Distribution-Free Conservative Bounds for QoS Measures

József J. Bíró, András Gulyás, Tamás Szénási, Zalán Heszberger
Budapest University of Technology and Economics
Department of Telecommunications and Media Informatics
H-1117, Magyar tudósok körútja 2., Budapest, Hungary
email: biro@tmit.bme.hu

Abstract

In this paper we set up a family of conservative upper bounds for QoS measures applicable in QoS guaranteed packet-based communication networks. The bounds are distribution free, that is they depend only on the number of traffic sources, the peak rates of the flows and the mean arrival intensity of the aggregated traffic stream. No other information on the traffic distribution is required. The bounds have been developed under the well known bufferless fluid flow multiplexing (bffm) modeling framework, nevertheless, an important property is also discussed in the paper, namely they can also be applicable to the buffered statistical multiplexer case when the input flows are regulated. The performance of the bounds are analyzed and comparative numerical investigations are also presented in both multiplexing models.

1 Introduction

As regards the available resources (e.g. transmission link capacities, buffers, processing capacities) at the nodes in packet-based networks, an important question is how often such resources are overloaded due to high volume of packets arrived. Within the family of resource based measures quantifying this overloading phenomena, one representative measure is the link saturation probability. This corresponds to the fraction of time when the sum of the instantaneous (or average over a sufficiently small time interval) arrival rate of traffic flows exceeds the transmission link capacity. The use of this measure assumes the bufferless fluid flow multiplexing framework which turned out to be powerful dimensioning tool in either the case of elastic traffic in certain access network scenarios or the case of stream-like traffic [6, 2].

Several asymptotic and approximate results have been formulated for buffer overflow probability in buffered statistical multiplexing models. Recent result in [7] incorporates bounds for buffer overflow probability provided the arrival traffic flows are regulated (the regulation is characterized by arrival curves) and the service offered to the traffic by the nodes is described by the so-called service curves. These bounds are based on Hoeffding’s result on the tail probability estimation of sum of bounded random variables [4].

Besides the fraction of resource overload periods, it is also important to identify quantities (measures) which are based on the amount of traffic becoming unconformant due to the resource overload. The class of these measures is referred to as stream-based measures. Although the unconformant packets can be either downgraded into lower level QoS class (e.g. best effort) or simply being lost, for simplicity the ratio between the unconformant traffic and the whole offered traffic is called as workload loss ratio (WLR). In the bffm modeling framework WLR corresponds to the fraction of traffic which can not be transmitted due to the link saturation. In buffered statistical multiplexers WLR means the fraction of packets which can not be placed into the buffer due to buffer overflow.

In this paper, we set up a family of conservative upper bounds for saturation probability and workload loss ratio, which can directly be applied in both (bufferless and buffered) multiplexing framework. For this purpose, the so-called Chernoff-Hoeffding bounding technique has been used which enables to treat the estimates for saturation probability and workload loss ratio in a common framework. The resulted bounds are expressed in closed-form formulae and distribution-free, that is they use few characteristic information (known a priori and/or measurable) on the traffic flows.

2 The Chernoff-Hoeffding Bounding Method

Bufferless fluid flow multiplexing is often used in the literature to analyze QoS measures, e.g., packet loss probability in a multiplexer [3, 2]. Because this approach assumes no buffer at burst time scales, it is able to provide conservative estimates for the QoS measures under question. For modeling purposes under bffm, let us assume that we have n
fluid flows to be multiplexed on a communication link with transmission capacity $C$. Let the instantaneous stationary (that is time dependence can be eliminated) arrival rate of flow $i$ be noted by $X_i$, as a random variable. Because every flow has a peak rate $p_i$ we also have $0 \leq X_i \leq p_i$. Further, let the aggregate flow arrival rate be $X = \sum_{i=1}^{n} X_i$.

