Extending Network Calculus to Deal with Min-Plus Service Curves in Multiple Flow Scenarios

Anja Hamscher
RPTU Kaiserslautern-Landau,Germany
hamscher@cs.uni-kl.de

Vlad-Cristian Constantin
RPTU Kaiserslautern-Landau,Germany
constantin@cs.uni-kl.de

Jens B. Schmitt
RPTU Kaiserslautern-Landau,Germany
jschmitt@cs.uni-kl.de

***Abstract*—Network Calculus (NC) is a versatile analytical methodology to efficiently compute performance bounds in networked real-time systems. The arrival and service curve abstractions allow to model diverse and heterogeneous distributed real-time systems. The operations to compute residual service curves and to concatenate sequences of systems enable an efficient and accurate calculation of per-flow timing guarantees. Yet, in some scenarios involving multiple concurrent flows at a system, the central notion of so-called *min-plus* service curves is too weak to still be able to compute a meaningful residual service curve. In these cases, one usually resorts to so-called *strict* service curves that enable the computation of per-flow bounds. However, strict service curves are restrictive: (1) there are service elements for which only min-plus service curves can be provided but not strict ones and (2) strict service curves generally have no concatenation property, i.e., a sequence of two strict systems does not yield a strict service curve. In this paper, we extend NC to deal with systems only offering aggregate min-plus service curves to multiple flows. The key to this extension is the exploitation of *minimal arrival curves*, i.e., lower bounds on the arrival process. Technically speaking, we provide basic performance bounds (backlog and delay) for the case of *negative* service curves. We also discuss their accuracy and show them to be tight in many cases and approximately tight up to a constant in the others. In order to illustrate their usefulness we also present patterns of application of these new results for: (1) heterogeneous systems involving computation and communication resources and (2) finite buffers that are shared between multiple flows.**

I. Introduction

Network Calculus (NC) has proven to be a useful analytical methodology in the worst-case performance analysis of networked systems. As a stateless method it is computationally efficient and allows for extensive design space explorations. As such, it has seen numerous usage in real-world systems (e.g., TSN [1]–[3], AFDX [4]–[6], Network-on-Chip [7], [8]).

A very closely related framework, Real-Time Calculus (RTC) [9], focuses on the modeling of distributed real-time systems (e.g., [10], [11]). Indeed, it has been shown that RTC and NC are largely equivalent [12]. Hence, while our results are cast in the NC framework, they are readily applicable in RTC as well. The RTC makes frequent usage of so-called *minimal arrival curves* [13], i.e., a lower bound on the input to a system, while in NC this has so far been somewhat neglected. This is due to the fact that, in general, it may be necessary that additional packets need to be generated in order to enforce a minimal arrival curve [12]. However, especially in real-time systems, this problem may not arise, as each task usually has a minimal amount of traffic that is entering the system over a certain period of time.

NC provides a rich set of results: it can deal with all kinds of arrival processes and service elements. Its strength lies in providing a (min-plus) system theory that enables a tight or at least accurate end-to-end delay analysis. It was pioneered by Cruz [14], [15] and Chang [16], a comprehensive and up-to-date account of NC results is given in [12]. A central notion in NC is the service curve, abstracting scheduling disciplines at communication and computational resources. Several definitions exist, the two main ones being min-plus and strict service curves. Strict service captures the system behavior in a relatively tight manner, whereas the min-plus service is a weaker approximation, but comes with nice mathematical properties. Many different NC analysis methods, from Total Flow Analysis [15] over PMOO [17], [18] to Deep Tandem Matching Analysis [19], [20], have been developed over the years to accommodate for different system topologies and provide different trade-offs between accuracy of the bounds and computational cost.

Yet, there is a *blind spot* of NC and thus also in RTC: when a residual per-flow service curve is calculated from an aggregate min-plus service curve offered to multiple flows, [12, p. 161] states the following

> ”WARNING.– This only has a theoretical interest, and we want to warn the reader against using it in practice, as the result cannot be applied to compute performance bounds.”

However, being able to calculate performance bounds using min-plus service curves, rather than strict ones, would be very interesting. This would allow to model systems that inherently cannot provide a strict service curve, such as variable delay computational components for which we just know a worst-case execution time (e.g. from WCET analysis [21]) instead of a service rate (as also discussed in [12, Section 6.3.1.2]). Further, the concatenation of individual components also only provides min-plus service curves, even if the individual components provide strict service curves. The same holds for hierarchical scheduling scenarios, e.g. [22]. Unfortunately, in these cases, existing NC results are unable to calculate performance bounds for the flow due to the residual service curve offered to it becoming (partially) negative. The problem of calculating delay bounds for negative service curves is caused by certain arrival patterns for which the min-plus service curve property allows the system to not provide any service at all. We provide a more in-depth discussion on this issue in Sect. II-C.

Reiterating on the notion of a minimal arrive curve, we find that it can avoid the inherent issues with these arrival patterns by having enough arrivals to "drive the system forward". Thus, the key idea of this paper is to use minimal arrival curves to enable a performance analysis using NC in multiple flow scenarios when strict service curves cannot be assumed, or more generally, when service curves can have negative values.

Let us briefly discuss an application scenario (further elaborated in Sect. IV-A), specifically a networked real-time system that includes both computational and communication components, such as systems employing in-network processing [23]. Here, computational components only provide a deadline for the completion of a task, i.e., they need to be modelled by min-plus service curves, thus resulting in *negative residual* service curves when shared by multiple tasks. Hence, for an analysis based on the current state of the art, we cannot make use of the concatenation theorem, but, instead, need to calculate an upper delay bound for each separate node, subsequently adding up all individual bounds. With our new results for negative service curves, we can exploit the concatenation theorem to calculate a potentially much more accurate end-to-end delay bound.

Overall, we make the following contributions in this paper:

- We extend NC such that the calculation of performance bounds is also possible for partially negative min-plus service curves in Sect. III. While this is completely novel for the delay bound, the conventional backlog bound remains largely the same with a slight adaptation.
- We discuss the accuracy of both delay and backlog bounds. For the backlog bound, we show that it is always tight by providing a non-trivial sample path argument in Sect. III-B. For the delay bound, tightness depends on the given parameters for arrival and service curves. However, in all cases the delay bound is at least approximately tight up to a constant (the maximum amount of time between consecutive arrivals), i.e., we can provide lower and upper bounds on the worst-case delay that are spaced apart by that constant. This is shown in Sect. III-A.
- We present patterns of application demonstrating the practical usefulness of the new results in Sect. IV. In fact, in several cases the novel bounds outperform state-of-the-art techniques, or even enable an analysis at all.

II. Background and Problem Statement

In this section, we introduce the necessary background on network calculus and recapitulate existing results regarding multiple flow scenarios. In addition, we discuss the problem of aggregate min-plus service curves when it comes to calculating the residual service curves in a multiple flow scenario. Based on this, we present the key idea on how minimal arrival curves can circumvent the issue at hand.

A. Some Mathematical Background

Let $a, b \in \mathbb{R}$. We call $\wedge$ the *minimum operator* with $a \wedge b := \min\{a, b\}$, and $\vee$ the *maximum operator* with $a \vee b := \max\{a, b\}$. The function $[a]^+ := \max\{0, a\}$ yields the *positive part* of the argument a.

We make use of certain properties of sets. Let $P, Q \subseteq \mathbb{R}$ be two non-empty sets of real numbers. It holds that

$$-\sup P = \inf P^-, -\inf P = \sup P^-, \tag{1}$$

where $P^- := \{-x \mid x \in P\}$. Infimum and supremum exhibit the following properties:

$$\begin{aligned} \sup(P \cup Q) &= (\sup P) \vee (\sup Q), \\ \inf(P \cup Q) &= (\inf P) \wedge (\inf Q). \end{aligned} \tag{2}$$

Moreover, if $P \subseteq Q$, it holds that

$$\sup P \leq \sup Q, \ \inf P \geq \inf Q. \tag{3}$$

B. Network Calculus Background

We begin by defining several function classes [12, p. 22] that are used throughout the paper. Let $\mathbb{R}^+$ be the set of non-negative real numbers. $\mathcal{F} := \{f : \mathbb{R}^+ \to \mathbb{R} \cup \{+\infty\}\}$ is the set of (min, plus) functions. Based on $\mathcal{F}$, we let $\mathcal{F}^\uparrow$ be the set of non-decreasing functions $f \in \mathcal{F}$, and $\mathcal{F}_0^\uparrow$ be the set of functions in $\mathcal{F}^\uparrow$ with $f(0) = 0$. Similarly, we introduce the following sets: $\mathcal{F}_{<0}^{+}$ is the set of functions in $\mathcal{F}^\uparrow$ with $f(0) < 0$ and $\mathcal{F}_{\leq 0}^\uparrow$ is the set of functions in $\mathcal{F}^\uparrow$ with $f(0) \leq 0$.

Definition 1. A function $f \in \mathcal{F}$ is right-continuous if $\forall t \in \mathbb{R}$,

$$f(t^+) := \lim_{s \searrow t} f(s) := \lim_{s \to t, s > t} f(s)$$

always exists and is equal to $f(t)$.

Definition 2 (Pseudo-inverse)**.** Let $f \in \mathcal{F}^\uparrow$ be a non-negative and non-decreasing function. Then, the pseudo-inverse f^{-1} is defined $\forall x \geq 0$ as

$$f^{-1}(x) = \inf\{t \mid f(t) \geq x\}. \tag{4}$$

Definition 3 (Shift Function)**.** The *shift function* is defined by

$$\delta_T(t) := \begin{cases} +\infty, & \text{if } t > T, \\ 0, & \text{otherwise.} \end{cases} \tag{5}$$

Definition 4 (Operators [12])**.** Let $f, g \in \mathcal{F}$ be two functions. The *(min, plus) convolution* of f and g is defined as $f \otimes g(t) := \inf_{0 \leq s \leq t}\{f(t-s) + g(s)\}$, the *(min, plus) deconvolution* is defined as $f \oslash g(t) := \sup_{s \geq 0}\{f(t+s) - g(s)\}$. The *(max, plus) convolution* is defined as $f \overline{\otimes} g(t) := \sup_{0 \leq s \leq t}\{f(t-s) + g(s)\}$, and the *(max, plus) deconvolution* is defined as $f \overline{\oslash} g(t) := \inf_{s \geq 0}\{f(t+s) - g(s)\}$.

We introduce several properties of these operators.

Remark 5 (Isotonicity of $\otimes$ [24]). Let $f, g, f', g' \in \mathcal{F}$. If $f \leq g$ and $f' \leq g'$, then $f \otimes f' \leq g \otimes g'$.

Proposition 6 (Composition of $\oslash$ and $\otimes$)**.** *Let* $f, g, h \in \mathcal{F}$:

$$(f \otimes g) \oslash h \leq f \otimes (g \oslash h). \tag{6}$$

Next, we define various notions that are used to model a network and derive its performance bounds. Let $A, D \in \mathcal{F}_0^{\uparrow}$ be the *cumulative arrival* and *departure process* of a flow in the network, assuming causality $A \geq D$. Furthermore, we assume all systems to be lossless. We define the most important performance measures for such a system:

Definition 7 (Backlog at Time t)**.** The *backlog* of system $\mathcal{S}$ at time t is the vertical distance between arrival process A and departure process D at time t,

$$q(t) \coloneqq A(t) - D(t). \tag{7}$$

Definition 8 (Virtual Delay at Time t)**.** The *virtual delay* of data arriving at system $\mathcal{S}$ at time t is the time until this data would be served, assuming FIFO order of service,

$$d(t) \coloneqq \inf \left\{ \tau \geq 0 : A(t) \leq D(t+\tau) \right\}. \tag{8}$$

Arrival and service curves are essential elements of the performance analysis using NC. We define arrival curves first.

