A Novel Direct Upper Approximation for Workload Loss Ratio in General Buffered Systems

József J. 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: {gulyas,biro}@tmit.bme.hu

Abstract. The workload loss ratio (WLR) is a key quantity from the point of Quality of Service (QoS) provisioning in packet-based communication, hence it's estimation is an important issue. The existing results in the area of WLR approximation usually interpret the workload loss as a product of some well assessable quantities. We call this approach as the indirect approximation of the WLR. The drawback of this approach is that each estimation has an error and the product of these errors could result in a highly inaccurate bound. This work deals with the upper approximation of the workload loss ratio based on it's original definition and proposes a new direct bound on the WLR applicable in general service curve network element. Extensive and systematic performance analysis of the formulae have been done, in which we found that the direct approach gives more accurate bound in several cases. We illustrate this analysis through some numerical examples.

Keywords scheduling, resource estimation, statistical multiplexing

1 Introduction

The increasing number of real-time Internet applications induce the preface of new services in telecommunication networks, besides best effort. These services have to meet some Quality of Service (QoS) requirements, which usually consist of prescriptions for QoS parameters. Thus the provision of QoS for packet switched networks generally means keeping the value of some quality related parameters at a level that fills these prescriptions. Since a significant part of the Internet applications are sensitive to the loss of the packets, the approximation of the workload loss ratio (WLR) parameter receives a significant attention. However the direct approximation of the prospective workload loss within network elements, multiplexing several inputs turns out to be a highly nontrivial problem. The probability of buffer overflow $Pr(Q > q)$ of an infinite queue is frequently used for WLR estimation [1] [2], nevertheless it is shown, that the ratio $\frac{WLR}{Pr(Q>q)}$ can be arbitrary [3]. Since the buffer overflow is closely related to the packet loss, it can be consumed for approximating the WLR indirectly as done in [3] [4] [5].

If the system is stationary and ergodic the following definition can be used[1]:

$$WLR = \frac{E[\text{number of lost bits}]}{E[\text{number of bits arriving}]} \tag{1}$$

We show, that estimating the packet loss in a direct manner using the definition (1) results in closed form bound which performs better than the existing WLR bounds. For the construction of the new bound only a little information is used about the input traffic (peak rate, upper estimated mean rate of the aggregate) so it might directly be applied in traffic management functions like call admission control (CAC) as well, without any complex measurement or information propagation.

Most of the existing bounds can be applied for constant rate servers only. We form our statements for more general queuing systems that can be described by a service curve property. The service curve property as defined in network calculus [6], with service curve β means, that at any time t, the observed output traffic in $[0, t]$ is at least equal to $A(s) + \beta(t - s)$ for some s in $[0, t]$, where $A(s)$ is the total input traffic in $[0, s]$. Using this definition, we derive new WLR formulae that can be used for a larger set of network elements, rather than for constant rate servers only.

Another problem of the existing closed form bounds for systems that satisfy the service curve property, that they assign different formulae for the case when the system is fed with inputs with the same characteristics (so called homogeneous case), and for the case when the properties on the inputs are different (heterogeneous case). We derive a universal bound that cover both cases.

Two different bounding approaches have been identified in [7] and [8]. The first one [7] based on the decomposition of the investigated network element, into virtual mini-nodes, that process one microflow as an input, and has a certain amount of processing capacity, usually a fraction of the entire server capacity. The summation of the lost packets in these mini-nodes gives an approximation of the lost packets within the original system. In the followings this approach is referenced as VNP (Virtual Node Partitioning). The other way to estimate the number of lost packets [8] will be named as Busy Period Partitioning (BPP), since it assigns a union bound for the lost packets on the time partition of the maximal possible busy period in which the packet loss could occur. Both of these approaches applied in [9] for buffer overflow and (indirect) workload loss ratio bounds in service curve network element. Because the BPP approach turned out to be more powerful in bounding buffer overflow as well as workload loss ratio [9], hereafter we concentrate on this approach.

In Section 2 we present different methods for the upper estimation of the probability generating function (PGF) used in bounding the WLR. Then in sec-

[1] Without loss of generality we consider fluid-like bit-processing system, since it can be shown, that the result can be applied for systems with rougher granularity (cells, packets).

tion 3.2 some existing results are presented in the area of indirect WLR approximation. Section 3.3 introduces a way to approximate the WLR in a definition based manner along the BPP methods, then we formalize our new bound. We illustrate through numerical examples that our direct formula performs better than the corresponding indirect one.

