described and bounded. Each of them leads to its own stochastic extension of the deterministic network calculus (see [3, 4, 1, 5]). So we cannot talk about *the* stochastic network calculus, but rather have to distinguish between several approaches, each with its own characteristics and strengths.

Overall we use a discrete time-scale, since it keeps formulas somewhat easier.

2.1 Stochastic Arrivals

2.2 Stochastic Service

In this chapter, we first discuss different alternatives for defining service guarantees in a determinstic context. Then we introduce the definition of stochastic service models before presenting the important results on how to concatenate stochastic servers in a network setting.

2.2.1 The Different Notions of Deterministic Service Curves

In this section, we make a short excursion into deterministic service curves and strict service curves. We talk about the differences between those two and generalize the concept of a service curve to fit our needs. For a more detailed survey on the zoo of different service guarantees [2] is a great point to start from.

We motivate stochastic service similarly to the stochastic arrivals and see service elements, which offer certainly *some* amount of service, but only achieve a deterministic lower bound equal to *zero*.

Throughout the rest of this chapter we consider a node working on an arrival flow A and denote the departure flow by D. Remember that in a backlogged period $[m, n]$ it holds that $A(k) > D(k)$ for all $k \in [m, n]$.

Definition 2.1. A service element offers a *strict service curve* $\beta : \mathbb{N}_0 \to \mathbb{R}_0^+$, if for all backlogged periods $[m, n]$ and all input flows A it holds that:

$$D(n) - D(m) \geq \beta(n - m).$$

Another definition for strict service curves, which incorporates the arrival flow A, is given by the following lemma:

Lemma 2.2. *Let $m \in \mathbb{N}_0$ be arbitrary and denote by $n^*(m)$ the beginning of the next(!) backlogged period. The service element offers a strict service curve β if and only if:*

$$D(n) \geq \beta(n-m) + D(m) \wedge A(n) \tag{2.1}$$

for all $n < n^*(m)$ and all input flows A.

Proof. Let us first assume S offers a strict service curve β and $D(m) = A(m)$, then for all $n < n^*(m)$ it also holds that $D(n) = A(n)$, and thus:

$$D(n) \geq A(n) \geq \beta(n-m) + D(m) \wedge A(n)$$

Now assume $D(m) < A(m)$, i.e. we start in a backlogged period. If also $D(n) < A(n)$ holds, then $[m, n]$ is a backlogged period and we get by our assumption $D(n) - D(m) \geq \beta(n-m)$. The case $D(n) = A(n)$ was already treated, again we have:

$$D(n) \geq A(n) \geq \beta(n-m) + D(m) \wedge A(n).$$

Assume now the service element fulfills equation (2.1), we need to show β is a strict service curve. Let $[m, n]$ be an arbitrary backlogged period, then surely $n < n^*(m)$. Assume $\beta(n-m) + D(m) > A(n)$, then by using equation (2.1) we would obtain:

$$D(n) \geq \beta(n-m) + D(m) \wedge A(n) = A(n),$$

which would imply $D(n) = A(n)$, which is a contradiction to our assumption that $[m, n]$ is a backlogged period. Hence $\beta(n-m) + D(m) \leq A(n)$ and again by equation (2.1):

$$D(n) \geq \beta(n-m) + D(m) \wedge A(n) = \beta(n-m) + D(m)$$

from which it follows that:

$$D(n) - D(m) \geq \beta(n-m).$$

□

There exists also the more general (non-strict) service curve[1].

Definition 2.3. A service element offers a *service curve* $\beta : \mathbb{N}_0 \to \mathbb{R}_0^+$ if for all $n \geq m \in \mathbb{N}_0$ and all input flows A it holds that:

$$D(n) \geq \min_{0 \leq k \leq n} \{A(k) + \beta(n-k)\}.$$

One might ask: What is the difference between a service curve and a strict service curve? It is easy to find examples, in which the departures of some node differ, depending if one uses a service curve or a strict service curve (you are asked to construct one in the exercises). However, we want to shed a bit more light on the differences between service curve definitions. Therefore we introduce Lindley's equation.

[1] The advantage of service curves over strict service curves is, that we can convolute service elements offering service curves into a single service element offering again a service curve. Such a property does in general not hold for strict service curves (see exercises).

Definition 2.4. A service element fulfills *Lindley's equation* if for all $q(n) := A(n)-D(n)$ and all input flows A it holds that:

$$q(n+1) = [q(n) + a(n+1) - s(n+1)]^+,$$

where $s(n)$ is the amount of data the service element can process at time n.

Lindley's equation states that the amount of data, which is queued at the service element, is the amount of data, which has been queued the time step before plus the amount of data which arrives in the meantime minus the amount of data the service element can process in that timestep. Of course, it is possible that the service element could serve more data than have been queued and newly arrived, hence we have to take the maximum with zero in Lindley's equation. This behaviour basically means that our service element is not "lazy" and always works as much as it can, i.e., it is work-conserving. The next lemma shows, that under this behaviour, a service curve becomes a strict service curve (and more).

Lemma 2.5. *Let a service element fulfill Lindley's equation. The following three items are equivalent:*

- *It offers a service curve* β*.*
- *It holds* $S(n) - S(k) \geq \beta(n-k)$ *for all* $n \geq k \in \mathbb{N}_0$*.*
- *It offers a strict service curve* β*.*

Here $S(n)$ *is defined by* $S(n) := \sum_{k=0}^{n} s(k)$ *with* $s(k)$ *from Lindley's equation.*

Proof. We first show that from the first item follows the second, from which in turn follows the third. In the last step we show, that the third item implies again the first.