Definition 9 (Maximal and Minimal Arrival Curve)**.** Let $\overline{\alpha}, \underline{\alpha} \in \mathcal{F}_0^{\uparrow}$. We say that $\overline{\alpha}$ is a *maximal arrival curve* for arrival process A, and $\underline{\alpha}$ is a *minimal arrival curve* for A, if it holds for all $0 \leq s \leq t$ that

$$\underline{\alpha}(t-s) \leq A(t) - A(s) \leq \overline{\alpha}(t-s). \tag{9}$$

A frequent example is the *token-bucket* arrival curve $\gamma_{r,b}(t) = b + rt$ for $t > 0$, $\gamma_{r,b}(0) = 0$. Note that $\gamma_{r_1,b_1} + \gamma_{r_2,b_2} = \gamma_{r_1+r_2,b_1+b_2}$. Next, we define service curves.

Definition 10 (Service Curve (SC))**.** Let a flow with arrival process A and departure process D traverse a system $\mathcal{S}$. The system offers a *min-plus service curve* β to the flow if $\beta \in \mathcal{F}$ and it holds for all $t \geq 0$ that

$$D(t) \geq A \otimes \beta\,(t) = \inf_{0 \leq s \leq t} \left\{ A(t-s) + \beta(s) \right\}. \tag{10}$$

Often, $\beta \in \mathcal{F}_0^{\uparrow}$ is assumed, yet we let $\beta \in \mathcal{F}$ as in [12].

Definition 11 (Strict Service Curve (SSC))**.** A system offers a *strict service curve* $\beta \in \mathcal{F}$ to a flow if, during any backlogged period $(s, t]$ (i.e. $\forall t' \in (s,t], q(t') > 0$), it holds that

$$D(t) - D(s) \geq \beta(t-s). \tag{11}$$

A frequently employed function for minimal arrival and service curves is the *rate-latency* curve $\beta_{R,T}(t) \coloneqq R \cdot [t-T]^{+}$.

We define two characteristic distances between functions.

Definition 12. Let $f, g \in \mathcal{F}$. The *vertical deviation* between f and g is defined as

$$v(f,g) \coloneqq \sup_{t \geq 0} \left\{ f(t) - g(t) \right\}, \tag{12}$$

and the *horizontal deviation* between f and g is defined as

$$h(f,g) \coloneqq \sup_{t \geq 0} \left\{ \inf \left\{ \tau \geq 0 \mid f(t) \leq g(t+\tau) \right\} \right\} \tag{13}$$

$$= \inf \left\{ \tau \geq 0 \mid \sup_{t \geq 0} \left\{ f(t) - g(t+\tau) \right\} \leq 0 \right\}. \tag{14}$$

There is a useful property of deviations [12, p. 115]:

Lemma 13 (Monotony of Deviations)**.** *For all $f, f', g, g' \in \mathcal{F}^{\uparrow}$, if $f \geq f'$ and $g \leq g'$, then*

$$v(f,g) \geq v(f',g') \quad \textit{and} \quad h(f,g) \geq h(f',g'). \tag{15}$$

Using these concepts, one can derive performance bounds for the measures defined previously [12, p. 115], [24, p. 118].

Theorem 14 (Performance Bounds)**.** *Assume an arrival process A, constrained by maximal arrival curve $\overline{\alpha} \in \mathcal{F}_0^{\uparrow}$, traverses a system $\mathcal{S}$. Let the system $\mathcal{S}$ offer a service curve $\beta \in \mathcal{F}_0^{\uparrow}$. The virtual delay $d(t)$ satisfies for all t*

$$d(t) \leq h(\overline{\alpha}, \beta). \tag{16}$$

The backlog $q(t)$ satisfies for all t

$$q(t) \leq v(\overline{\alpha}, \beta). \tag{17}$$

Note that Thm. 14 requires $\beta \in \mathcal{F}_0^{\uparrow}$.

We can also calculate a bound on the departure process D of a system offering a min-plus service curve $\beta \in \mathcal{F}$:

$$D \leq \overline{\alpha} \oslash \beta. \tag{18}$$

A central result of NC is the concatenation theorem.

Theorem 15 (Concatenation Theorem)**.** *Let a flow with arrival process A traverse systems $\mathcal{S}_1$ and $\mathcal{S}_2$, offering service curves $\beta_1, \beta_2 \in \mathcal{F}$, in sequence. Then, the concatenation of the two systems, $S_{1,2} = \langle S_1, S_2 \rangle$, offers an end-to-end service curve $\beta_{1,2} = \beta_1 \otimes \beta_2$ to the arrival process.*

Definition 16 (Sub-additive and Super-additive Functions)**.** Let $f \in \mathcal{F}$. Then f is *sub-additive* if for all $s, t \geq 0$

$$f(t+s) \leq f(t) + f(s). \tag{19}$$

On the other hand, f is *super-additive* if for all $s, t \geq 0$

$$f(t+s) \geq f(t) + f(s). \tag{20}$$

Definition 17 (Sub-additive Closure [24])**.** Let $f \in \mathcal{F}$. The sub-additive closure of f is defined by

$$f^{*} \coloneqq \inf_{n \geq 0} \left\{ f^{(n)} \right\}, \tag{21}$$

where $f^{(n)}$ is the n-fold self-convolution of f, i.e., $f^{(0)} = \delta_0$, $f^{(1)} = f$ and $f^{(n)} = \bigotimes_{i=0}^{n} f^{(i)}$ for $n \geq 2$.

With respect to tightness, we remark that maximal arrival curves that are not sub-additive and minimal arrival curves that are not super-additive can be improved by replacing them by their sub-additive and super-additive closures, respectively (see [12], Propositions 5.2 and 5.3).

Moreover. we also note that both arrival curves may be further improved by combining their respective information [25] (see also [12, Theorem 5.1]).

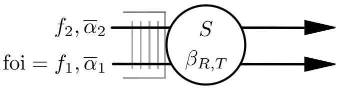


Fig. 1: Two flows crossing a single system.

C. The Issue with Min-Plus SCs in Multiple Flow Scenarios

In the following, we give a motivational example on the limits of current state-of-the-art modeling using network calculus. Consider the basic system in Fig. 1. Two tasks f_1 and f_2 with packet arrival processes A_1 and A_2 (an example shape of A_2 is illustrated in Fig. 2a), constrained by maximal token-bucket arrival curves $\overline{\alpha}_1$ and $\overline{\alpha}_2$, respectively, are using the component $\mathcal{S}$ offering a rate-latency service curve $\beta_{R,T}$.

For now, we assume that $\mathcal{S}$ employs a static priority scheduling between the tasks and offers a *strict* service curve, where task f_1 has a lower priority than task f_2. We want to calculate a delay bound for f_1, our flow of interest (foi). To that end, we calculate the residual service curve $\beta_{\text{res}}^{\text{SSC}}$, which represents the residual capacity $\mathcal{S}$ can offer f_1 after it has served f_2, as [12, Theorem 7.1]

$$\beta_{\text{res}}^{\text{SSC}} := [\beta_{R,T} - \overline{\alpha}_2]^+. \tag{22}$$

Its shape is illustrated in Fig. 2a. We compute the horizontal deviation $h\big(\overline{\alpha}_1, \beta_{\text{res}}^{\text{SSC}}\big)$ as delay bound for f_1.

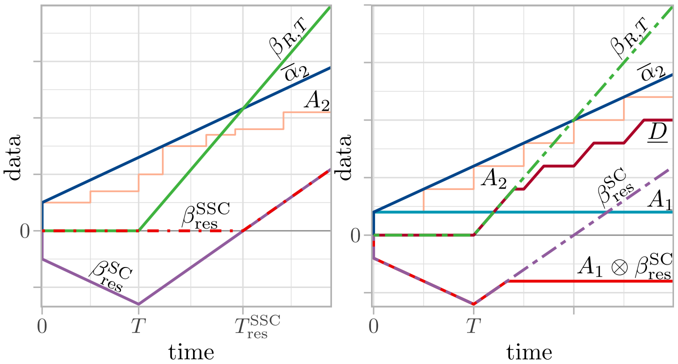
As mentioned above, assuming that $\beta_{R,T}$ is strict is restrictive, though, since not all resources can be modeled by an SSC; for instance, if a component only offers a deadline to a task instead of a service rate, or if the service curve is a residual service curve itself (potentially resulting from hierarchical scheduling). Furthermore, we cannot make use of the concatenation property (Thm. 15) anymore, as the convolution of SSCs is not an SSC [12, Section 9.3]. That is unfortunate, since the worst-case end-to-end delay bound using this property is generally smaller than the sum of delay bounds at each component along the path a task takes. We further note that even if a component can provide a strict service curve, it may offer a min-plus service curve that is *larger* than the strict one, e.g., for residual service curves under arbitrary multiplexing [12, Table 7.1].

Coming back to the example, we encounter problems if we are only given a min-plus service curve for $\mathcal{S}$. We can still calculate the residual service curve $\beta_{\text{res}}^{\text{SC}}$ [12, Theorem 7.3] as

$$\beta_{\text{res}}^{\text{SC}} := \beta_{R,T} - \overline{\alpha}_2. \tag{23}$$

While the service curve property (Def. 10) for $\beta_{\text{res}}^{\text{SC}}$ holds, there is an interval over which it is negative and even decreasing (see Fig. 2a), because we cannot apply the positive part as in Eq. (22). Consequently, as mentioned in the WARNING in [12, Section 7.2.3], the assumptions of Thm. 14 (i.e., $\beta \in \mathcal{F}_0^{\uparrow}$) to compute performance bounds are not met anymore. As such, with Thm. 14, a delay bound using $\beta_{\text{res}}^{\text{SC}}$ cannot be computed. In fact, without further assumptions, there is no finite delay bound in this system.

The underlying issue is illustrated in Fig. 2b and represents a possible actual behavior of the system. Consider that f_1 sends a single packet at some reference instant 0, while f_2 sends a packet at the same instant, and then continuously sends packets in a fixed interval (shown as A_2). We again assume that f_2 has a higher priority than f_1. If $\mathcal{S}$ offers a min-plus service curve, it is allowed to only have an output $\underline{D} := (A_1 + A_2) \otimes \beta_{R,T}$. Simply speaking, the component is "lazy" and does not even have to offer the service to fully serve f_2, and as a result, can delay the packet of f_1 indefinitely. This is further illustrated by calculating the residual service curve $\beta_{\text{res}}^{\text{SC}}$ using Eq. (23), and subsequently calculating the departure guarantee for f_1 as $A_1 \otimes \beta_{\text{res}}^{\text{SC}}$. We see that this never becomes positive, i.e., there is not enough residual service to output the packet of f_1. In contrast, if $\mathcal{S}$ conformed to a strict service curve $\beta_{R,T}$, the server must have an output with rate R as long as there are packets in the buffer, and will service the packet of f_1 in a finite amount of time.



a) Residual service curves from an SSC and SC.

b) The problem with a negative residual SC.

Fig. 2: Illustration of residual service curves.

Understanding the root cause of the problem contains the key to its mitigation: making sure that there are enough arrivals to drive the departure guarantee of the system $A_1 \otimes \beta_{\text{res}}^{\text{SC}}$ into the positive. We need to ensure that it is not an acceptable behavior of task f_1 to input only a finite amount of packets into the system over an infinite time horizon. We achieve this by assuming that f_1 is also subject to a *minimal* arrival curve, in addition to being subject to a maximal arrival curve. In the context of real-time systems, the minimum amount of packets sent over an interval can be used as a minimal arrival curve. Using this insight, we generalize the performance bounds for negative service curves in Sect. III.

The problem of "lazy" service curves has been observed and tackled before by so-called Adaptive Service Curves (ASC) [24], [26]. The ASC addresses this behavior by providing an intermediate service guarantee between a min-plus and strict service curve. Yet, this cannot solve the issue at hand, since we are confronted with the same situation of obtaining a negative service curve from the calculation of the residual service curve.

III. Extension of NC Performance Bounds for Negative Service Curves

In this section, we extend the performance bounds presented in Thm. 14 to the case where the service curve $\beta \notin \mathcal{F}_0^{\uparrow}$. As discussed above, we need to assume the existence of a minimal arrival curve in order to provide the generalization of the delay bound in Thm. 14. However, as we will see, for the backlog bound, there is no need to assume that the arrival process conforms to a minimal arrival curve.