2 Theoretical background

Buffer overflow and WLR approximation in service curve network element with regulated inputs [9] relies on bounding the tail distribution of sum of random variables with finite support.

One of the widely used approximation techniques for this tail distribution is the Chernoff-Hoeffding bounding method, which looks like as

$$P(X > q) \leq \inf_{\theta>0} \frac{G_X(\theta)}{e^{\theta q}} \leq \inf_{\theta>0} \frac{\tilde{G}_X(\theta)}{e^{\theta q}}, \tag{2}$$

where $G_X(\theta) = E[\exp(\theta X)]$ is the probability generating function (PGF) of X and $\tilde{G}_X(\theta)$ is a kind of reasonable bound of $G_X(\theta)$ (i. e. $G_X(\theta) \leq \tilde{G}_X(\theta)$). From (2) it can be seen, that giving a better bound on the PGF, gives a better bound on the buffer overflow as well.

The following two lemmas presents Hoeffding's results (using our notations) on the PGF approximation [10] for the homogeneous and heterogeneous cases.

Lemma 1 ([10]). *Let $X_1, X_2, ..., X_I$, independent random variables with $X = \sum_{i=1}^{I} X_i$, $M = E[X]$ and $0 \leq X_i \leq p$. Then for $\theta > 0$*

$$G_X(\theta) \leq \left(1 - \frac{M}{Ip} + \frac{M}{Ip} e^{(\theta p)}\right)^I. \tag{3}$$

Lemma 2 ([10]). *Let $X_1, X_2, ..., X_I$, independent random variables with $X = \sum_{i=1}^{I} X_i$, $M = E[X]$ and $0 \leq X_i \leq p_i$. Then for $\theta > 0$*

$$G_X(\theta) \leq e^{\theta M} e^{\frac{\theta^2 \sum_{i=1}^{I} p_i^2}{8}}. \tag{4}$$

One can see, that Lemma 2 does not coincide with Lemma 1 for the special setting of $p_1 = p_2 = ... = p_n$. The following PGF approximation leads to a formula that covers both cases.

Lemma 3 ([11],[12]). *Let $X_1, X_2, ..., X_I$ be I independent (and not necessarily identically distributed) random variables with $0 \leq X_i \leq p_i, X = \sum_{i=1}^{I} X_i$ and $M = E[X]$. Then for $s > 0$*

$$G_X(\theta) \leq \left(1 - \frac{M}{I^* p} + \frac{M}{I^* p} e^{\theta p}\right)^{I^*}, \tag{5}$$

where the right side is the PGF of the sum of $X_1^*, X_2^*, ..., X_{I^*}^*$, $I^* = \lceil \sum_{i=1}^{n} p_i / p \rceil$ independent homogeneous Bernoulli random variables, with the identical peak value $p = \max(p_1, p_2, ..., p_I)$ and identical mean value $E[X_i^*] = M/I^*$.

Although this PGF bound is suitable to cover both the homogeneous and heterogeneous cases in bounding the buffer overflow probability, we need further results for obtaining uniform WLR bound. For the relation of random variables X and X^* a much more general results is valid (with the apparent combination of Lemma 1 and Theorem 3 in [13] and also found in [14]), namely

$$\frac{E[(X-q)^+]}{E[X]} \leq \frac{E[(X^*-q)^+]}{E[X^*]} \ , \forall q \geq 0 \tag{6}$$

where $X^* = \sum_{i=1}^{I^*} X_i^*$. Note that the random variable X^* has binomial distribution, hence,

$$\frac{E[(X^*-q)^+]}{E[X^*]} = \frac{p}{M} \sum_{l=\lceil q \rceil}^{I^*} (l - \lceil q \rceil) \binom{I^*}{l} \left(\frac{M}{I^* p}\right)^l \left(1 - \frac{M}{I^* p}\right)^{I^*-l} \ . \tag{7}$$

It can be shown by straightforward calculation that the derivative of this function with respect to M is always non-negative, that is

$$\partial_M \frac{E[(X^*-q)^+]}{E[X^*]} \geq 0 \ , \ \forall q > 0 \ . \tag{8}$$

A very practical consequence of this result is that if only an upper bound of M is known, then the inequality still holds in (6) when this known upper bound of M is used in the computation or an upper approximation of the right hand side of (6).