(1)⇒(2):

First we claim that

$$q(n) = \max_{0\leq k\leq n}\{A(n) - A(k) - S(n) + S(k)\} \tag{2.2}$$

and prove this by induction over n: For $n = 1$, we have by Lindley's equation (assuming $q(0) = a(0) = s(0) = 0$):

$$q(1) = [a(1) - s(1)]^+ = \max_{0\leq k\leq 1}\{A(1) - A(k) - S(1) + S(k)\}$$

Now assume (2.2) holds for n, then we have again using Lindley's equation:

$$\begin{aligned} q(n+1) &= [q(n) + a(n+1) - s(n+1)]^+ \\ &= [\max_{0\leq k\leq n}\{A(n) - A(k) - S(n) + S(k)\} + a(n+1) - s(n+1)]^+ \\ &= [\max_{0\leq k\leq n}\{A(n+1) - A(k) + S(n+1) + S(k)\}]^+ \\ &= \max_{0\leq k\leq n+1}\{A(n+1) - A(k) + S(n+1) + S(k)\}. \end{aligned}$$

Assume now S offers a service curve β and let $n, k \in \mathbb{N}_0$ be arbitrary such that $n \geq k$. From (2.2) we know already that for all flows A

$$q(n) = A(n) - D(n) \geq A(n) - A(k) - S(n) + S(k).$$

And by the definition of service curves:

$$D(n) \geq \min_{0 \leq k' \leq n} \{A(k') + \beta(n - k')\}$$

Combining these, we obtain:

$$\begin{aligned}
S(n) - S(k) &\geq D(n) - A(k) \\
&\geq \min_{0 \leq k' \leq n} \{A(k') - A(k) + \beta(n - k')\} && (2.3) \\
&= \min_{0 \leq k' \leq k} \{A(k') - A(k) + \beta(n - k')\} \wedge && (2.4) \\
&\quad \min_{k < k' \leq n} \{A(k') - A(k) + \beta(n - k')\}
\end{aligned}$$

Now use the flow A^*, defined by:

$$\begin{aligned}
a^*(1) &\in \mathbb{R}_0^+ \\
a^*(2) &= a^*(3) = \ldots = a^*(k) = 0
\end{aligned}$$

$$\begin{aligned}
a^*(k+1) &= \beta(n-k) - \beta(n-k-1) \\
a^*(k+2) &= \beta(n-k-1) - \beta(n-k-2) \\
a^*(k+3) &= \beta(n-k-2) - \beta(n-k-3) \\
&\vdots \\
a^*(n) &= \beta(1) - \beta(0) = \beta(1)
\end{aligned}$$

Note that by this construction we have $A^*(k') = A^*(k)$ for all $k' \leq k$. Inserting that flow into (2.4) we eventually have:

$$\begin{aligned}
S(n) - S(k) &\geq \min_{0 \leq k' \leq k} \{0 + \beta(n - k')\} \wedge \min_{k < k' \leq n} \left\{ \beta(n - k') + \sum_{l=k+1}^{k'} a^*(l) \right\} \\
&= \beta(n-k) \wedge \min_{k < k' \leq n} \left\{ \beta(n - k') + \sum_{l=k+1}^{k'} \beta(n - l + 1) - \beta(n - l) \right\} \\
&= \beta(n-k) \wedge \min_{k < k' \leq n} \{\beta(n - k') + \beta(n-k) - \beta(n - k'\} = \beta(n-k)
\end{aligned}$$

for all $n \geq k \in \mathbb{N}_0$.

Hence the first item implies second item.

(2)⇒(3)

If we assume the second item, we have by Lindley's equation for each k in an arbitrary backlogged period $[m, n]$

$$\begin{aligned} A(k) - D(k) = q(k) &= [q(k-1) + a(k) - s(k)]^+ \\ &= q(k-1) + a(k) - s(k) = \\ &= A(k-1) - D(k-1) + a(k) - s(k) \\ &= A(k) - D(k) + d(k) - s(k) \end{aligned}$$

and hence $d(k) = s(k)$ for all $k \in [m, n]$. Using the second item, it follows:

$$D(n) - D(m) = \sum_{k=m+1}^{n} d(k) = \sum_{k=m+1}^{n} s(k) = S(n) - S(m) \geq \beta(n-m)$$

Since we have chosen $[m, n]$ to be an arbitrary backlogged period, we have shown that β is a strict service curve. It is left to show, that from the third item follows again the first item.

(3)⇒(1)

Let n be arbitrary and consider first the case of $D(n) < A(n)$, i.e. n lies in a backlog period and define m^* by the last time before n with $A(m^*) = D(m^*)$. It holds then:

$$\beta(n - m^*) \leq D(n) - D(m^*) = D(n) - A(m^*)$$

Hence:

$$D(n) \geq \beta(n - m^*) + A(m^*) \geq \min_{0 \leq k \leq n} \{A(k) + \beta(n-k)\}$$

Consider now the case of $D(n) = A(n)$, then follows immediately:

$$D(n) = A(n) \geq \min_{0 \leq k \leq n} \{A(k) + \beta(n-k)\}$$

□

Notice that a strict service curve still gives the service element some space to be "lazy", since it only has to work as much as needed to fulfill the constraint given by β. If there is more service available even the strict service curve does not necessarily claim it. Following this train of thoughts one can sort servers by their "lazyness" in the following increasing order:

- A server fulfills Lindley's equation.
- A server offers a strict service curve.
- A server offers a service curve.

However keep in mind, that Lindley's equation does not hold any information about *how much* data the service element processes. In a certain sense one can see the sum of Lindley's equation and a (strict) service curve, as the "strictest" service guarantee possible. Since it gives us information about how much service is available at least ($S(n) - S(m) \geq \beta(n-m)$) and also states that the service element is never lazy, i.e., we always use the whole available service (as seen in the above proof it holds under Lindley's equation: $d(k) = s(k)$ in each backlogged period). This is exactly the second item in the previous lemma. What you should keep in mind is, that if some service element fulfills Lindley's equation, the three kind of service curves immediately fall together.

2.2.2 Definition of Stochastic Service Models

As announced we encounter the same problems with stochastic service, as for stochastic arrivals. As an example consider a service element which offers a constant rate of service (denoted by c) and Lindley's equation. There are two flows entering, which we call A_1 and A_2 and we set A_1 as prioritized flow. This means, that for the second flow only the amount of service, which is leftover by A_1 is offered. We further assume, that A_1 does not queue. This means if A_1 sends in one time step an amount of data larger c, we drop all exceeding data. If we consider one time step and denote by $s_l(n)$ the amount of service A_2 receives in this time step we have $s_l(n) = [c - a_1(n)]^+$(Note: If A_1 queued we would have $s_l(n) = [c - a_1(n) - q_1(n-1)]^+$, where $q_1(n-1)$ would be the queue length of A_1 at time $n-1$). Now we ask for some service curve β, which the server might offer for A_2. Depending on the arrivals, this might not be possible. If we take for example i.i.d. exponentially distributed increments $a_1(n)$ we get:

$$\mathbb{P}(s_l(n) = 0) = \mathbb{P}([c - a_1(n)]^+) = \mathbb{P}(a_i(n) \geq c) = e^{-\lambda c} > 0$$

Hence the best service curve we can expect is $\beta(n) = 0$ for all $n \in \mathbb{N}_0$.