For the service curve under consideration we make the very general starting assumption that $\beta \in \mathcal{F}_{\leq 0}$. Note that $\beta(0) \leq 0$ means no reduction in generality [12, p. 107]. Next, as a preprocessing step, we "safely" replace the original service curve $\beta \in \mathcal{F}$ by $\xi = \beta_{\downarrow} := \beta \overline{\oslash} 0$ [12, p. 107]. $\beta_{\downarrow}$ is the largest non-decreasing function with $\beta_{\downarrow} \leq \beta$, which is why we call it the *lower* non-decreasing closure. Then, by isotonicity of the (min,plus) convolution (see Remark 5), ξ is also a service curve. Note that this is different from the (upper) non-decreasing closure as defined in [12, p. 45].

It is clear that $\xi \in \mathcal{F}^{\uparrow}_{\leq 0}$. In particular, $\xi \in \mathcal{F}^{\uparrow}_{0}$ if and only if $\beta \geq 0$, and $\xi \in \mathcal{F}^{\uparrow}_{<0}$ if and only if $\exists s \geq 0$ with $\beta(s) < 0$.

While the lower non-decreasing closure is safe to use as $\xi \leq \beta$, it possibly runs the risk to be conservative. However, in many practical cases we do not measure any quantity of interest (horizontal or vertical deviation) on a decreasing part of the original service curve β. Hence, the calculated performance bounds would be identical for ξ and β (for an illustration in case of the backlog, see also Fig. 6 below).

A. Generalizing the Delay Bound

We start with generalizing the delay bound and discuss its tightness thereafter. Let us first state a useful technical lemma.

Lemma 18. *Let $f, g \in \mathcal{F}$ be non-increasing. Then,*

$$\begin{aligned} &\inf\{\tau \geq 0 \mid f(\tau) \leq 0\} \vee \inf\{\tau \geq 0 \mid g(\tau) \leq 0\} \\ &= \inf\{\tau \geq 0 \mid (f \vee g)(\tau) \leq 0\}. \end{aligned} \tag{24}$$

Proof. Let $\tau_f := \arg\inf\{\tau \geq 0 \mid f(\tau) \leq 0\}$. We define τ_g similarly. Note that for $f > 0$, $\tau_f = \infty$. For the moment, let f be right-continuous at τ_f, then we have $f(\tau) \leq 0, \forall \tau \geq \tau_f$, since f is non-increasing. This clearly also holds for g. Then,

$$\begin{aligned} &\inf\{\tau \geq 0 \mid f(\tau) \leq 0\} \vee \inf\{\tau \geq 0 \mid g(\tau) \leq 0\} \\ &= \inf[\tau_f, +\infty) \vee \inf[\tau_g, +\infty) \\ &= \inf[\tau_f \vee \tau_g, +\infty) \\ &= \inf[\tau_f, +\infty) \cap [\tau_g, +\infty) \\ &= \inf\{\tau \geq 0 \mid f(\tau) \leq 0\} \cap \{\tau \geq 0 \mid g(\tau) \leq 0\} \\ &= \inf\{\tau \geq 0 \mid f(\tau) \leq 0 \textit{ AND } g(\tau) \leq 0\} \\ &= \inf\{\tau \geq 0 \mid (f \vee g)(\tau) \leq 0\}. \end{aligned}$$

Now, in case f is not right-continuous at τ_f, we have that $\{\tau \geq 0 \mid f(\tau) \leq 0\} = (\tau_f, +\infty)$, and the proof proceeds along the same lines. □

Theorem 19 (Generalized Delay Bound)**.** *Let an arrival process A traverse a system $\mathcal{S}$. Further, let the arrivals be constrained by maximal arrival curve $\overline{\alpha} \in \mathcal{F}^{\uparrow}_{0}$ and minimal arrival curve $\underline{\alpha} \in \mathcal{F}^{\uparrow}_{0}$, and let the system offer a service curve $\xi \in \mathcal{F}^{\uparrow}_{\leq 0}$. The virtual delay $d(t)$ satisfies for all $t \geq 0$*

$$d(t) \leq z(\underline{\alpha}, \xi) \vee h(\overline{\alpha}, \xi), \tag{25}$$

with $z(\underline{\alpha}, \xi) := \inf\{\tau \geq 0 \mid \underline{\alpha} \otimes \xi(\tau) \geq 0\}$.

Proof. First, consider the case when $\xi \in \mathcal{F}^{\uparrow}_{0}$. This is the classical case from Thm. 14, for which we know that $d(t) \leq h(\overline{\alpha}, \xi)$. It suffices to show that $d(t) \leq z(\underline{\alpha}, \xi) \vee h(\overline{\alpha}, \xi)$. We have that

$$z(\underline{\alpha}, \xi) = \inf\{\tau \geq 0 \mid \underline{\alpha} \otimes \xi(\tau) \geq 0\} = 0, \tag{26}$$

since $\underline{\alpha}(0) = \xi(0) = 0$. Therefore,

$$d(t) \leq z(\underline{\alpha}, \xi) \vee h(\overline{\alpha}, \xi) = h(\overline{\alpha}, \xi).$$

Next, consider the case when $\xi \in \mathcal{F}^{\uparrow}_{<0}$. We derive

$$\begin{aligned} d(t) &\overset{(8)}{=} \inf\{\tau \geq 0 \mid D(t+\tau) \geq A(t)\} \\ &\overset{(10),(3)}{\leq} \inf\{\tau \geq 0 \mid A \otimes \xi(t+\tau) \geq A(t)\} \\ &= \inf\left\{\tau \geq 0 \mid \inf_{0 \leq s \leq t+\tau}\{A(t+\tau-s) + \xi(s)\} \geq A(t)\right\} \\ &= \inf\left\{\tau \geq 0 \mid A(t) - \inf_{0 \leq s \leq t+\tau}\{A(t-(s-\tau)) + \xi(s)\} \right. \\ &\qquad \left. \leq 0\right\} \\ &\overset{(1)}{=} \inf\left\{\tau \geq 0 \mid \sup_{0 \leq s \leq t+\tau}\{A(t) - A(t-(s-\tau)) - \xi(s)\} \right. \\ &\qquad \left. \leq 0\right\} \\ &\overset{(2)}{=} \inf\left\{\tau \geq 0 \mid \sup_{0 \leq s \leq \tau}\{A(t) - A(t-(s-\tau)) - \xi(s)\} \right. \\ &\qquad \left. \vee \sup_{\tau < s \leq t+\tau}\{A(t) - A(t-(s-\tau)) - \xi(s)\} \leq 0\right\} \\ &\overset{(9),(3)}{\leq} \inf\left\{\tau \geq 0 \mid \sup_{0 \leq s \leq \tau}\{-\underline{\alpha}(\tau-s) - \xi(s)\} \right. \\ &\qquad \left. \vee \sup_{\tau < s \leq t+\tau}\{\overline{\alpha}(s-\tau) - \xi(s)\} \leq 0\right\} \\ &\overset{(1)}{=} \inf\left\{\tau \geq 0 \mid -\underline{\alpha} \otimes \xi(\tau) \vee \sup_{0 < s' \leq t}\{\overline{\alpha}(s') - \xi(s'+\tau)\} \right. \\ &\qquad \left. \leq 0\right\} \quad (27) \\ &\overset{(24)}{=} \inf\{\tau \geq 0 \mid -\underline{\alpha} \otimes \xi(\tau) \leq 0\} \\ &\qquad \vee \inf\left\{\tau \geq 0 \mid \sup_{0 < s' \leq t}\{\overline{\alpha}(s') - \xi(s'+\tau)\} \leq 0\right\} \\ &\overset{(3)}{\leq} \inf\{\tau \geq 0 \mid \underline{\alpha} \otimes \xi(\tau) \geq 0\} \\ &\qquad \vee \sup_{s' \geq 0}\{\inf\{\tau \geq 0 \mid \overline{\alpha}(s') - \xi(s'+\tau) \leq 0\}\} \quad (28) \\ &= z(\underline{\alpha}, \xi) \vee h(\overline{\alpha}, \xi). \end{aligned}$$

In line 8 (Eq. (27)) we make the substitution $s' := s - \tau$. In line 10 (Eq. (28)) we rewrite the supremum as in Eq. (14) and take the supremum over a larger set. It is left to check that the conditions of Lem. 18 in line 8 (Eq. (27)) apply:

- due to the closedness of the min-plus convolution for the set of non-decreasing functions [12, p. 22], and both $\underline{\alpha}$

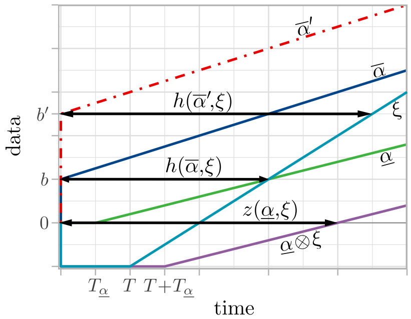


Fig. 3: Different cases of the generalized delay bound theorem.

and ξ being non-decreasing, we see that $-\underline{\alpha}\otimes\xi(\tau)$ is non-increasing in τ;

- since ξ is non-decreasing, $\sup_{0<s\leq t}\{\overline{\alpha}(s)-\xi(s+\tau)\}$ is clearly non-increasing in τ.

□

For an illustrative example, showing the different cases governing the generalized delay bound, see Fig. 3. Here, we assume a maximal token-bucket arrival curve $\overline{\alpha}=\gamma_{r,b}$, a minimal rate-latency arrival curve $\underline{\alpha}=\beta_{R,T_{\underline{\alpha}}}$ and a simple negative service ξ. We show two cases of maximal arrival curves with different burst sizes such that both cases ($h(\overline{\alpha},\xi)$ and $z(\underline{\alpha},\xi)$) of the generalized delay bound are provoked. One can observe that in case of $\overline{\alpha}$ having a smaller burst, the delay bound is given by $z(\underline{\alpha},\xi)$, whereas when we have a burstier maximal arrival curve $\overline{\alpha}'$ then $h(\overline{\alpha}',\xi)$ dominates.

So, we have extended the delay bound analysis to functions which are not in $\mathcal{F}_0^{\uparrow}$. But is the delay bound proved above tight? We show below that at least in an approximate sense it is tight. Before that we provide a helpful lemma.

Lemma 20. *Let $f\in\mathcal{F}_0^{\uparrow}$, $f(t)>0,\forall t>0$ and $g\in\mathcal{F}_{\leq 0}^{\uparrow}$.* Assume that f and g are right-continuous. *Then*

$$h(f,[g]^+)=h(f,g). \tag{29}$$

Proof. We have $\forall\epsilon>0$

$$\begin{aligned}
h(f,[g]^+)&\overset{(13)}{=}\sup_{t\geq 0}\left\{\inf\left\{\tau\geq 0\mid f(t)\leq[g(t+\tau)]^+\right\}\right\}\\
&\overset{(2)}{=}\sup_{t>0}\left\{\inf\left\{\tau\geq 0\mid f(t)\leq[g(t+\tau)]^+\right\}\right\}\\
&\quad\vee\inf\{\tau\geq 0\mid f(0)\leq[g(\tau)]^+\}\\
&=\sup_{t>0}\left\{\inf\left\{\tau\geq 0\mid f(t)\leq[g(t+\tau)]^+\right\}\right\} && (30)\\
&=\sup_{t>0}\left\{\inf\left\{\tau\geq 0\mid f(t)\leq 0\vee g(t+\tau)\right\}\right\}\\
&=\sup_{t>0}\left\{\inf\left\{\tau\geq 0\mid f(t)\leq g(t+\tau)\right\}\right\} && (31)\\
&\overset{(2)}{=}\sup_{t>0}\left\{\inf\left\{\tau\geq 0\mid f(t)\leq g(t+\tau)\right\}\right\}\\
&\quad\vee\inf\{\tau\geq 0\mid f(\epsilon)\leq g(\epsilon+\tau)\}. && (32)
\end{aligned}$$

In line 3 (Eq. (30)) we use that $f(0)=0$, and hence the infimum is 0. In the fifth step (Eq. (31)), we use the fact that $f(t)\leq 0,\forall t>0$ is false, via assumption. By idempotency, we go from Eq. (31) to Eq. (32), which holds for all $\epsilon>0$. Thus, we can conclude:

$$\begin{aligned}
h(f,[g]^+)&=\sup_{t>0}\left\{\inf\left\{\tau\geq 0\mid f(t)\leq g(t+\tau)\right\}\right\}\\
&\quad\vee\inf\{\tau\geq 0\mid\lim_{t\to 0,t>0}f(t)\leq\lim_{t\to 0,t>0}g(t+\tau)\}\\
&\overset{(2)}{=}\sup_{t\geq 0}\left\{\inf\left\{\tau\geq 0\mid f(t)\leq g(t+\tau)\right\}\right\} && (33)\\
&\overset{(13)}{=}h(f,g).
\end{aligned}$$

In step (33), we use the right-continuity of f and g. □

We move on to prove the (approximate) tightness of the generalized delay bound. By approximate we mean that we can at least provide lower and upper bounds on the worst-case delay of the system that are only spaced apart by a constant. Here, that constant is the latency of the minimal arrival curve, i.e., the maximum amount of time between two packet arrivals.