3 Upper Approximations of WLR

3.1 Notation and assumptions

In this paper the following notations of network calculus [6] are used: $\mathcal{I} = \{1, 2, ..., I\}$ is a set of input flows in a network element. $A_i(s,t], i \in \mathcal{I}$ denotes the number of bits arrived in input i in the interval (s,t]. $A_i^*(s,t], i \in \mathcal{I}$ means the same for the output of the ith flow. Let $A(s,t] := \sum_{i=1}^{I} A_i(s,t]$, and $A^*(s,t] := \sum_{i=1}^{I} A_i^*(s,t]$. The notation $v(f,g) = \sup_{t \geq 0}\{f(t) - g(t)\}$ stands for the maximal vertical, and the notation $h(f,g) = \sup_{t \geq 0}\{\inf\{u \geq 0 : f(t) \leq g(t+u)\}\}$ for the maximal horizontal deviation between f and g. We define $\bar{\alpha} = \sum_{i=0}^{I} \bar{\alpha}_i$, where $\lim_{u \to \infty} A_i(0,u]/u \leq \lim_{u \to \infty} \alpha_i(u)/u = \bar{\alpha}_i$, and $\alpha = \sum_{i=0}^{I} \alpha_i$.

The following assumptions are also needed for the derivation.

- $(A1)$ $A_1, A_2, ..., A_I$-s are independent
- $(A2)$ For all $i \in \mathcal{I}, A_i$ has α_i as an arrival curve, where α_i is a non-negative wide-sense increasing function.
- $(A3)$ For each $i \in \mathcal{I}$, and any $s, t \in R$, $E[A_i(s,t]] \leq \bar{\alpha}_i \cdot (t-s)$, where $\bar{\alpha_i} = \lim_{t \to \infty} \alpha_i(t)/t$.

- $(A4)$ There exists a sequence of random points: ... $< S_{-2} < S_{-1} < S_0 \leq 0 < S_1 < S_2 <$ ..., such that $\lim_{n \to -\infty} S_n = -\infty$, and $\lim_{n \to \infty} S_n = \infty$, and for all $n \in Z, A(S_n, S_{n+1}] = A^*(S_n, S_{n+1}]$
- $(A5)$ If $S(t) = \{S_n, n \in Z : S_n \leq t\}$, and β is the aggregate service curve for the flows, than for all $t \in R$, and any $u \in S(t)$, $\exists s \in [u,t] : A^*(u,t] - A(u,s] \geq \beta(t-s)$, where β is a non-negative wide sense increasing function.
- $(A6)$ There exists[2] $\tau < \infty$ such that for all $s \geq \tau$, $\beta(s) \geq \alpha(s)$.
- $(A7)$ Let A_i and A_i^* be stationary and ergodic.

3.2 Indirect bounds on the WLR

In our explanation the indirect derivation in the approximation of the WLR means, that the given method does not estimate quantity that defines the workload loss ratio, but interprets it as a product of other quantities, and defines upper bounds on each of these related quantities. For systems that satisfy the service curve property such bound is proposed in [4], where the WLR estimator formula is the product of the bound on the buffer overflow probability and an additional term, which is a hard deterministic bound on the loss ratio over any time interval of length t. The following Theorem recalls that result.

Theorem 1 ([9]) *Assume $(A1) - (A3)$, and that $v(\alpha, \beta) < \infty$, $h(\alpha, \beta) < \infty$, and $\alpha_i = \alpha_1$ for all $i \in I$. Then*

$$WLR \leq \frac{\hat{l}(1)\alpha(1)}{\bar{\alpha}} P(Q^\infty(0) > q) \tag{9}$$

where $\hat{l}(t) = 1 - \inf_{s \leq t} \frac{\beta(s)+q}{\alpha(s)}$, and $Q^\infty(t)$ is the buffer occupancy of a virtual system identical to the original system, but with a buffer size sufficient to ensure no losses.

