Fast Algorithms for Computing Parsimonious Estimates of QoS Measures

József Bíró[1], Non-member

ABSTRACT

In recent publications concerning quality of service (QoS) measures in communication networks several attempts have been made to improve the techniques of computing or estimating the quality of data transfer. Many of the promising results are based on the knowledge of only very few parameters on the traffic situation, some of which are known *a priory*, others are measured. The paper deals with such parsimonious estimation techniques of QoS measures under the bufferless fluid flow multiplexing (bffm) framework, and concentrates on the efficient computation of them. The proposed fixed-point recursive algorithms are key elements in applying either the implicit formulae or ones containing optimization tasks in real-time environment.

1. INTRODUCTION

Parsimonious estimation of QoS measures is becoming a popular approach of introducing premium services into traditional traffic management of broadband communication networks. The motivation behind the idea is that the available or measurable characteristics of a traffic situation of a network is often very limited in real environment. This is especially true when fast decisions have to be made according to the findings, when e.g. stream-type traffic flows [1] are handled having strict timing constraints. Typical parsimonious techniques assume knowing only first or second order statistics of traffic flows.

This paper deals with the estimation of the overflow probability as an important QoS measure and the corresponding equivalent capacity referring to the bandwidth requirement of traffic flows introducing targeted link saturation. The investigation is made under the bufferless fluid flow multiplexing (bffm) framework, which is a suitable modeling approach for scenarios with high network utilization with large number of traffic flows multiplexed. In the framework, the effect of buffers throughout data transmission is disregarded, which is proved to be a suitable modeling approach in numerous traffic scenarios [2] [3], where buffering is limited to a packet level as opposed to burst time scales due to e.g. strict delay con-

straints. The use of bffm to dimension transmission link capacity or design decision rules for connection admission control algorithms requires the estimation of the overflow probability or the expected workload loss ratio as possible measures of the quality of service. The estimate is then to be compared to a predefined QoS level. A related highly non-trivial question is how to obtain a direct and optimized estimate for the minimum bandwidth required (i.e. the equivalent capacity) to achieve the predefined QoS level [4], [5].

More specifically, this paper focuses on a single link abstraction, where a number of independent and stationary traffic sources are served, each represented by its instantaneous transmission speed. The link performance can be investigated by considering the aggregate data rates of the flows and the available capacity, in the context of e.g. the link overflow probability. The presented QoS measure estimates, in the paper, require only the knowledge of the aggregate mean rate of the traffic on the link, the number of flows, and the maximum instantaneous rate of each individual flow.

After a short summary of the used terms and definitions, in the next section, two conservative upper bounds on the probability generating function (PGF) of the aggregate traffic rate distribution is presented. These upper bounds are then used to build link-overflow probability and corresponding bandwidth requirement estimators applying the Chernoff bounding method [6], which introduces an optimization task in the formulation to be performed. In a few cases, this task can be simplified by solving an explicit equation, in others, it is done through suitable reformulation and further approximation, however, in many cases none of these actions can be done in a reasonable way. In the last section, efficient, fast converging recursive algorithms are proposed to solve the implicit formulae or ones containing optimization tasks making it possible to apply them in real-time traffic management.

2. APPROXIMATIONS OF PROBABILITY GENERATING FUNCTION (PGF)

For modeling purposes under bffm, let us assume that there are 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

[1] The author is with Department of Telecommunication and Media Informatics Budapest University of Technology and Economics, E-mail: biro@tmit.bme.hu

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 $X = \sum_{i=1}^n X_i$. A frequently investigated and estimated measure of transmission links is the saturation probability:

$$P_{sat} \overset{def}{=} P(X > C). \qquad (1)$$

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.

From traffic management (e.g. connection admission control) point of view two important questions can arise. First, whether the ongoing session (possibly together with a newcomer) satisfies a predefined QoS constraint i.e. a targeted overflow probability. In a more formal way, in this case the inequality

$$P(X > C) \leq e^{-\gamma}, \qquad (2)$$

represents the fulfillment of the pre-given constraint ($e^{-\gamma}$) on overflow probability. Secondly, a more significant question in practice is to identify the minimum capacity that must be allocated for the traffic flows, in order to satisfy the corresponding constraint in (2). Formally, it can be written as

$$C_{equ} \overset{def}{=} \inf \left\{ C : P_{sat} \leq e^{-\gamma} \right\}. \qquad (3)$$

One of the widely used and accepted techniques to approximate P_{sat} is the Chernoff bounding method as follows [7]:

$$P(X > C) \leq \inf_{s>0} \frac{G_X(s)}{e^{sC}} = \inf_{s>0} \exp(\Lambda_X(s) - sC), \quad (4)$$

where $G_X(s) \overset{def}{=} E[\exp(sX)]$ and $\Lambda_X(s) \overset{def}{=} \log G_X(s)$ are the probability generating function (PGF) and the cumulant generating function (CGF) of X, respectively and $E[.]$ is the expectation value operator. In several occasions the form

$$P(X \geq C_{equ}) \leq \inf_{s>0} e^{s(\alpha(s) - C)}, \qquad (5)$$

is also used, where $\alpha(s) = s^{-1} \log G_X(s)$ referred to as the effective bandwidth of aggregate traffic X.

