A Novel Probabilistic Extension of Network Calculus for Workload Loss Examinations

József Bíró, András Gulyás, and Zalán Heszberger
Budapest University of Technology and Economics
Department of Telecommunications and Media Informatics
1117 Budapest, Magyar tudósok körútja 2. Hungary
Email: {biro,gulyas,heszi}@tmit.bme.hu

Abstract—In this paper we define a calculus for communication networks which is suitable for workload loss estimation based on the original definition of stationary loss ratio. Our novel calculus is a probabilistic extension of the deterministic network calculus, and takes an envelope approach to describe arrivals and services for the quantification of resource requirements in the network. We introduce the effective w-arrival curve and the effective w-service curve for describing the inputs and the service and we show that the per-node results can be extended to a network of nodes with the definition of the effective network w-service curve. The derivation of effective w-arrival curves and effective w-service curves for typical arrivals processes and schedulers is also an important contribution of this paper.

Keywords Network calculus, resource estimation, statistical multiplexing

I. INTRODUCTION

Real time applications in today and future heterogeneous networking environment require simple and efficient Quality of Service provisioning. The expected traffic (packet) loss ratio at network nodes is one of the key QoS parameters which should always be considered and controlled in almost all kind of traffic. Traffic management functions (like connection admission control, packet scheduling algorithms) strongly rely on loss performance analysis.

During the past few years significant attention has been paid for bounding the workload loss ratio within the framework of deterministic network calculus [1]. In [2], [3] some long run loss ratio bounds have been presented, which are founded on buffer saturation probability approximations, hence we call them indirect bounds\(^1\). More recently in [7] [8] a definition based stochastic workload loss bounding technique has been proposed for deterministic network calculus.

Since the worst-case view of the deterministic network calculus results in an overestimation of the actual resource requirements of traffic flows in a packet network, the extension of the network calculus to a probabilistic setting receives a significant attention nowadays [2], [9], [10], [11], [12], [13]. The existing probabilistic extensions share a common property that they assign some kind of violation probability to the definitions of the arrival and service curves. This property makes the estimation of the the overflow type quantities much easier as is shown in [14], however such extensions are not suitable for the direct estimation of the workload loss ratio which still has to be done in an indirect way. These complications indicate, that the workload loss ratio bounds cannot be deduced from the current stochastic versions of network calculus in a straightforward manner [8]. This fact urged us to compose the problem in a more natural way.

Our paper is organized as follows: In section III a short overview of deterministic network calculus is given followed by the most important results of a recently introduced min-plus algebra [15] [1] based stochastic extension [10] to the deterministic network calculus. After that, a novel calculus is defined which is designed for direct (definition based) workload loss ratio approximations. We introduce the effective w-arrival curve and the effective w-service curve for describing the inputs and the service and we prove fundamental per-node statements for the backlog, delay and the effective w-arrival curve of the output traffic. It will be shown that the per-node results can be extended to a network of nodes with the definition of the effective network w-service curve in section IV. The connection between the effective w-arrival curve and effective bandwidth [16], is pointed out in section V and effective w-arrival curves and effective w-service curves are derived for typical arrivals processes and schedulers in section VI and VII.

\(^1\)It is true in general, that most of the papers concerning loss ratio apply buffer overflow probability for WLR estimation [4], [5], nevertheless, it is shown, that the ratio \(\frac{\text{WLR}}{\text{P}(Q>q)}\) can be arbitrary under certain circumstances [6].
In section VIII we compare the derived workload loss bound with the closest existing probabilistic direct bound [7] and some simulation results.

