Workload Loss Examinations with a Novel Probabilistic Extension of Network Calculus

András Gulyás and József Bíró

Budapest University of Technology and Economics,
Department of Telecommunications and Media Informatics,
1117 Budapest, Magyar tudósok körútja 2., Hungary
{gulyas, biro}@tmit.bme.hu

Abstract. The estimation of the expected traffic loss ratio (workload loss ratio, WLR) is a key issue in provisioning Quality of Service in packet based communication networks. Despite of its importance, the stationary (long run) loss ratio in queueing analysis is usually estimated through other assessable quantities, typically based on the approximates of the buffer overflow probability. 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.

Keywords: Network calculus, resource estimation, statistical multiplexing.

1 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.

[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 $\frac{WLR}{Pr(Q>q)}$ can be arbitrary under certain circumstances [6].

F. Boavida et al. (Eds.): NETWORKING 2006, LNCS 3976, pp. 533–544, 2006.

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 3 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 4. The connection between the effective w-arrival curve and effective bandwidth [16], is pointed out in section 5. In section 6 we compare the derived workload loss bound with the closest existing probabilistic direct bound [7] and some simulation results.

2 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 $A_I(t) := \sum_{i=1}^{I} A_i(t)$, and $D_I(t) := \sum_{i=1}^{I} D_i(t)$. 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 $f \otimes g(t) = \inf_{0 \leq s \leq t}\{f(t-s) + g(s)\}$ denote the min-plus convolution and $f \oslash g(t) = \sup_{0 \leq u}\{f(t+u) - g(u)\}$ 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 $A_1, A_2, ..., A_I$ are independent and A_i and D_i are stationary and ergodic.

[2] Without 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).

3 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.

3.1 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 α as an arrival curve if for all $t > s$:*

$$A(t) - A(s) \leq \alpha(t - s) \tag{1}$$

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 β if and only if*

$$D(t) \geq A \otimes \beta(t). \tag{2}$$

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. If the hth node within the route ($h = 1, 2, ..., H$) of nodes offers to a flow a service curve β_h, then the network service curve can be expressed as $\beta_{net} = \beta_1 \otimes \beta_2 \otimes ... \otimes \beta_H$.

However the deterministic network calculus is a powerful and expressive tool for describing the properties of communication networks, its worst-case system view cannot take the effects of the statistical multiplexing into consideration. This fact usually leads to the overestimation of the resource requirements of multiplexed traffic sources.

3.2 Probabilistic Extensions of the Deterministic Network Calculus

In order to benefit from the statistical multiplexing several probabilistic extensions of the deterministic network calculus have been elaborated in the past few years. The common property of these studies, that they assign a bound on the violation probability that the incoming traffic exceeds its statistical envelope. 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 recall 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 τ:*

$$P\{A(t+\tau) - A(t) \leq G^{\varepsilon}(\tau)\} > 1 - \varepsilon \tag{3}$$

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^{ε} that satisfies for all $t \geq 0$:*

$$P\{D(t) \geq A \otimes S^{\varepsilon}(t)\} \geq 1 - \varepsilon \tag{4}$$

The following theorems recall the statistical bounds for the delay, the output envelope and the backlog using the terminology of the min-plus algebra on effective envelopes and effective service curves. As we referred earlier, in order to derive such results appropriate time scale limit assumptions are needed, it is assumed, that the node offers a service curve S^{ε_s} which satisfies the additional requirement that there exists a time scale T such that for all $t \geq 0$:

$$P\left\{D(t) \geq \inf_{\tau \leq T}\{A(t-\tau) + S^{\varepsilon_s}(\tau)\}\right\} \geq 1 - \varepsilon_s \tag{5}$$

For all theorems we assume that G^{ε} is an effective envelope for the arrivals A to a node and we have a $T < \infty$ in (5). Define $\varepsilon_{\omega} := \varepsilon_s + T\varepsilon$.