Theorem 1 defines a framework for WLR approximation in an indirect manner. The substitution of the existing buffer overflow bounds for service curve network elements according to the VNP and the BPP methods results in different formulae for the WLR estimation. Since these result need a buffer overflow estimation, the bound inherits the undesirable property, that the buffer overflow bounds presented in [4] splits into two different formulae, for the homogeneous and the heterogeneous cases. The reason for this is the use of two different PGF approximation ((3), (4)) for these two cases and as a corollary two bounds raise for each bounding approach (VNP, BPP). The presentation of these formulae is omitted here, one can easily recover them from [9] and [4]. Among these bounds in this paper we use the indirect BPP-based WLR bound assuming heterogeneous inputs as a reference and for comparison purposes to our direct BPP-based WLR bound.

[2] The maximum possible busy period in such a system is τ.

3.3 WLR estimation with direct formula

According to the definition of the WLR for stationary and ergodic systems [3],[15]:

$$WLR = \frac{E[\text{number of lost bits in a unit time interval}]}{E[\text{number of bits arriving in a unit time interval}]} \qquad (10)$$

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 infinite buffer system [3]:

$$WLR \le \frac{E[(Q^{\infty} - q)^{+}]}{E[A]}$$

where Q^{∞} represents the stationary buffer occupancy of the system with infinite buffer, and $E[A] = E[A(0,1)]$ is the number of bits arriving in a unit time interval.

3.4 Bounds using he BPP approach

In what follows we derive bound on the WLR according to the BPP approach. To this end the following two lemmas are needed.

Lemma 4 ([9]). *If the assumptions (A2),(A5) hold and furthermore β is super-additive then*

$$Q(0) \le \sup_{0 \le s \le \tau} \{(A(-s,0) - \beta(s))\} \qquad (11)$$

where $Q(0)$ represents the buffer occupancy at an arbitrary time instant 0.

According to assumption (A7) Lemma 4 is also valid for the stationary buffer occupancy Q^{∞}, hence

$$WLR \le \frac{E[\sup_{0 \le s \le \tau}\{(A(-s,0) - \beta(s)\} - q)^{+}]}{E[A]}. \qquad (12)$$

Before the presentation of our WLR bound based on equation (12), another lemma has to be considered. For any $K \in N$, and any $t \ge 0$, let $T_K(t)$ be the set of partitions of $[0,t)$ in K intervals: $T_K = \{(t_0, t_1, ..., t_K) : 0 = t_0 \le t_1 \le ... \le t_K = t\}$.

Lemma 5 ([9]). *For any $K \in N$, $\boldsymbol{t} \in T_K(\tau)$ and $q \ge 0$,*

$$Pr[\sup_{0 < s \le \tau} \{(A(-s,0) - \beta(s)\} > q] \le \sum_{k=0}^{K-1} Pr[(A(0,t_{k+1}) > q + \beta(t_k))] \qquad (13)$$

The next theorem gives a direct approximation of the WLR, based on inequality (12):

Theorem 2 *Assume* $(A1)-(A7)$*. Then*

$$WLR \leq \inf_{K\in N,\boldsymbol{t}\in T_K(\tau)} \tfrac{1}{\bar{\alpha}} \textstyle\sum_{k=0}^{K-1} \\ \theta_k^* \left(\frac{m_k}{c_k}\right)^{\frac{c_k}{\hat{\alpha}_{t_{k+1}}}} \left(\frac{I_{k+1}^*\hat{\alpha}_{t_{k+1}}-m_k}{I_{k+1}^*\hat{\alpha}_{t_{k+1}}-c_k}\right)^{I_{k+1}^*-\frac{c_k}{\hat{\alpha}_{t_{k+1}}}} \tag{14}$$

where $m_k = \bar{\alpha}t_{k+1}$, $c_k = \beta(t_k)+q$, $I_{k+1}^* = \left\lceil \frac{\sum_{i=1}^{I}\alpha_i(t_{k+1})}{\hat{\alpha}_{t_{k+1}}} \right\rceil$, $\hat{\alpha}_{t_{k+1}} = \max_{i\in I}(\alpha_i(t_{k+1}))$, *and* $\theta_k^* = \frac{1}{\hat{\alpha}_{t_{k+1}}} \log \frac{c_k}{m_k} \frac{I_{k+1}^*\hat{\alpha}_{t_{k+1}}-m_k}{I_{k+1}^*\hat{\alpha}_{t_{k+1}}-c_k}$.