The link saturation probability can now be defined as $P_{\text{sat}} \overset{\text{def}}{=} P(X > C)$. This probability reflects the fraction of time when the link is overloaded (provided the system is ergodic), that is the combined arrival rate exceeds the link capacity. This resource-based congestion measure could be important from network operation point of view. The estimation of this quantity can provide more accurate bounds for $P_{\text{loss}}$ performance analysis. This measure better characterizes the expected loss rate and could also contribute to determining the users’ satisfaction. The well-known Chernoff bounds for $P_{\text{sat}}$ and $WLR$ are as follows:

$$P(X > C) \leq \inf_{s > 0} G_X(s) e^{sC} = \inf_{s > 0} (\Lambda_X(s) - sC) , \quad (1)$$

$$WLR \leq \exp (\Lambda_X(s^*) - s^*C - \log(s^*M)) , \quad (2)$$

where

$$s^* = \arg\inf_{s}(\Lambda_X(s) - sC) , \quad (3)$$

$G_X(s) \overset{\text{def}}{=} \mathbb{E}[e^{sX}]$ and $\Lambda_X(s) \overset{\text{def}}{=} \log G_X(s)$ are the probability generating function (PGF) and the cumulant generating function (CGF) of $X$, respectively. The direct computation of these bounds is usually not possible, because the underlying generating functions would require all the moments of $X$ to be known. Instead, the CGF’s are to be further bounded based on the available information (moments) on $X$ and embedded into the Chernoff bound. This is called the Chernoff-Hoeffding bounding method.

Previously known and newly developed bounds for $P_{\text{sat}}$ and $WLR$ are presented under a common framework in Section 3. After that, analysis and comparisons based on extensive numerical investigations are highlighted. The applicability of the performance bounds for buffered multiplexers is also briefly discussed.

### 3 Conservative Estimates for the QoS Measures

First, we provide three conservative bounds for the PGF of aggregate traffic rate distribution, provided only the following pieces of information are available on $X$: the number of traffic flows multiplexed ($n$), the peak rates of the traffic flows ($p_i$) and the aggregate mean arrival rate ($M \overset{\text{def}}{=} \mathbb{E}[X]$).

**Lemma 1** (4) Let $X_i$, $i = 1\ldots n$ be independent random variables with $X = \sum_{i=1}^{n} X_i$, $M = \mathbb{E}[X]$ and $0 \leq X_i \leq p_i$. Then for $s > 0$,

$$G_X(s) \leq \exp(sM) \exp \left( \frac{s^2 \sum_{i=1}^{n} p_i^2}{8} \right) . \quad (4)$$

**Theorem 1** (3) Let $X_i$ be independent bounded random variables with $0 \leq X_i \leq p_i$, $X = \sum_{i=1}^{n} X_i$ and $M = \mathbb{E}[X]$. Then for $s > 0$,

$$G_X(s) \leq \prod_{i=1}^{n} \left( \frac{e^{sp_i} - 1}{p_i} \right) \left( M + \sum_{k=1}^{n} p_k \exp(s - 1) \right)^{nY} . \quad (5)$$

Now, the PGF bound based on increasing convex stochastic ordering can be formulated in the following theorem:

**Theorem 2** (11) Let $X_1, \ldots, X_n$ indicate $n$ independent random variables with $0 \leq X_i \leq p_i$, $X = \sum_{i=1}^{n} X_i$ and $M = \mathbb{E}[X]$. Then for $s > 0$,

$$G_X(s) \leq \left( 1 - \frac{M}{nY} + \frac{M}{nY} e^{sp} \right)^{nY} . \quad (6)$$

The following bounds can be obtained by the substitution of the right hand side of equation (4), (5) and (6) into the Chernoff bound of $P_{\text{sat}}$ (11).