Before we give a method to bound stochastic service elements, we need a generalisation of the above concepts.

Definition 2.6. Define the index set $\Lambda(\mathbb{N}_0) := \{(i,j) \in \mathbb{N}_0 \times \mathbb{N}_0 : i \leq j\}$ triangular subset of the lattice $\mathbb{N}_0 \times \mathbb{N}_0$. We call a set indexed by $\Lambda(\mathbb{N}_0)$ a *triangular array* (over $\mathbb{R}$) and denote it to be an element of $\mathcal{D}$.

We can list the indices of $\Lambda(\mathbb{N}_0)$ by:

$$\begin{array}{llll} (0,0); & & & \\ (0,1); & (1,1); & & \\ (0,2); & (1,2); & (2,2); & \\ (0,3); & \cdots & \cdots & (3,3); \\ \vdots & & & \ddots \end{array}$$

The set of such doubly indexed stochastic processes is denoted by $\mathcal{S}_2$.

Definition 2.7. If A is a flow we define the doubly indexed stochastic process $A(\cdot,\cdot) : \Lambda(\mathbb{N}_0) \to \mathbb{R}_0^+$ by:

$$A(m,n) := A(n) - A(m) \qquad \forall\, n \geq m \geq 0$$

The set of such doubly indexed stochastic processes *resulting from flows* is denoted by $\mathcal{F}_2$.

Clearly $\mathcal{F}_2 \subsetneq \mathcal{S}_2 \subset \mathcal{D}$. It is important to note the difference between $A(m,n)$ and $A(n-m)$. In the former case we talk about the arrivals in the interval $[m+1,n]$ (an interval of length $n-m$), in contrast to the latter case, in which we consider the arrivals up to time $n-m$. We also need a min-plus-convolution and a min-plus-deconvolution for the bivariate case. Remember and compare the univariate operations:

$$A \otimes B(n) = \min_{0 \le k \le n} \{A(k) + B(n-k)\}$$
$$A \oslash B(n) = \max_{0 \le k} \{A(n+k) - B(k)\}.$$

Definition 2.8. Let $A, B \in \mathcal{D}$. Then the *bivariate convolution* $\otimes : \mathcal{D} \times \mathcal{D} \to \mathcal{D}$ is defined by:

$$A \otimes B(m,n) := \min_{m \le k \le n} \{A(m,k) + B(k,n)\}$$

Further the *bivariate deconvolution* $\oslash : \mathcal{D} \times \mathcal{D} \to \mathcal{D}$ is defined by:

$$A \oslash B(m,n) := \max_{0 \le k \le m} \{A(k,n) - B(k,m)\}$$

One should mention that this bivariate operations are no longer commutative. You are further asked in the exercises, to show that the bivariate definitions are a generalization of the univariate case. We now present the bivariate "service curve".

Definition 2.9. Assume a service element has a flow A as input and the output is denoted by D. Let $S \in \mathcal{S}_2$ be a doubly indexed stochastic process. The service element is a *dynamic S-server*[2] if for all $n \ge 0$ it holds that:

$$D(0,n) \ge A \otimes S(0,n) \qquad \text{(a.s.)}$$

Note that the process $S(m,n)$ does generally *not* lie in $\mathcal{F}_2$ and hence we cannot assume $S(m,n) = S(n) - S(m)$, although we encounter situations in which this equation is met.

Example 2.10. We first consider a deterministic rate-latency server, with service curve $\beta_{R,N}(n) = R \cdot (n-N)$ if $n > N$ and equal to zero otherwise. For $S(k,n) := \beta_{R,N}(n-k)$ we have that a dynamic S-server offers $\beta_{R,N}$ as service curve in the univariate sense:

$$D(n) \ge A \otimes S(0,n) = \min_{0 \le k \le n} \{A(k) + S(k,n)\} = \min_{0 \le k \le n} \{A(k) + \beta_{R,N}(n-k)\} = A \otimes \beta_{R,N}(n)$$

Example 2.11. We return to our example of a service element satisfying Lindley's equation and serving two flows A_1 and A_2, of which the first is prioritized. But we now allow that flow A_1 can queue. As before the service element offers a constant rate of c. In this case, the service element is a dynamic S-server for A_2 with:

$$S(m,n) := [c \cdot (n-m) - A_1(m,n)]^+$$

[2] *In literature there is often the additional assumption of $S(m,n) \ge 0$ for all m and n. This assumption is not necessary and in some situations (see e.g. [REF]) burdens us with unwanted restrictions. Note that, in fact, the case $S(m,n) < 0$ causes no trouble, since $D(0,n) \ge A \otimes S(0,n)$ represents a lower bound. The worst thing, which can happen here, is that the lower bound is useless (nevertheless correct): $D(0,n) \ge A \otimes S(0,n)$ with $A \otimes S(0,n) < 0$.*

To see this assume $n \geq 0$ arbitrary and choose $m \leq n$ maximal such that $A_1(m) = D_1(m)$ and $A_2(m) = D_2(m)$ (such an m exists since all flows are zero for $m = 0$). Our server satisfies:

$$D_1(n) + D_2(n) = D_1(m) + D_2(m) + c \cdot (n - m) = A_1(m) + D_2(m) + c \cdot (n - m)$$

Since $D_1(n) \leq A_1(n)$ we can continue with:

$$D_2(n) \geq D_2(m) + c \cdot (n - m) - (A_1(0, n) - A_1(0, m))$$

We have for sure that $D_2(n) \geq D_2(m) = D_2(m) + 0$ and therefore obtain:

$$\begin{aligned} D_2(n) &\geq D_2(m) + [c \cdot (n - m) - A_1(m, n)]^+ = A_2(m) + [c \cdot (n - m) - A_1(m, n)]^+ \\ &\geq \min_{0 \leq k \leq n} \{A_2(k) + S(k, n)\} = A_2 \otimes S(0, n) \end{aligned}$$

