Novel techniques for assessing resource requirements in packet-based networks

MÁTYÁS MARTINECZ, JÓZSEF BÍRÓ, ZALÁN HESZBERGER

martinecz@tmit.bme.hu

The authors are supported by Inter-University Centre of Telecommunications and Informatics (www.etik.hu)

Reviewed

Key words: equivalent capacity, QoS, call admission control

The lack of quality of service (QoS) guarantees is the classic problem of packet switching networks. Despite the access technologies (e.g. DSL) providing sufficient transmission speed are already available, without such QoS guarantees the rapid spread of novel, value-added services can not be imagined. In this article a novel technique capable to approximate the minimum bandwidth that should be provided for an aggregated network traffic flow in order to maintain a predefined QoS level is introduced. This new method can form the basis of load control (e.g. call admission control) algorithms to be applied in future packet-based networks.

1. Introduction

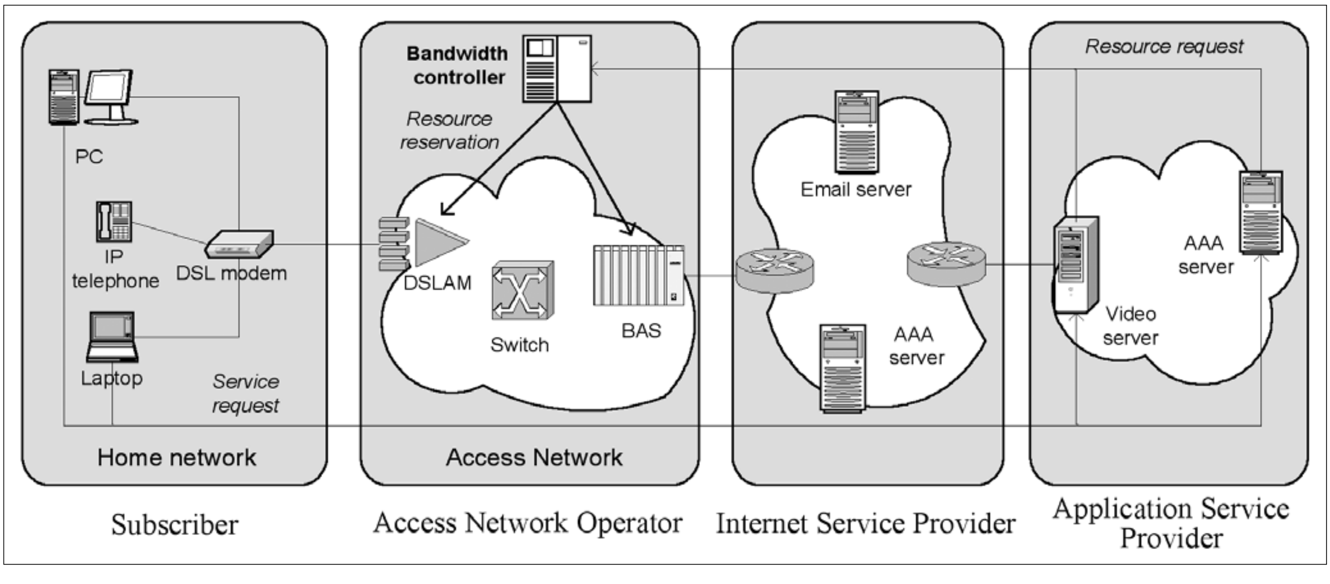
The number of DSL subscribers increases rapidly these days. This fact can be explained by the reasonable price and the relatively high reachable data rate this type of access technology offers. The reason of low price is that for the DSL access technology the already-in-use symmetric copper wires can be used by exploiting their higher (>144 kHz) frequency domain. As these copper wires can already be found at telephone users, in many cases the installation of DSLs may be the cheapest and best choice.

Amongst the services provisioned through DSLs the Internet access is the most popular [1]. The digital subscriber lines provide sufficient data rate for the majority of currently available services offered via the Internet. The previously available slow-speed access technologies severely hindered the development and spread of web applications. With the show-up of DSLs however the evolution of modern, broadband network services gained a new momentum. The spread of these premium applications (e.g. VoIP, VoD) is also encouraged by access network operators as they may attract new subscribers into their domain.