The exact computation of these bounds is usually not possible, because the underlying generating functions would require the distribution of X to be known. Instead, the CGFs are to be further bounded based on the available information (moments) on X and embedded into the Chernoff bound. This is usually called the Chernoff-Hoeffding bounding method [9]. In the following, two conservative bounds of the PGF of aggregate traffic rate distribution is provided. The common property of the formulae that the necessary information about the link-traffic is limited to n, the number of traffic flows multiplexed, p_i, the

peak rates of the traffic flows, and $M \overset{def}{=} E[X]$, the aggregate mean arrival rate.

Theorem 1 ([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 $M = E[X]$. Then for $s > 0$,

$$G_X(s) \leq \left(\frac{M + \sum_{k=1}^n \dfrac{p_k}{e^{sp_k} - 1}}{n} \right)^n \prod_{i=1}^n \left(\frac{e^{sp_i} - 1}{p_i} \right). \qquad (6)$$

It is also shown in [12] that under mild conditions, the above upper bound, denoted by $G_{X,\text{ih}}(s)$ in the rest of the paper, is optimal, that is, no better can be given:

Theorem 2 ([12]) $G_{X,\text{ih}}(s)$, the PGF approximation of a sum of positive real valued random variables is optimal on a S set of s if:

1. All the information known are n the number of the random variables, the aggregate mean M and the individual maximum values $p_1, p_2, \ldots, p_n$ of the random variables.

2. The following inequality holds

$$0 \leq \frac{M - \sum_{k=1}^n \left(\dfrac{p_i}{e^{sp_i} - 1} - \dfrac{p_k}{e^{sp_k} - 1} \right)}{n} \leq p_i, \ \forall i, \ s \in S. \qquad (7)$$

Another PGF approximation technique is proposed in [13]:

Theorem 3 ([13]) 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 $s > 0$,

$$G_X(s) \leq \left(1 - \frac{M}{n_Y p} + \frac{M}{n_Y p} e^{sp} \right)^{n_Y} \overset{def}{=} G_{X,\text{so}} \qquad (8)$$

The optimality of $G_{X,\text{so}}$ with knowing no further information of the traffic is assured if the peak rates are equal.

According to the above theorems, $\Lambda_{X,ih}$, $\alpha_{X,ih}$, $\Lambda_{X,so}$ and $\alpha_{X,so}$ are defined correspondingly.

3. PARSIMONIOUS BANDWIDTH REQUIREMENT ESTIMATORS

At this point, combining the Chernoff formula in (4) and the two PGF approximations discussed in the previous section, we can construct two estimation technique for the overflow probability. The fist one using $G_{X,ih}$ is given in the following theorem:

Theorem 4 ([11]) 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

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

(9)

where s^* is the solution of the following equation.

$$\sum_{k=1}^{n} \frac{e^{s p_k} p_k}{e^{s p_k} - 1} - \frac{n \sum_{j=1}^{n} \frac{e^{s p_j} p_j^2}{(e^{s p_j} - 1)^2}}{M + \sum_{j=1}^{n} \frac{p_j}{e^{s p_j} - 1}} - C = 0. \quad (10)$$

The one related to $G_{X,so}$ is derived as:

Theorem 5: Let $X_1, \ldots, X_n$ indicate n independent random variables (e.g. transmission rates of communication sources) with $0 \le X_i \le p_i$, $X = \sum_{i=1}^{n} X_i$ and $M = E[X]$. Then for $s > 0$,

$$P(X > C) \le \left(\frac{M - n_Y p}{C - n_Y p} \right)^{n_Y - \frac{C}{p}} \left(\frac{M}{C} \right)^{\frac{C}{p}}, \quad (11)$$

where $p = \max(p_i, i = 1, \ldots, n)$, $n_Y = \lceil \sum_{i=1}^{n} p_i / p \rceil$. With respect to the computation of the equivalent capacity estimators that are related to a targeted overflow probability, the definition according to (3):

$$\widetilde{C}_{\text{equ}} \overset{def}{=} \inf \{ C : \inf_{s>0} \exp(\widetilde{\Lambda}_X(s) - sC) \le e^{-\gamma} \}. \quad (12)$$