Similar to the previous section, we next present two different ways to bound dynamic S-servers stochastically. Again, we have a tail-bound and an MGF-bound. There is a slight difference to the way we bounded arrivals, though, and we are going to investigate it in some detail. This time we start with the MGF-bound, which is straightforward:

Definition 2.12. A dynamic S-server is $(\sigma(\theta), \rho(\theta))$-*bounded* for some $\theta > 0$, if $\phi_{S(m,n)}(-\theta)$ exists and

$$\phi_{S(m,n)}(-\theta) \leq e^{\theta\rho(\theta)(n-m)+\theta\sigma(\theta)} \qquad \forall\, (m, n) \in \Lambda(\mathbb{N}_0).$$

A dynamic S-server is $f(\theta, n)$-*bounded* for some $\theta > 0$, if $\phi_{S(m,n)}(-\theta)$ exists and

$$\phi_{S(m,n)}(-\theta) \leq f(\theta, n - m) \qquad \forall\, (m, n) \in \Lambda(\mathbb{N}_0).$$

Note that usually $\rho(\theta)$ is a negative quantity. This basically parallels the definition of the MGF-bound for arrivals.

Next we present the tail-bound for dynamic S-servers:

Definition 2.13. A dynamic S-server is tail-bounded by envelope α with error η, if for all arrival flows A and all $n \in \mathbb{N}_0$, $\varepsilon \in \mathbb{R}^+$ it holds that:

$$\mathbb{P}(D(0, n) < A \otimes (S - \alpha(\varepsilon))(0, n)) < \eta(n, \varepsilon) \tag{2.5}$$

In the above definition the expression $A \otimes (S - \alpha(\varepsilon))(0, n)$ has to be interpreted as: $\min_{0 \leq k \leq n}\{A(k) + S(k, n) - \alpha(n - k, \varepsilon)\}$. As mentioned, the above definition looks a bit different from the MGF-bound. Instead of bounding the function S directly, rather the output D of the dynamic S-server is bounded - and that for all possible inputs A. We leave the reason for that open and turn back to it, after presenting an example, which concentrates on the leftover-service:

Example 2.14. Consider the same situation as in example 2.11. We have already seen that this service element is an S_l-server with $S_l(m, n) = [c \cdot (n - m) - A_1(m, n)]^+$ with respect to A_2. If we assume the increments of the prioritized flow to be i.i.d. exponentially distributed with parameter λ, then we have for all $\theta \in (0, \lambda)$:

$$\begin{aligned}\phi_{S_l(m,n)}(-\theta) &= \left(\mathbb{E}(e^{-\theta[c\cdot(n-m)-A_1(m,n)]^+}\right) = \left(\mathbb{E}(e^{\min\{0,\theta A_1(m,n)-\theta c\cdot(n-m)\}})\right) \\ &\leq \left(\mathbb{E}(e^{\theta A_1(m,n)-\theta c\cdot(n-m)})\right) = e^{-\theta c(n-m)}\left(\frac{\lambda}{\lambda-\theta}\right)^{n-m}.\end{aligned}$$

Hence we have a dynamic S_l-Server, which is bounded by $f(\theta, n) = e^{-\theta cn}\left(\frac{\lambda}{\lambda-\theta}\right)^n$. We can also express this as a $(\sigma(\theta), \rho(\theta))$-bound with $\sigma(\theta) = 0$ and $\rho(\theta) = -c + 1/\theta \log\left(\frac{\lambda}{\lambda-\theta}\right)$.

Can we also find a tail-bound, given, that A_1 is tail-bounded by the envelope α_A? We need to find a fitting envelope α_S and some error function to establish (2.5) for all flows A_2. Inspired by the MGF-bound the envelope could look like $\alpha_S(n, \varepsilon) = \alpha_A(n, \varepsilon)$, which is in fact true. To see this choose $n \in \mathbb{N}$ and $\varepsilon > 0$ arbitrary. We know from example 2.11:

$$D_2(0, n) \geq \min_{0\leq k\leq n}\{A_2(k) + c(n-k) - A_1(k, n)\}$$

Now, let m be the index in $k = 0, \ldots, n$, which minimizes the right-hand sided expression and assume for a while that $A_1(m, n) \leq \alpha_A(n-m, \varepsilon)$ holds. Then the above can be continued with:

$$\begin{aligned}D_2(0, n) &\geq A_2(m) + c(n-m) - A_1(m, n) \\ &\geq A_2(m) + c(n-m) - \alpha_A(n-m, \varepsilon) \\ &\geq \min_{0\leq k\leq n}\{A_2(k) + c(n-k) - \alpha(n-k, \varepsilon)\} \\ &= A \otimes (S - \alpha_S(\varepsilon))(0, n)\end{aligned}$$

So we have proven the following statement:

$$A_1(m, n) \leq \alpha_A(n-m, \varepsilon) \quad \Rightarrow \quad D_2(0, n) \geq A \otimes (S - \alpha_S(\varepsilon))(0, n)$$

Or more precisely, based on definition **??**, we have:

$$\{\omega' \in \Omega \,|\, A_1(m, n) \leq \alpha_A(n-m, \varepsilon)\} \subset \{\omega' \in \Omega \,|\, D_2(0, n) \geq A \otimes (S - \alpha_S(\varepsilon))(0, n)\}$$

and hence:

$$\mathbb{P}(A_1(m, n) \leq \alpha_A(n-m, \varepsilon)) \leq \mathbb{P}(D_2(0, n) \geq A \otimes (S - \alpha_S(\varepsilon))(0, n)$$

From which we eventually obtain

$$\mathbb{P}(D_2(0, n) < A\otimes(S-\alpha_S(\varepsilon))(0, n)) \leq \mathbb{P}(A_1(m, n) > \alpha_A(n-m, \varepsilon)) \leq \eta(n-m, \varepsilon) \leq \eta(0, \varepsilon)$$

by the monotonicity of the error function.

We see in the above example that constructing the tail-bound for leftover service is a bit harder, than the corresponding MGF-bounds. Fortunately, the generalization of the above example follows pretty much along the same lines, such that most of the work is done here already.