II. NOTATION AND ASSUMPTIONS

In this paper the following notations are used: $A_i(s,t)^2$ denotes the number of bits arrived to a a node from flow $i$ and $D_i(s,t)$ the output of flow $i$ from the node within the interval $(s,t]$. If we use $A_i(t)$ and $D_i(t)$ that will mean $A_i(0,t]$ and $D_i(0,t]$ respectively. If a node has $I$ inputs $A_i(t)$ and $D_i(t)$ we define:

$$
A(t) := \sum_{i=1}^I A_i(t) \text{ and } D(t) := \sum_{i=1}^I D_i(t).
$$

The backlog at time $t$ is given by $B(t) = A(t) - D(t)$ and the delay at time $t$ is given by $W(t) = \inf\{d \geq 0 : A(t-d) \leq D(t)\}$. In a network context we denote by $A^N(t)$ and $D^N(t)$ the arrivals and departures in node $N$. Subscripts and superscripts are dropped whenever possible to simplify the notation. Let $\int g(t) = \inf_{0 \leq s \leq t} \{f(t-s) + g(s)\}$ denote the min-plus convolution and $\int g(t) = \sup_{0 \leq u \leq t} \{f(t+u) - g(u)\}$ the min-plus deconvolution of functions $f$ and $g$ as it is defined in the min-plus algebra [15] [1]. We define the positive part operator as $(expr)^+ := \max[expr,0]$. For the theorems assume that $A_1, A_2, \ldots, A_I$ are independent and $A_i$ and $D_i$ are stationary and ergodic for all $i \in I$.

III. THEORETICAL BACKGROUND

Network calculus is a method to determine resource requirements of traffic flows by taking an envelope approach to describe arrivals and services in the network. One of the first applications of this type of analysis to computer networks was given in [17] and extensions can be found in [15] [1]. In the followings we recall the fundamental results.

A. Deterministic network calculus

In the deterministic network calculus the characteristics of the input sources are described in terms of arrival curves and the offered service from the nodes are given by the so called service curves. In the followings we recall the exact definitions of these notions from [1]:

**Definition 1 (Arrival curve [1]):** We say that a given arrival process $A(t)$ has $\alpha$ as an arrival curve if for all $t > s$:

$$
A(t) - A(s) \leq \alpha(t-s) \quad (1)
$$

**Definition 2 (Service curve [1]):** Consider a node $N$ and a flow through $N$ with input and output function $A(t)$ and $D(t)$. We say that $N$ offers to the flow a service curve $\beta$ if and only if

$$
D(t) \geq A \otimes \beta(t) \quad (2)
$$

The greatest advantage of the deterministic network calculus is the applicability of the per node results to the concatenation of several nodes. This happens through the definition of the network service curve which express the offered service from a network of nodes. For example in [13] we found assumptions that the inputs have stochastically bounded burstiness, in [11] the authors assume that the moment generating functions of the inputs are exponentially bounded. Probabilistic extensions of the network calculus are usually referred as statistical network calculus. Since our novel calculus relies on the min-plus algebra we mention here the results of the only statistical network calculus approach that is based on the min-plus algebra [10]. This calculus defines the effective envelope for the arrival processes.

**Definition 3 (Effective envelope [10]):** An effective envelope for an arrival process $A$ is a non-negative function $G$ such that for all $t$ and $\tau$:

$$
P\{A(t+\tau) - A(t) \leq G^*(\tau)\} > 1 - \epsilon \quad (3)
$$

To characterize the available service to a flow or to multiplexed flows the effective service curve is used which can be seen as a probabilistic measure of the available service.

**Definition 4 (Effective service curve [10]):** Given an arrival process $A$, an effective service curve is a non-negative function $S^e$ that satisfies for all $t \geq 0$:

$$
P\{D(t) \geq A \otimes S^e(t)\} \geq 1 - \epsilon \quad (4)
$$

2Without loss of generality we consider a bit-processing system, since it can be shown, that the result can be applied for systems with higher granularity (cells, packets).
We mention, that the statements for backlog delay etc. in [10] can expressed with a straightforward calculation from the defined effective envelopes and service curves, however quantifying packet loss with the existing probabilistic extensions of network calculus is a highly non-trivial problem even in an indirect way [2] [7] [8]. It is also worth mentioning, that the results based on effective envelopes and effective service curves rely on an accurate busy period analysis for estimating the appropriate time scale. In the next section we define a statistical network calculus, which is designed for direct packet loss calculations and which application does not require additional assumptions for the time scale.