One of the gravest problems of TCP/IP based packet switching networks is the lack of transmission quality guarantees. Without these guarantees however the introduction and spread of value-added services is unimaginable. QoS guarantees and network management algorithms are primarily needed where resources may be scarce: in the local loop and the access network.

The quality of data transmission depends on the actual load of the network, which may be characterized by the saturation probabilities or packet loss ratios measured over the links of the network domain. The former metric (in case there are no buffers attached to links) is the probability of the instantaneous data rate

Figure 1. QoS guaranteed service provisioning with a bandwidth controller located in the access network



of the aggregated traffic flowing through a link exceeds its capacity. Unfortunately the link saturation probability does not tell anything about the amount of lost information, so it is more useful to prescribe the desired packet loss ratio, which is the ratio of lost and sent packets.

In this article a novel technique capable to approximate the expected load status of link while requiring only few a priori parameters is introduced. This approximation can be used by a bandwidth controller to supervise a network domain and making efficient and reliable admission control decisions.

The formulae to be presented approximate the required bandwidth need of a certain aggregated traffic flow in contrast with those that determine the expected level of QoS for a given link capacity. The advantage of our method is that it is enough to periodically refresh the actual amount of required bandwidth by background computations, while in the other case the expected QoS level has to be checked each time a new service request arises, which of course considerably slows down making admission decisions.

The rest of this article is organized as follows. In the next part the applied mathematical model is explained briefly. In the third section techniques capable to approximate the moment generating function of the aggregated traffic's rate distribution are presented. In the fourth part methods to convert QoS level approximations into bandwidth requirement values are introduced. In the fifth section the efficiency of previously and newly developed methods are compared through numerical examples. Our concluding remarks can be found in the last section.

2. QoS metrics and their approximation in packet-based networks

For the approach to the problem outlined in the introduction we used the popular BFFM (Bufferless Fluid Flow Multiplexing) framework. As in this model there are no buffers that may reduce the packet loss ratio it can be used to approximate important QoS metrics in a conservative manner.

Let us suppose that n fluid flows are aggregated on a link with capacity C. Let X_i be a random variable denoting the instantaneous data rate of the ith stationary flow. Let us suppose that for each source a p_i peak data rate can be determined, that is $0 \le X_i \le p_i$. Let X be a random variable denoting the instantaneous data rate of the aggregated traffic flow: $X = \sum_{i=1}^{n} X_i$.

Thus the link saturation probability can be defined as:

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

This probability means the fraction of time when the instantaneous data rate of the aggregated traffic exceeds the link capacity and so information loss occurs. This metric can be determined relatively easily, but may be useful only for the network operators, as the saturation probability does not give any reliable information regarding the amount of lost data. It is easy to imagine that beside the same saturation probability the number of lost packets may totally differ. Thus the level of users' satisfaction should be characterized with the packet loss ratio instead. It is by definition:

Abbreviations		
BFFM	–	Bufferless Fluid Flow Multiplexing
DSL	–	Digital Subscriber Line
LMGF	–	Logarithmic Moment Generating Function
PLR	–	Packet Loss Ratio
QoS	–	Quality of Sevice

$$PLR \overset{def}{=} \frac{E\left[(X-C)^{+}\right]}{E[X]} \qquad (2)$$

where *E[.]* is the expected value operator, and $(X-C)^{+}=\max(X-C,0)$. So in other words the packet loss ratio can be computed as the expected value of the instantaneous data rates exceeding the link capacity (and so causing packet loss) divided by the mean instantaneous data rate of the aggregated flow.

The call admission decision is based on the relation of the expected and prescribed level of the QoS metric:

$$P(X > C) \le e^{-\gamma} \quad \text{vagy} \quad \frac{E\left[(X-C)^{+}\right]}{E[X]} \le e^{-\gamma} \qquad (3)$$

Practically it is more tractable to compare the equivalent capacity of the aggregated traffic to the link capacity. The equivalent capacity is the minimum required bandwidth that the aggregated traffic needs for attaining the predefined QoS level. The definition of *equivalent capacity* can be written in the following forms in case the guaranteed QoS level is composed in terms of saturation probability or packet loss ratio:

$$C_{equ,sat} \overset{def}{=} \inf\left\{C : P_{sat} \le e^{-\gamma}\right\} \quad \text{or}$$

$$C_{equ,PLR} \overset{def}{=} \inf\left\{C : PLR \le e^{-\gamma}\right\} \qquad (4)$$

For the approximation of the expected link saturation probability or packet loss ratio the Bahadur-Rao extension of the well-known Chernoff bound can be used.