Theorem 2.15. *Assume a dynamic S-server with two inputs A_1 and A_2, of which the former one is prioritized. Further assume A_1 to be tail-bounded with envelope α_A and error η. Then the dynamic S-server for A_2 is tail-bounded by envelope α_A and error $\eta_S(n,\varepsilon) := \eta_A(0,\varepsilon)$.*

Proof. Exercise 2.25. □

2.2.3 Concatenation of Stochastic Servers

Next we present the concatenation theorem. Imagine two servers in a tandem, i.e. the output of the first service element is fed into the second service element. The concatenation theorem states, that we can abstract the two service elements as a new service element, which describes the system as a whole.

Theorem 2.16. *Let there be two service elements, such that the output of the first serivce element is the input for the second service element. Assume the first element to be a dynamic S-server and the second element to be a dynamic T-server. Then holds:*

- *The whole system is a dynamic $S \otimes T$-server.*
- *If the dynamic servers are stochastically independent and bounded by $(\sigma_S(\theta), \rho_S(\theta))$ and $(\sigma_T(\theta), \rho_T(\theta))$, respectively, with $\rho_T(\theta), \rho_S(\theta) < 0$ and $\rho_S(\theta) \neq \rho_T(\theta)$ for some $\theta > 0$, the whole system is bounded by*

  $$(\sigma_S(\theta) + \sigma_T(\theta) + B, \max\{\rho_T(\theta), \rho_S(\theta)\})$$

  with:

  $$B := -\frac{1}{\theta}\log(1 - e^{-\theta|\rho_S(\theta)-\rho_T(\theta)|})$$

- *If the dynamic servers are tailbounded by envelopes α_1 and α_2 with errors η_1 and η_2, respectively. Further assume $\alpha_1(n,\varepsilon)$ and $\alpha_2(n,\varepsilon)$ are monotone increasing in n. Then the whole system is tailbounded with envelope $\alpha_1 + \alpha_2$ and error $\eta_1 + \eta_2$.*

Proof. We prove first that the whole system is a dynamic $S \otimes T$-server. Denote the input flow at the first service element by A, the output of the first service element by I and the output of the second service element by D. We have then:

$$\begin{aligned}
D(0,n) \geq I \otimes T(0,n) \geq (A \otimes S) \otimes T(0,n) &= \min_{0\leq k\leq n}\{(A\otimes S)(0,k) + T(k,n)\} \\
&= \min_{0\leq k\leq n}\{\min_{0\leq k'\leq k}\{A(0,k') + S(k',k)\} + T(k,n)\} \\
&= \min_{0\leq k'\leq k\leq n}\{A(0,k') + S(k',k) + T(k,n)\} \\
&= \min_{0\leq k'\leq n}\{A(0,k') + \min_{k'\leq k\leq n}\{S(k',k) + T(k,n)\}\} \\
&= \min_{0\leq k'\leq n}\{A(0,k') + S\otimes T(k',n)\} = A \otimes (S\otimes T)(0,n)
\end{aligned}$$

Hence the whole system is a dynamic $S \otimes T$-server.

Next we validate the MGF-bound: Assume first $|\rho_T(\theta)| < |\rho_S(\theta)|$. With the use of **??** holds for all $(m,n) \in \Lambda(\mathbb{N}_0)$:

$$\begin{aligned}
\phi_{S\otimes T(m,n)}(-\theta) &\leq \sum_{k=m}^{n} e^{\theta\rho_S(\theta)(k-m)+\theta\sigma_S(\theta)} e^{\theta\rho_T(\theta)(n-k)+\theta\sigma_T(\theta)} \\
&= e^{\theta\rho_T(\theta)(n-m)+\theta(\sigma_S(\theta)+\sigma_T(\theta))} \sum_{k=m}^{n} e^{\theta\rho_S(\theta)(k-m)} e^{\theta\rho_T(\theta)(m-k)} \\
&= e^{\theta\rho_T(\theta)(n-m)+\theta(\sigma_S(\theta)+\sigma_T(\theta))} \sum_{k=m}^{n} e^{\theta(k-m)(\rho_S(\theta)-\rho_T(\theta))} \\
&= e^{\theta\rho_T(\theta)(n-m)+\theta(\sigma_S(\theta)+\sigma_T(\theta))} \sum_{k'=0}^{n-m} (e^{\theta(\rho_S(\theta)-\rho_T(\theta))})^{k'}
\end{aligned}$$

and since we have $e^{\theta(\rho_S(\theta)-\rho_T(\theta))} < 1$ and can approximate the sum by its corresponding geometric series:

$$\phi_{S\otimes T(m,n)}(-\theta) \leq e^{\theta\rho_T(\theta)(n-m)+\theta(\sigma_S(\theta)+\sigma_T(\theta))} \frac{1}{1 - e^{\theta(\rho_S(\theta)-\rho_T(\theta))}}$$

The case of $|\rho_T(\theta)| > |\rho_S(\theta)|$ is very similar:

$$\begin{aligned}
\phi_{S\otimes T(m,n)}(-\theta) &\leq \sum_{k=m}^{n} e^{\theta\rho_S(\theta)(k-m)+\theta\sigma_S(\theta)} e^{\theta\rho_T(\theta)(n-k)+\theta\sigma_T(\theta)} \\
&= e^{\theta\rho_S(\theta)(n-m)+\theta(\sigma_S(\theta)+\sigma_T(\theta))} \sum_{k=m}^{n} e^{\theta\rho_S(\theta)(k-n)} e^{\theta\rho_T(\theta)(n-k)} \\
&= e^{\theta\rho_S(\theta)(n-m)+\theta(\sigma_S(\theta)+\sigma_T(\theta))} \sum_{k=m}^{n} e^{\theta(n-k)(\rho_T(\theta)-\rho_S(\theta))} \\
&= e^{\theta\rho_S(\theta)(n-m)+\theta(\sigma_S(\theta)+\sigma_T(\theta))} \sum_{k'=0}^{n-m} (e^{\theta(\rho_T(\theta)-\rho_S(\theta))})^{k'}
\end{aligned}$$