Proof. Using a well-known computation of the expected value of non-negative random variable:

$$\frac{E[\sup_{0\leq s\leq\tau}\{(A(-s,0)-\beta(s)\}-q)^+]}{E[A]} = \\ = \frac{\int_{x=0}^{\infty} Pr[\sup_{0\leq s\leq\tau}\{(A(-s,0)-\beta(s)\} > q+x]dx}{E[A]}$$

In accordance with Lemma 5:

$$\frac{\int_{x=0}^{\infty} Pr[\sup_{0\leq s\leq\tau}\{(A(-s,0)-\beta(s)\} > q+x]dx}{E[A]} \leq \\ \leq \frac{\int_{x=0}^{\infty} \sum_{k=0}^{K-1} Pr[A(0,t_{k+1}) > q+\beta(t_k)+x]dx}{E[A]}.$$

By the commutation of the integration and summation we get:

$$\frac{\int_{x=0}^{\infty} \sum_{k=0}^{K-1} Pr[A(0,t_{k+1}) > q+\beta(t_k)+x]dx}{E[A]} \leq \\ \leq \frac{\sum_{k=0}^{K-1} \int_{x=0}^{\infty} Pr[A(0,t_{k+1}) > q+\beta(t_k)+x]dx}{E[A]}.$$

One element in the summation on the right hand side can be written as

$$\int_{x=0}^{\infty} Pr[A(0,t_{k+1}) > q+\beta(t_k)+x]dx = E[(A(0,t_{k+1})-q-\beta(t_k))^+] \tag{15}$$

Due to the assumed stationary increment of $A(0,t)$, $E[A]$ can be rewritten as

$$E[A] = \frac{E[A(0,t_{k+1})]}{t_{k+1}} \ , \ k=0,\ldots,K-1 \ . \tag{16}$$

Using this it follows that

$$WLR \leq \sum_{k=0}^{K-1} \frac{E[(A(0,t_{k+1})-q-\beta(t_k))^+]}{E[A(0,t_k+1)]} t_{k+1} \tag{17}$$

Since $A(0,t_{k+1}) = \sum_{i=1}^{I} A_i(0,t_{k+1})$, $A_i(0,t_{k+1}) \leq \alpha_i(t_{k+1})$ and $E[A_i(0,t_{k+1}] \leq \bar{\alpha}_i t_{k+1}$ we can apply the results presented in Section 2. That is let $A_i^*(0,t_{k+1})$, $i = 1,\ldots,I^*$ independent and identically distributed Bernoulli random variables with mean $E[A_i^*(0,t_{k+1})] = \bar{\alpha}t_{k+1}/I^*$, $A_i^*(0,t_{k+1}) \leq \hat{\alpha}_{t_{k+1}}$, where

$$\hat{\alpha}_{t_{k+1}} = \max_{1\leq i\leq I}(\alpha_i(t_{k+1})) \ , \ I^* = \left\lceil \frac{\sum_{i=1}^{I}\alpha_i(t_{k+1})}{\hat{\alpha}_{t_{k+1}}} \right\rceil \ . \tag{18}$$

According to (6) and (8) one can deduce that

$$\frac{E[(A(0,t_{k+1}) - q - \beta(t_k))^+]}{E[A(0,t_k+1)]} t_{k+1} \leq \frac{E[(A^*(0,t_{k+1}) - q - \beta(t_k))^+]}{\bar{\alpha}} \tag{19}$$

where $A^*(0,t_{k+1}) = \sum_{i=1}^{I^*} A_i^*(0,t_{k+1})$ and hence $E[A^*(0,t_{k+1})] = \bar{\alpha}t_{k+1}$. Applying a well-known Chernoff bound [16] to the right hand side of (19) gives

$$E[(A^*(0,t_{k+1}) - q - \beta(t_k))^+] \leq \frac{1}{\theta_k^*}\frac{E\left[\exp(\theta_k^* A^*(0,t_{k+1}))\right]}{\exp(\theta_k^*(q+\beta(t_k)))} \tag{20}$$

where[3]

$$\theta_k^* = \mathrm{arginf}_{\theta_k}\left(\log E\left[\exp(\theta_k A^*(0,t_{k+1}))\right] - \theta_k(q+\beta(t_k))\right) \ . \tag{21}$$