The Bahadur-Rao approximation of the given QoS metric is more accurate than the Chernoff bound, however it is not necessarily conservative (i.e. it may underestimate the true value) [5,6]:

$$P(X > C) \approx \frac{1}{s^{*}\sqrt{2\pi\sigma^{2}(s^{*})}} \exp(\Lambda_X(s^{*}) - s^{*}C) \quad \text{or} \qquad (6)$$

$$PLR \approx \frac{1}{M(s^{*})^{2}\sqrt{2\pi\sigma^{2}(s^{*})}} \exp(\Lambda_X(s^{*}) - s^{*}C) \qquad (7)$$

where $\Lambda_X(s)$ is the logarithmic moment generating function *(LMGF)* of X,

$$M \overset{def}{=} E[X],\ \sigma^{2}(s) = \frac{\partial^{2}}{\partial s^{2}}\Lambda_X(s),\ s^{*} = \arg\inf_{s}\left\{\Lambda_X(s) - sC\right\}$$

It can be seen that for the presented approximations (6) and (7) the LMGF of the distribution function of *X* is needed. The LMGF can be computed if all the moments of *X* are known, which is usually not the case. To overcome this problem three methods for approximating the moment generating function of *X* are presented in the next section. These techniques are easy-to-implement as they require only three parameters: the number of flows, the peak data rates of flows and the mean data rate of the aggregated flow.

3. Parsimonious upper bounds of the moment generating function

The first approximation method with which an upper bound for the moment generating function of *X* can be determined is a corollary of the results published by Hoeffding in 1963 [2]. Let X_i, $i=1...n$ denote independent, bounded random variables, for which

$$X = \sum_{i=1}^{n} X_i, \quad M \overset{def}{=} E[X], \quad 0 \le X_i \le p_i .$$

Then for $s>0$

$$G_X(s) \le \exp(sM) \exp\left(\frac{s^2 \sum_{i=1}^{n} p_i^2}{8} \right), \tag{8}$$

where $G_X(s)$ is the moment generating function of *X*.

Using Hoeffding's results Heszberger et al [3] formed the following conservative bound which can be applied to bound the moment generating function of the sum of bounded ($0 \le X_i \le p_i$) random variables ($X = \sum_{i=1}^{n} X_i$) :

$$G_X(s) \le \left(\frac{M + \sum_{k=1}^{n} \frac{p_k}{e^{sp_k} - 1}}{n} \right)^{n} \prod_{i=1}^{n} \left(\frac{e^{sp_i} - 1}{p_i} \right). \tag{9}$$

The already presented two bounds are based on the results of Hoeffding, however another approach may also be used for obtaining an upper bound for the moment generating function. For the construction of this third bound the concept of a certain type of stochastic ordering of random variables will be used. Let us assume that we have two random variables *X* and *Y*, whose distribution functions are denoted by F_X and F_Y, respectively. Then *X* is said to be smaller than *Y* with respect to increasing convex ordering [4], written as $X <_{icx} Y$, if the condition

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

holds for each increasing convex function $\phi(x)$, for which the integral exists.

From the definition it can be deduced that if $X <_{icx} Y$, then for $s>0$, $G_X(s) \le G_Y(s)$ holds. This can easily be verified by substituting $\phi(x)$ with e^{sx}.

Using the following lemma a new approximation for the upper bound of the moment generating function can be constructed [4]. Let $X_1^{onoff},...,X_n^{onoff}$ random variables denote *n* independent, heterogeneous (i.e. non-uniformly bounded) on-off sources whose peak data rates are $p_1,...,p_n$, and mean data rates are $m_1,...,m_n$, respectively. Let $Y_1^{onoff},...,Y_{n_Y}^{onoff}$ random variables denote n_Y independent homogeneous on-off sources, whose peak data rates are identically $p=max(p_i, i=1,...,n)$, and $n_Y = \text{int}\left\{ \sum_{i=1}^{n} p_i / p \right\}$ (i.e. the upper integer value of the expression between the braces), their mean data rates are identically $m = \sum_{i=1}^{n} m_i / n_Y$. Then

$$X_{onoff} <_{icx} Y_{onoff}, \qquad X_{onoff} \overset{def}{=} \sum_{i=1}^{n} X_i^{onoff}, \qquad Y_{onoff} \overset{def}{=} \sum_{i=1}^{n_Y} Y_i^{onoff}.$$

Using this lemma and the consequence of the definition of increasing convex ordering the upper bound of the moment generating function of the *X* can be written as follows [8].