Theorem 1 (Output traffic envelope [10]). *The function $G^{\varepsilon} \oslash S^{\varepsilon_s}$ is an effective envelope for the output traffic.*

Theorem 2 (Backlog bound [10]). *$G^{\varepsilon} \oslash S^{\varepsilon_s}(0)$ is a probabilistic bound on the backlog, in the sense that, for all $t \geq 0$,*

$$P\{B(t) \leq G^{\varepsilon} \oslash S^{\varepsilon_s}(0)\} \geq 1 - \varepsilon_{\omega} \tag{6}$$

Theorem 3 (Delay bound [10]). *If $d \geq 0$ satisfies that $\sup_{\tau \leq T}\{G^{\varepsilon}(\tau - d) - S^{\varepsilon_s}(\tau)\} \leq 0$, then d is a probabilistic delay bound in the sense that, for all $t \geq 0$:*

$$P\{W(t) \leq d\} \geq 1 - \varepsilon_{\omega} \tag{7}$$

Similar to the deterministic calculus the effective service curve of a network can be expressed as the convolution of the service at each node. Consider a network of nodes where the hth node offers an effective service curve $S_h^{\varepsilon_s}$ to a flow. It is assumed that:

$$P\left\{D^h(t) \geq \inf_{\tau \leq T_h}\{A^h(t-\tau) + S_h^{\varepsilon_s}(\tau)\}\right\} \geq 1 - \varepsilon_s \tag{8}$$

Theorem 4 (Effective network service curve [10]). *If the service offered at each node $h = 1, ..., H$ on the path of a flow is given by a service curve $S_h^{\varepsilon_s}$, then an effective network service curve $S_{net}^{\varepsilon_\omega}$ for the flow is given by $S_{net}^{\varepsilon_\omega} = S_1^{\varepsilon_s} \otimes S_2^{\varepsilon_s} \otimes ... \otimes S_H^{\varepsilon_s}$ with a violation probability bounded above by $\varepsilon_\omega = \varepsilon_s \sum_{h=1}^{H} \left(1 + (h-1)T^h\right)$.*

We can see, that these statements for backlog delay etc. are 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]. One can also observe, that these statements above rely on an accurate busy period analysis for estimating the appropriate time scale and require that the infimum in (5) and (8) is taken within a finite interval. In the next section we define a statistical network calculus, which is designed for direct packet loss calculations and which application does not require such assumptions for the time scale.

4 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 however quantifying packet loss in a direct way turns out to be non-trivial.

In the followings a novel calculus is defined which is suitable for loss examinations. We set out from the definition of the workload loss ratio which looks like this for stationary and ergodic systems:

$$WLR = \frac{E[\#\text{ of lost bits in a unit time interval}]}{E[\#\text{ of bits arriving in a unit time interval}]} \leq \frac{E[(B-q)^+]}{E[A]} \,. \quad (9)$$

where B represents the stationary backlog of the system with infinite buffer, q is the buffer threshold and $E[A] = E[A(0,1)]$ is the number of bits arriving in a unit time interval[3]. Based on (9) we assign Z^φ and S^{φ_s} 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^φ the effective w-arrival curve of the flow with arrival process A if for all t and τ:*

$$E[(A(t+\tau) - A(t) - Z^\varphi(\tau))^+] \leq \varphi \quad (10)$$

[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 overflown in the system with infinite buffer.

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

$$E[(A \otimes S^{\varphi_s}(t) - D(t))^+] \leq \varphi_s \tag{11}$$

We note that by letting φ and φ_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:*

$$E[(X_1 - X_2 + X_3 - X_4)^+] \leq E[(X_1 - X_2)^+] + E[(X_3 - X_4)^+] \tag{12}$$

The proof of this lemma is left to the reader.

Theorem 5 (Statement for the backlog). *$Z^{\varphi} \oslash S^{\varphi_s}(0)$ is a probabilistic bound on the backlog, in the sense that, for all $t \geq 0$,*

$$E[(B(t) - Z^{\varphi} \oslash S^{\varphi_s}(0))^+] \leq \varphi + \varphi_s \tag{13}$$