All left to do is to prove the tail-bound. Let $\varepsilon > 0$, $n \in \mathbb{N}_0$ be arbitrary and A some flow. Assume for a while that $I(n) \geq A \otimes S - \alpha_1(\varepsilon)(0,n)$ and $D(n) \geq I \otimes T - \alpha_2(\varepsilon)(0,n)$. Then follows as above:

$$\begin{aligned}
&D(0,n) \\
\geq& \min_{0 \leq k' \leq k \leq n} \{A(k') + S(k',k) + T(k,n) - \alpha_1(k-k',\varepsilon) - \alpha_2(n-k,\varepsilon)\} \\
\geq& \min_{0 \leq k' \leq n} \{A(k') + \min_{k' \leq k \leq n} \{S(k',k) + T(k,n)\} - \max_{k' \leq k \leq n} \{\alpha_1(k-k',\varepsilon) + \alpha_2(n-k,\varepsilon)\}\} \\
\geq& \min_{0 \leq k' \leq n} \{A(k') + S \otimes T(k',n) - \alpha_1(n-k',\varepsilon) - \alpha_2(n-k',\varepsilon)\}
\end{aligned}$$

Define $\alpha = \alpha_1 + \alpha_2$, we have then:

$$\begin{aligned}
&\mathbb{P}(D(0,n) < A \otimes (S \otimes T) - \alpha(\varepsilon)(0,n) \\
\leq& \mathbb{P}(\{I(0,n) < A \otimes S - \alpha_1(\varepsilon)(0,n)\} \cup \{D(0,n) < I \otimes T - \alpha_2(\varepsilon)(0,n)\} \\
\leq& \mathbb{P}(I(0,n) < A \otimes S - \alpha_1(\varepsilon)(0,n)) + \mathbb{P}(D(0,n) < I \otimes T - \alpha_2(\varepsilon)(0,n)) \\
\leq& \eta_1(n,\varepsilon) + \eta_2(n,\varepsilon)
\end{aligned}$$

□

We want to give some notes about the above theorem, since some obstacles may rise at this stage. First: note that for the MGF-bound we need both performance bounds to exist for the same θ. Since the service bounds are usually valid for all θ in some interval $(0,b)$, we can ensure, by intersecting the corresponding intervals, to find an interval on which both performance bounds hold. Second: The assumption of stochastic independence may not be given (see the exercises). The trick is to use again Hölder's inequality, however this may - as in exercise **??** - lead to poorer bounds. Third: We used $\rho_S(\theta) \neq \rho_T(\theta)$. If we have $\rho_S(\theta) = \rho_T(\theta)$ there is no way to use adavantage of the geometric series to achieve a $(\sigma(\theta), \rho(\theta))$-bound. Instead the sum degenerates to $n - m + 1$ and we need the more general $f(\theta, n)$-bounds.

The tail-bound seems to cause less trouble, but note that we have chosen the same ε for the two original tailbounds. One can easily generalise the above proof to the case, where one chooses a different ε for each tailbound. How to choose the ε is not clear, like the question if there lies a worthwhile optimization potential in doing so. Further using the inequality $\max_{k' \leq k \leq n}\{\alpha_1(k-k',\varepsilon) + \alpha_2(n-k,\varepsilon)\} \leq \alpha(n-k',\varepsilon)$ is not necessary and one might continue here instead with the "max-plus-convolution" $\max_{k' \leq k \leq n}\{\alpha_1(k-k',\varepsilon) + \alpha_2(n-k,\varepsilon)\} =: \alpha_1(\varepsilon)\bar{\otimes}\alpha_2(\varepsilon)(k,n)$, however in this case $\alpha(k,n,\varepsilon) := \alpha_1(\varepsilon)\bar{\otimes}\alpha_2(\varepsilon)(k,n)$ is not longer an envelope in the sense of **??**, since not univariate.

We shed now a bit of light on the question why the tail-bound is formulated for the output D, while the MGF-bound is concerned directly with S. In fact the tail-bound looks a bit "odd" and a more intuitive tail-bound would look like:

$$\mathbb{P}(S(m,n) < \alpha(n-m,\varepsilon)) < \eta(n-m,\varepsilon) \qquad \forall m \leq n,\, \varepsilon > 0 \tag{2.6}$$

What is the difference between this tail-bound and the one used in literature and definition 2.13?

Theorem 2.17. *Assume a dynamic S-server, with some envelope α and error fucntion η fulfilling (2.6). Then the server can be rewritten as a dynamic α-server (naturally extend α to its bivariate $\alpha(m,n,\varepsilon) := \alpha(n-m,\varepsilon)$), with envelope 0 and error $\eta_0(n,\varepsilon) = \eta(0,\varepsilon)$.*

Proof. Assume an arbitrary flow A and let α and η be given. Fix an arbitrary $\varepsilon > 0$ and $n \in \mathbb{N}_0$. We know from the definition of the dynamic S-server:

$$\begin{aligned} D(0,n) &\geq A \otimes S(0,n) = \min_{0\leq k\leq n}\{A(k) + S(k,n)\} \\ &= A(k^*) + S(k^*,n) \end{aligned}$$

Here k^* is the index minimizing the right handed side of the first line. Assume now, that $S(k^*,n) \geq \alpha(n-k^*,\varepsilon)$ holds. We can continue with:

$$\begin{aligned} D(0,n) &\geq A(k^*) + \alpha(n-k^*,\varepsilon) \\ &\geq \min_{0\leq k\leq n}\{A(k) + \alpha(n-k,\varepsilon)\} \\ &= A \otimes \alpha(\varepsilon)(0,n) \end{aligned}$$

Hence we have (with the same argument as in example 2.14) that:

$$\mathbb{P}(D(0,n) < A \otimes \alpha(\varepsilon)(0,n)) \leq \mathbb{P}(S(k^*,n) < \alpha(n-k^*,\varepsilon)) < \eta(n-k^*,\varepsilon) \leq \eta(0,\varepsilon)$$

□

This theorem practically means, that the condition in equation (2.6) is at least as strict, as the one given in definition 2.13, since we can always construct a tail-bound from (2.6). The question rises, if the converse is also true: Given a dynamic S-server for which we have

$$\mathbb{P}(D(0,n) < A \otimes \alpha(\varepsilon)(0,n)) < \eta(n,\varepsilon) \tag{2.7}$$

for all A, $\varepsilon > 0$ and $n \in \mathbb{N}_0$, can we infer equation (2.6)? In general this is not the case, as the following example shows:

Example 2.18. Let $S(m,n) = \sum_{k=m+1}^{n} s(k)$ with $s(k) = \frac{1}{k}$ and define $\alpha(n,\varepsilon) = \frac{1}{n}$ for all $n \geq 1$ and $\alpha(0,\varepsilon) := \infty$. Then holds for all A, $\varepsilon > 0$ and $n \in \mathbb{N}_0$:

$$\begin{aligned} D(0,n) \geq A \otimes S(0,n) &= \min_{0\leq k\leq n}\{A(k) + S(k,n)\} \geq \min_{0\leq k\leq n}\left\{A(k) + \tfrac{1}{n}\right\} = \frac{1}{n} \\ &= \min_{0\leq k\leq n}\{A(k) + \alpha(n-k,\varepsilon)\} = A \otimes \alpha(\varepsilon)(0,n) \end{aligned}$$

again with the natural expansion of α to its bivariate version. From this and the definitoin of dynamic S-servers one easily sees, that (2.7) is fulfilled, i.e.:

$$0 = \mathbb{P}(D(0,n) < A \otimes \alpha(\varepsilon)(0,n)) < \varepsilon =: \eta(n,\varepsilon)$$

for all $\varepsilon > 0$, $n \in \mathbb{N}_0$.

However we also have

$$S(n-m,n) = \frac{1}{n} + \dots \frac{1}{n-m+1} < \frac{1}{m} = \alpha(m,\varepsilon)$$

for $n > m^2 + m - 1$. Hence we can find m, n, ε such that:

$$\mathbb{P}(S(m,n) < \alpha(n-m,\varepsilon)) = 1 > \varepsilon = \eta(n-m,\varepsilon)$$

contradicting equation (2.6).

This gives us a strong argument why the tail-bound looks like it does in the literature: Using a less strict assumption on the service and still being able to calculate performance bounds (which we will show in the following chapter 3) is of course desirable. This gives us one part of the answer, why MGF-bound and tail-bound differ in their appearance. The other part is the question: Why not bound the MGF of $D(0,n) = A \otimes S(0,n)$ and what is its relation to definition 2.12?

Theorem 2.19. *Assume a dynamic S-server with*

$$\phi_{A \otimes S(0,n)}(-\theta) \leq e^{\theta\sigma(\theta)+\theta\rho(\theta)n}$$

for some $\theta > 0$, $\rho(\theta) < 0$ and all flows A and $n \in \mathbb{N}_0$. Then S is $(\sigma(\theta), \rho(\theta))$-bounded.

Proof. Let $m, n \in \mathbb{N}_0$ be arbitrary, such that $m \leq n$ and define a flow A with $a(k) = 0$ for all $k \leq m$. Then holds:

$$\begin{aligned} S(m,n) &\geq \min_{0 \leq k \leq m} \{S(k,n)\} = \min_{0 \leq k \leq m} \{A(k) + S(k,n)\} \\ &\geq \min_{0 \leq k \leq n} \{A(k) + S(k,n)\} = A \otimes S(0,n) \end{aligned}$$

and hence

$$-\theta S(m,n) \leq -\theta A \otimes S(0,n).$$

We end with:

$$\mathbb{E}(e^{-\theta S(m,n)}) \leq \mathbb{E}(e^{-\theta A \otimes S(0,n)}) \leq e^{\theta\sigma(\theta)+\theta\rho(\theta)n} \leq e^{\theta\sigma(\theta)+\theta\rho(\theta)(n-m)}$$

□

Again the converse is not true as the following example shows:

Example 2.20. Assume a dynamic S-server being $(\sigma(\theta), \rho(\theta))$-bounded for some $\theta > 0$. We need to find an A and n, such that the MGF of $A \otimes S(0,n)$ can not be bounded for any $\rho'(\theta) < 0$. Just consider the zero-flow $a(m) = 0$ for all $m \in \mathbb{N}_0$. For arbitrary n we have then by the definition of dynamic S-servers:

$$A \otimes S(0,n) \leq D(0,n) = 0$$

and hence:

$$\mathbb{E}(e^{-\theta A \otimes S(0,n)}) \geq \mathbb{E}(e^{-\theta \cdot 0}) = 1$$

Consider now any $\rho'(\theta) < 0$, arbitrary $\sigma(\theta)$ and some $n > \frac{\sigma(\theta)}{-\rho'(\theta)}$:

$$\mathbb{E}(e^{-\theta A \otimes S(0,n)}) \geq e^{\theta\sigma(\theta)+\theta\rho(\theta)n} > 1$$

This example shows, that the MGF-bound as presented, is again the formulation, shich assumes less and is hence preferable. In a larger context one can see it like this: the tail-bound asks for the probability, that an unwanted event happens (a flow exceeds an envelope, an output is too small). If we ignore the expectation and exponentation in the MGF-bounds we see the opposite: Here the inequalities $A(m,n) \leq \sigma(\theta) + \rho(\theta)$ and $S(m,n) > \sigma(\theta) - \rho(\theta)$ appear, describing a situation we would like to have (a flow does *not* exceed an envelope and the service *is* large enough). Although this is a very rough argument, it gives some intuition, why different expressions appear, when moving from MGF-bounds to tail-bounds.

Exercises

Exercise 2.21. Show that for $S(m,n) = S(n-m)$ we have:

$$A \otimes S(0,n) = A \otimes S(n)$$

where the right handed side is the usual univariate convolution.

Exercise 2.22. Assume $S_i(m,n) = S_i(n) - S_i(m)$ for $i \in \{1,2\}$. Give an example for S_i and $m \leq n$, such that

$$S(m,n) := S_1 \otimes S_2(m,n) \neq S(0,n) - S(0,m)$$

Exercise 2.23. In this exercise we investigate how *strict* a server can be, after subtracting a cross-flow from it. The answer is simple[3]: After subtracting no strict service can be offered, regardless of how strict our original service had been. To show this two statements must be proven. Throughout this we assume a node serving two arrivals, of which A_1 is prioritized over A_2 (see also example 2.11).

- Assume first the service element is a dynamic S-Server with respect to $A := A_1 + A_2$. Show that it is also a dynamic S_l-Server with respect to A_2 with $S_l(m,n) = S(m,n) - A_1(m,n)$.
- Assume now the service element fulfills Lindley's equation with respect to A (i.e. it is a "strictest" server). Construct input flows A_1, A_2 and a service S such that for some interval $[m,n]$ with $D_2(m-1) = A_2(m-1)$ holds $S_l(m,n) > D_2(m,n)$, i.e. S_l is not a strict service curve in sense of definition (2.1).[4] (Hint: You can choose a very simple S)

[3]and might surprise someone who is familiar with deterministic network calculus. In the univariate determinstic calculus it is proven, that you always need a strict service curve, when subtracting some crossflow, otherwise there exists none non-trivial service curve for the leftover service. This is an important difference between stochastic network calculus - as presented here - and deterministic network calculus!