IV. A NOVEL STATISTICAL NETWORK CALCULUS FOR WORKLOAD LOSS ESTIMATIONS

We can see in (3) and (4) that the definition of the effective envelope and the effective service curve happens by assigning some violation probability to the deterministic arrival and service curves (1) (2). As it was pointed out earlier this approach is favourable for overflow type quantities like buffer overflow probability was pointed out earlier this approach is favourable for deterministic arrival and service curves (1) (2). As it happens by assigning some violation probability to the effective envelope and the effective service curve

\[ \text{WLR} = \frac{E}\{Z\text{# of lost bits in a unit time interval}\} \leq \frac{E}\{(\mu_{\text{short}})^\ast\}, \]  (5)

where \( B \) represents the stationary backlog of the system with infinite buffer, \( q \) is the the buffer threshold and \( E[A] = E[A(0,1)] \) is the number of bits arriving in a unit time interval\(^3\). Based on (5) we assign \( Z^p \) and \( S^p \) functions to the input and the service respectively and we call them effective w-arrival curve and effective w-service curve hereafter.

**Definition 5 (Effective w-arrival curve):** We call \( Z^p \) the effective w-arrival curve of the flow with arrival process \( A \) if for all \( t \) and \( \tau \):

\[ E[(A(t+\tau) - A(t) - Z^p(\tau))^\ast] \leq \varphi \]  (6)

**Definition 6 (Effective w-service curve):** For an input with arrival process \( A \) a node offers an effective w-service curve \( S^p \) if for all \( t \geq 0 \):

\[ E[(A \otimes S^p(t) - D(t))^\ast] \leq \varphi_s \]  (7)

\(^3\)It is proven (e.g. in [6] and [18]) that the expected value of the number of lost bits in a finite buffer system, can be bounded from above by the number of packets overflowed in the system with infinite buffer.

We note that by letting \( \varphi \) and \( \varphi_s \) to zero the arrival and service curves of the deterministic network calculus can be recovered.

Within the framework of the following theorems we formalize stochastic bounds on some fundamental system characteristics like backlog, delay and output traffic envelope, with min-plus calculus operations on effective w-arrival curves and effective w-service curves. For the proofs the following lemma is needed about the positive part operator:

**Lemma 1:** For given \( X_1, X_2, X_3, X_4 \) random variables:

\[ E[(X_1-X_2\pm X_3\pm X_4)^\ast] \leq E[(X_1-X_2)^\ast]+E[(X_3-X_4)^\ast] \]

(8)

The proof of this lemma is left to the reader.

**Theorem 1 (Statement for the backlog):** \( Z^p \otimes S^p(0) \) is a probabilistic bound on the backlog, in the sense that, for all \( t \geq 0 \),

\[ E[(B(t) - Z^p \otimes S^p(0))^\ast] \leq \varphi + \varphi_s \]  (9)

**Proof:** It follows from the definition of the backlog that

\[ E[(B(t) - Z^p \otimes S^p(0))^\ast] = E[(A(t) - D(t) - Z^p \otimes S^p(0))^\ast] = E[(A(t) + A \otimes S^p(t) - D(t) - A \otimes S^p(t) - Z^p \otimes S^p(0))^\ast]. \]

If \( s^* \) attains the infimum in \( A \otimes S^p(t) \), then the whole expression is increased by the substitution of \( s^* \) into the min-plus deconvolution in \( Z^p \otimes S^p \), so we get that:

\[ E[(A(t) + A \otimes S^p(t) - D(t) - A \otimes S^p(t) - Z^p \otimes S^p(0))^\ast] \leq E[(A(t) + A \otimes S^p(t) - D(t) - A(t-s^*) - S^p(s^*) - Z^p(s^*) + S^p(s^*))^\ast]. \]

After simplification we obtain that:

\[ E[(A(t) - (A(t-s^*) - Z^p(s^*) + A \otimes S^p(t) - D(t))^\ast] \leq E[(A(t) - A(t-s^*) - Z^p(s^*))^\ast] + E[(A \otimes S^p(t) - D(t))^\ast]. \]

From the definition of the effective w-arrival curve and the effective w-service curve we recover that:

\[ E[(A(t) - A(t-s^*) - Z^p(s^*) + A \otimes S^p(t) - D(t))^\ast] \leq E[(A(t) - A(t-s^*) - Z^p(s^*))^\ast] + E[(A \otimes S^p(t) - D(t))^\ast] \leq \varphi + \varphi_s, \] which completes the proof.