**Theorem 3** (4) Let $X_i$ be independent bounded random variables with $0 \leq X_i \leq p_i$, $X = \sum_{i=1}^{n} X_i$ and $M = \mathbb{E}[X]$, then

$$P(X > C) \leq \exp \left( \frac{-2(C - M)^2}{\sum_{i=1}^{n} p_i^2} \right) . \quad (7)$$

**Theorem 4** (3) If $X_1, X_2, \ldots, X_n$ are independent random variables, for which $0 \leq X_i \leq p_i$ holds, then

$$P(X \geq C) \leq e^{-s^*C} \left( \frac{M + \sum_{j=1}^{n} e^{sp_j}}{n} \right)^{n} \prod_{k=1}^{n} \frac{e^{sp_k} - 1}{p_k} , \quad (8)$$

where $s^*$ is as follows:

$$s^* = \frac{C - M}{\frac{1}{2} \sum_{i=1}^{n} p_i^2 - \frac{1}{n} (M - \frac{1}{2} \sum_{i=1}^{n} p_i)^2} . \quad (9)$$

**Theorem 5** (11) Let $X_i$ be independent bounded random variables with $0 \leq X_i \leq p_i$, $X = \sum_{i=1}^{n} X_i$ and $M = \mathbb{E}[X]$. Further, let $p = \max(p_i, i = 1, \ldots, n)$, $nY = \lceil \sum_{i=1}^{n} p_i/p \rceil$, and $m = \sum_{i=1}^{n} m_i/nY$, then

$$P(X > C) \leq \left( \frac{M}{C} \right)^{nY} \left( \frac{nYp - M}{nYp - C} \right)^{nY - \frac{nY}{p}} . \quad (10)$$
In the following theorem we summarize the closed form conservative bounds for $WLR$ based on (2) and the pgf bounds:

**Theorem 6** Let $X_i$ be independent bounded (and not necessarily identically distributed) random variables with $0 \leq X_i \leq p_i$, $X = \sum_{i=1}^{n} X_i$ and $M = E[X]$. Further, let $K = \frac{1}{4} \sum_{i=1}^{n} p_i^2 - \frac{1}{4} (M - \frac{1}{2} \sum_{i=1}^{n} p_i)^2$, $p = \max(p_i, i = 1, \ldots, n)$, $n_Y = \| \sum_{i=1}^{n} p_i / p \|$, and $m = \sum_{i=1}^{n} m_i / n_Y$, then the following three inequalities hold for $WLR$:

$$WLR \leq \frac{\sum_{i=1}^{n} p_i^2}{4(C-M)} \exp \left( \frac{-2(C-M)^2}{\sum_{i=1}^{n} p_i^2} \right), \quad (11)$$

$$WLR \leq \left( \frac{M + \sum_{j=1}^{n} \frac{p_j}{n}}{n} \right)^n \frac{K e^{(M-C)C}}{(C-M)M} \prod_{k=1}^{n} \frac{e^{C-M} - 1}{p_k}, \quad (12)$$

$$WLR \leq \left( \frac{M}{C} \right) \frac{e^{n_Y p - M}}{n_Y p - C} \frac{p}{M \log \frac{C}{M} \frac{n_Y p - M}{n_Y p - C}}. \quad (13)$$

In the set of bounds presented above the ones in (7) and (8) are already known from [4] and [3], but, to the authors best knowledge, the bounds in (11), (12) and (13) are neither presented nor analyzed previously.

### 4 Performance Analysis

In this section the performance of the bounds presented are analyzed and illustrated through numerical examples. For this purpose a simple two-class on-off traffic mix has been defined. The number of sources within the classes are represented by $n_1$ and $n_2$, respectively. The mean arrival rate and the peak rate of a source within a class are assumed to be identical and indicated by $m_i$, $p_i$, $i = \{1, 2\}$. The representative traffic scenarios considered in the paper for illustrating the numerical investigations are summarized in Table 1. The first traffic mix (Mix 1) resembles the aggregation of uncompressed and compressed video flows. The second (Mix 2) and third one (Mix 3) represent the multiplexing of uncompressed and compressed voice traffic with lower and higher peak to mean ratios, respectively.