Proof. It follows from the definition of the backlog that
$E[(B(t) - Z^{\varphi} \oslash S^{\varphi_s}(0))^+] = E[(A(t) - D(t) - Z^{\varphi} \oslash S^{\varphi_s}(0))^+] = E[(A(t) + A \otimes S^{\varphi_s}(t) - D(t) - A \otimes S^{\varphi_s}(t) - Z^{\varphi} \oslash S^{\varphi_s}(0))^+]$.
For any choice of an arbitrarily small $\delta > 0$, there exists a finite s^* such that $A(t - s^*) + S^{\varphi_s}(s^*) < A \otimes S^{\varphi_s}(t) + \delta$ and the whole expression is increased by the substitution of this s^* into the min-plus deconvolution in $Z^{\varphi} \oslash S^{\varphi_s}$, so we get that:
$E[(A(t) + A \otimes S^{\varphi_s}(t) - D(t) - A \otimes S^{\varphi_s}(t) - Z^{\varphi} \oslash S^{\varphi_s}(0))^+] \leq E[(A(t) + A \otimes S^{\varphi_s}(t) - D(t) - A(t - s^*) - S^{\varphi_s}(s^*) + \delta - Z^{\varphi}(s^*) + S^{\varphi_s}(s^*))^+]$.
After simplification we obtain that:
$E[(A(t) - A(t - s^*) + \delta - Z^{\varphi}(s^*) + A \otimes S^{\varphi_s}(t) - D(t))^+]$.
By using Lemma 1 twice we get:
$E[(A(t) - A(t - s^*) + \delta - Z^{\varphi}(s^*) + A \otimes S^{\varphi_s}(t) - D(t))^+] \leq E[(A(t) - A(t - s^*) - Z^{\varphi}(s^*))^+] + E[(A \otimes S^{\varphi_s}(t) - D(t))^+] + E[\delta^+]$.
From the definition of the effective w-arrival curve and the effective w-service curve we recover that:
$E[(A(t) - A(t - s^*) - Z^{\varphi}(s^*))^+] + E[(A \otimes S^{\varphi_s}(t) - D(t))^+] + E[\delta^+] \leq \varphi + \varphi_s + \delta$.
Since δ can be arbitrarily small, letting it shrink to 0 recovers the desired result, which completes the proof. Q.E.D.

The alert reader may notice that the left hand side of (13) express the expected value of the number of bits above a certain buffer level $Z^{\varphi} \oslash S^{\varphi_s}(0)$ in an infinite buffer system. In other words if we imagine a buffered system with a buffer size $Z^{\varphi} \oslash S^{\varphi_s}(0)$ the statement in (13) establishes an upper bound on the loss rate. Dividing this upper bound of the loss rate with the expected value of the bits arriving to the node gives an upper bound on the workload loss ratio.

Theorem 6 (W-arrival curve for the output). *The function $Z^{\varphi} \oslash S^{\varphi_s}$ is an effective w-arrival curve for the output traffic from the node in the sense that for all t and τ:*

$$E[(D(t+\tau) - D(t) - Z^{\varphi} \oslash S^{\varphi_s}(\tau))^+] \leq \varphi + \varphi_s \tag{14}$$

Proof. $E[(D(t+\tau) - D(t) - Z^{\varphi} \oslash S^{\varphi_s}(\tau))^+] = E[(D(t+\tau) + A \otimes S^{\varphi_s}(t) - D(t) - A \otimes S^{\varphi_s}(t) - Z^{\varphi} \oslash S^{\varphi_s}(\tau))^+]$.
Using Lemma 1 and the fact that $A(t+\tau) \geq D(t+\tau)$ we obtain that:
$E[(D(t+\tau) + A \otimes S^{\varphi_s}(t) - D(t) - A \otimes S^{\varphi_s}(t) - Z^{\varphi} \oslash S^{\varphi_s}(\tau))^+] \leq E[(A(t+\tau) - A \otimes S^{\varphi_s}(t) - Z^{\varphi} \oslash S^{\varphi_s}(\tau))^+] + E[(A \otimes S^{\varphi_s}(t) - D(t))^+]$.
For any choice of an arbitrarily small $\delta > 0$, there exists a finite s^* such that $A(t-s^*) + S^{\varphi_s}(s^*) < A \otimes S^{\varphi_s}(t) + \delta$ and the whole expression is increased by the substitution of this s^* into the min-plus deconvolution in $Z^{\varphi} \oslash S^{\varphi_s}$, so we obtain:
$E[(A(t+\tau) - A \otimes S^{\varphi_s}(t) - Z^{\varphi} \oslash S^{\varphi_s}(\tau))^+] + E[(A \otimes S^{\varphi_s}(t) - D(t))^+] \leq E[(A(t+\tau) - A(t-s^*) - S^{\varphi_s}(s^*) + \delta - Z^{\varphi}(\tau + s^*) + S^{\varphi_s}(s^*))^+] + E[(A \otimes S^{\varphi_s}(t) - D(t))^+]$.
After some simplification and applying Lemma 1 we get:
$E[(A(t+\tau) - A(t-s^*) - Z^{\varphi}(\tau + s^*))^+] + E[(A \otimes S^{\varphi_s}(t) - D(t))^+] + E[\delta^+] \leq \varphi + \varphi_s + \delta$. The last step follows from the definition of the effective w-arrival curve and the effective w-service curve. Since δ can be arbitrarily small, letting it shrink to 0 recovers the desired result, which completes the proof. Q.E.D.