Theorem 21 (Approximate Tightness of the Generalized Delay Bound)**.** *Let an arrival process A traverse a system $\mathcal{S}$. Further, let the arrivals be constrained by a sub-additive maximal arrival curve $\overline{\alpha}\in\mathcal{F}_0^{\uparrow}$ and a super-additive minimal arrival curve $\underline{\alpha}\in\mathcal{F}_0^{\uparrow}$, and assume these cannot be further improved by combining their respective information (see [12, Theorem 5.1]). Let the system offer a service curve $\xi\in\mathcal{F}_{\leq 0}^{\uparrow}$. We also assume that $\underline{\alpha}$,$\overline{\alpha}$ and ξ are right-continuous.*
If $h(\overline{\alpha},\xi)\geq z(\underline{\alpha},\xi)$, we let $A^{\mathrm{WC}}:=\overline{\alpha}$. and $D^{\mathrm{WC}}:=[\overline{\alpha}\otimes\xi]^+$, then the worst-case delay (WCD) *is*

$$\mathrm{WCD}=h(A^{\mathrm{WC}},D^{\mathrm{WC}})=h(\overline{\alpha},\xi), \tag{34}$$

and thus the generalized delay bound is perfectly tight.
If $h(\overline{\alpha},\xi)<z(\underline{\alpha},\xi)$, with $T_{\underline{\alpha}}:=\sup\{t\geq 0\mid\underline{\alpha}(t)=0\}$, we have upper and lower bounds on the worst-case delay (WCD)*:*

$$z(\underline{\alpha},\xi)-T_{\underline{\alpha}}\leq\mathrm{WCD}\leq z(\underline{\alpha},\xi), \tag{35}$$

and thus the generalized delay bound is approximately tight.

Proof. We start with the first case, where $h(\overline{\alpha},\xi)\geq z(\underline{\alpha},\xi)$, with $A^{\mathrm{WC}}(t):=\overline{\alpha}(t)$ and $D^{\mathrm{WC}}:=[\overline{\alpha}\otimes\xi]^+$, then

$$\begin{aligned}
h(A^{\mathrm{WC}},D^{\mathrm{WC}})&=h(\overline{\alpha},[\overline{\alpha}\otimes\xi]^+)\\
&\overset{(29)}{=}h(\overline{\alpha},\overline{\alpha}\otimes\xi) && (36)\\
&\overset{(15)}{\geq}h(\overline{\alpha},\xi).
\end{aligned}$$

In the second step (Eq. (36)), we use the fact that $\overline{\alpha}\in\mathcal{F}_0^{\uparrow}$ and then apply Lem. 20, based on the fact that, in our case, the convolution remains right-continuous [27]. In the last step, we use the monotony of deviations (Lem. 13). By using Thm. 19, we have that $h(A^{\mathrm{WC}},D^{\mathrm{WC}})=h(\overline{\alpha},\xi)=h(\overline{\alpha},\xi)\ \vee z(\underline{\alpha},\xi)$.

Let us now consider the second case, $h(\overline{\alpha},\xi)<z(\underline{\alpha},\xi)$.

We set $A:=\underline{\alpha}\oslash\delta_{T_{\underline{\alpha}}}$ and $D:=\left[\left(\underline{\alpha}\oslash\delta_{T_{\underline{\alpha}}}\right)\otimes\xi\right]^+$. We note that the deconvolution by $\delta_{T_{\underline{\alpha}}}$ is a left-shift by $T_{\underline{\alpha}}$ (Prop. 3.2 in [12, p. 40]), i.e. $\underline{\alpha}\oslash\delta_{T_{\underline{\alpha}}}(t)=\underline{\alpha}(t+T_{\underline{\alpha}})$. Then,

$$h(A,D)=h\left(\underline{\alpha}\oslash\delta_{T_{\underline{\alpha}}},\left[\left(\underline{\alpha}\oslash\delta_{T_{\underline{\alpha}}}\right)\otimes\xi\right]^+\right)$$

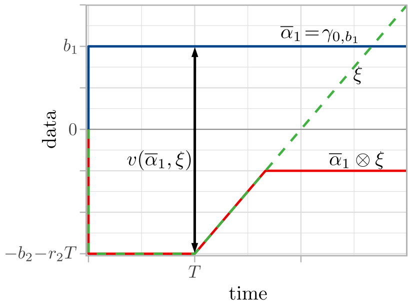


Fig. 4: Vertical deviation in case of a negative service curve.

$$
\begin{aligned}
&\overset{(29)}{=} h\left(\underline{\alpha}\oslash\delta_{T_{\underline{\alpha}}}, \left(\underline{\alpha}\oslash\delta_{T_{\underline{\alpha}}}\right)\otimes\xi\right) && (37)\\
&\overset{(13)}{=} \sup_{t\geq 0}\left\{\inf\left\{\tau\geq 0 \mid \left(\underline{\alpha}\oslash\delta_{T_{\underline{\alpha}}}\right)(t)\leq \right.\right.\\
&\qquad\qquad \left.\left.\left(\left(\underline{\alpha}\oslash\delta_{T_{\underline{\alpha}}}\right)\otimes\xi\right)(t+\tau)\right\}\right\}\\
&\overset{(3)}{\geq} \inf\left\{\tau\geq 0 \mid 0\leq \left(\left(\underline{\alpha}\oslash\delta_{T_{\underline{\alpha}}}\right)\otimes\xi\right)(\tau)\right\}\\
&\overset{(6)}{\geq} \inf\left\{\tau\geq 0 \mid 0\leq \left(\left(\underline{\alpha}\otimes\xi\right)\oslash\delta_{T_{\underline{\alpha}}}\right)(\tau)\right\}\\
&= \inf\left\{\tau\geq 0 \mid 0\leq \left(\underline{\alpha}\otimes\xi\right)(\tau+T_{\underline{\alpha}})\right\}\\
&= \inf\left\{\tau'-T_{\underline{\alpha}}\geq 0 \mid 0\leq \left(\underline{\alpha}\otimes\xi\right)(\tau')\right\} && (38)\\
&= \inf\left\{\tau'\geq 0 \mid 0\leq \left(\underline{\alpha}\otimes\xi\right)(\tau')\right\} - T_{\underline{\alpha}} && (39)\\
&= z(\underline{\alpha},\xi) - T_{\underline{\alpha}}.
\end{aligned}
$$

In the second line (Eq. (37)), we apply Lem. 20, since the convolution remains again right-continuous [27]. In the seventh step (Eq. (38)), we perform a variable substitution $\tau' := \tau + T_{\underline{\alpha}}$. In the second to last line (Eq. (39)), we use the distributivity of the addition over the infimum.

As we have shown for an actual sample path that $h(A,D)\geq z(\underline{\alpha},\xi)-T_{\underline{\alpha}}$, we also have that WCD $\geq z(\underline{\alpha},\xi)-T_{\underline{\alpha}}$, and from Thm. 19 we have that WCD $\leq z(\underline{\alpha},\xi) = z(\underline{\alpha},\xi)\vee h(\overline{\alpha},\xi)$.

Moreover, the created sample paths (A,D) and (A^{WC},D^{WC}) are conforming to their arrival and service curves. The system is also causal, i.e. $A^{\mathrm{WC}}\geq D^{\mathrm{WC}}$, since $\xi(0)\leq 0$ and, thus, for instance $D^{\mathrm{WC}} = \left[A^{\mathrm{WC}}\otimes\xi\right]^{+} \leq \left[A^{\mathrm{WC}}\right]^{+} = A^{\mathrm{WC}}$. □

B. Backlog Bound

While it is explicitly mentioned in [12, p. 115] that service curves have to be an element of $\mathcal{F}_0^{\uparrow}$ for finite delay bounds to exist, the same assumption is implicitly made for the backlog bound. However, as we show in the following, the backlog bound from Thm. 14 can be applied to negative service curves with a slight technical adaptation and, more importantly, without the need to assume a minimal arrival curve. The latter becomes clear when looking at Fig. 4. We can see that the backlog remains finite even for this notorious example of a maximal arrival curve and thus without the departure guarantee ever becoming positive.

Theorem 22 (Backlog Bound)**.** *Let an arrival process A traverse a system $\mathcal{S}$. Further, let the arrivals be constrained by maximal arrival curve $\overline{\alpha}\in\mathcal{F}_0^{\uparrow}$, and let the system offer a service curve $\xi\in\mathcal{F}_{\leq 0}^{\uparrow}$. The backlog $q(t)$ satisfies for all t*

$$q(t)\leq v(\overline{\alpha},\xi)\wedge\sup_{s\geq 0}\left\{\overline{\alpha}(s)\right\}. \tag{40}$$

Proof. We have that

$$
\begin{aligned}
q(t) &\overset{(7)}{=} A(t) - D(t)\\
&\overset{(10)}{\leq} A(t) - A\otimes\xi\,(t)\\
&\overset{(1)}{=} \sup_{0\leq s\leq t}\left\{A(t)-A(t-s)-\xi(s)\right\}\\
&\overset{(9)}{\leq} \sup_{0\leq s\leq t}\left\{\overline{\alpha}(s)-\xi(s)\right\}\\
&\overset{(3)}{\leq} \sup_{t\geq 0}\left\{\overline{\alpha}(s)-\xi(s)\right\} = v(\overline{\alpha},\xi)\,. && (41)
\end{aligned}
$$

In the last line (Eq. (41)), we took the supremum over a larger set, so it can potentially increase. On the other hand, we also have that

$$
\begin{aligned}
q(t) &\overset{(7)}{=} A(t) - D(t)\\
&\leq A(t) && (42)\\
&\overset{(9)}{\leq} \overline{\alpha}(t)\\
&\leq \sup_{s\geq 0}\left\{\overline{\alpha}(s)\right\}
\end{aligned}
$$

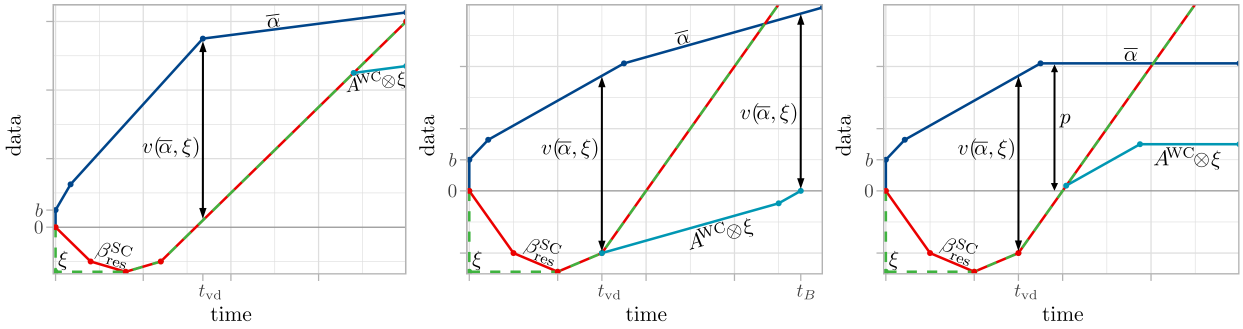
In the second line (Eq. (42)) we used the fact that $D\geq 0$. Therefore, the backlog is less than the minimum of the two bounds. □

So, the usual backlog bound from Thm. 14 is almost recovered. Note, however, that the special case of a bounded arrival curve needs to be treated explicitly in the case of negative service curves, since the vertical deviation can be conservative for the case that the arrival curve never reaches $v(\overline{\alpha},\xi)$ (see also Fig. 5c).