The generating function of $A^*(0,t_{k+1})$ is

$$E[\exp(\theta_k A^*(0,t_{k+1}))] = \left(1 - \frac{\bar{\alpha}t_{k+1}}{I^*\hat{\alpha}_{t_{k+1}}} + \frac{\bar{\alpha}t_{k+1}}{I^*\hat{\alpha}_{t_{k+1}}}\exp(\theta_k\hat{\alpha}_{t_{k+1}})\right)^{I^*} \ . \tag{22}$$

Substituting it into the Chernoff bound in (20) and after straightforward calculations the closed form formula of (14) is obtained. Q.E.D.

4 Evaluation

For illustrating the evaluation and comparison of the bounds presented the following scenario is used. We have 100 input flows, which are token bucket constrained with some arrival curve ($\alpha(t) = \bar{\alpha}+\sigma$), and the packet forwarder satisfies a rate latency service curve property, with $\beta = c \cdot max(t-e,0)$, in a work-conserving manner. We take parameter values that are close to many practical, common applications. The service rate of the server will be 150Mbps and let the packets size be 1500 bytes. This means the node can serve 12500 packets during a second (pps). The latency time (e) is $8 \cdot 10^{-5}$ sec. We set up four configurations for the evaluation:

[3] Note that θ_k^* does not provide the optimal Chernoff bound, however, it guarantees closed form WLR bound as opposed to the optimal $\mathrm{arginf}_{\theta_k} \ \log E\left[\exp(\theta_k A^*(0,t_{k+1}))\right] - \theta_k(q+\beta(t_k) - \log\theta_k)$.

Configuration 1: $\alpha_1(t) = 33.3pps + 8p$, $\alpha_2(t) = 16.6pps + 5p$. If we have 50 microflows with $\alpha_1(t)$ and $\alpha_2(t)$ each, it results $\alpha(t) = 50 \cdot \alpha_1(t) + 50 \cdot \alpha_2(t) = 2500pps + 650p$ as an aggregate arrival curve. This configuration represent a utilization of 0.2 for the server.

Configuration 2: $\alpha_1(t) = 133.3pps + 8p$, $\alpha_2(t) = 66.6pps + 5p$. If we have 50 microflows with $\alpha_1(t)$ and $\alpha_2(t)$ each, it results $\alpha(t) = 50 \cdot \alpha_1(t) + 50 \cdot \alpha_2(t) = 10000pps + 650p$ as an aggregate arrival curve. This configuration represent a utilization of 0.8 for the server.

Configuration 3: $\alpha_1(t) = 26pps + 8p$, $\alpha_2(t) = 24pps + 8p$. 100 microflows results in $\alpha(t) = 2500pps + 800p$ as an aggregate arrival curve. This configuration represent a utilization of 0.2 for the server.

Configuration 4: $\alpha_1(t) = 102pps + 8p$, $\alpha_2(t) = 88pps + 8p$. 100 microflows results in $\alpha(t) = 10000pps + 800p$ as an aggregate arrival curve. This configuration represent a utilization of 0.8 for the server. A summary of the configurations can be seen on Table 1.

Conf.	$\alpha_1(t)$	$\alpha_2(t)$	$\alpha(t)$	utilization
Conf. 1.	$33.3pps + 8p$	$16.6pps + 5p$	$2500pps + 650p$	0.2
Conf. 2.	$133.3pps + 8p$	$66.6pps + 5p$	$10000pps + 650p$	0.8
Conf. 3.	$26pps + 8p$	$24pps + 8p$	$2500pps + 800p$	0.2
Conf. 4.	$102pps + 8p$	$88pps + 8p$	$10000pps + 800p$	0.8

Table 1. The summary of configurations

The choice of configuration 3 and 4 is to evaluate the performance of the bounds, when the inputs have almost same characteristics. Therefore the scenario is close to the homogeneous case, but since it is heterogeneous the heterogeneous form of the PGF approximation have to be used.