Theorem 7 (Statement for the delay). *If $d : Z^{\varphi}(\tau - d) \leq S^{\varphi_s}(\tau)$ for all τ then:*

$$E[A(t-d) - D(t)] \leq \varphi + \varphi_s \tag{15}$$

Proof. $E[A(t-d) - D(t)] \leq E[A(t-d) - A \otimes S^{\varphi_s}(t) + A \otimes S^{\varphi_s}(t) - D(t)] \leq E[(A(t-d) - A \otimes S^{\varphi_s}(t) + A \otimes S^{\varphi_s}(t) - D(t))^+]$.
For any choice of an arbitrarily small $\delta > 0$, there exists a finite s^* such that $A(t-s^*) + S^{\varphi_s}(s^*) < A \otimes S^{\varphi_s}(t) + \delta$ and the whole expression is increased by the substitution of this s^* into the first min-plus convolution:
$E[(A(t-d) - A \otimes S^{\varphi_s}(t) + A \otimes S^{\varphi_s}(t) - D(t))^+] \leq E[(A(t-d) - A(t-s^*) - S^{\varphi_s}(s^*) + \delta + A \otimes S^{\varphi_s}(t) - D(t))^+]$
From Lemma 1 it follows that:
$E[(A(t-d) - A(t-s^*) - S^{\varphi_s}(s^*) + \delta + A \otimes S^{\varphi_s}(t) - D(t))^+] \leq E[(A(t-d) - A(t-s^*) - S^{\varphi_s}(s^*))^+] + E[(A \otimes S^{\varphi_s}(t) - D(t))^+] + E[\delta^+]$.
It follows from the additional assumption of the theorem that:
$E[(A(t-d) - A(t-s^*) - S^{\varphi_s}(s^*))^+] + E[(A \otimes S^{\varphi_s}(t) - D(t))^+] + E[\delta^+] \leq E[(A(t-d) - A(t-s^*) - Z^{\varphi}(s^*-d))^+] + E[(A \otimes S^{\varphi_s}(t) - D(t))^+] + E[\delta^+] \leq \varphi + \varphi_s + \delta$
The last step follows from the definition of the effective w-arrival curve and the effective w-service curve. Since δ can be arbitrarily small, letting it shrink to 0 recovers the desired result, which completes the proof. Q.E.D.

One can notice that Theorem 7 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 $S_N^{\varphi_i}$ mean the effective w-arrival curve of input process A_i at node N.

Theorem 8 (Concatenation of nodes). *Assume that a flow traverses nodes N_1 and N_2 in sequence. If $E[(A^{N1} \otimes S_{N1}^{\varphi_1}(t) - A^{N2}(t))^+] \leq \varphi_1$ and $E[(A^{N2} \otimes S_{N2}^{\varphi_2}(t) - D^{N2}(t))^+] \leq \varphi_2$, then*

$$E[(A^{N1} \otimes S_{N1}^{\varphi_1} \otimes S_{N2}^{\varphi_2}(t) - D^{N2}(t))^+] \leq \varphi_1 + \varphi_2 \tag{16}$$

which means, that $S_{N1}^{\varphi_1} \otimes S_{N2}^{\varphi_2}$ is a stochastic w-service curve for the system which consists of the concatenation of these two nodes with $\varphi_1 + \varphi_2$ parameter.