The alert reader may notice that the left hand side of (9) express the expected value of the number of bits above a certain buffer level \( Z^p \otimes S^p(0) \) in an infinite buffer system. In other words if we imagine a buffered system with a buffer size \( Z^p \otimes S^p(0) \) the statement in (9) establishes an upper bound on the loss rate. Dividing this upper bound on the loss rate with the expected value of the bits arriving to the node gives an upper bound on the workload loss ratio.

**Theorem 2 (W-arrival curve for the output):** The function \( Z^p \otimes S^p \) is an effective w-arrival curve for
E[(D(t) - D(t))] ≤ Φ(τ)

Theorem 3 (Statement for the delay): If \( t \geq D(t) \) for all \( τ \) then:

\[
E[A(t) - D(t)] \leq Φ(τ) + Φ_s
\]

Proof: \( E[A(t) - D(t)] \leq E[A(t) - D(t)] + A \odot S(τ) - D(t)] \leq E[(A(t) - D(t)] \leq Φ(τ) + Φ_s
\)

Assuming that \( s^* \) attains the infimum in \( A \odot S(τ) \), then the whole expression is increased by the substitution of \( s^* \) into the first min-plus convolution:

\[
E[(A(t) - D(t)] \leq E[(A(t) - D(t)] + A \odot S(τ) - D(t)] \leq E[(A(t) - D(t)] \leq Φ(τ) + Φ_s
\]

From Lemma 1 it follows that:

\[
E[(A(t) - D(t)] \leq E[(A(t) - D(t)] - E[(A(t) - D(t)] \leq Φ(τ) + Φ_s
\]

It follows from the additional assumption of the theorem that:

\[
E[(A(t) - D(t)] \leq E[(A(t) - D(t)] - E[(A(t) - D(t)] \leq Φ(τ) + Φ_s
\]

One can notice that Theorem 3 establishes a bound on the expected value of the number of bits that suffers from a delay larger than \( d \). In order to establish end-to-end bounds from the single node results we are going to express the effective w-service curve of a network of nodes. In the following theorem the effective w-service curve of two concatenated nodes is given. Let \( S_N \) mean the effective w-arrival curve of input process \( A_t \) at node \( N \).

Theorem 4 (Concatenation of nodes): Assume that a flow traverses nodes \( N_1 \) and \( N_2 \) in sequence. If \( E[(A_{N_1} \odot S_N(t)) - A_{N_2}(t)] \leq Φ_1 \) and \( E[(A_{N_2} \odot S_N(t)) - D_N(t)] \leq Φ_2 \), then:

\[
E[(A_{N_1} \odot S_N(t)) - D_N(t)] \leq Φ_1 + Φ_2
\]

which means that, \( S_N \) is a stochastic w-service curve for the system which consists of the concatenation of these two nodes with \( Φ_1 + Φ_2 \) parameter.

Proof: \( E[(A_{N_1} \odot S_N(t)) - A_{N_2}(t)] \leq E[(A_{N_1} \odot S_N(t)) - D_N(t)] \]

From Lemma 1 it follows that:

\[
E[(A_{N_1} \odot S_N(t)) - A_{N_2}(t)] \leq E[(A_{N_1} \odot S_N(t)) - D_N(t)] \]

Using the definition of the min-plus convolution and the effective w-service curve we recover that:

\[
E[(A_{N_1} \odot S_N(t)) - A_{N_2}(t)] \leq E[(A_{N_1} \odot S_N(t)) - D_N(t)] \]

If \( s^* \) attains the infimum in the second int, then:

\[
E[(A_{N_1} \odot S_N(t)) - A_{N_2}(t)] \leq E[(A_{N_1} \odot S_N(t)) - D_N(t)] \]

The application of Theorem 4 iteratively to a network of nodes gives the following corollary.