For each configuration we have two graphs. The one on the left side shows the logWLR approximation as a function of the buffer size (the continuous line represents our new bound, while the dotted line represents the old bound) and the graph to the right indicates the relative buffer requirement gain according to the two bounds. This latter quantity refers to the amount of bandwidth which can be saved when our new bound is used to keep the loss ratio under a certain level. This is defined in the following way. Let $Q_{\text{req},1}$ and $Q_{\text{req},2}$ be the buffer requirements formulated as

$$Q_{\text{req},1}(-\gamma) = \min(q, WLR_1(q) \leq 10^{-\gamma}) \; , \; Q_{\text{req},2}(-\gamma) = \min(q, WLR_2(q) \leq 10^{-\gamma}) \tag{23}$$

where $WLR_1(q)$ and $WLR_2(q)$ are the WLR approximations according to Theorem 1 and Theorem 2 . The relative gain in buffer requirement drawn in the graphs is computed as

$$\text{gain}(-\gamma) = \frac{Q_{\text{req},1}(-\gamma) - Q_{\text{req},2}(-\gamma)}{Q_{\text{req},1}(-\gamma)} \tag{24}$$

All figures well illustrate our experience in extensive comparison that our new bound significantly outperforms the old WLR approximation, especially in case of low traffic intensity. For higher traffic loads the relative buffer requirement gain decreases, but still remains high, around 40-50% in the examples. The relative gain is usually higher for less stringent QoS constraints on the WLR (smaller γ). In Fig. 1 for conf. 1 one can observe that to achieve no more than 10^{-6} loss ratio the required buffer size is smaller than $50p$ when WLR_2 is used and it is around $180p$ when WLR_1.

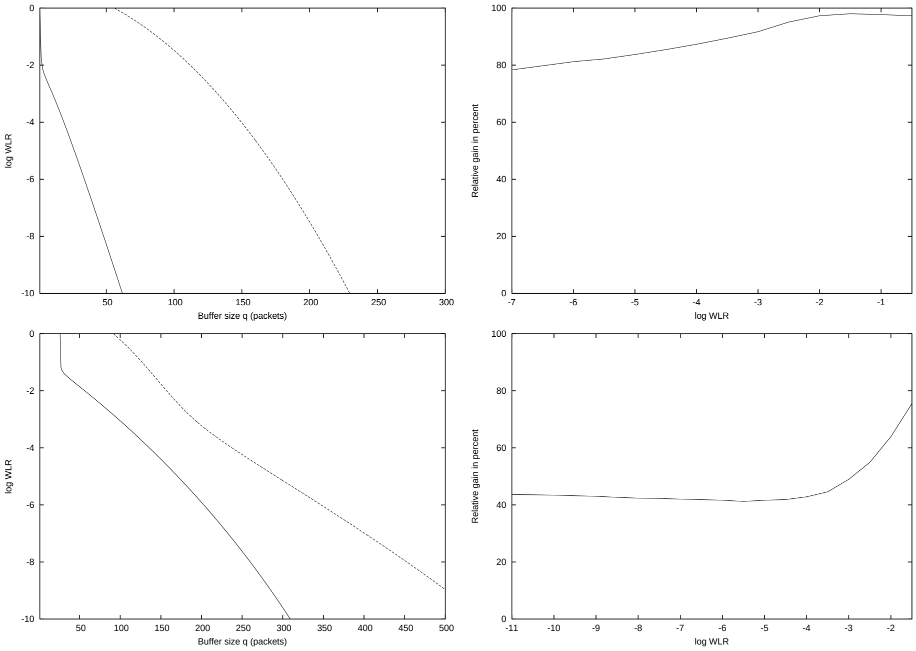

Fig. 1. The results of Theorem 1 and Theorem 2 for conf. 1 (up) and conf. 2 (down).

5 Conclusions

The scope of this paper was to present our new direct (definition based) formula for the workload loss ratio applicable in general buffered systems characterized as service curve network element. Our new bound is significantly better then the corresponding existing one, and can ensure to save tremendous amount of buffer space when it is used to guarantee QoS level for the workload loss ratio.