Proof. $E[(A^{N1} \otimes S_{N1}^{\varphi_1} \otimes S_{N2}^{\varphi_2}(t) - D^{N2}(t))^+] = E[(A^{N1} \otimes S_{N1}^{\varphi_1} \otimes S_{N2}^{\varphi_2}(t) - A^{N2} \otimes S_{N2}^{\varphi_2}(t) + A^{N2} \otimes S_{N2}^{\varphi_2}(t) - D^{N2}(t))^+]$.
From Lemma 1 it follows that:
$E[(A^{N1} \otimes S_{N1}^{\varphi_1} \otimes S_{N2}^{\varphi_2}(t) - A^{N2} \otimes S_{N2}^{\varphi_2}(t) + A^{N2} \otimes S_{N2}^{\varphi_2}(t) - D^{N2}(t))^+] \leq E[(A^{N1} \otimes S_{N1}^{\varphi_1} \otimes S_{N2}^{\varphi_2}(t) - A^{N2} \otimes S_{N2}^{\varphi_2}(t))^+] + E[(A^{N2} \otimes S_{N2}^{\varphi_2}(t) - D^{N2}(t))^+]$.
Using the definition on the min-plus convolution and the effective w-service curve we recover that :
$E[(A^{N1} \otimes S_{N1}^{\varphi_1} \otimes S_{N2}^{\varphi_2}(t) - A^{N2} \otimes S_{N2}^{\varphi_2}(t))^+] + E[(A^{N2} \otimes S_{N2}^{\varphi_2}(t) - D^{N2}(t))^+] \leq E[(\inf_{0 \leq s \leq t}\{\inf_{0 \leq u \leq t-s}\{A^{N1}(t-s-u) + S_{N1}^{\varphi_1}(u)\} + S_{N2}^{\varphi_2}(s)\} - \inf_{0 \leq s \leq t}\{A^{N2}(t-s) + S_{N2}^{\varphi_2}(s)\})^+] + \varphi_2$.

For any choice of an arbitrarily small $\delta > 0$, there exists a finite s^* such that $A^{N2}(t-s^*) + S_{N2}^{\varphi_2}(s^*) < \inf_{0 \leq s \leq t}\{A^{N2}(t-s) + S_{N2}^{\varphi_2}(s)\} + \delta$ we get:
$E[(\inf_{0 \leq s \leq t}\{\inf_{0 \leq u \leq t-s}\{A^{N1}(t-s-u) + S_{N1}^{\varphi_1}(u)\} + S_{N2}^{\varphi_2}(s)\} - \inf_{0 \leq s \leq t}\{A^{N2}(t-s) + S_{N2}^{\varphi_2}(s)\})^+] + \varphi_2 \leq E[(\inf_{0 \leq u \leq t-s^*}\{A^{N1}(t-s^*-u) + S_{N1}^{\varphi_1}(u)\} + S_{N2}^{\varphi_2}(s^*) - A^{N2}(t-s^*) - S_{N2}^{\varphi_2}(s^*) + \delta)^+] + \varphi_2 = E[(A^{N1} \otimes S_{N1}^{\varphi_1}(t-s^*) - A^{N2}(t-s^*) + \delta)^+] + \varphi_2$.
Applying Lemma 1 and using the definition of the effective w-service curve we get:
$E[(A^{N1} \otimes S_{N1}^{\varphi_1}(t-s^*) - A^{N2}(t-s^*) + \delta)^+] + \varphi_2 \leq \varphi_1 + \varphi_2 + \delta$.
Since δ can be arbitrarily small, letting it shrink to 0 recovers the desired result, which completes the proof. Q.E.D.

The application of Theorem 8 iteratively to a network of nodes the gives the following corollary.