Corollary 1 (Effective network w-service curve): If the service offered at each node \( h \) is independent and \( \eta \) on the path of a flow is given by an effective w-service curve \( S^\eta_{N_1} \), then an effective w-service curve \( S^\eta_{N_1} \) for the flow is given by:

\[
S^\eta_{N_1} = S^1 \odot S^2 \odot \ldots \odot S^\eta_H
\]

with a parameter:

\[
Φ_ω = \sum_{h=0}^H Φ_{sh}
\]

Using Corollary 1 we are able to draw up end-to-end workload loss ratio bounds according to Theorem 9. Finally we prove a statement about the relationship between effective envelopes and effective w-service curves.

Theorem 5: Assume that \( G^e \) is an effective envelope (3) for the arrivals \( A \) for input processes with maximal arrival (peak) rate \( \eta_0 \):

\[
E[(A(t) - A(t) - G^e(τ))] ≤ ε (\eta_0 - G^e(τ))
\]
Proof: Using a well-known computation of the expected value of a non-negative random variable:

\[ E[(A(t + \tau) - A(t) - G_r^x(\tau))^+] = \int_{x=0}^{\infty} P(A(t + \tau) - A(t) > G_r^x(\tau) + x) dx \]  

(16)

From the definition of the effective arrival curve and using that \( h \) is the maximal arrival rate we obtain:

\[ \int_{x=0}^{\infty} P(A(t + \tau) - A(t) > G_r^x(\tau) + x) dx \leq \int_{x=0}^{h \tau - G_r^x(\tau)} \varepsilon dx = \varepsilon (h \tau - G_r^x(\tau)) \]  

(17)

Q.E.D.

One can see, that the effective envelopes are functioning as effective w-arrival curves with a special parameter in the right side of (15), therefore Theorem 5 gives a framework for workload loss approximations with effective envelopes.

V. THE EFFECTIVE W-ARRIVAL CURVE AND THE EFFECTIVE BANDWIDTH

The theory of effective bandwidth [16] defines a framework for service provisioning, which describes the minimum bandwidth requirement of a traffic source in terms of the effective bandwidth, that is a probabilistic quantity between the average and peak rate of the input source. This concept provides a measure of resource usage which takes proper account of the varying statistical characteristics and QoS requirements of traffic sources. A widely referenced definition of effective bandwidth is the following.

Definition 7 (Effective bandwidth [16]): The effective bandwidth of the source with arrival process \( A(t) \) is defined as:

\[ \alpha_e(s, \tau) = \sup_{t \geq 0} \left\{ \frac{1}{st} \log E[e^{s(A(t + \tau) - A(t))}] \right\}, \]  

(18)

where \( 0 < s \) and \( \tau < \infty \).

The following theorem makes contact between the effective w-arrival curve and the effective bandwidth.

Theorem 6:

\[ Z_e^s(\tau) = \inf_{s > 0} \left\{ \tau \alpha_e(s, \tau) - \frac{\log(\varphi s)}{s} \right\} \]  

(19)

Proof:

\[ E[(A(t + \tau) - A(t) - Z_e^s(\tau))^+] \leq \frac{e^{s[-Z_e^s(\tau) + \tau \alpha_e(s, \tau)]} - s}{s} \]  

(20)

for all values of \( s \). Let \( \varphi \) defined as:

\[ \frac{e^{s[-Z_e^s(\tau) + \tau \alpha_e(s, \tau)]}}{s} := \varphi. \]  

(21)

For \( Z_e^s(\tau) \) we obtain:

\[ Z_e^s(\tau) = \tau \alpha_e(s, \tau) - \frac{\log(\varphi s)}{s}. \]  

(22)

By taking the infimum over \( s \) we obtain the smallest effective w-arrival curve:

\[ Z_e^s(\tau) = \inf_{s > 0} \left\{ \tau \alpha_e(s, \tau) - \frac{\log(\varphi s)}{s} \right\}. \]  

(23)

Since the effective bandwidth expressions of various traffic sources have been developed in the last decade the effective w-arrival curve for those sources can be calculated according to Theorem 6. For demonstration the effective w-arrival curve of multiplexed regulated input flows is shown in Figure 1. The w-arrival curve is normalized by the number of flows and the per flow deterministic arrival curve is also shown for easier interpretation of the figure. One can see that the effective w-arrival curve exploits a significant statistical multiplexing gain.