This observation indicates that proving the tightness of the backlog bound is more involved than in the standard case, where we achieve the vertical deviation by simply setting $A=\overline{\alpha}$ (”greedy arrivals”) and $D=\overline{\alpha}\otimes\beta$ (”lazy server”) [12]. The complication arises due to the fact that the vertical deviation is taken on when $\xi<0$, yet for the actual departures we have, of course, $D\geq 0$. Hence, we need to find a worst-case sample path that actually provokes the backlog bound from Thm. 22.

Next, we prove the tightness of the backlog bound. Here, we need to distinguish cases corresponding to the minimum $v(\overline{\alpha},\xi)\wedge\sup_{s\geq 0}\left\{\overline{\alpha}(s)\right\}$ in Thm. 22. Further, for ease of presentation in the proof, we make the assumption of the maximal arrival curve $\overline{\alpha}$ being continuous $\forall t>0$.

Theorem 23 (Tightness of the Backlog Bound)**.** *Let an arrival process A traverse a system $\mathcal{S}$. Further, let the arrivals be constrained by a sub-additive maximal arrival curve $\overline{\alpha}\in\mathcal{F}_0^{\uparrow}$, $\overline{\alpha}(t)$ being continuous $\forall t>0$. The system offers a service*



a) Case I-A, the standard case.

b) Case I-B, the interesting case.

c) Case II, the plateau case.

Fig. 5: Graphical illustration of different cases in Thm. 23.

curve $\xi \in \mathcal{F}^{\uparrow}_{\leq 0}$. Let $t_{\mathrm{vd}} := \arg\sup_{s\geq 0} \{\overline{\alpha}(s) - \xi(s)\}$. We have to treat the following cases:

<u>Case I</u> ("No plateau"): $\exists t \geq 0 : \overline{\alpha}(t) \geq v(\overline{\alpha}, \xi)$.

That means we have an arrival curve which grows large enough such that it is possible for the backlog to attain $v(\overline{\alpha}, \xi)$.

<u>Case I-A</u> ("The standard case", see Fig. 5a): $\xi(t_{\mathrm{vd}}) \geq 0$. Set $A^{\mathrm{WC}} := \overline{\alpha}$ and $D^{\mathrm{WC}} := [\overline{\alpha} \otimes \xi]^+$, then,

$$q(t_{\mathrm{vd}}) = v(\overline{\alpha}, \xi) \wedge \sup_{s\geq 0} \{\overline{\alpha}(s)\}. \tag{43}$$

In this case, the negativity of ξ essentially plays no role (as the vertical deviation is attained when $\xi \geq 0$, see again Fig. 5a) and the worst-case sample path is the conventional one with greedy arrivals and lazy server.

<u>Case I-B</u> ("The interesting case", see Fig. 5b): $\xi(t_{\mathrm{vd}}) < 0$. Set $t_B := \overline{\alpha}^{-1}(v(\overline{\alpha}, \xi))$,

$$A^{\mathrm{WC}}(t) := \begin{cases} \overline{\alpha}(t_B) - \overline{\alpha}(t_B - t), & \text{if } t \leq t_B, \\ \overline{\alpha}(t_B), & \text{otherwise,} \end{cases}$$

and $D^{\mathrm{WC}} := \left[A^{\mathrm{WC}} \otimes \xi\right]^+$.

Then,

$$q(t_B) = v(\overline{\alpha}, \xi) \wedge \sup_{s\geq 0} \{\overline{\alpha}(s)\}. \tag{44}$$

This is the interesting case where $v(\overline{\alpha}, \xi)$ is attained at a later point in time on the worst-case sample path than for arrival and service curve, because ξ is still negative at time t_{vd}. Here, the worst-case sample path is not just greedy arrivals and lazy server.

<u>Case II</u> ("The plateau case", see Fig. 5c):
$\overline{\alpha}(t) < v(\overline{\alpha}, \xi), \forall t \geq 0$ and $\exists t_p \geq 0$ such that *$\forall t \geq t_p : \overline{\alpha}(t) = p := \sup_{s\geq 0} \overline{\alpha}(s)$. Set*

$$A^{\mathrm{WC}}(t) := \begin{cases} \overline{\alpha}(t_p) - \overline{\alpha}(t_p - t), & \text{if } t \leq t_p, \\ \overline{\alpha}(t_p), & \text{otherwise,} \end{cases}$$

and $D^{\mathrm{WC}} := \left[A^{\mathrm{WC}} \otimes \xi\right]^+$. Then,

$$q(t_p) = v(\overline{\alpha}, \xi) \wedge \sup_{s\geq 0} \{\overline{\alpha}(s)\}. \tag{45}$$

The special case of an arrival curve with a plateau needs to be dealt with separately, since the backlog may never attain $v(\overline{\alpha}, \xi)$, when the plateau is not high enough.

Proof. Let us consider <u>Case I-A</u>, then:

$$\begin{aligned} q(t_{\mathrm{vd}}) \overset{(7)}{=} & A^{\mathrm{WC}}(t_{\mathrm{vd}}) - D^{\mathrm{WC}}(t_{\mathrm{vd}}) \\ =& \overline{\alpha}(t_{\mathrm{vd}}) - [\overline{\alpha} \otimes \xi(t_{\mathrm{vd}})]^+ \\ =& \overline{\alpha}(t_{\mathrm{vd}}) - \overline{\alpha} \otimes \xi(t_{\mathrm{vd}}) && (46) \\ \geq& \overline{\alpha}(t_{\mathrm{vd}}) - \xi(t_{\mathrm{vd}}) && (47) \\ =& v(\overline{\alpha}, \xi) \\ =& v(\overline{\alpha}, \xi) \wedge \sup_{s\geq 0} \overline{\alpha}(s). \end{aligned}$$

Due to $\xi(t_{\mathrm{vd}}) \geq 0$, the third Eq. (46) holds. In the fourth step (Eq. (47)) we used the fact that $\xi = \delta_0 \otimes \xi \geq \overline{\alpha} \otimes \xi$, since $\overline{\alpha}(0) = 0$ and the isotonicity of the convolution (see Remark 5). Then, by the upper bound on the backlog from Thm. 22 and $\sup_{s\geq 0} \overline{\alpha}(s) \geq v(\overline{\alpha}, \xi)$, the claim follows.

The arrival and service curve properties as well as causality are obviously fulfilled since we are in the standard case.

For <u>Case I-B</u>, we first check that the sample path A^{WC} is conforming to the maximal arrival curve $\overline{\alpha}$.

To that end, it suffices to verify the maximal arrival curve property in the interval $[0, t_B]$, since for $t > t_B$, $A^{\mathrm{WC}}(t)$ is trivially conforming. Hence, $\forall s, t \in [0, t_B]$ with $s \leq t$:

$$\begin{aligned} A^{\mathrm{WC}}(t) - A^{\mathrm{WC}}(s) =& \overline{\alpha}(t_B) - \overline{\alpha}(t_B - t) - \\ & (\overline{\alpha}(t_B) - \overline{\alpha}(t_B - s)) \\ =& \overline{\alpha}(t_B - s) - \overline{\alpha}(t_B - t) \\ \overset{(19)}{\leq}& \overline{\alpha}(t - s). \end{aligned}$$

Next, we show that

$$D^{\mathrm{WC}}(t_B) = 0. \tag{48}$$

For this, it suffices to show that $A^{\mathrm{WC}} \otimes \xi\,(t_B) = 0$:

$$\begin{aligned} A^{\mathrm{WC}} \otimes \xi\,(t_B) =& \inf_{0\leq s\leq t_B} \left\{A^{\mathrm{WC}}(t_B - s) + \xi(s)\right\} \\ =& \inf_{0\leq s\leq t_B} \{\overline{\alpha}(t_B) - \overline{\alpha}(s) + \xi(s)\} \end{aligned}$$

$$\begin{aligned}&\overset{(1)}{=}\overline{\alpha}(t_B) - \sup_{0\leq s\leq t_B}\left\{\overline{\alpha}(s) - \xi(s)\right\}\\&=v(\overline{\alpha},\xi) - v(\overline{\alpha},\xi) = 0.\end{aligned}$$

In the last line, we used the definition of t_B and the fact that $t_{\mathrm{vd}} \leq t_B$ in this case, since $\xi(t_{\mathrm{vd}}) < 0$ and thus $\overline{\alpha}(t_{\mathrm{vd}}) \leq \overline{\alpha}(t_{\mathrm{vd}}) - \xi(t_{\mathrm{vd}}) = v(\overline{\alpha},\xi) = \overline{\alpha}(t_B)$, and $\overline{\alpha}$ being non-decreasing.

Then, we obtain

$$\begin{aligned}q(t_B) &\overset{(7)}{=} A^{\mathrm{WC}}(t_B) - D^{\mathrm{WC}}(t_B) \overset{(48)}{=} A^{\mathrm{WC}}(t_B) = \overline{\alpha}(t_B)\\&=\overline{\alpha}(\overline{\alpha}^{-1}(v(\overline{\alpha},\xi))\\&=v(\overline{\alpha},\xi) \qquad (49)\\&=v(\overline{\alpha},\xi)\wedge\sup_{s\geq 0}\overline{\alpha}(s),\end{aligned}$$

where in the second to last step (Eq. (49)) the pseudo-inverse (Def. 2) is exact, due to the continuity of $\overline{\alpha}$, and in the last step we use the same argument as in the last step of Case I-A. We note that due to Th. 22, $\forall t \geq 0$:

$$q(t) \leq v(\overline{\alpha},\xi)\wedge\sup_{s\geq 0}\left\{\overline{\alpha}(s)\right\} = q(t_B).$$

D^{WC} is clearly conforming to the service curve ξ. Further, we created a system which is causal, i.e. $A^{\mathrm{WC}} \geq D^{\mathrm{WC}}$ since $A^{\mathrm{WC}} \in \mathcal{F}_0^{\uparrow}$ and thus $A^{\mathrm{WC}} \geq \left[A^{\mathrm{WC}}\otimes\xi\right]^+$ (using again the special case of the isotonicity of the convolution).

Lastly, we treat Case II: clearly $p < v(\overline{\alpha},\xi)$. Again, A^{WC} is conforming to the maximal arrival curve $\overline{\alpha}$ (due to the sub-additivity of $\overline{\alpha}$, see Case I-B).