For the computation of $\widetilde{C}_{\text{equ}}$ the following theorem [11] provides the guidelines:

Theorem 6: Let us assume, that we have a given (aggregate) traffic characterized so, that the effective bandwidth of X or any conservative bound on it is obtainable, which bound is denoted by $\hat{\alpha}_X(s)$. For a given link capacity C and saturation probability $e^{-\gamma}$, it is stated that

$$\inf_{s,s>0} e^{s(\hat{\alpha}_X(s) - C)} < e^{-\gamma} \Leftrightarrow \inf_{s,s>0} \hat{\alpha}_X(s) + \frac{\gamma}{s} < C \quad (13)$$

and

$$\inf_{s,s>0} e^{s(\hat{\alpha}_X(s) - C)} = e^{-\gamma} \Leftrightarrow \inf_{s,s>0} \hat{\alpha}_X(s) + \frac{\gamma}{s} = C. \quad (14)$$

Based on the above theorem the equivalent capacity estimator can be computed in a more efficient way as:

$$\widetilde{C}_{\text{equ}} = \inf_{s>0} \frac{\hat{\Lambda}_X(s) + \gamma}{s} \quad (15)$$

or

$$\widetilde{C}_{\text{equ}} = \inf_{s>0} \hat{\alpha}_X(s) + \frac{\gamma}{s}, \quad (16)$$

where $\hat{\Lambda}_X(s)$ ($\hat{\alpha}_X(s)$) is any kind of upper bound on the CGF (effective bandwidth) of X. Using $\Lambda_{X,ih}$ or $\Lambda_{X,so}$ ($\alpha_{X,ih}$ or $\alpha_{X,so}$) we get two different estimations for C_{equ}

4. COMPUTATIONAL ISSUES OF P_{SAT} AND C_{EQU}

In the previous section, a number of formulae have been discussed for the computation of the overflow probability and respective equivalent capacity values. It was also pointed out, that under certain conditions, the formulae provide the best available (Chernoff-type) upper bounds. Taking a closer look at the results, however, it can be seen, that apart from (5), the formulae are in implicit form or contain optimization tasks to be performed, which is a significant obstacle in using them in real-time environment. Although for the overflow probability there exists a few solution to overcome the difficulties in the evaluation [14], the attempts to yield upper bounds for C_{equ} in closed form, or even to make appropriate approximations of them encounter several difficulties. In these cases, to reach acceptable results, numerical approximation techniques are inevitable to use, yielding sometimes prohibitively increased processing time. In the following, first, a novel recursive algorithm is proposed to fasten numerical evaluation of the C_{equ} estimation formulae discussed. The focus is on the replacement of the optimization task with a fixed-point equation, constructed in a way that the number of necessary steps towards a good estimation is reduced significantly. The effectiveness of the algorithm is also demonstrated through numerical examples.

4.1 Fast calculation of equivalent capacity estimators

The proposed fixed-point algorithms in this section, specifically designed to solve Chernoff-based formulae connected to the saturation probability (that is derived from (4)). The ultimate purpose is to provide means of computing equivalent bandwidth according to (15), but without the compromise of finding appropriate closed form, yet suboptimal, approximations. As discussed previously, such type of approximations turned out to be rather difficult to find, due to the nature of the formulation. For saturation probability estimators, like (9), computationally feasible (closed-form) solutions are not too difficult to find, and so the proposed fixed-point algorithms are of less significance, nevertheless, in certain cases, the increased exactness of the new methods may have increased importance also. From the large number of existing techniques of obtaining an optimum point numerically, fixed-point type algorithms are chosen for several reasons. Most importantly this type of method in general provides very fast convergence, which is, as discussed previously, a fundamental requirement

in real-time decision making processes. Also an important condition is that due to the approximating nature of the bounds to be optimized, more than one derivation on the objective function would yield unreasonable loss of exactness, for which reason the fixed-point type optimum search algorithm is also suitable. Finally, as it will be seen, for almost all cases to be examined the solvable fixed-point equation comes naturally.

The proposed algorithms are presented in two steps: First, the straightforward and more simple versions will be derived. Then, due to the experienced bad convergence properties of the algorithms, next a modified technique is shown, that is capable of providing the same results significantly faster, and also it shows substantial improvement in stability.