Let X_i, $i=1...n$ denote independent, bounded random variables,

$$X = \sum_{i=1}^{n} X_i, \quad M \overset{def}{=} E[X], \quad 0 \le X_i \le p_i .$$

Then for $s>0$

$$G_X(s) \le \left(1 - \frac{M\left(1 + e^{sp}\right)}{n_Y p} \right)^{n_Y}. \tag{10}$$

From now on the logarithms of the moment generating function bounds (8), (9) and (10) (i.e. the corresponding LMGFs) will be denoted by $\tilde{\Lambda}_{X,hoe}(s)$, $\tilde{\Lambda}_{X,ih}(s)$ and $\tilde{\Lambda}_{X,so}(s)$ respectively.

4. Direct equivalent capacity estimation methods

By putting the previously introduced moment generating function bounds into formulae (6) and (7), an upper bound of the expected QoS level (saturation probability or PLR) can be obtained. This value then can be compared with the prescribed QoS level – as it was indicated in (3) – and the admission decision can be made according to the result of this comparison.

If we take another look on (6) and (7) we see that the original Bahadur-Rao formulae contain not only the LMGF, but also the second derivative of the LMGF.

Investigations show that as the exact moment generating function is not known (only an upper bound of it can be obtained), for its second derivative only a very imprecise approximation can be given, and this eventually makes the Bahadur-Rao formulae inapplicable. Thus it is desirable to eliminate the second derivative from the formulae somehow. It can be managed by using the results of Montgomery and de Veciana [7]:

$$P_{sat} \approx \exp(-I - \frac{1}{2}\log 4\pi I), \tag{11}$$

$$PLR \approx \exp(-I - \frac{1}{2}\log 4\pi I - \log s^{*} M), \text{ where} \tag{12}$$

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

	n_1	m_1 [kbit/s]	p_1 [kbit/s]	n_2	m_2 [kbit/s]	p_2 [kbit/s]	P/M
M1	100	51	64	10	200	500	2,24
M2	100	51	64	1000	4,8	5,8	1,34

Table 1. Characteristics of the investigated traffic mixes

It was mentioned earlier, that it is more tractable to compute the equivalent capacity of an aggregated flow instead of the expected QoS level, because the equivalent capacity need to be refreshed only periodically while the expected value of the appropriate QoS metric should be recomputed each time a new service requests arrives. In case the equivalent capacity is tracked, a new flow can be admitted if its peak data rate plus the actual equivalent capacity of the aggregated flow is less than or equal to the link capacity.

If formula (11) or (12) is used in the appropriate part of formula (4), an indirect method for computing the equivalent capacity can be obtained. However, in this case a double optimization should be performed (with respect to parameters *s* and *C*), which considerable increases the computational complexity of this method.

To overcome this problem we have developed direct formulae which are capable to determine the equivalent capacity in one step. These can be written in the following forms (13 and 14):

$$\widetilde{C}^{B-R}_{equ,sat} \overset{def}{=} \inf_{s>0}\left\{\frac{\widetilde{\Lambda}_X(s)}{s}+\frac{\gamma}{s}-\frac{\gamma\log 4\pi\gamma}{s(1+2\gamma)}\right\}$$

$$\widetilde{C}^{B-R}_{equ,WLR} \overset{def}{=} \inf_{s>0}\left\{\frac{\widetilde{\Lambda}_X(s)+\gamma-1+\log M+\frac{2\gamma}{1+2\gamma}\log\frac{1+2\gamma}{4M\sqrt{\pi}\gamma^{\frac{3}{2}}}}{-\frac{1}{M}+s}\right\}$$

where $\widetilde{\Lambda}_X(s)$ can be any appropriate approximation of $\Lambda_X(s)$ (e.g. the bounds presented in Section 3 are such). A more detailed discussion of these new results can be found in [8].