Corollary 1 (Effective network w-service curve). *If the service offered at each node $h = 1, ..., H$ on the path of a flow is given by an effective w-service curve $S_h^{\varphi_{sh}}$, then an effective network w-service curve $S_{net}^{\varphi_\omega}$ for the flow is given by:*

$$S_{net}^{\varphi_\omega} = S_1^{\varphi_{s1}} \otimes S_2^{\varphi_{s2}} \otimes ... \otimes S_H^{\varphi_{sH}} \tag{17}$$

with a parameter:

$$\varphi_\omega = \sum_{h=1}^{H} \varphi_{sh} \tag{18}$$

Using corollary 1 we are able to draw up end-to-end workload loss ratio bounds according to Theorem 13.

5 The Effective w-Arrival Curve and the Effective Bandwidth

The theory of effective bandwidth [16] defines a framework for service provisioning, that describes the minimum bandwidth requirement of a traffic source in terms of the effective bandwidth, which 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:*

$$\alpha_e(s,\tau) = \sup_{t\geq 0}\left\{\frac{1}{st}\log E[e^{s(A(t+\tau)-A(t))}]\right\}, 0 < s,\tau < \infty. \tag{19}$$

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

Theorem 9

$$Z^{\varphi}(\tau) = \inf_{s>0}\{\tau\alpha_e(s,\tau) - \frac{\log(\varphi s)}{s}\} \tag{20}$$

Proof.

$$E[(A(t+\tau) - A(t) - Z^{\varphi}(\tau))^{+}] \leq \frac{e^{s(-Z^{\varphi}(\tau)+\tau\alpha_e(s,\tau))}}{s} \tag{21}$$

for all values of s. Let φ defined as:

$$\frac{e^{s(-Z^{\varphi}(\tau)+\tau\alpha_e(s,\tau))}}{s} := \varphi. \tag{22}$$

For $Z^{\varphi}(\tau)$ we obtain:

$$Z^{\varphi}(\tau) = \tau\alpha_e(s,\tau) - \frac{\log(\varphi s)}{s}. \tag{23}$$

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

$$Z^{\varphi}(\tau) = \inf_{s>0}\left\{\tau\alpha_e(s,\tau) - \frac{\log(\varphi s)}{s}\right\}. \tag{24}$$

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 9. 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.

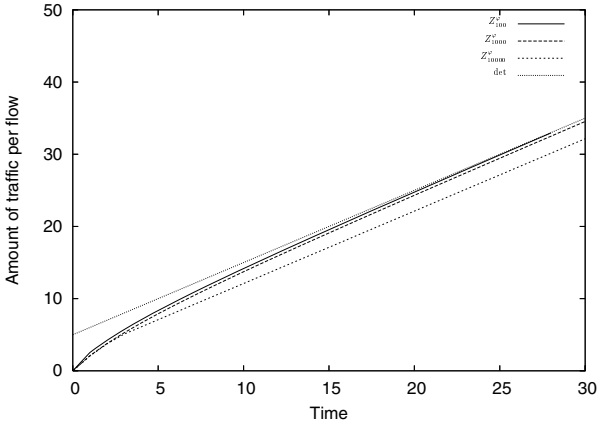

Fig. 1. The statistical multiplexing gain

6 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 $\alpha_i(t) = \bar{\alpha}_i t + \sigma_i$ ($\bar{\alpha}_{1..50} = 133.3$, $\sigma_{1..50} = 8$, $\bar{\alpha}_{51..100} = 66.6$, $\sigma_{51..100} = 5$), 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[4]. 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 (24):

$$Z^{\varphi}(t) = \inf_{s>0} \left\{ \sum_{i \in \mathcal{I}} \frac{1}{s} \log\left(1 + \frac{\bar{\alpha}_i t}{\alpha_i(t)} \left(e^{(s\alpha_i(t))} - 1\right)\right) - \frac{\log(\varphi s)}{s} \right\}. \tag{25}$$

The calculation of the workload loss ratio happens according to Theorem 5.

For simulation purposes we made an implementation of the evaluation scenario under the NS2 network simulator [19]. 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

[4] For the proper comparison of the performance of the arrival and w-arrival curves the same deterministic service curve is used for the server.