We show that

$$D^{\mathrm{WC}}(t_p) = 0, \qquad (50)$$

for which it is sufficient to show that $A^{\mathrm{WC}}\otimes\xi\,(t_p) < 0$:

$$\begin{aligned}A^{\mathrm{WC}}\otimes\xi\,(t_p) &= \inf_{0\leq s\leq t_p}\left\{A^{\mathrm{WC}}(t_p - s) + \xi(s)\right\}\\&= \inf_{0\leq s\leq t_p}\left\{\overline{\alpha}(t_p) - \overline{\alpha}(s) + \xi(s)\right\}\\&\overset{(1)}{=}\overline{\alpha}(t_p) - \sup_{0\leq s\leq t_p}\left\{\overline{\alpha}(s) - \xi(s)\right\}. \qquad (51)\end{aligned}$$

To continue with Eq. (51), we need to distinguish two cases: (a) if $t_p \geq t_{\mathrm{vd}}$, we have

$$\overline{\alpha}(t_p) - \sup_{0\leq s\leq t_p}\left\{\overline{\alpha}(s) - \xi(s)\right\} = p - v(\overline{\alpha},\xi) < 0;$$

(b) if $t_p < t_{\mathrm{vd}}$, we have $\xi(t_p) \leq \xi(t_{vd})$ (as ξ is non-decreasing) and $\alpha(t_p) = \alpha(t_{vd}) = p$. This implies

$$\begin{aligned}p - \xi(t_p) &= \alpha(t_p) - \xi(t_p)\\&\geq \alpha(t_{vd}) - \xi(t_{vd}) = v(\overline{\alpha},\xi)\,.\end{aligned}$$

For $p < v(\overline{\alpha},\xi)$, we see $\xi(t_p) < 0$, and thus $\forall t \leq t_p, \xi(t) < 0$ (as ξ is non-decreasing). We continue with Eq. (51):

$$\overline{\alpha}(t_p) - \sup_{0\leq s\leq t_p}\left\{\overline{\alpha}(s) - \xi(s)\right\} < \overline{\alpha}(t_p) - \sup_{0\leq s\leq t_p}\left\{\overline{\alpha}(s)\right\} = 0.$$

Thus, we obtain

$$\begin{aligned}q(t_p) &\overset{(7)}{=} A^{\mathrm{WC}}(t_p) - D^{\mathrm{WC}}(t_p) \overset{(50)}{=} A^{\mathrm{WC}}(t_p)\\&=\overline{\alpha}(t_p) = p = v(\overline{\alpha},\xi)\wedge p\\&=v(\overline{\alpha},\xi)\wedge\sup_{s\geq 0}\left\{\overline{\alpha}(s)\right\},\end{aligned}$$

where in the second to last step we use $p < v(\overline{\alpha},\xi)$. We note that due to Th. 22, $\forall t \geq 0$:

$$q(t) \leq v(\overline{\alpha},\xi)\wedge\sup_{s\geq 0}\left\{\overline{\alpha}(s)\right\} = q(t_p).$$

D^{WC} is clearly conforming to the service curve ξ. Further, we created a system which is causal, i.e. $A^{\mathrm{WC}} \geq D^{\mathrm{WC}}$ since $A^{\mathrm{WC}} \in \mathcal{F}_0^{\uparrow}$ and thus $A^{\mathrm{WC}} \geq \left[A^{\mathrm{WC}}\otimes\xi\right]^+$ (again by the special case of the isotonicity of the convolution).

□

IV. Patterns of Application

With a broader set of service curves that we can derive performance bounds from, the question of potential applications arises. The extended NC results remove a previous blind spot, where a system with multiple flows but no *strict* service curve could not be adequately modeled and analyzed. We are now also able to exploit the concatenation theorem (see Thm. 15) while still obtaining performance bounds in a system that would normally rely on a strict service curve. This is desirable, as a node-by-node analysis often cannot capture certain properties of the overall system [18], [24, Section 1.4.3], resulting in less accurate performance bounds.

In this section, we highlight two possible patterns of applications that have found previous research interest in the real-time domain and where the novel results provide an interesting insight into the system performance analysis. Due to the assumptions typically made in these systems, a minimal arrival curve usually exists. We show that the new analysis can improve on results of state-of-the-art techniques and may even enable a system analysis for certain areas of the parameter space where existing techniques deliver no solution at all.

A. Computation-Communication Systems

When conducting a system performance analysis, a frequent convenient assumption is that of a homogeneous network where each node fulfills the same purpose. This allows us to exploit the central NC theorems and efficiently calculate performance bounds. However, if we cannot make this assumption, the system analysis quickly deteriorates into a node-by-node analysis not utilizing the concatenation property (see Thm. 15). Of special interest are applications whose components fulfill different tasks, e.g., loosely time-triggered architectures [28] or in-network processing [23], mixing computation and communication resources. In general, whenever we encounter a system with both computational and communication components, our new analysis gives interesting insights and potential improvements over conventional NC analyses.

We consider a pattern of a mixed Computation-Communication (C/C) system consisting of n components and $n+2$ flows (see Fig. 6 on the next page), and proceed with deriving formulas for calculating the end-to-end delay bound across the n components. Let f_1 be the flow of interest and

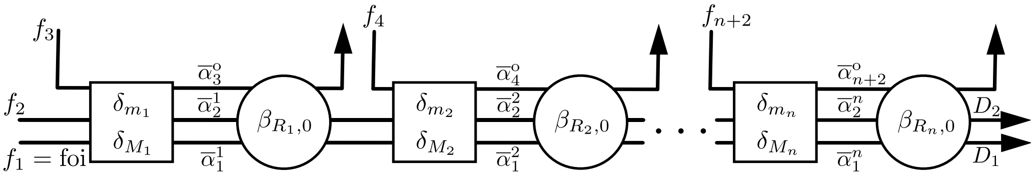


Fig. 6: General form of a C/C system with three flows crossing each C/C component.

f_2 a cross-flow. Both flows traverse components $1, \ldots, n$ as an aggregate. Assume that there are n additional cross-flows $f_i, i \in \{3, n+2\}$, passing through each C/C component $i-2$, respectively. We assume static priority scheduling at each component, and assign the highest priority to flows $f_i, i \in \{3, n+2\}$. Flow f_2 is assigned the second highest priority, and the foi the lowest. Each flow is constrained by a maximal token-bucket arrival curve $\overline{\alpha}_i = \gamma_{r_i,b_i}$. The foi is additionally restricted by a minimal arrival curve $\underline{\alpha}_1 = \beta_{\underline{r}_1,T_{\underline{\alpha}_1}}$. We assume that the delay at each computational element $i \in \{1, n\}$ in the system is lower bounded by m_i and upper bounded by M_i (modeled in Fig. 6 as service curves using the shift function δ as in [12, Theorem 6.2]). Let $T_i := M_i - m_i$ be the delay variance at each computational element i. Each communication element $i \in \{1, n\}$ provides a simple constant-rate service curve $\beta_{R_i,0}$. Using the results proposed in Sect. III for our new analysis (na), we first calculate the residual service curve for flow f_1 as

$$\begin{aligned}\beta_{\text{res}}^{\text{na}} &:= \left(\left(\bigotimes_{i=1}^{n} (\beta_{R_i,T_i} - \overline{\alpha}_{i+2})\right)_{\downarrow} - \overline{\alpha}_2\right)_{\downarrow} \\ &= \left(\xi_{b_{i+2}+r_{i+2}\sum_{i=1}^n T_i, \bigwedge_{i=1}^n (R_i - r_{i+2}), \sum_{i=1}^n T_i} - \overline{\alpha}_2\right)_{\downarrow} \\ &= \xi_{b_2+b_{i+2}+(r_2+r_{i+2})\sum_{i=1}^n T_i, \bigwedge_{i=1}^n (R_i - r_{i+2}) - r_2, \sum_{i=1}^n T_i},\end{aligned}$$

with

$$\xi_{b_N,R,T}(t) := \beta_{R,T}(t) - b_N.$$

An end-to-end delay bound for flow f_1 using the new analysis is then calculated as

$$d_{\text{e2e}}^{\text{na}} = \max\{h(\overline{\alpha}_1, \beta_{\text{res}}^{\text{na}}), z(\underline{\alpha}_1, \beta_{\text{res}}^{\text{na}})\}. \tag{52}$$

For the conventional analysis (ca), we determine the input to the C/C components by using the output bound (see Eq. (18)), as the input to each component is the output of the respective flow at the preceding component. For the communication element of component i, the input flows are thus given as

$$\begin{aligned}&\overline{\alpha}_2^1 = \overline{\alpha}_2 \oslash \delta_{T_1}, \overline{\alpha}_1^1 = \overline{\alpha}_1 \oslash \delta_{T_1}, \\ &\overline{\alpha}_2^i = \overline{\alpha}_2^{i-1} \oslash [\beta_{R_{i-1},0} - (\overline{\alpha}_{i+1} \oslash \delta_{T_{i-1}})]^+ \oslash \delta_{T_i} = \gamma_{r_2^i, b_2^i}, \\ &\overline{\alpha}_1^i = \overline{\alpha}_1^{i-1} \oslash [\beta_{R_{i-1},0} - ((\overline{\alpha}_2^{i-1} + \overline{\alpha}_{i+1}) \oslash \delta_{T_{i-1}})]^+ \oslash \delta_{T_i}, \\ &\overline{\alpha}_{i+2}^o = \gamma_{r_{i+2}, b_{i+2}+r_{i+2}\cdot T_i} = \gamma_{r_{i+2}^o, b_{i+2}^o}.\end{aligned}$$

Next, the residual service curve for flow f_1 employing a static priority policy can be calculated for each component i as

$$\beta_{\text{res}}^{\text{ca},i} := \left[\beta_{R_i,0} - \overline{\alpha}_2^i - \overline{\alpha}_{i+2}^o\right]^+ = \beta_{R_{\text{res}}^{\text{ca},i}, T_{\text{res}}^{\text{ca},i}}.$$

A delay bound for communication component i is obtained by

$$h(\overline{\alpha}_1^i, \beta_{\text{res}}^{\text{ca},i}) = \frac{b_1 + r_1 \sum_{j=1}^{i} T_j + r_1 \sum_{j=1}^{i-1} \frac{b_2^j + b_{j+2}^o}{R_{\text{res}}^{\text{ca},j}}}{R_{\text{res}}^{\text{ca},i}} + T_{\text{res}}^{\text{ca},i}$$

A delay bound for the whole C/C component i is simply $d_i = T_i + h(\overline{\alpha}_1, \beta_{\text{res}}^{\text{ca},i})$. An end-to-end delay bound for flow f_1 using the conventional analysis is then calculated as

$$d_{\text{e2e}}^{\text{ca}} = \sum_{i=1}^{n} d_i. \tag{53}$$

Equipped with these formulas, we proceed with a small case study, evaluating the two analyses. To this end, we consider the general system previously described (see Fig. 6). We calculate the end-to-end delay bound for the foi f_1 in this system. To evaluate the effect of the minimal arrival curve, we define a general parameter set and vary the minimal rate $\underline{r}_1$ over a range of values. Let $b_1 = b_2 = b_3 = 1\,\text{Mbit}$, $r_1 = r_2 = r_3 = 5\,\frac{\text{Mbit}}{\text{s}}$, $R_i = 20\,\frac{\text{Mbit}}{\text{s}} =: R$, and $T_i = 50\,\text{ms}$, $i \in \{1, \ldots, n\}$. We set $T_{\underline{\alpha}_1} = \frac{b_1}{R}$ and choose $\underline{r}_1 \in \{0.5, 1.25, 2.5, 3.75, 5\}\,\frac{\text{Mbit}}{\text{s}}$. The delay bound is calculated for different numbers of C/C components $n \in \{2 \ldots, 20\}$. For each value of $\underline{r}_1$ and n, we calculate the end-to-end delay bounds for the new analysis using Eq. (52), and for the conventional analysis using Eq. (53). The results are shown in Fig. 7. We can see that for $\underline{r}_1 \in \{3.75\,\frac{\text{Mbit}}{\text{s}}, 5\,\frac{\text{Mbit}}{\text{s}}\}$, we always achieve a much more accurate end-to-end delay bound. For $\underline{r}_1 \in \{1.25\,\frac{\text{Mbit}}{\text{s}}, 2.5\,\frac{\text{Mbit}}{\text{s}}\}$, the new delay bound is below the conventional delay bound from 5 resp. 3 C/C components in the system onwards. For $0.5\,\frac{\text{Mbit}}{\text{s}}$, however, the new end-to-end delay bound becomes more conservative