Before proceeding to the derivation of the algorithms, first let us take a look at the exact problem to be solved. In the previous section, two type of formulae were composed for the problem of finding equivalent capacity belonging to fixed saturation probability, depending on the used characteristic feature of an aggregated traffic source (15) and (16). The importance of the distinction stems from the fact that in practice, in certain cases there is a possibility of measuring $\Lambda_X(s)$ as the CGF or $\alpha_X(s)$ as the effective bandwidth of X directly [4]. Depending on the two choices, as it will be seen that the results are different. First let us investigate the case of (15). Performing the optimization, we get

$$\frac{\partial_s \widetilde{C}_{\text{equ}}}{\partial s} = \frac{\frac{\partial_s \Lambda(s)}{\partial s} s - \Lambda(s) + \gamma}{s^2} = 0, \qquad (17)$$

yielding s_{opt} is the solution of the following equation

$$s = \frac{\Lambda(s) - \gamma}{\frac{\partial_s \Lambda(s)}{\partial s}} \quad . \qquad (18)$$

This last form naturally induces the following fixed-point equation

$$s_{n+1} = \frac{\Lambda(s_n) - \gamma}{\frac{\partial_s \Lambda(s)}{\partial s}\big|_{s=s_n}} \quad . \qquad (19)$$

Regarding to the stability of the algorithm above–due to $\Lambda(s)$ being undefined–general statements cannot be derived, only constraints can be shown with respect to $\Lambda(s)$ to be fulfilled. On the other hand using the PGF approximations proposed previously, a thorough investigation have been made through numerical examples using a wide set of parameter settings. According to the results, in many cases (19) is proved to be unstable, however, when being stable, the minimum has always been found.

To get an impression of the performance of the algorithm first let us take the example of the aggregation of ten theoretical traffic sources with diverse traffic parameters given in Table 1. [1]

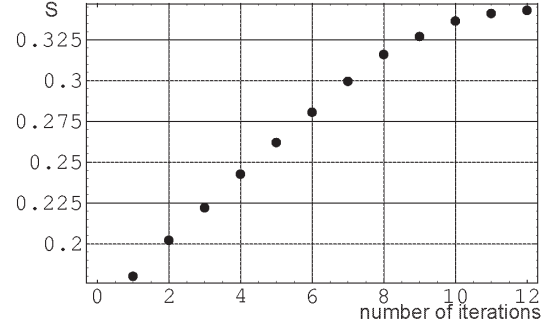m_i denotes the average arrival rate of flow i.

Fig.1: *Steps of fixed-point algorithm in scenario 2*

Let us use (6) as the bases for $\widetilde{\Lambda}_{X,\text{ih}}(s)$. We try to find $\widetilde{C}_{\text{equ}}$, when the maximum acceptable saturation probability is fixed at a moderate level of 10^{-3}. In this relatively simple case it is easy to compute that the minimum is reached at $s = 0.1886$ and the corresponding best (i.e. minimum) $\widetilde{C}_{\text{equ}} = 56.41$ Mbps. Now if starting the recursion according to (19) at an initial good guess of $s_1 = \frac{1}{n}$, we find that the results of the recursion at each steps are: $s_2 = 0.1010$, $s_3 = 0.1021$, $s_4 = 0.1032$, $s_5 = 0.1043$, $s_6 = 0.1054$. Putting out that iteration stops when the value at a subsequent step does not change more than 1%, we find that over 40 steps are needed for the algorithm.

Now for a more practical scenario let us take the traffic mix of the aggregation of uncompressed voice ($p_1 = 64$ kbit/s, $m_1 = 25.6$ kbit/s) and compressed video ($p_2 = 5$ Mbit/s, $m_2 = 2$ Mbit/s) flows with $n_1 = 100$, $n_2 = 10$. The data are summarized in Table 2.

Fixing the target saturation probability at 10^{-4} using (8) as the PGF approximation we can compute the optimal $\widetilde{C}_{\text{equ}} = 46.579$ at $s = 0.350$. The results of the fixed-point algorithm in each steps is illustrated in Fig. 1. As it is clearly seen, even in this relatively simple case, at least 12 steps were needed to reach the optimal point with acceptable accuracy. Unfortunately, as extensive numerical evaluation shows that in general this behaviour turns out to be typical for the fixed-point algorithm given in (19), which property is not acceptable in real-time usage. In complex cases, often even a few hundred steps are still insufficient to find the optimal value within an acceptable accuracy. Moreover, the algorithm frequently shows an unstable behavior, it manifests that s_n tends to infinity in the recursion.

To improve the low convergence rate and instability of the algorithm above a novel technique is proposed for (15). The main idea is to use a recursive technique, where we have the value of $\Lambda(s)$ and $\frac{\partial_s \Lambda(s)}{\partial s}$ at given points and at every step we fit these values onto the cumulant generation function of the Gaussian distribution (that has two free parameters: the mean and the variance) and then the optimization process is performed accordingly.

