Towards a Framework of QoS Measure Estimates for Packet-Based Networks

József J. Biro, Zalán Heszberger, Mátyás Martinecz, Felicián Németh, Octavian Pop
Budapest University of Technology and Economics
Department of Telecommunications and Media Informatics
Email: biro@tmit.bme.hu

Abstract—In the widely studied and powerful bufferless fluid flow multiplexing model (bffm) important QoS measures are often considered such as saturation probability and workload loss ratio. When regulated traffic are multiplexed into a buffer, similar QoS measures can also be identified such as buffer overflow probability and buffer workload loss ratio. In this paper, we set up a family of closed-form bounds which can be applied for QoS measures arising in both multiplexing models. The bounds have been compared based on numerical investigations.

I. BACKGROUND AND MOTIVATION

Bufferless fluid flow multiplexing is often used in the literature to analyze QoS measures, e.g., packet loss probability in a multiplexer [1], [2], [3]. 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 \( P_{sat} = \mathbb{P}(X > C) \). 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 workload loss ratio can be identified as \( WLR = \mathbb{E}[(X - C)^+] / \mathbb{E}[X] \), where \( \mathbb{E}[] \) stands for the expectation value operator and \( (X - C)^+ = \max(X - C, 0) \). The estimation of this quantity can provide more accurate loss performance analysis. This measure better characterizes the expected loss rate and could also contribute to determining the users’ satisfaction. From traffic management (e.g., connection admission control) point of view an important question can arise: Whether the ongoing session (possibly together with a newcomer) satisfies a predefined QoS constraint related to some quality of service measure? In a more formal way, the inequalities

\[
\mathbb{P}(X > C) \leq e^{-\gamma} , \quad \frac{\mathbb{E}[(X - C)^+]}{\mathbb{E}[X]} \leq e^{-\gamma} \tag{1}
\]

represent the fulfillment of the constraint on saturation probability and workload loss ratio, respectively.

The Chernoff bounds for \( P_{sat} \) and \( WLR \) are as follows [4]:

\[
\mathbb{P}(X > C) \leq \inf_{s>0} \frac{G_X(s)}{e^{sC}} = \inf_{s>0} \exp (A_X(s) - sC) , \tag{2}
\]

\[
WLR \leq \exp (A_X(s^*) - s^*C - \log(s^*M)) , \tag{3}
\]

where

\[
s^* = \arg\inf_s (A_X(s) - sC) , \tag{4}
\]

\( G_X(s) \) and \( A_X(s) \) are the probability generating function (PGF) and the cumulant generating function (CGF) of \( X \), respectively. The 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.

In Section II the underlying probability generating function approximations are described. Previously known and newly developed bounds are presented under a common framework in Section III. 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.

II. APPROXIMATIONS OF PROBABILITY GENERATING FUNCTIONS (PGF)

In this section, 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 \( (M = \mathbb{E}[X]) \).

A. Approximations Based on Hoeffding’s results

The following lemmas are based on Hoeffding’s results (1963) [5] on the PGF approximation of bounded random variables.

Lemma I: Let \( X_i, i = 1 \ldots n \) be independent random variables with \( X = \sum_{i=1}^{n} X_i, M = \mathbb{E}[X] \) and \( 0 \leq X_i \leq p_i \).

This work was supported by the Ministry of Education, Hungary under the reference No. IKTA-0092/2002.
Then, for $s > 0$

$$G_X(s) \leq \left( 1 - \frac{M}{np} + \frac{M}{np} \exp(sp) \right)^n. \quad (5)$$

**Lemma 2:** Let $X_i$, $i = 1 \ldots n$ be independent random variables with $X = \sum_{i=1}^{n} X_i$, $M = E[X]$ and $0 \leq X_i \leq p_i$. Then for $s > 0$,

$$G_X(s) \leq \exp(sM) \exp \left( \frac{s^2 \sum_{i=1}^{n} p_i^2}{8} \right). \quad (6)$$

The PGF bound in (5) is applicable for random variables bounded uniformly (but with not necessarily identical distributions), while that in (6) covers a more general case with non-uniformly bounded (that is having different $p_i$ values) random variables. Nevertheless, the latter one does not coincide the former one in the special case of $p_1 = p_2 = \ldots = p_n$. This fact motivated the construction of an improved upper bound to be presented in the next subsection.