In all of the figures in this paper the 10-based logarithm of the exact values of saturation probability and workload loss ratio and their corresponding bounds are drawn in the function of the transmission capacity $C$. Since the bounds presented give reasonable values when $M < C < P$ ($P \overset{\text{def}}{=} \sum_{i=1}^{n} p_i$), parts of the interval $(M, C)$ is considered in the drawing in such a way that the exact values of $P_{\text{sat}}$ and $WLR$ should not be smaller than $10^{-8}$. The exact values are drawn with continuous lines, while the bounds based on the Hoeffding (4), improved Hoeffding (5) and stochastic ordering-based (6) CGF bounds are represented by dotted, dash-dotted and dash-dot-dotted lines, respectively.

Common observations and remarks based on our extensive numerical analysis are given, which are partly illustrated by the numerical examples.

**Observations:**

The bounds (7), (11) based on the CGF approximation (4) has usually the poorest performance, due to the underlying coarse bound on the cumulant generating function.
The differences between the improved Hoeffding and stochastic ordering-based $P_{\text{sat}}$ bounds are usually small, furthermore, it turned out to be negligible when the number of sources are higher than 100 in each traffic class and the peak rates of the traffic classes are in similar order of magnitude (e.g. in Mix2 and Mix 3).

The superiority of the stochastic ordering-based WLR bound can be observed in several cases, especially when the aggregate peak to mean ratio ($P/M$) is small (e.g. in Mix 2).

In case of high peak to mean ratio and high differences between the peak rates of the traffic classes, the improved Hoeffding-based $P_{\text{sat}}$ bound can outperform (such figures are not presented) the stochastic ordering-based $P_{\text{sat}}$ bound.

The horizontal and vertical distances between the curves are usually increases with increasing $C$.

![Figure 3. Bounds on $P_{\text{sat}}$ and WLR, Mix 3](image)

**Remarks:** Although all the bounds presented require the same amount of information on the the traffic flows, the complexity of their closed-form formulae are different. The bounds based on the Hoeffding-based CGF approximation appear in the simplest way, however, these have the poorest accuracy. The bounds [8], [12] based on the CGF approximation [5] have the most complicated appearance in the formulae, the implementation of their computation might encounter serious problems due to the presence of the several exponential-like terms. The formulae of the stochastic ordering-based bounds are relatively simple. They seem to be implementable (especially the logarithm of the bounds) in a straightforward manner. Consequently, the application of these bounds (especially the WLR bound) are strongly encouraged, also because of the good performance in accuracy.

## 5 Bounds for Buffer Overflow Probability

Within this section, the probability of buffer overflow will be investigated, in network elements represented by a service curve property. The service curve property is defined in network calculus [7], with the notion of arrival curve. A flow is said to be regulated by an arrival curve $\alpha$, if the number of bits arrived from the flow is at most $\alpha(t-s)$, within a time interval $(s, t)$. A service curve property with service curve $\beta$ 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 $[0, s]$.

One of the existing approaches to estimate backlog in network nodes is the method of Kesidis and Konstantopoulos. They presented a technique [5], that decomposes the aggregate backlog into a sum of backlogs generated by virtual nodes. These virtual nodes process one microflow as an input with a service curve, which is derived dividing the aggregate service curve equally between the inputs. The bound for the aggregate backlog is determined by bounding the sum of these virtual backlogs.

The original bound for the backlog in [5] is valid for constant rate servers, and the combination of two leaky buckets as arrival curves. Milan Vojnovic, and Jean-Yves Le Boudec [7] extended these results to a node that offers an arbitrary service curve, and for arbitrary arrival curves. Our motivation for defining new bounds was the drawback of the results in [7], that they give different bound in case of homogeneous and heterogeneous feeding of the node. Moreover, the bound for the heterogeneous case does not give back the homogeneous one as a special case. We present a bound for the arbitrary case, which improves the heterogeneous bound in [7], and recovers the homogeneous one in case of homogeneous input flows. The same notation is used as in [7], which is the following:

$I = \{1, 2, \ldots, I\}$ is a set of input flows in a network element. $A_i(s, t), i \in I$ denotes the number of bits arrived in input $i$ in the interval $(s, t]$. $A_i^\ast(s, t), i \in I$ means the same for the output of the $i$th flow. Let $A(s, t) := \sum_{i=1}^I A_i(s, t)$, and $A^\ast(s, t) := \sum_{i=1}^I A_i^\ast(s, t)$. The notation $v(f, g) = \sup_{t>0} \{f(t) - g(t)\}$ stands for the maximal vertical, and the notation $h(f, g) = \sup_{t>0} \{\inf\{u \geq 0 : f(t) \leq g(t + u)\}\}$ for the maximal horizontal deviation between $f$ and $g$.

We define $\bar{\alpha} = \sum_{i=0}^I \bar{\alpha}_i$, where $\lim_{u \to \infty} A_i(0, u)/u \leq \lim_{u \to \infty} \bar{\alpha}_i(u)/u = \bar{\alpha}_i$, and $\alpha = \sum_{i=0}^I \alpha_i$.

The bounds hold under the following assumptions:
Now we set up a new bound for the buffer overflow can be derived for the arbitrary case, using the stochastic ordering based PGF approximation, mentioned in section 3. The bound creation process is based on partitioning the aggregate backlog, and bounding it with the sum of backlogs in virtual nodes \( \alpha_i \), input, and \( \gamma_i \beta \) service curve: \( Q_i(t) = \sup_{-\infty < s \leq t} \{ A_i(s, t) - A(u, s) \geq \beta (t - s) \} \) For these \( Q_i \)-s the following properties hold:

1. \( (T1) \) \( P(Q(0) > q) \leq P \left( \sum_{i=1}^{I} Q_i(0) > q \right) \)

2. \( (T2) \) For any \( t \in R, Q_1(t), Q_2(t), ..., Q_I(t) \) are independent.

3. \( (T3) \) For any \( t \in R, and each i \in I \), \( 0 \leq Q_i(t) \leq v(\alpha_i, \gamma_i, \beta) \).

4. \( (T4) \) For any \( t \in R, E[Q(t)] \leq \alpha h(\alpha, \beta) \).

Since, these properties hold, the backlog bounds can be obtained by using different types of Hoeffding’s inequalities directly on the \( Q_i(t) \) random variables, for the homogeneous, and the heterogeneous case. This procedure eventuates the inequalities of \( \cite{7} \) Theorem 1 and \( \cite{7} \) Theorem 2. Now we set up a new bound for \( P(Q > q) \) on the basis of the stochastic ordering based PGF approximation. This bound significantly differs from the heterogeneous bound in \( \cite{7} \) Theorem 2, and recovers the homogeneous bound in \( \cite{7} \) Theorem 1.

**Theorem 7** Heterogeneous and homogeneous (arbitrary) case: Assume \( (A1) \) – \( (A3) \), and that for each \( i \in I \), \( (A4) \) holds for a virtual node, that offers the service curve \( \gamma_i \beta \) for the arrival process \( A_i \). If

\[
I^* = \left[ \frac{\sum_{i=1}^{I} v(\alpha_i, \gamma_i \beta)}{v(\hat{\alpha}, \hat{\gamma} \beta)} \right]
\]

where \( v(\hat{\alpha}, \hat{\gamma} \beta) = \max_{i \in I}(v(\alpha_i, \gamma_i \beta)) \), then for \( \bar{h}(\alpha, \beta) < q < v(\alpha, \beta) \):

\[
P(Q(0) > q) \leq \exp \left( -I^* \frac{q}{v} \ln \frac{q}{\bar{h}} + I^* \left( 1 - \frac{q}{v} \right) \ln \frac{v - \bar{h}}{v - q} \right)
\]  

(14)

where for brevity \( v = v(\alpha, \beta) \) and \( h = h(\alpha, \beta) \).