Table 1: *Traffic scenario 1: Data for a 10-source traffic aggregation (bandwidth is given in Mbps)*

	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10
p_i	2	3	5	7	7	9	9	11	11	15
m_i	0.6	0.9	1.5	2.1	1.4	2.7	3.4	3.0	3.6	4.5

The cumulant generating function of a Gaussian distributed random variable (m: mean, σ^2: variance) is

$$\Lambda^{\text{Gauss}}(s) = sm + \frac{s^2\sigma^2}{2} \ . \qquad (20)$$

Derivating the cumulant generating function with respect to s we get:

$$\frac{\partial_s \Lambda^{\text{Gauss}}(s)}{\partial s} = m + s\sigma^2. \qquad (21)$$

Solving the two-equation system to m and σ^2 (ie. adding (20) to $-\frac{s}{2}$ times (21) to get the mean and adding -1 times (20) to s times (21) to get the variance) we can obtain

$$m = \frac{2}{s}\Lambda^{\text{Gauss}}(s) - \frac{\partial_s \Lambda^{\text{Gauss}}(s)}{\partial s} \qquad (22)$$

and

$$\sigma^2 = -\frac{2}{s^2}\Lambda^{\text{Gauss}}(s) + \frac{2}{s}\frac{\partial_s \Lambda^{\text{Gauss}}(s)}{\partial s}. \qquad (23)$$

Performing the optimization task in the formula of the equivalent capacity of the Gaussian distributed random variable:

$$\widetilde{C}_{\text{equ}}^{\text{Gauss}} = \inf_s \frac{\Lambda^{\text{Gauss}}(s) + \gamma}{s} = \inf_s m + \frac{s\sigma^2}{2} + \frac{\gamma}{s} \quad (24)$$

and

$$\frac{\partial_s \widetilde{C}_{\text{equ}}^{\text{Gauss}}}{\partial s} = \frac{\sigma^2}{2} - \frac{\gamma}{s^2} = 0. \qquad (25)$$

The results are

$$s_{\text{opt}} = \sqrt{\frac{2\gamma}{\sigma^2}} \qquad (26)$$

and

$$\widetilde{C}_{\text{equ}}^{\text{opt}} = m + \sqrt{\frac{2\gamma}{\sigma^2}}\frac{\sigma^2}{2} + \frac{\gamma}{\sqrt{\frac{2\gamma}{\sigma^2}}} = m + \sqrt{2\sigma^2\gamma}. \quad (27)$$

Now applying the fitting conditions at s_1, the first point of a recursive numerical process $\Lambda^{\text{Gauss}}(s_1) = \Lambda(s_1)$ and $\frac{\partial_s \Lambda^{\text{Gauss}}(s)}{\partial s}\mid_{s=s_1} = \frac{\partial_s \Lambda(s)}{\partial s}\mid_{s=s_1}$, we obtain the parameters of the Gaussian distributed random variable at the first step:

$$m_1 = \frac{2}{s_1}\Lambda(s_1) - \frac{\partial_s \Lambda(s)}{\partial s}\mid_{s=s_1} \qquad (28)$$

and

$$\sigma_1^2 = -\frac{2}{s_1^2}\Lambda(s) + \frac{2}{s_1}\frac{\partial_s \Lambda(s)}{\partial s}\mid_{s=s_1} \ . \qquad (29)$$

For the following steps of the computation, from (26) we get our resulting fixed-point type recursive equation:

$$s_{n+1} = \sqrt{\frac{2\gamma}{\sigma_1^2}} = \sqrt{\frac{\gamma s_n^2}{-\Lambda(s_n) + s_n\frac{\partial_s \Lambda(s)}{\partial s}\mid_{s=s_n}}}. \quad (30)$$

For the investigation of the performance of (30), let us take e.g. the 10-source traffic aggregation example set up previously. We find that starting from $s_1 = 0.1$, we get $s_2 = 0.1741$ and $s_3 = 0.1883$, which, according to our stop-rule above, can be considered as the last step of the algorithm as opposed to the result of over 40 steps using (19). For a more complex scenario now, we take a traffic mix of 10 traffic classes with widely diverse traffic parameters, beginning with compressed VBR audio at 16kbps and H263 video teleconferencing at 64 kbps to MPEG2 HDTV and Full HDTV at 14.5 and 19.4 Mbps. The configuration data is detailed in Table 3. Applying (6) and fixing the saturation probability at $\gamma = 4$, the results in Fig. 2 clearly show that the novel fixed-point algorithm in (30) provides a fairly accurate optimum point even on the third step. On the other hand (19) does not even provide a stable result. Now investigating the equivalent capacity formula (16) with the given effective bandwidth $\alpha(s)$ we have $\widetilde{C}_{\text{equ}} = \inf_s \alpha(s) + \frac{\gamma}{s}$, and the resulting fixed-point equation to be solved becomes:

$$s_2 = \sqrt{\frac{\gamma}{\frac{\partial_s \alpha(s)}{\partial s}\mid_{s=s_1}}} \ . \qquad (31)$$