**B. An Improved Hoeffding-type Approximation**

Applying the Chernoff-Hoeffding bounding method on $X$ in a different way, we have obtained the following conservative bound for $G_X(s)$.

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

$$G_X(s) \leq \prod_{i=1}^{n} \left( 1 + \frac{m_i e^{sp} - 1}{p_i} \right), \quad (7)$$

The proof of this theorem is not detailed here, it can be found in [2]. Nevertheless, an important step in the chain of bounding formulae is worth repeating here:

$$G_X(s) = E[e^{sX}] \leq \prod_{i=1}^{n} \left( 1 + \frac{m_i e^{sp} - 1}{p_i} \right), \quad (8)$$

where $m_i \overset{def}{=} E[X_i]$.  

**Corollary 1:** The PGF bound on the right-hand side in (8) is the exact generating function of the sum of heterogeneous on-off random variables with the distribution

$$P\left(X_{i}^{\text{onoff}} = p_i\right) = \frac{m_i}{p_i}, \quad P\left(X_{i}^{\text{onoff}} = 0\right) = 1 - \frac{m_i}{p_i}.$$  

This is because

$$E[e^{sX_{i}^{\text{onoff}}}] = \left( 1 - \frac{m_i}{p_i} + \frac{m_i e^{sp_i}}{p_i} \right).$$

It can also be seen that the formula on the right hand side in (7) gives back the bound in (5) in the case of uniformly bounded random variables. In this way, this improved Hoeffding-type approximation is a consistent extension of the result presented in Lemma 1.

**C. A PGF Approximation Based on Stochastic Ordering**

In this subsection let us recall the essential definitions and properties of a certain type of stochastic ordering of random variables to be applied for PGF approximation.

**Definition 1 ([11]):** Let $X$ and $Y$ denote two random variables with distribution function $F_X$ and $F_Y$, respectively. $X$ is said to be smaller than $Y$ with respect to increasing convex ordering, written as $X <_{\text{icx}} Y$, if the condition

$$\int_{-\infty}^{\infty} \phi(x)dF_X(x) \leq \int_{-\infty}^{\infty} \phi(x)dF_Y(x)$$

holds for all increasing convex function $\phi$, for which the integral exists.

An important consequence of this definition for probability generating functions of random variables is the following:

**Lemma 3:** Let $X$ and $Y$ be two random variables with the relation $X <_{\text{icx}} Y$. Then for $s > 0$, $G_X(s) \leq G_Y(s)$.

This can be justified by the substitution $\phi(x) = \exp(sx)$.

The following results presented in [1] lead us to construct a new PGF bound.

**Lemma 4:** Let the random variables $X_1^{\text{onoff}}, \ldots, X_n^{\text{onoff}}$ represent $n$ independent heterogeneous on-off sources with peak rates $p_1, \ldots, p_n$ and mean rates $m_1, \ldots, m_n$. Let $Y_1^{\text{onoff}}, \ldots, Y_n^{\text{onoff}}$ be $n$ independent homogeneous on-off sources with the identical peak rate $p = \max(p_i, i = 1, \ldots, n)$, $nY = \lceil \sum_{i=1}^{n} p_i / p \rceil$, and identical mean rate $m = \sum_{i=1}^{n} m_i / n$. Then $X_1^{\text{onoff}} <_{\text{icx}} Y_1^{\text{onoff}}$, where

$$X_1^{\text{onoff}} \overset{def}{=} \sum_{i=1}^{n} X_i^{\text{onoff}}$$

and $Y_1^{\text{onoff}} \overset{def}{=} \sum_{i=1}^{n} Y_i^{\text{onoff}}$. \quad (9)

For a proof of Lemma 4 see [1].

Now, the PGF bound based on increasing convex stochastic ordering can be formulated in the following theorem [6]:

**Theorem 2:** 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 $M = E[X]$. Then for $s > 0$,

$$G_X(s) \leq \left( 1 - \frac{M}{np} + \frac{M}{np} e^{sp} \right)^n. \quad (10)$$

**Proof of Theorem 2:** By Corollary 1 we have $G_X(s) \leq G_{X^{\text{onoff}}}(s)$, $\forall ~ s > 0$. Further, by combining Lemma 4 and Lemma 3 the following relation also holds: $G_{X^{\text{onoff}}}(s) \leq G_{Y^{\text{onoff}}}(s)$, $\forall ~ s > 0$. Because

$$G_{Y^{\text{onoff}}}(s) = \left( 1 - \frac{M}{np} + \frac{M}{np} e^{sp} \right)^n,$$

the two inequalities above give the statement of the theorem. Q.E.D.