Proof: The \( i \in I, Q_i(t)-s \) are random variables, which are bounded by \( v(\alpha_i, \gamma_i \beta) \), and \( E[Q(t)] \leq \alpha h(\alpha, \beta) \). Let \( Q_{i1}, Q_{i2}, ..., Q_{iI} \), represent \( I^* \) independent homogeneous on-off random variables with peak rate \( v(\hat{\alpha}, \hat{\gamma} \beta) = \max_{i \in I}(v(\alpha_i, \gamma_i \beta)) \), where

\[
I^* = \left[ \frac{\sum_{i=1}^{I} v(\alpha_i, \gamma_i \beta)}{v(\hat{\alpha}, \hat{\gamma} \beta)} \right].
\]

The \( (T4) \) property also holds, since the aggregate characteristic of the new homogeneous traffic, is the same as the original heterogeneous traffic. Since

\[
P(Q(0) > q) \leq P \left( \sum_{i=1}^{I} Q_i(0) > q \right) \leq \inf_{s>0} \frac{G_{Q}(s)}{e^{sq}}, \quad (15)
\]

and from Theorem 2 it can be revealed that:

\[
\inf_{s>0} \frac{G_{Q}(s)}{e^{sq}} \leq \inf_{s>0} \frac{G_{Q^*}(s)}{e^{sq}}
\]  

(16)

which means that, the bound on the new variables, is also a bound on the original system. The right side of \( (16) \) can be expressed in closed form and immediately gives the right side of \( (14) \).

In what follows we present numerical comparisons between this new bound and the bounds presented in \( \cite{7} \).

Input flows are token bucket constrained \( (\alpha_i(t) = \alpha_i + \sigma_i) \), and the packet forwarder satisfies a rate latency service curve property, with \( \beta = c \cdot \max(t - e, 0) \), in a work-conserving manner. The service rate of the server is 150Mbps. Let the packets size be 1500 bytes. This means the node can serve 12500 packets during a second (pps).

**Configuration 1:** \( \alpha_1(t) = 133.3 \text{pps} + 8p, \alpha_2(t) = 66.6 \text{pps} + 5p \). If we have 50 microflows with \( \alpha_1(t) \) and \( \alpha_2(t) \) each, it results \( \alpha(t) = 50 \cdot \alpha_1(t) + 50 \cdot \alpha_2(t) = 10000 \text{pps} + 650p \) as an aggregate arrival curve. This configuration represent a utilization of 0.8 for the server.

**Configuration 2:** \( \alpha_1(t) = \alpha_2(t) = 100 \text{pps} + 8p \). 100 microflows results in \( \alpha(t) = 10000 \text{pps} + 800p \) as an aggregate arrival curve. This configuration represent a utilization of 0.8 for the server.

In the followings we investigate the bound, and compare them with the results in \( \cite{7} \). The results for configuration
1 and configuration 2 can be found on Figure 4. One can see, that the new bound of Theorem 7 produces significantly better result, even for large queue sizes. From our numerical investigations it can be revealed, that the performance of the new bounds increases with the load (not seen in the Figure).

![Figure 4: Probability bounds for Theorem 7 and [7] Theorem 2. Configuration 1 with heterogeneous, Configuration 2 with homogeneous inputs.](image)

The lower drawing in Figure 4 shows the results given by the two theorems for homogeneous input traffic. The alert reader can easily find, that the formula of Theorem 7 turns out to be same as the one of [7] Theorem 1 in case of homogeneous feeding. From Figure 4 we can also see that the formula in [7] Theorem 2, can not give a good estimation for nearly identical input flows, especially for high load values (not seen in the Figure).

The main advantage of the new bound, that it does not require the check of the property of homogenity before application, and a single bound can be used for all input scenarios. However, the new bound gives significantly better performance, then the previous ones in many cases.

6 Conclusion

In this paper a family of distribution-free conservative bounds on saturation probability and workload loss ratio has been set up, both for the bufferless and buffered fluid flow multiplexing framework. According to the analysis the stochastic ordering-based bounds usually have the best performance.

References