Making the Gaussian distributed traffic substitution as discussed above the respective equation system is:

$$\alpha(s) = m + \frac{s\sigma^2}{2} \qquad (32)$$

and

$$\frac{\partial_s \alpha(s)}{\partial s} = \frac{\sigma^2}{2} \ . \qquad (33)$$

Table 2: *Traffic mix of uncompressed voice and compressed video*

	n_1	m_1	p_1	n_2	m_2	p_2
Scenario 2	100	0.0256	0.064	10	2	5

Table 3: *Traffic mix consisting 10 traffic classes (bandwidth given in Mbps)*

	Class 1	Class 2	Class 3	Class 4	Class 5
n_i	200	100	100	50	50
m_i	0.008	0.0256	0.064	0.7	1
p_i	0.016	0.064	0.256	1.2	2.5
	Class 6	Class 7	Class 8	Class 9	Class 10
n_i	30	20	10	10	10
m_i	2	3.5	5	7	8.5
p_i	5	8	12	14.5	19.4
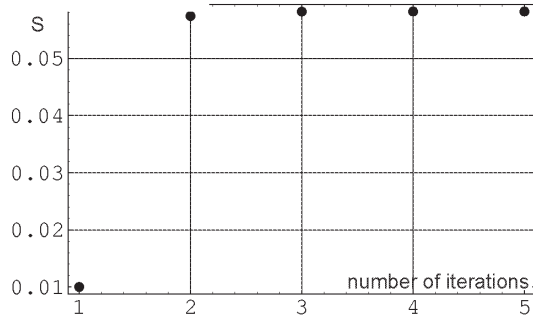
Fig.2: *Steps of novel fixed-point algorithm on traffic mix of 10 classes*

The resulting mean and variance in this case become:

$$m = \alpha^{\text{Gauss}}(s) - s\frac{\partial_s \alpha^{\text{Gauss}}(s)}{\partial s} \qquad (34)$$

and

$$\sigma^2 = 2\frac{\partial_s \alpha^{\text{Gauss}}(s)}{\partial s} \ . \qquad (35)$$

Finally, using the effective bandwidth of the Gaussian distributed traffic

$$\widetilde{C}_{\text{equ}} = \inf_s \alpha(s) + \frac{\gamma}{s} = \inf_s m + \frac{s\sigma^2}{2} + \frac{\gamma}{s} \qquad (36)$$

with $s_{\text{opt}} = \sqrt{\frac{2\gamma}{\sigma^2}}$, we obtain the new fixed-point equation

$$s_{n+1} = \sqrt{\frac{\gamma}{\frac{\partial_s \alpha(s)}{\partial s}\ |_{s=s_n}}} \quad , \qquad (37)$$

which interestingly turns out to be the same as in the original algorithm. In this case, our advanced method does not yield a different result.

4.2 Fixed-point equations for the computation of the overflow probability

As it was mentioned, for the (Chernoff-type) formulae estimating the link overflow probability, there exists closed-form solution in several cases depending on the PGF approximation used. Also, even, when no closed-form solution can be found appropriate actions can be taken to reach reasonable estimations of it.

Nevertheless it is not pointless to investigate numerical evaluation methods for the formulae for two main reasons. Firstly, with numerically computed exact results the accuracy in several occasions can be significantly improved. Secondly, if the cumulant generating function or the effective bandwidth of the aggregate traffic is directly established/measured (instead of approximating from a few given traffic parameters) numerical evaluation, in general, is obviously necessary. Here we briefly discuss the fixed-point equations for the computation of the saturation probability for both cases of given $\Lambda(s)$ and $\alpha(s)$.

Considering the first case we have $\Lambda(s)$, and the solvable equation for $P_{\text{sat}} = e^{-\gamma}$ is $\gamma = \inf_s \Lambda(s) - sC$.