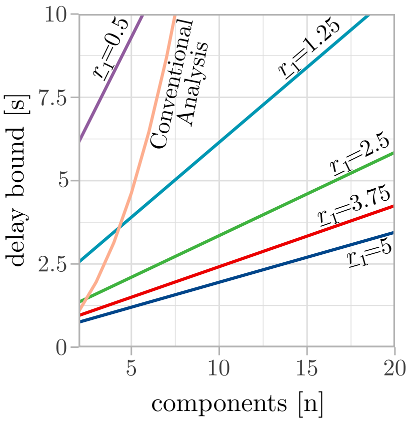
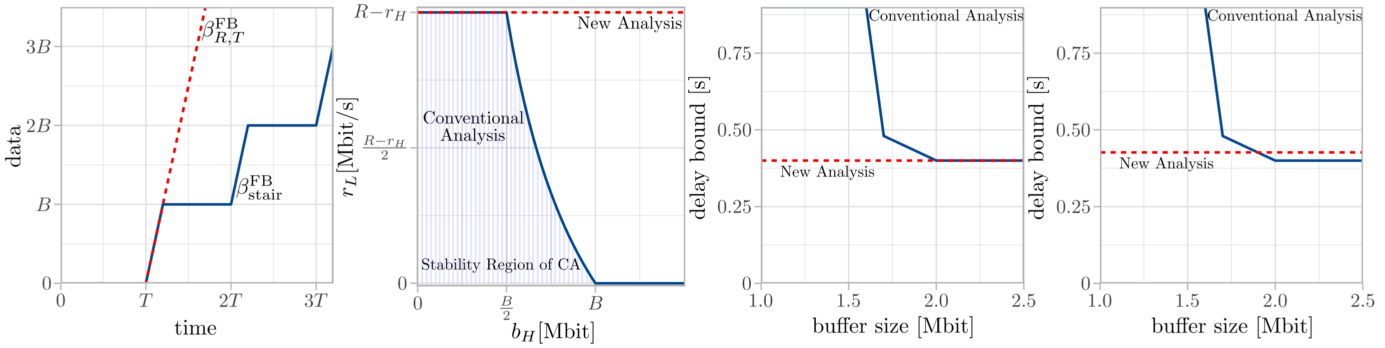


Fig. 7: Comparison of delay bounds for varying rates of $\underline{r}_1$.



a) Shapes of $\beta^{\text{FB}}_{R,T}$ and $\beta^{\text{FB}}_{\text{stair}}$. b) Stability regions of the analyses. c) Delay bounds for $\underline{r}_L = 4.5\,\frac{\text{Mbit}}{\text{s}}$. d) Delay bounds for $\underline{r}_L = 3.75\,\frac{\text{Mbit}}{\text{s}}$.

Fig. 9: Illustration of finite buffer service curves, stability regions and delay bounds for the case study.

for small numbers of components. Expectedly, with our newly proposed approach, we rely on the guarantees provided by a minimal arrival curve. Consequently, the better the guarantees, i.e., the higher $\underline{r}_1$, the better the calculated end-to-end delay bound becomes. However, we can observe in Fig. 7 that even for the smallest minimum arrival rate of $0.5\,\frac{\text{Mbit}}{\text{s}}$ we have a better scaling of the delay bound than for the conventional analysis, which exhibits a super-linear scaling in the number of components. This means that, for large enough systems, the new approach will eventually outperform the conventional analysis, even with low minimal arrival guarantees.

B. *Finite Shared Buffers*

Systems with finite buffers and their sizing are of high relevance in various application areas, such as Network-on-Chip [29]–[31]. Especially systems employing window-based flow control in the event of the input exceeding the capacity of a component have seen previous work [32]–[34]. All of these have treated the case of a *single* flow (aggregate). However, oftentimes multiple flows are sharing buffers in such systems, potentially with different priorities, and it is necessary to size buffers for each priority adequately.

In this application pattern, we consider two priority queues, one for a high and the other for a low priority flow. First, we derive the required buffer sizes for each flow and find that conventional NC analyses cannot properly express and analyze all feasible system designs. Next, we calculate delay bounds for the low priority foi f_L in this finite shared buffer system.

Before we can derive performance bounds in the system for the low priority flow f_L, we need to determine the residual service curve for both the conventional and new analysis.

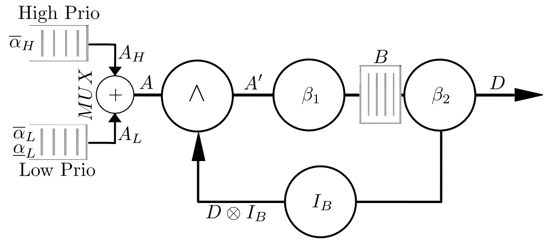


Fig. 8: System with a finite shared buffer.

Consider the system in Fig. 8. Let I_B be the service curve of the feedback control for arrivals exceeding the finite buffer with capacity B at β_2. We define $I_B(t) = +\infty$ for $t > 0$ and $I_B(0) = B$, as in [12], [34], [16, Section 2.3.7]. It holds that $D \otimes I_B(t) = D(t) + B$, and, hence, A' cannot be more than B data units ahead of D (as $A' = A \wedge (D \otimes I_B)$). We let $\beta_i = \beta_{R_i,0}, i = 1, 2$, $R = R_1 \wedge R_2$, and $T = T_1 + T_2$. For this closed-loop feedback system, we have

$$A' \geq A \wedge (D \otimes I_B), D \geq A' \otimes \beta_1 \otimes \beta_2,$$

where $A = A_H + A_L$. Combining both inequalities, we obtain

$$D \geq A \otimes (\beta_1 \otimes \beta_2) \wedge D \otimes (I_B \otimes \beta_1 \otimes \beta_2),$$

which can be turned into an open-loop system [16, Section 2.3.7]

$$D \geq A \otimes (\beta_1 \otimes \beta_2) \otimes (I_B \otimes \beta_1 \otimes \beta_2)^*,$$

where $(I_B \otimes \beta_1 \otimes \beta_2)^*$ is the sub-additive closure (see Def. 17). Hence, the system offers a service curve $\beta^{\text{FB}} = \beta_1 \otimes \beta_2 \otimes (I_B \otimes \beta_1 \otimes \beta_2)^*$. In general, it holds that, for $RT \leq B$, the service curve offered to the flow is equal to $\beta^{\text{FB}}_{R,T} = \beta_1 \otimes \beta_2$. If, however, the bandwidth-delay product RT is greater than the available buffer B, it holds that the service curve is a staircase function $\beta^{\text{FB}}_{\text{stair}}$, since there is not enough buffer space available to serve the flow without delaying it at the entrance to the feedback loop. Both service curves are illustrated in Fig. 9a.

In the following, we assume that both flows are upper-constrained by token buckets $\overline{\alpha}_H = \gamma_{r_H,b_H}$ and $\overline{\alpha}_L = \gamma_{r_L,b_L}$, respectively. We let $T_i = 0$ for β_1 and β_2. Consequently, it always holds that $\beta^{\text{FB}} = \beta^{\text{FB}}_{R,T} = \beta_{R,0}$ for the aggregate of the flows. Thus, the buffer requirement for the high priority queue is equal to

$$v(\overline{\alpha}_H, \beta^{\text{FB}}) = v(\gamma_{r_H,b_H}, \beta_{R,0}) = b_H,$$

independent of which analysis we choose. This is not the case for flow f_L, though. For this, we first have to calculate the residual service curves for each analysis. For the new analysis, the residual service curve is calculated as

$$\beta^{\text{na}}_{\text{res}} = \left(\beta^{\text{FB}} - \overline{\alpha}_H\right)_{\downarrow} = \left(\beta_{R,0} - \overline{\alpha}_H\right)_{\downarrow} = \xi_{b_H,R-r_H,0}.$$

Note that $\beta_{\text{res}}^{\text{na}}$ is independent of the feedback control, i.e., its shape does not depend on the relation of RT and B. In contrast, the residual service curve for the conventional analysis does depend on this relation. We calculate it as

$$\begin{aligned}\beta_{\text{res}}^{\text{ca}} &= [\beta_1 \otimes \beta_2 - \overline{\alpha}_H]^+ \otimes ([\beta_1 \otimes \beta_2 - \overline{\alpha}_H]^+ \otimes I_{B-v(\overline{\alpha}_H,\beta)})^* \\ &= \beta_{R-r_H,\frac{b_H}{R-r_H}} \otimes (\beta_{R-r_H,\frac{b_H}{R-r_H}} \otimes I_{B-b_H})^*.\end{aligned}$$

Using each residual service curve, we determine the buffer requirement for the low priority queue. For the new analysis, we obtain

$$v(\overline{\alpha}_L, \beta_{\text{res}}^{\text{na}}) = b_H + b_L. \tag{54}$$

For the conventional analysis, we need to consider the relation of the bandwidth-delay product for the residual feedback system $R^{\text{res}}T^{\text{res}} = (R - r_H)\frac{b_H}{R-r_H} = b_H$, and the buffer available in it for the low priority flow $B^{\text{res}} = B - b_H$. If $R^{\text{res}}T^{\text{res}} \leq B^{\text{res}}$, i.e., $b_H \leq B - b_H$, then the residual service curve $\beta_{\text{res}}^{\text{ca}}$ follows the shape of $\beta_{R,T}^{\text{FB}}$ (see Fig. 9a). In this case, i.e., $b_H \leq \frac{B}{2}$, we obtain that

$$v(\overline{\alpha}_L, \beta_{\text{res}}^{\text{ca}}) = v(\gamma_{r_L,b_L}, \beta_{R-r_H,\frac{b_H}{R-r_H}}) = b_L + r_L\frac{b_H}{R - r_H}.$$

For $R^{\text{res}}T^{\text{res}} > B^{\text{res}}$, i.e. $b_H > \frac{B}{2}$, we follow the shape of $\beta_{\text{stair}}^{\text{FB}}$, and discover an interesting restriction regarding the rate r_L of flow f_L in order to not diverge from $\beta_{\text{res}}^{\text{ca}}$:

$$r_L \leq \frac{B^{\text{res}}}{T^{\text{res}}} = \left(\frac{B}{b_H} - 1\right)(R - r_H). \tag{55}$$

We give a brief intuition for Eq. (55). $\frac{B^{\text{res}}}{T^{\text{res}}}$ is the long-term rate of $\beta_{\text{res}}^{\text{ca}}$. We can only calculate finite bounds if α_L and $\beta_{\text{res}}^{\text{ca}}$ do not diverge. If $r_L > \frac{B^{\text{res}}}{T^{\text{res}}}$, i.e., the rate of α_L is larger than the long-term rate of $\beta_{\text{res}}^{\text{ca}}$, then the stability of the system is not ensured and infinite performance bounds result.

As $B^{\text{res}} = B - b_H$, we recognize that for $b_H \geq B$, we cannot compute a backlog bound for f_L using the conventional residual service curve $\beta_{\text{res}}^{\text{ca}}$. Furthermore, as we see in Eq. (55), the feasible rate r_L decreases hyperbolically in b_H over the interval $\left(\frac{B}{2}, B\right)$, further limiting the ability to calculate a backlog bound. In Fig. 9b, the so-called stability region of the conventional analysis is shown. Here, the stability region is the parameter space for which we can compute finite backlog bounds.

Clearly, the closer the buffer is to being full with traffic of f_H, the less f_L can send in each window interval $[(i-1)T^{\text{res}}, iT^{\text{res}}), i > 1$, eventually diverging from $\beta_{\text{res}}^{\text{ca}}$. As a result, we cannot determine the vertical deviation for arbitrary r_L that would be valid under the "normal" stability condition $r_L \leq R - r_H$, but violate Eq. (55). In conclusion, the conventional analysis is not able to provide a backlog bound for arbitrary flows f_H, f_L. In contrast, the calculation of the buffer requirement based on the new analysis in Eq. (54) is only restricted by $r_L \leq R - r_H$, thus resulting in a much larger stability region (see again Fig. 9b).