5. Numerical investigations

In this section the comparative analysis of the presented moment generating function bounds and equivalent capacity estimators will be carried out through numerical examples. For this investigation let us define a two-class, on-off traffic mix. The numbers of sources within the classes are represented by n_1 and n_2, respectively. The peak and mean data rate of the sources belonging to the same class are identical, these are denoted by m_i and p_i, $i \in \{1,2\}$. The important characteristics of the investigated traffic mixes are summarized in *Table 1*.

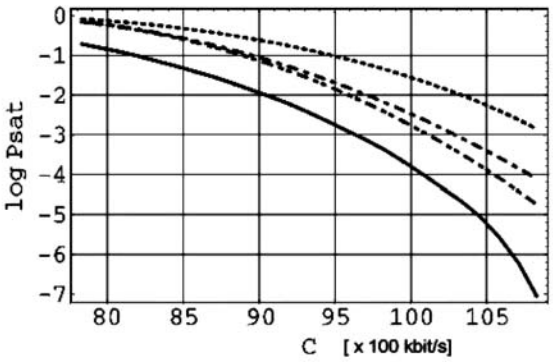


Figure 2.
Link saturation probability estimations, M1

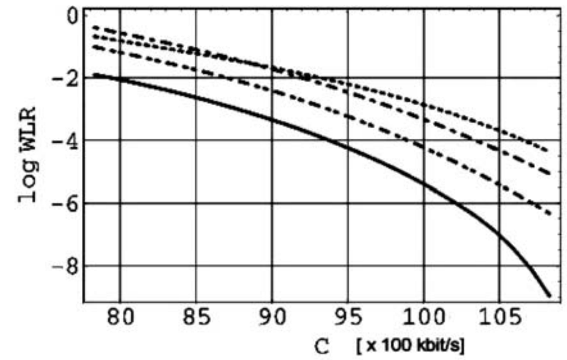


Figure 3.
Packet loss ratio estimations, M1

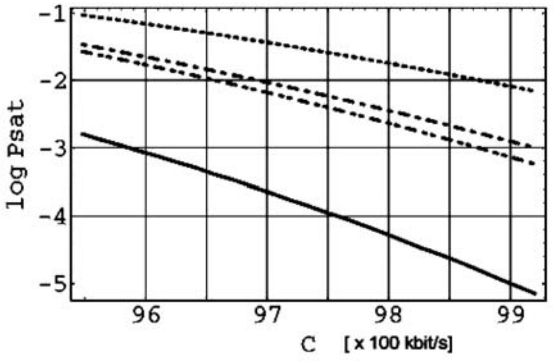


Figure 4.
Link saturation probability estimations, M2

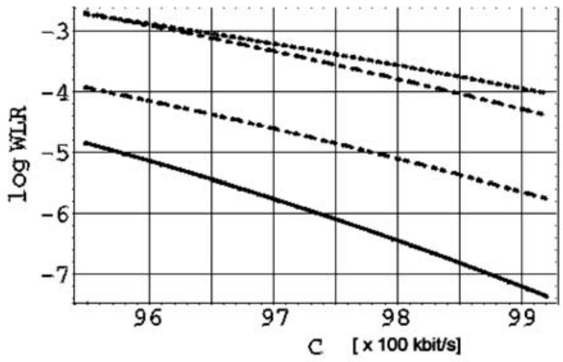


Figure 5.
Packet loss ratio estimations, M1

The first traffic mix (M1) can be considered as the aggregate of uncompressed voice and compressed video flows, while the second traffic mix (M2) resembles to the aggregation of compressed and uncompressed voice flows. The difference between the two mixes lies in the difference of the aggregate peak to mean ratio (shown in the last column of Table 1).

On *Figures 2-5* (on the previous page), the 10-based logarithms of the exact and approximated values of the link saturation probability or packet loss ratio are drawn as a function of the link capacity *C*. As the presented bounds give applicable results in the

$M < C < P$ ($P \overset{def}{=} \sum_{i=1}^{n} p_i$) interval,

only a part of the *(M,C)* interval is plotted. The exact values are drawn with continuous, while the bounds are drawn with dotted ($\tilde{\Lambda}_{X,hoe}(s)$), dash-dotted ($\tilde{\Lambda}_{X,ih}(s)$) and dash-dot-dotted ($\tilde{\Lambda}_{X,so}(s)$) lines.