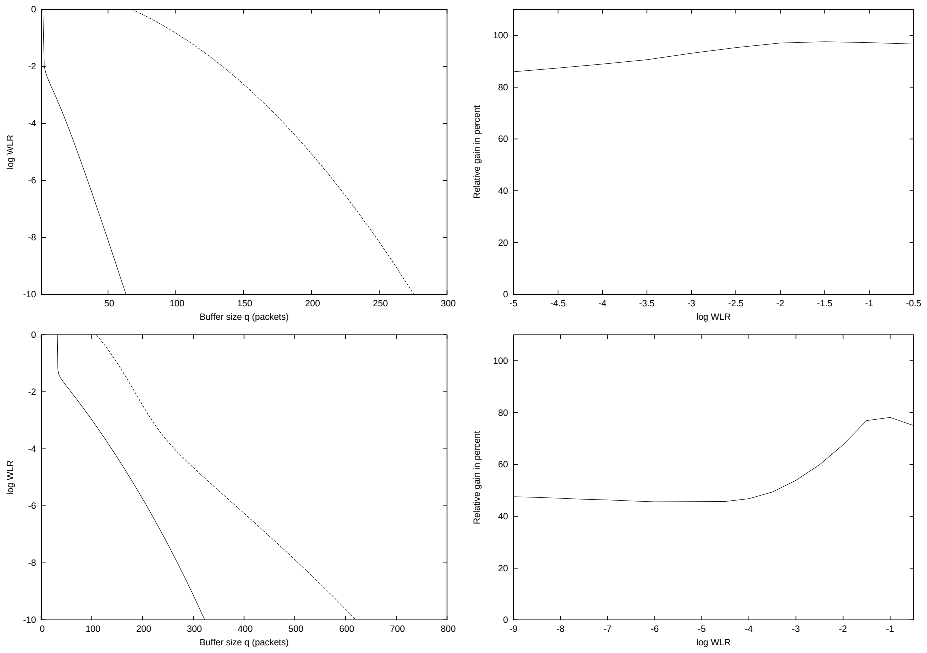

Fig. 2. The results of Theorem 1 and Theorem 2 for conf. 3 (up) and conf. 4 (down).

References

1. 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*, 13, March 1995.
2. 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*, 2, July 2000.
3. András György and Tamás Borsos. Estimates on the packet loss ration via queue tail probabilities. *IEEE Globecom*, March 2001.
4. Milan Vojnovic and Jean-Yves Le Boudec. Stochastic analysis of some expedited forwarding networks. *IEEE INFOCOM New York*, June 2002.
5. Nikolay Likhanov and Ravi R. Mazumdar. Cell loss asymptotics in buffers fed with a large number of independent stationary sources. *Journal of Applied Probability*, 36, March 1999.
6. Jean-Yves Le Boudec and Patrick Thiran. Network calculus, 2002.
7. George Kesidis and Takis Konstantopoulos. Worst-case performance of a buffer with independent shaped arrival processes. *IEEE Communication Letters.*
8. C.-S. Chang, W. Song, and Y. Ming Chiu. On the performance of multiplexing independent regulated inputs. *Proceedings of Sigmetrics*, May 2001.
9. Milan Vojnovic and Jean-Yves Le Boudec. Bounds for independent regulated inputs multiplexed in a service curve network element. *IEEE Trans. on Communications*, 51(5):449–451, May 2003.

10. Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association, 58:13-30*, March 1963.
11. J. Bíró, A. Gulyás, and M. Martinecz. Parsimonious estimates of bandwidth requirement in quality of service packet networks. *Performance Evaluation*, 59(2–3):159–178, February 2005.
12. J. Bíró, Z. Heszberger, and M. Martinecz. A family of performance bounds for qos measures in packet-based networks. In *IFIP Networking 2004*, pages 1108–1119, 2004.
13. G. Mao and D. Habibi. Loss performance analysis for heterogeneous on-off sources with application to connection admission control. *IEEE Transactions on Networking*, July 2001.
14. R. Szekli. *Stochastic ordering and dependence in applied probability (Lecture Notes in Statistics).*
15. M. Vojnovic and J. Y. Le Boudec. Stochastic analysis of some expedited forwarding networks. Technical Report DSC/2001/039, EPFL-DI-ICA, July 2001.
16. F. P. Kelly. Notes on effective bandwidth. *Stochastic Networks: Theory and Applications*, 4, Sep 1995.