Let the PGF approximations presented in (6), (7) and (10), be designated by $G_{X,\text{hoe}}(s)$, $G_{X,\text{ih}}(s)$ and $G_{X,\text{so}}(s)$, respectively. The corresponding cumulant generation functions (CGF’s) are $A_{X,\text{hoe}}(s)$, $A_{X,\text{ih}}(s)$ and $A_{X,\text{so}}(s)$.

**III. CONSERVATIVE ESTIMATES FOR THE QOS MEASURES**

**A. Bounds for the saturation probability $P_{\text{sat}}$**

Applying the Chernoff bound for the saturation probability of uniformly bounded random variables with the CGF bound in (5) the following bound can be obtained [5]:

**Theorem 3:** Let $X_i$, $i = 1 \ldots n$ be independent random variables with $X = \sum_{i=1}^{n} X_i$, $M = E[X]$ and $0 \leq X_i \leq p$. Then, for $C > M$,

$$P(X > C) \leq \left( \frac{M}{C} \right)^{\frac{C}{np}} \left( \frac{np - M}{np - C} \right)^{n - \frac{C}{np}}. \quad (11)$$
In this case the optimal $s$ parameter can also be expressed as

$$s^* = \frac{1}{p} \log \frac{C}{n^{-p} - M}.$$

$$s^* = \frac{1}{p} \log \frac{C}{M}.$$

In the more interesting case of non-uniformly bounded random variables the following bounds can be obtained by the substitution of $\tilde{A}_{\text{hoe}}(s)$, $\tilde{A}_{\text{inh}}(s)$ and $\tilde{A}_{\text{iso}}(s)$ into the Chernoff bound of $P_{\text{sat}}$ (2).

**Theorem 4 ([5]):** Let $X_i$ be independent bounded random variables with $0 \leq X_i \leq p_i$, $X = \sum_{i=1}^{n} X_i$, and $M = E[X]$, then

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

The optimizing parameter $s$ can be formulated here as

$$s^* = \frac{4(C - M)}{\sum_{i} p_i}.$$

**Theorem 5 ([2]):** If $X_1, X_2, \ldots, X_n$ are independent (and not necessarily identically distributed) random variables, for which $0 \leq X_i \leq p_i$ holds, then

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

where $s^*$ is the solution of the following equation:

$$\sum_{k=1}^{n} e^{s^* p_k} - 1 - \frac{n \sum_{j=1}^{n} e^{s^* p_j}}{M + \sum_{j=1}^{n} e^{s^* p_j}} = C = 0.$$

Unfortunately, in this case neither the optimizing parameter $s^*$ nor the bound of $P_{\text{sat}}$ can be expressed in closed form. In [2] closed-form solutions have been developed for $P_{\text{sat}}$ through finding closed-form suboptimal solutions of the equation above with respect to $s$. One of those bounds is as follows:

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

where

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

As a new contribution of this paper, applying the CGF approximation $\tilde{A}_{\text{iso}}(s)$ based on stochastic ordering, inherently a closed form upper bound can be obtained for $P_{\text{sat}}$.

**Theorem 6:** Let $X_i$ be independent bounded random variables with $0 \leq X_i \leq p_i$, $X = \sum_{i=1}^{n} X_i$ and $M = E[X]$. Further, let $p = \max(p_i, i = 1, \ldots, n)$, $n_Y = \lfloor \sum_{i=1}^{n} p_i / p \rfloor$, and $m = \sum_{i=1}^{n} m_i / n_y$, then

$$P(X > C) \leq \left( \frac{M}{C} \right)^{n_Y} \left( \frac{n_Y p - M}{n_Y p - C} \right)^{n_Y - \frac{n_Y}{p}}.$$

**Proof sketch of Theorem 6:** Combining the results of Theorem 2 and Theorem 3 gives the required statement.

In this case the optimal $s$ parameter can also be expressed as

$$s^* = \frac{1}{p} \log \frac{C}{M}.$$

**B. Bounds for the workload loss ratio WLR**

Turning to the WLR approximation, here it is worth using formula (3), because in this case the optimizing parameter $s$ and hence the resulted conservative upper bounds can be expressed directly in closed-form, when the CGF approximations $\tilde{A}_{\text{X,hoe}}(s)$ and $\tilde{A}_{\text{X,iso}}(s)$ are embedded in (3). When the CGF bound $\tilde{A}_{\text{X,inh}}(s)$ is used similar sub-optimal solutions can be obtained as in the corresponding $P_{\text{sat}}$ bound in (17).