It is easy to see that in this case fixed-point equation cannot be established as previously:

$$0 = C - \frac{\partial_s \Lambda(s)}{\partial s} \ . \qquad (38)$$

On the other hand using the new method proposed in the previous section we obtain $s_{\text{opt}} = \frac{C-m}{\sigma^2}$, and using (22) and (23) we get the resulting recursive equation:

$$s_{n+1} = \frac{s_n{}^2(C + \frac{\partial_s \Lambda(s)}{\partial s}\ |_{s=s_n}) - 2s_n \Lambda(s_n)}{2s_n\frac{\partial_s \Lambda(s)}{\partial s}\ |_{s=s_n} - 2\Lambda(s_n)} \ . \qquad (39)$$

Following the same considerations, using $\alpha(s)$ in the computation of the saturation probability as $\gamma = \inf_s s(\alpha(s) - C)$, we obtain the following fixed-point equation:

$$s_{n+1} = \frac{C - \alpha(s_n)}{\frac{\partial_s \alpha(s)}{\partial s}\ |_{s=s_n}} \ . \qquad (40)$$

Now, the result for the Gaussian distributed traffic substituted version, using (34) and (35) would be:

$$s_2 = \frac{C - \alpha(s_1) + s\frac{\partial_s \alpha(s)}{\partial s}\ |_{s=s_1}}{2\frac{\partial_s \alpha(s)}{\partial s}\ |_{s=s_1}} \ . \qquad (41)$$

For a summary Table 4 provides the obtained fixed-point equations for the different cases discussed in the paper. The two columns are for the type of measure, which the formula is targeted (that is γ, referring to the overflow probability or $\widetilde{C}_{\text{equ}}$, as the equivalent capacity), and the four rows stand for the type of function (i.e. $\Lambda(s)$ or $\alpha(s)$) the formula uses. $\Lambda_{\text{Gauss}}(s)$ and $\alpha_{\text{Gauss}}(s)$ indicate the Gaussian distributed traffic substitution method.

4.3 Suitability for real-time traffic management

In real-time traffic management the accurate and simple enough estimations of QoS measures play central role. In other words, a good engineering trade-off between accuracy and simplicity is the prerequisite of the estimators' applicability. For example, connection admission control for loss sensitive traffic requires a decision rule based on either the estimation of the

Table 4: *Summary of the fixed-point equations for γ and $\widetilde{C}_{equ}$*

	γ	$\widetilde{C}_{\text{equ}}$
$\Lambda(s)$	$0 = C - \dfrac{\partial_s \Lambda(s)}{\partial s}$	$s_{n+1} = \dfrac{\Lambda(s_n)+\gamma}{\frac{\partial_s \Lambda(s)}{\partial s}\big|_{s=s_n}}$
$\Lambda_{\text{Gauss}}(s)$	$s_{n+1} = \dfrac{s_n{}^2(C+\frac{\partial_s\Lambda(s)}{\partial s}\big|_{s=s_n})-2s_n\Lambda(s_n)}{2s_n\frac{\partial_s\Lambda(s)}{\partial s}\big|_{s=s_n}-2\Lambda(s)}$	$s_{n+1} = \sqrt{\dfrac{\gamma s_n^2}{-\Lambda(s_n)+s_n\frac{\partial_s\Lambda(s)}{\partial s}\big|_{s=s_n}}}$
$\alpha(s)$	$s_{n+1} = \dfrac{C-\alpha(s_n)}{\frac{\partial_s\alpha(s)}{\partial s}\big|_{s=s_n}}$	$s_{n+1} = \sqrt{\dfrac{\gamma}{\frac{\partial_s\alpha(s)}{\partial s}\big|_{s=s_n}}}$
$\alpha_{\text{Gauss}}(s)$	$s_{n+1} = \dfrac{C-\alpha(s_n)+s\frac{\partial_s\alpha(s)}{\partial s}\big|_{s=s_n}}{2\frac{\partial_s\alpha(s)}{\partial s}\big|_{s=s_n}}$	$s_{n+1} = \sqrt{\dfrac{\gamma}{\frac{\partial_s\alpha(s)}{\partial s}\big|_{s=s_n}}}$

equivalent capacity of the ongoing flows together with the newcomer or the estimation of the link saturation probability caused by the statistical fluctuation of traffic flows.

As argued in previous subsections the Chernoff bounding method under the bufferless fluid flow multiplexing can provide accurate enough formulae for estimating the saturation probability or the equivalent capacity, even if very few pieces of information is available on traffic flows. In previous works the simplicity is provided by closed form formulae which resulted in further inaccuracy, the more accurate numerical evaluations of QoS measures were not preferred due to their computationally more involved implementations [8], [11], [12].

An important step towards the applicability of the numerically evaluated estimators in on-line decision problems have been the statistically reliable on-line estimations of logarithmic moment generating function and its first derivative of traffic flows [15], [3]. Embedding this estimators in our dramatically fast converging and stable recursive formulae, real-time estimation of saturation probability or equivalent capacity becomes possible. This result can be considered as an efficient building block in and an important step towards real-time traffic management, however, this does not solve every implementation of complexity problem in traffic management.