![Figure 1. The statistical multiplexing gain](image)

VI. DETERMINING EFFECTIVE W-ARRIVAL CURVES

Now we are going to illustrate how to derive effective w-arrival curves for inputs. For this purpose the w-arrival curves for regulated arrivals and on-off arrivals are investigated.

A. Regulated arrivals

We may imagine regulated inputs as any kind of inputs shaped by a general traffic shaper⁴ e.g a token bucket controller. This traffic shaper ensures, that the output flow has \( \alpha \) as an arrival curve (see Definition 1), where \( \alpha \) is a non-negative wide-sense increasing

⁴Traffic shaping is frequently used for example in QoS architectures e.g. in DiffServ [19].
function. Consider the aggregate traffic of $I$ inputs and assume, that these inputs are independent and for each $i \in I$, and for any $s, t \in \mathcal{R}$, $E[A_i(s, t)] \leq \alpha_i \cdot (t - s)$, where $\alpha_i = \lim_{t \to \infty} \alpha_i(t)/t < \infty$ is the average rate of the flow. The effective bandwidth of such an aggregate satisfies [16]:

$$\alpha(s, t) \leq \frac{1}{s} \sum_{i \in I} \log \left(1 + \frac{\alpha_i(t)}{\alpha_i(t)} \left(e^{(\alpha_i(t))} - 1 \right) \right). \quad (24)$$

Therefore in the meaning of Theorem 6 the effective w-arrival curve is given by:

$$Z^w(t) = \inf_{s > 0} \left\{ \frac{1}{s} \sum_{i \in I} \log \left(1 + \frac{\alpha_i(t)}{\alpha_i(t)} \left(e^{(\alpha_i(t))} - 1 \right) \right) - \frac{\log(\varphi_s)}{s} \right\}. \quad (25)$$

B. On-off sources

Since on-off sources play key role in traffic modelling and in deriving performance bounds, the effective w-arrival curve of the collection of $I$ on-off input is considered here. The on-off sources have two states, in the so called "off" state there is no traffic produced, and in the "on" state the source generates traffic at some peak rate $p$. Let denote the average traffic rate by $\rho$. The effective bandwidth of the aggregate of $I$ on-off sources with $p_i$ peak rates and $\rho_i$ average rates is given by [16]:

$$\alpha(s, t) \leq \frac{1}{s} \sum_{i \in I} \log \left(1 + \frac{\rho_i}{\rho_i} \left(e^{(\rho_i s)} - 1 \right) \right). \quad (26)$$

Using Theorem 6 we get:

$$Z^w(t) = \inf_{s > 0} \left\{ \frac{1}{s} \sum_{i \in I} \log \left(1 + \frac{\rho_i}{\rho_i} \left(e^{(\rho_i s)} - 1 \right) \right) - \frac{\log(\varphi_s)}{s} \right\}. \quad (27)$$

for the effective w-arrival curve.

VII. DETERMINING EFFECTIVE W-SERVICE CURVES FOR SCHEDULERS

In the followings the effective w-service curve is deduced for static priority and GPS (General Processor Sharing) schedulers. For the proof of the results the following lemma is needed.

Lemma 2: For given $X_1, X_2, X_3, X_4$ random variables:

$$E[(X_1-X_2)+(X_3-X_4)] \leq E[(X_1-X_2)+(X_3-X_4)]_+ \quad (28)$$

The proof of this lemma is left to the reader.

A. Static priority scheduler

Consider a static priority scheduler with $I$ classes. Let $i = 1, \ldots, I$ denote the priority of the classes, where the lower number represents higher priority.