In the following theorem we summarize these closed form conservative bounds:

**Theorem 7:** 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 $K = \frac{1}{2} \sum_{i=1}^{n} p_i^2 - \frac{1}{2} (M - \frac{1}{2} \sum_{i=1}^{n} p_i)^2$, $p = \max(p_i, i = 1, \ldots, n)$, $n_Y = \lfloor \sum_{i=1}^{n} p_i / p \rfloor$, and $m = \sum_{i=1}^{n} m_i / n_y$, then the following three inequalities hold for WLR.

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

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

**Proof sketch of Theorem 7:** Substituting the three CGF approximations $\tilde{A}_{\text{X,hoe}}(s)$, $\tilde{A}_{\text{X,inh}}(s)$ and $\tilde{A}_{\text{X,iso}}(s)$ and the corresponding optimization parameters $s^*$ performed in equations (14), (18), (20) into the Chernoff bound of WLR in (3), the three bounds above are obtained.

In the set of bounds presented above the ones in (13) and (17) are already known from [5] and [2], but, to the authors best knowledge, the bounds in (19), (21), (22) and (23) are neither presented nor analyzed previously.

**IV. 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 I. The first traffic mix (Mix 1) resembles the aggregation of uncompressed voice and compressed video flows. The second (Mix 2) and third one (Mix 3) represent the multiplexing of uncompressed and compressed voice traffic with low and high 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, improved Hoeffding and stochastic ordering-based 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 (13), (21) based on the CGF approximation (6) 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 $\gamma$ (with tightening the QoS constraint).

Remarks: Although all the bounds presented require the same amount of information on 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. Nevertheless, these bounds can result in saving considerable amount of capacity compared to the outrageous peak rate reservation scheme (see the horizontal distances between the corresponding curves), and the application of them could be recommended when the simplicity is an exclusive criterion.

The bounds (17), (22) based on the CGF approximation (7) 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 improved Hoeffding-based saturation probability bound can have better performance in some cases (not seen in the figures) than the stochastic ordering-based one (see Observations), but the gain in capacity savings is not in proportion to the higher complexity of implementation.

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.

V. CONCLUSION

In this paper a family of bounds on saturation probability and workload loss ratio has been set up under the bufferless fluid flow multiplexing framework. This family comprises previously known as well as newly developed bounds. According to the analysis the stochastic ordering-based bounds have the best performance, especially in the case of workload loss ratio approximation. Nevertheless, the simple Hoeffding-based bounds could still form viable alternative from implementation point of view.
APPENDIX: POSSIBLE IMPROVED BOUNDS FOR INDEPENDENT REGULATED FLOWS MULTIPLEXED INTO A SERVICE CURVE NETWORK ELEMENT

In [7] and [8] performance bounds (bounds on buffer overflow probability) have been derived when independent and regulated traffic flows have been multiplexed into a buffer with fixed service rate. These bounds have been extended in [9] to the case of a more general multiplexer model, i.e. the service is characterized by a general service curve 1.

All the buffer overflow bounds presented in [7], [8] and [9] relies on the use of Hoeffding’s inequalities [5], also presented in this paper in equation (11) and (13) for the homogeneous and heterogeneous case, respectively. Our closed form bounds presented in (17) and (19) are also based on and improve one of the Hoeffding’s inequalities (13), and use the same amount of information. Therefore, they can apparently be used for further improving the results of Vojnovic and Le Boudec in [9] in the case of heterogeneously regulated traffic multiplexed into a service curve network element. For the same reason, improved bounds on buffer workload loss ratio can also be set up by the use of (22) and (23). The performance evaluation of the improved bounds in the buffered multiplexer context is a matter of future work.

1In this model the multiplexer is referred to as service curve network element.

REFERENCES