5. CONCLUSIONS

In the paper, parsimonious link performance measure estimators have been discussed. More specifically the overflow probability and the equivalent capacity were investigated. The analyzed estimators are based on very limited information, requiring only the knowledge of the aggregate mean arrival rate of the link-traffic, the number of flows on the link, and the individual peak rates. As the main contribution of the paper, extremely fast converging optimum-finding algorithms were derived to eliminate

the evaluation complexities of the provided formulae. According to thorough numerical investigations, the proposed novel fixed-point type algorithms seem to be an important step towards enabling the (Chernoff-type) estimators to be used in real-time traffic management.

References

[1] J. Roberts, "Engineering for Quality of Service," J. W. Roberts. Engineering for quality of service. In K. Park and W. Willinger, editors, *Self-Similar Network Trac and Performance Evaluation*. Wiley-Interscience, 2000.

[2] P Camarda, D Striccoli, "Bandwidth Estimation of Smoothed VBR Video Streams Exploiting Queueing Networks Theory," *Proceeding of SPECTS'02*, San Diego, California, USA, July 2002.

[3] M. Reisslein, "Measurement-Based Admission Control: A Large Deviations Approach for Bufferless Multiplexers," *Proceeding of ISCC 2000*, Antibes, France, July 2000.

[4] G. Seres, Á. Szlávik, J. Zátonyi, J. Bíró, "Quantifying resource usage - a large deviation-based approach," *IEICE Transact. Communicat. (Jpn.)* E85-B (1), pp. 938–952, 2002.

[5] S. Imre, K. Hankó, P. Petrás, R. Táncsics, "Optimized effective bandwidth based admission control for multi-service cdma cellular networks," *Proceeding of the 10th SoftCOM2001*, Split, Dubrovnik (Croatia), Ancona, Venice (Italy), pp. 299–304, 2001.

[6] H.Chernoff,"A Measure of Asymptotic Efficiency of Tests of a Hypothesis Based on the Sum of Observations," *Ann. Math. Statist.* 23, pp. 493–507, 1952.

[7] F.P. Kelly, "Notes on effective bandwidths," in: F. Kelly, S. Zachary, I. Ziedins (Eds.), *Stochastic Networks: Theory and Applications,* Oxford University Press, Oxford, pp. 141–168, 1996.

[8] M. Grossglauser and D. Tse. A framework for robust measurement-based admission control. In *ACM SIGCOMM Symposium on Communications Architectures and Protocols*, pages 237–248, Cannes, France, September 1997.

[9] J. Schmidt, A. Siegel, A. Srinivasan, "Chernoff-Hoeffding bounds for applications with limited independence," *Proceeding of the ACM-SIAM Symposium on Discrete Algorithms,* 1995.

[10] C. Courcoubetis, V. A. Siris, and G. D. Stamoulis. Application of the many sources asymptotic and effective bandwidths to traffic engineering. *Telecommunication Systems*, 12:167–191, 1999.

[11] Z. Heszberger, J. Zátonyi, and J. Bíró, "Efficient Chernoff-based resource assessment techniques in multi-service networks," *Telecommunication Systems*, Volume 20, Issue 1,2, Special Issue: Wide Area Networks Design and Analysis, pp. 59–80, 2002.

[12] Z. Heszberger, J. Zátonyi, and J. Bíró, "Performance bounds for rate envelope multiplexing," *Performance Evaluation*, Volume 48, Issue 1, Special Issue on ATM & IP Networks: Performance Modelling and Analysis, pp. 87–101, May 2002.

[13] J. Bíró, "Novel Equivalent Capacity Approximation Through Asymptotic Loss Analysis," *Computer Communications*, Special Issue on ATM & IP Networks: Performance Modelling and Analysis, in press.

[14] Z. Heszberger, J. Zátonyi, J. Bíró, and T. Henk, "Efficient bounds for bufferless statistical multiplexing", Proceeding of *IEEE Global Telecommunications Conference - GLOBECOM 2000*, volume 1, pp. 641–646, San Francisco, CA, USA, Nov./Dec. 2000.

[15] J. Bíró, "A New Scalable Estimation for the Logarithmic Moment Generating Function of Network Traffic Flows", *International Journal of Computer Science*, Vol. 7. , No. 12. pp. 140-145., 2007.

József Bíró received his M.Sc. and Ph.D. degrees in electrical engineering and computer science from Budapest University of Technology and Economics (BUTE), Budapest, Hungary in 1993 and 1998, respectively. In 1996 he joined to Department of Telecommunications and Media Informatics, where he is currently working as an Associate Professor. His research interests include network performance modeling and evaluation, and the support of multimedia applications over heterogeneous networks.