[4]In fact any backlogged interval would work here. But in taking the additional assumption $D_2(m-1) = A_2(m-1)$ into account, we can also exclude that the leftover service could be *weak strict*. A dynamic S-Server is defined to be *weak strict* if for any backlogged period $[m,n]$ with $D(m-1) = A(m-1)$ holds: $D(m,n) \geq S(m,n)$.

Exercise 2.24. Consider two rate-latency server with rate and latency equal to 1

$$S_1(m,n) = S_2(m,n) = \beta(n-m) = [n-m-1]^+$$

Show that the concatenated server $S = S_1 \otimes S_2$ does not offer a strict service curve. In expression construct an input flow A, such that a backlogged period $[m,n]$ emerges from it and $A \otimes S(m,n) < S(m,n)$.

Exercise 2.25. Proof theorem 2.15.

Exercise 2.26. Assume a dynamic S_1-server and a dynamic S_2-server, which is stochastically *dependent* of the first one. Assume further that the S_1-server is $(\sigma_1(p\theta), \rho_1(p\theta))$-bounded and the S_2-server is $(\sigma_2(q\theta), \rho_2(q\theta))$-bounded for some $\theta > 0$, p and q with $\frac{1}{p} + \frac{1}{q} = 1$ and $\rho_1(p\theta) \neq \rho_2(q\theta)$. Show that the concatenation of S_1 and S_2 is $(\sigma_1(p\theta) + \sigma_2(q\theta) + B, \min\{|\rho_1(p\theta)|, |\rho_2(q\theta)|\})$-bounded, with $B = -\frac{1}{\theta}\log(1 - e^{-\theta|\rho_1(p\theta) - \rho_2(q\theta)|})$

Exercise 2.27. Assume the situation as in theorem 2.16. Show that for $\rho_S(\theta) = \rho_T(\theta)$ we the dynamic $S \otimes T$-server is $(\sigma(\theta), \rho(\theta))$-bounded with:

$$\rho(\theta) := \rho_S(\theta) + \frac{1}{\theta}$$

and

$$\sigma(\theta) := \sigma_S(\theta) + \sigma_T(\theta)$$

(Hint: Use the simple inequality $n + 1 \leq e^n$ for all $n \in \mathbb{N}_0$)

Exercise 2.28. Let a dynamic S-server be $(\sigma(\theta), \rho(\theta))$-bounded. Consider the following expression:

$$s_S^*(\theta) = \limsup_{m\to\infty} -\frac{1}{\theta m} \log \mathbb{E}(e^{-\theta S(n,n+m)})$$

Show that $s^*(\theta) \leq -\rho(\theta)$. Assume now the situation as in theorem 2.16. Show that for the concatenated server holds:

$$s_{S\otimes T}^*(\theta) \leq -\rho_T(\theta)$$

for all $\rho_T(\theta)$ with $|\rho_T(\theta)| < |\rho_S(\theta)|$. (Hint: Use the geomtric sum, instead of the geometric series: $\sum_{k'=0}^{m} p = \frac{1-p^{m+1}}{1-p}$ for $0 < p < 1$). Next show that also for the case $\rho_T(\theta) = \rho_S(\theta)$ holds:

$$s_{S\otimes T}^*(\theta) \leq -\rho_T(\theta)$$

Exercise 2.29. In this exercise we generalise example 2.14. Consider a dynamic S-server, which is bounded by $(\sigma_S(\theta), \rho_S(\theta))$. Further assume there is a flow A which is $(\sigma_A(\theta), \rho_A(\theta))$-bounded. Give $(\sigma(\theta), \rho(\theta))$-bounds for the leftover Service and the case that

- A is stochastically independent of S.

- A is stochastically dependent of S. What assumptions are further needed for this case?

Exercise 2.30. In this programming exercise we want to visualize how differently *service guarantees* (weak strict and normal service elements) can behave, compared to the *real* output of a node. For this we assume again the scenario of a node serving two flows with a fixed rate $c = 1$. The two flows behave very similar, assume them both to be exp(2.5)-distributed up to time $n = 49$. At time $n = 50$, however, the prioritized flow breaks this pattern and produces a large burst of value b. Compute the following values, to compare the different service guarantees for the output D_2 of the second flow and this "bad" behaving prioritized flow:

- (real output) Simulate the output for the second flow and calculate the cumulative output $D_2^{(1)}(n)$ of it.
- (weak strict service element) $D_2^{(2)}(n) = S_l(m,n) + A_2(m)$ and $S_l(m,n) := [c(n-m) - A_1(n) + A_1(m)]^+$, with m being the last time A_2 was not backlogged (i.e. $m = \max_{k \geq 0}\{k : A_2(k) = D_2^{(2)}(k)\}$, to compute m use $D_2^{(2)}(k) = D_2^{(1)}(k)$ for all k).
- (normal service element) $D_2^{(3)}(n) = A_2 \otimes S_l(0,n)$

Compare the three guarantees for different values of $b \in \{1, 5, 10, 50\}$. Analyse the above formulas to discover, why the service guarantees differ from each other and the real output.

Bibliography

[1] *R.R. Boorstyn, A. Burchard, J. Liebeherr, and C. Oottamakorn. Statistical service assurances for traffic scheduling algorithms.* Selected Areas in Communications, IEEE Journal on*, 18(12):2651–2664, 2000.*

[2] *Anne Bouillard, Laurent Jouhet, and Eric Thierry. Service curves in Network Calculus: dos and don'ts. Research Report RR-7094, INRIA, 2009.*

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

[4] *M. Fidler. An end-to-end probabilistic network calculus with moment generating functions. In* Quality of Service, 2006. IWQoS 2006. 14th IEEE International Workshop on*, pages 261 –270, 2006.*

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

[6] *J. B. Schmitt. Lecture notes: Performance modeling of distributed systems. http://disco.informatik.uni-kl.de/content/PDS_ WS1112, October 2011.*