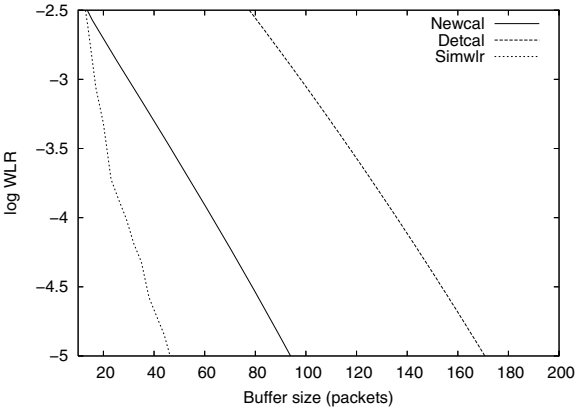

Fig. 2. The comparison of the bounds and the simulation results

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 5 gives a considerably well bound on the workload loss ratio.

7 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. Besides these fundamental results our novel calculus raises

a lot of questions that have to be answered. The determination of the effective w-service curve for various packet schedulers is a possible topic of further research.

References

1. Jean-Yves Le Boudec and Patrick Thiran. *Network Calculus: A theory of deterministic queuing systems for the Internet*. Springer, 2002.
2. Milan Vojnovic and Jean-Yves Le Boudec. Stochastic analysis of some expedited forwarding networks. *IEEE INFOCOM New York*, June 2002.
3. M. Vojnovic and J. Y. Le Boudec. Stochastic analysis of some expedited forwarding networks. Technical Report DSC/2001/039, EPFL-DI-ICA, July 2001.
4. N. G. Duffield, J. T. Lewis, N. O'Connel, R. Russel, and F. Foomey. Entropy of atm traffic streams: tool for estimating qos parameters. *IEEE Journal of Selected Areas in Communications vol. 13*, March 1995.
5. M. Krunz and A. M. Ramasamy. The correlation structure for a class of scene-based video models and its impact on the dimensioning of video buffers. *IEEE Trans. Multimedia vol. 2*, July 2000.
6. A. György and T. Borsos. Estimates on the packet loss ratio via queue tail probabilities. *IEEE Globecom*, March 2001.
7. András Gulyás, J. Bíró, and Z. Heszberger. A novel direct upper approximation for workload loss ratio in general buffered systems. In *IFIP Networking 2005*, page 718, Waterloo, Canada, May 2005.
8. András Gulyás and J. Bíró. Direct and indirect methods for packet loss estimation in buffered systems. In *EuroNGI 2005*, Rome, Italy, April 2005.
9. A. Burchard R. R. Boorstyn, J. Liebeherr, and C. Oottamakorn. Statistical service assurances for traffic scheduling algorithms. *IEEE Journal on Selected Areas in Communications*, December 2000.
10. A. Burchard, J. Liebeherr, and S. D. Patek. A calculus for end-to-end statistical service guarantees. Technical Report CS-2001-19, University of Virginia, May 2002.
11. C. S. Chang. On deterministic traffic regulation and service guarantees: A systematic approach by filtering. *IEEE Transactions on Information Theory, Vol. 44, pp. 1097-1110*, May 1998.
12. S. Ayyorgun and R. Cruz. A service curve model with loss. Technical Report LA-UR-03-3939, Los Alamos National Laboratory, June 2003.
13. D. Starobinski and M. Sidi. Stochastically bounded burstiness for communication networks. *IEEE Transactions on Information Theory*, January 2000.
14. Chengzhi Li, A. Burchard, and J. Liebeherr. A network calculus with effective bandwidth. Technical Report CS-2003-20, University of Virginia, November 2003.
15. C. S. Chang. Stability, queue length and delay of deterministic and stochastic queueing networks. *IEEE Transactions on Automatic Control*, May 1994.
16. F. P. Kelly. Notes on effective bandwidth. *Stochastic Networks: Theory and Applications vol. 4*, Sep 1995.
17. Rene Cruz. Quality of service guarantees in virtual circuit switched networks. *IEEE Journal on Selected Areas in Communications, 13(6):1048-1056*, Aug 1995.
18. H. Kim and N. B. Shroff. Loss probability calculations and asymptotic analysis for finite buffer multiplexers. *IEEE/ACM Trans. on Networking*, 9(6):755–768, Dec 2001.
19. Network Simulator v2. http://www.isi.edu/nsnam/ns/.