Theorem 7 (Static priority): The curve

$$S^w_i = [\beta(t) - Z^p_{j < i}(t)]^+ \quad (29)$$

is an effective w-service curve for the input of class $i$ with $\varphi_s = \varphi$, where $\beta(t)$ is the deterministic service curve offered by the scheduler for the aggregate inputs from all classes, and $Z^p_{j < i}(t)$ is the effective w-arrival curve of the aggregate traffic from classes $j < i$.

Proof:

Fix $t \geq 0$ and let $t_i = \max\{s \leq t : B_{j \leq i}(s) = 0\}$, where $B_{j \leq i}(s)$ means the backlog from classes $j < i$. The time instant $t_i$ is considered as the beginning of the busy period that contains $t$ from the point of flow $i$. Let $A_{j < i}$ and $D_{j < i}$ denote the arrivals and departures from classes $j < i$ respectively. If $B(t_i) > 0$ from the characteristics of the static priority scheduler we obtain that:

$$D_i(t) = D_i(t_i) + (D_{j \leq i}(t_i) - D_{j < i}(t_i)) - (D_{j > i}(t_i) - D_{j > i}(t)) \quad (30)$$

where we exploited that $D_{j_i}(t) = A_{j_i}(t_i)$ for all $j \leq i$, that $D(t) - D(t_i) \geq \beta(t - t_i)$ if the scheduler is work-conserving, and that $D_j(t) \leq A_j(t)$ for all $j$. From the definition of the effective w-service curve we obtain:

$$E[(A_i \otimes S^p_i(t) - D_i(t))]^{+} \leq E[(A_i(t_i) + S^p_i(t - t_i) - D_i(t))]^{+} = E[(A_i(t_i) + [\beta(t - t_i) - Z^p_{j < i}(t - t_i)]^{+} - D_i(t))]^{+}. \quad (31)$$

By applying (31) we recover that:

$$E[(A_i(t_i) + [\beta(t - t_i) - Z^p_{j < i}(t - t_i)]^{+} - D_i(t))]^{+} \leq E[(A_i(t_i) + [\beta(t - t_i) - Z^p_{j < i}(t - t_i)]^{+} - A_i(t_i)) - [\beta(t - t_i) - (A_{j < i}(t) - D_{j < i}(t))]^{+}]. \quad (32)$$

With the usage of Lemma 2 and doing some simplification:

$$E[(A_i(t_i) + [\beta(t - t_i) - Z^p_{j < i}(t - t_i)]^{+} - A_i(t_i)) - [\beta(t - t_i) - (A_{j < i}(t) - D_{j < i}(t))]^{+}] \leq E[(A_{j < i}(t) - A_{j < i}(t_i))^{+}] \leq \varphi$$

The last step follows from the definition of the effective w-arrival curve.

B. GPS scheduler

Let $Z^w_i$ be an effective w-arrival curve of the input from class $i$, $\psi_i$ be the weight of input $i$, and assume that the $Z^w_i$ functions are concave.

Theorem 8 (GPS): The curve

$$S^w_i = \lambda_i(Ct + \sum_{i \neq j} [A_j Ct - Z^p_{j < i}(t)])^{+} \quad (32)$$

where $\lambda_i$ is the priority of the classes, where the lower number represents higher priority.
is an effective w-service curve for the input of class i with \( \varphi_s = \lambda_i I_p \), where C is the capacity of the scheduler, and \( \lambda_i = \psi_i / \sum \psi_j \) is the guaranteed rate for class i.

**Proof:**

Fix \( t \geq 0 \) and let \( t_i = \max \{ s \leq t : B_i(s) = 0 \} \) and \( t_{ij} = \max \{ s \leq t_i, i \neq j : B_j(s) = 0 \} \). Here \( t_i \) can be considered as the latest backlog clearing time point in the system before \( t_i \). From the proof of Lemma 3 in [10]:

\[
D_i(t) \geq A_i(t_i) + \lambda_i(C(t - t_i) + \sum_{j \neq i} \lambda_j C(t - t_{ij}) - (A_j(t) - A_j(t_{ij})))^+.
\]

(33)