We move on to the delay bound calculation. For flow f_H, the delay bound calculation is the same for both analyses:

$$h(\overline{\alpha}_H, \beta^{\text{FB}}) = h(\gamma_{r_H,b_H}, \beta_{R,0}) = \frac{b_H}{R}.$$

For flow f_L, this looks different, as we have different residual service curves. For the new anaysis, assuming $\underline{\alpha}_L = \beta_{r_L,T_{\underline{\alpha}_L}}$, we calculate

$$\begin{aligned}d_{\text{e2e}}^{\text{na}} &= h(\overline{\alpha}_L, \xi_{b_H,R-r_H,0}) \vee z(\underline{\alpha}_L, \xi_{b_H,R-r_H,0}) \\ &= \left(\frac{b_H + b_L}{R - r_H}\right) \vee \left(T_{\underline{\alpha}_L} + \frac{b_H}{\underline{r}_L}\right).\end{aligned} \tag{56}$$

For the conventional analysis, if we have a staircase residual service curve, we calculate the number of stairs that are needed in the delay bound calculation as $i^* := \lceil \frac{b_L}{B-b_H} \rceil$, and obtain

$$d_{\text{e2e}}^{\text{ca}} = \begin{cases} \frac{b_H+b_L}{R-r_H}, & R^{\text{res}}T^{\text{res}} \leq B^{\text{res}}, \\ \frac{b_L-(i^*-1)(B-b_H)}{R-r_H} + i^*T^{\text{res}}, & \text{otherwise}, \end{cases} \tag{57}$$

where Eq. (55) and $b_H < B$ have to hold (see also Fig. 9b again), otherwise, $h(\overline{\alpha}_L, \beta_{\text{res}}^{\text{ca}}) = \infty$.

We proceed with a brief case study on the two approaches to calculating delay bounds. Consider again the system in Fig. 8. Let $b_L = 2\,\text{Mbit}, b_H = 1\,\text{Mbit}$, and $r_L = r_H = 5\,\frac{\text{Mbit}}{\text{s}}$. For $\underline{\alpha}_L$, we let $T_{\underline{\alpha}_L} = \frac{b_L}{R}$ and $\underline{r}_L \in \{3.75\,\frac{\text{Mbit}}{\text{s}}, 4.5\,\frac{\text{Mbit}}{\text{s}}\}$. Each system offers a service curve $\beta_i = \beta_{R_i,0}$ with $R_i = 12.5\,\frac{\text{Mbit}}{\text{s}}$. We calculate the delay bound using these parameter values for both approaches, varying the size of the finite buffer B. The results are given in Fig. 9. For $\underline{r}_L = 4.5\,\frac{\text{Mbit}}{\text{s}}$ (Fig. 9c), we see that the delay bound of the new analysis is always either equal to or more accurate than the conventional analysis. For $\underline{r}_L = 3.75\,\frac{\text{Mbit}}{\text{s}}$ (Fig. 9d), the delay bound calculation using Eq. (56) falls into the second case of the maximum operator. As a result, for $b_H \leq \frac{B}{2}$, the conventional analysis achieves slightly more accurate bounds. However, for $b_H > \frac{B}{2}$, this changes, as the conventional analysis now calculates its delay bound using the second case of Eq. (57), instead. Now, the delay bound becomes much larger than for the new analysis.

V. Conclusion

In this paper, we extended the NC framework to deal with scenarios in which an aggregate min-plus service curve is given and we want to calculate residual service curves in order to compute per-flow performance bounds. In this case, partially negative service curves arise and existing NC results on performance bounds cannot be applied. We remove this blind spot with the aid of minimal arrival curves, which allow us to calculate tight or at least approximately tight performance bounds even for negative service curves.

This generalization of the performance bounds for negative service curves leads to more flexibility in the modeling of applications, though also requiring more assumptions towards the system. However, this assumption, the minimal arrival curve, is often given in real-time systems. Using the new NC results, we have shown that we can improve the performance analysis of interesting application patterns that occur in real-time systems; we are even able to analyze systems for which a conventional analysis fails to provide performance bounds.

Acknowledgment

We are grateful to the anonymous reviewers for their insightful and constructive feedback. Special thanks to our shepherd for his guidance.

References

[1] L. Maile, K.-S. Hielscher, and R. German, "Network calculus results for TSN: An introduction," in *2020 Information Communication Technologies Conference (ICTC)*. IEEE, 2020, pp. 131–140.

[2] J. A. R. De Azua and M. Boyer, "Complete modelling of AVB in network calculus framework," in *Proceedings of the 22nd International Conference on Real-Time Networks and Systems*, 2014, pp. 55–64.

[3] L. Zhao, P. Pop, and S. S. Craciunas, "Worst-case latency analysis for IEEE 802.1 Qbv time sensitive networks using network calculus," *IEEE Access*, vol. 6, pp. 41 803–41 815, 2018.

[4] H. Charara, J.-L. Scharbarg, J. Ermont, and C. Fraboul, "Methods for bounding end-to-end delays on an AFDX network," in *18th Euromicro Conference on Real-Time Systems (ECRTS'06)*. IEEE, 2006, pp. 10–19.

[5] F. Frances, C. Fraboul, and J. Grieu, "Using network calculus to optimize the AFDX network," in *Conference ERTS'06*, 2006.

[6] M. Boyer and C. Fraboul, "Tightening end to end delay upper bound for AFDX network calculus with rate latency FIFO servers using network calculus," in *2008 IEEE International Workshop on Factory Communication Systems*. IEEE, 2008, pp. 11–20.

[7] M. Bakhouya, S. Suboh, J. Gaber, and T. El-Ghazawi, "Analytical modeling and evaluation of on-chip interconnects using network calculus," in *2009 3rd ACM/IEEE International Symposium on Networks-on-Chip*. IEEE, 2009, pp. 74–79.

[8] M. Boyer, A. Graillat, B. D. De Dinechin, and J. Migge, "Bounding the delays of the MPPA network-on-chip with network calculus: Models and benchmarks," *Performance Evaluation*, vol. 143, pp. 102–124, 2020.

[9] L. Thiele, S. Chakraborty, and M. Naedele, "Real-time calculus for scheduling hard real-time systems," in *2000 IEEE International Symposium on Circuits and Systems (ISCAS)*, vol. 4. IEEE, 2000, pp. 101–104.

[10] E. Wandeler, L. Thiele, M. Verhoef, and P. Lieverse, "System architecture evaluation using modular performance analysis: a case study," *International Journal on Software Tools for Technology Transfer*, vol. 8, pp. 649–667, 2006.

[11] E. Wandeler and L. Thiele, "Workload correlations in multi-processor hard real-time systems," *Journal of Computer and System Sciences*, vol. 73, no. 2, pp. 207–224, 2007.

[12] A. Bouillard, M. Boyer, and E. Le Corronc, *Deterministic Network Calculus: From Theory to Practical Implementation*. John Wiley & Sons, 2018.

[13] S. Chakraborty, S. Künzli, L. Thiele, A. Herkersdorf, and P. Sagmeister, "Performance evaluation of network processor architectures: Combining simulation with analytical estimation," *Computer Networks*, vol. 41, no. 5, pp. 641–665, 2003.

[14] R. L. Cruz, "A calculus for network delay. I. Network elements in isolation," *IEEE Transactions on Information Theory*, vol. 37, no. 1, pp. 114–131, 1991.

[15] ——, "A calculus for network delay. II. Network analysis," *IEEE Transactions on Information Theory*, vol. 37, no. 1, pp. 132–141, 1991.

[16] C.-S. Chang, *Performance Guarantees in Communication Networks*. Springer Science & Business Media, 2000.

[17] J. B. Schmitt, F. A. Zdarsky, and I. Martinovic, "Improving performance bounds in feed-forward networks by paying multiplexing only once," in *14th GI/ITG Conference-Measurement, Modelling and Evalutation of Computer and Communication Systems*. VDE, 2008, pp. 1–15.

[18] J. B. Schmitt, F. A. Zdarsky, and M. Fidler, "Delay bounds under arbitrary multiplexing: When network calculus leaves you in the lurch..." in *IEEE INFOCOM 2008-The 27th Conference on Computer Communications*. IEEE, 2008, pp. 1669–1677.

[19] F. Geyer and S. Bondorf, "DeepTMA: Predicting effective contention models for network calculus using graph neural networks," in *IEEE INFOCOM 2019-IEEE Conference on Computer Communications*. IEEE, 2019, pp. 1009–1017.

[20] S. Bondorf, P. Nikolaus, and J. B. Schmitt, "Quality and cost of deterministic network calculus: Design and evaluation of an accurate and fast analysis," *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, vol. 1, no. 1, pp. 1–34, 2017.

[21] R. Wilhelm, J. Engblom, A. Ermedahl, N. Holsti, S. Thesing, D. Whalley, G. Bernat, C. Ferdinand, R. Heckmann, T. Mitra *et al.*, "The worst-case execution-time problem—overview of methods and survey of tools," *ACM Transactions on Embedded Computing Systems (TECS)*, vol. 7, no. 3, pp. 1–53, 2008.

[22] I. Stoica, H. Zhang, and T. E. Ng, "A hierarchical fair service curve algorithm for link-sharing, real-time, and priority services," *IEEE/ACM Transactions on Networking*, vol. 8, no. 2, pp. 185–199, 2000.

[23] J. B. Schmitt, F. A. Zdarsky, and L. Thiele, "A comprehensive worst-case calculus for wireless sensor networks with in-network processing," in *28th IEEE International Real-Time Systems Symposium (RTSS 2007)*. IEEE, 2007, pp. 193–202.

[24] J.-Y. Le Boudec and P. Thiran, *Network Calculus: A Theory of Deterministic Queuing Systems for the Internet*. Springer, 2001. [Online]. Available: https://leboudec.github.io/netcal/

[25] M. Moy and K. Altisen, "Arrival curves for real-time calculus: the causality problem and its solutions," in *International Conference on Tools and Algorithms for the Construction and Analysis of Systems*. Springer, 2010, pp. 358–372.

[26] R. Agrawal, R. Cruz, C. M. Okino, and R. Rajan, "A framework for adaptive service guarantees," in *Proceedings of the Annual Allerton Conference on Communication Control and Computing*, vol. 36, 1998, pp. 693–702.

[27] D. Guidolin-Pina and M. Boyer, "Looking for equivalences of the services between left and right continuity in the Network Calculus theory," Sep. 2022, working paper or preprint. [Online]. Available: https://hal.science/hal-03772867

[28] A. Benveniste, P. Caspi, P. L. Guernic, H. Marchand, J.-P. Talpin, and S. Tripakis, "A protocol for loosely time-triggered architectures," in *International Workshop on Embedded Software*. Springer, 2002, pp. 252–265.

[29] M. Coenen, S. Murali, A. Rădulescu, K. Goossens, and G. De Micheli, "A buffer-sizing algorithm for networks on chip using TDMA and credit-based end-to-end flow control," in *Proceedings of the 4th International Conference on Hardware/Software Codesign and System Synthesis*, 2006, pp. 130–135.

[30] S. Kumar, A. Jantsch, J.-P. Soininen, M. Forsell, M. Millberg, J. Oberg, K. Tiensyrja, and A. Hemani, "A network on chip architecture and design methodology," in *Proceedings IEEE Computer Society Annual Symposium on VLSI. New Paradigms for VLSI Systems Design. ISVLSI 2002*. IEEE, 2002, pp. 117–124.

[31] Y. Qian, Z. Lu, and W. Dou, "Analysis of worst-case delay bounds for best-effort communication in wormhole networks on chip," in *2009 3rd ACM/IEEE International Symposium on Networks-on-Chip*. IEEE, 2009, pp. 44–53.

[32] A. Bose, X. Jiang, B. Liu, and G. Li, "Analysis of manufacturing blocking systems with network calculus," *Performance Evaluation*, vol. 63, no. 12, pp. 1216–1234, 2006.

[33] A. Bouillard, L. T. Phan, and S. Chakraborty, "Lightweight modeling of complex state dependencies in stream processing systems," in *2009 15th IEEE Real-Time and Embedded Technology and Applications Symposium*. IEEE, 2009, pp. 195–204.

[34] R. Agrawal, R. L. Cruz, C. Okino, and R. Rajan, "Performance bounds for flow control protocols," *IEEE/ACM Transactions on Networking*, vol. 7, no. 3, pp. 310–323, 1999.