From the definition of the effective w-service curve we obtain:

\[
E[(A_i(t_i) + \lambda_i(C(t - t_i) + \sum_{j \neq i} \lambda_j C(t - t_{ij}) - Z_i^p(t - t_i))^+) - D_i(t)]^+ = E[(A_i(t_i) + \lambda_i(C(t - t_i) + \sum_{j \neq i} \lambda_j C(t - t_{ij}) - (A_j(t) - A_j(t_{ij})))^+)]
\]

By applying (33) we recover that:

\[
E[(A_i(t_i) + \lambda_i(C(t - t_i) + \sum_{j \neq i} \lambda_j C(t - t_{ij}) - Z_i^p(t - t_i))^+) - D_i(t)]^+ = E[(A_i(t_i) + \lambda_i(C(t - t_i) + \sum_{j \neq i} \lambda_j C(t - t_{ij}) - (A_j(t) - A_j(t_{ij})))^+)]
\]

(34)

The calculation of the workload loss ratio happens according to Theorem 1.

For simulation purposes we made an implementation of the evaluation scenario under the NS2 network simulator [20]. We used random packet generators as inputs, which send packet to the server through a token bucket traffic regulator. For the token bucket regulator we used the Differentiated Services module of the NS2 and set the bucket size and the token generating rate according to the values of the input scenario. The server was a non-preemptive constant rate server with the appropriate service rate. Besides the 100 inputs we set up another packet generator, which sends lower priority packets to the server with the same packet size. This way we ensured the given rate-latency service curve for the input flows among realistic conditions, since there is no service for the higher priority packets, while the server finishes the inchoate. The interesting case from the point of the packet loss is when the inputs exploit the entire input profile, so we set up the packet generators to generate different traffic bursts of alternating sizes with exponentially distributed random inter arrival times. We also controlled the average rate of the generators in order to meet the maximum input rate requirement. We run the simulation ten times for some queue sizes and took the average of the results. Figure 2 show the results of the bounds and the simulation.

We can observe that the novel bound provides a significant improvement of the existing closest result. Comparing with the simulation we state that within the range of interest \((10^{-3} - 10^{-6})\) the result of Theorem

VIII. NUMERICAL RESULTS

In this section we investigate the novel workload loss ratio bound deduced from our novel statistical calculus and compare it with the best existing deterministic calculus based probabilistic bound [7] and also with simulation results under NS2. For analysis the following scenario is used. We have 100 input flows, which are token bucket constrained with some deterministic arrival curve \( \alpha_i(t) = \sigma_i t + \bar{\sigma}_i (\bar{\sigma}_{1,50} = 133.3, \sigma_{1,50} = 8, \bar{\sigma}_{1,100} = 66.6, \sigma_{1,100} = 5) \), and the packet forwarder satisfies a rate latency service curve property, with \( \beta(t) = 12500 \cdot \max(t - 8 \cdot 10^{-5}, 0) \), in a work-conserving manner. The sustainable rate of the inputs and the size of the bucket is given in packets and the service rate is given in packets during a second (pps). These parameter values are close to many practical, common applications.

Based on the effective bandwidth for regulated inputs in [16] we use the following formula for the calculation of the effective w-arrival curves in accordance with equation (25):

\[
Z^w_S(t) = \inf \left\{ \sum_{i \in I} \frac{1}{s_i} \log \left( 1 + \frac{\alpha_i(t)}{\lambda_i} \left( e^{(\sigma_i(t))} - 1 \right) \right) - \frac{\log(\varphi_S)}{s} \right\}.
\]

The proper comparison of the performance of the arrival and w-arrival curves the same deterministic service curve is used for the server.
I gives a considerably well bound on the workload loss ratio.

IX. CONCLUSIONS

In the focus of this paper was to establish a novel probabilistic calculus for packet networks which is designed for direct workload loss ratio approximations. We introduced the effective w-arrival curve and the effective w-service curve and proved fundamental statements about the backlog, delay and output traffic envelope. We also showed that the per-node results can be carried over to a network of nodes with the definition of the effective network w-service curve and a performance evaluation was given on the workload loss ratio bound that follows from our new theory.

REFERENCES