On the figures it can be seen that in most cases the $\tilde{\Lambda}_{X,hoe}(s)$ bound is the less accurate, while the other two bounds' accuracy is acceptable. The vertical and horizontal distances between the curves usually increase with increasing γ (as the prescribed QoS level gets more stringent). The difference between the bounds of $\tilde{\Lambda}_{X,ih}(s)$ and $\tilde{\Lambda}_{X,so}(s)$ is sometimes negligible, however the computational complexity of those are fairly different: the stochastic ordering based bound can be obtained more easily. The application of the $\tilde{\Lambda}_{X,hoe}(s)$ bound can be recommended only if the computational complexity is the most important factor.

The performances of the equivalent capacity estimator formulae have also been compared. The numerical analysis was carried out the following way. First the exact values of the saturation probability and the packet loss ratio were determined for a given *C* value. Then from the $P_{sat}=e^{-\gamma}$ or $PLR=e^{-\gamma}$ formulae the corresponding γ values were determined. These γ values and the previously obtained $\tilde{\Lambda}_{X,hoe}(s)$, $\tilde{\Lambda}_{X,ih}(s)$ and $\tilde{\Lambda}_{X,so}(s)$ bounds were finally substituted into (13) or (14). The relation between the exact (i.e. in this case the link capacity *C*) and approximated value of the equivalent capacity has been then investigated.

On *Figures 6-7,* the $(\tilde{C}^{B\text{-}R}_{equ,sat}-C)/C$ relative error was drawn for M1 and M2 traffic mixes.

The equivalent capacity estimation for which $\tilde{\Lambda}_{X,hoe}(s)$ bound was used is plotted with continuous line, while the dotted and dash-dotted lines refer to the equivalent capacity approximations for which $\tilde{\Lambda}_{X,ih}(s)$ or $\tilde{\Lambda}_{X,so}(s)$ was used. It can be seen that the $\tilde{\Lambda}_{X,hoe}(s)$ based approximation severely underestimates, while for traffic mix M2 the $\tilde{\Lambda}_{X,so}(s)$ based approximation partly underestimates the exact equivalent capacity.

On *Figures 8-9,* the $(\tilde{C}^{B\text{-}R}_{equ,WLR}-C)/C$ relative error was drawn for the two traffic mixes. The equivalent capacity approximation for which $\tilde{\Lambda}_{X,hoe}(s)$ bound was used is

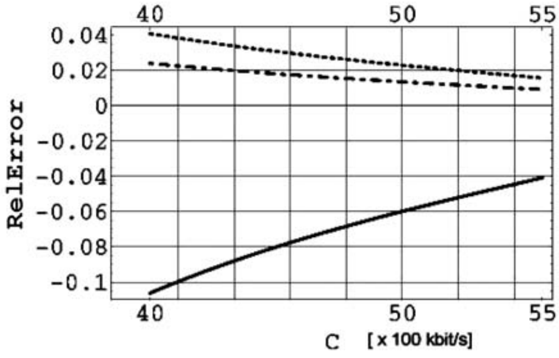


Figure 6.
The relative error of $\tilde{C}^{B\text{-}R}_{equ,sat}$*, M1*

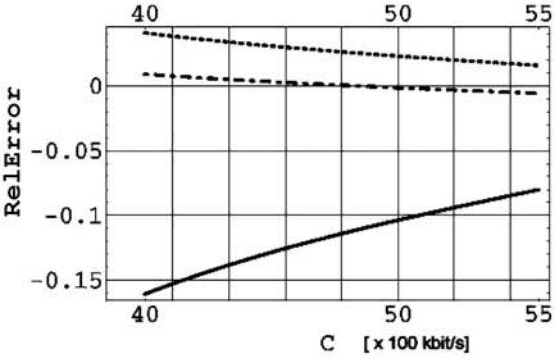


Figure 7.
The relative error of $\tilde{C}^{B\text{-}R}_{equ,sat}$*, M2*

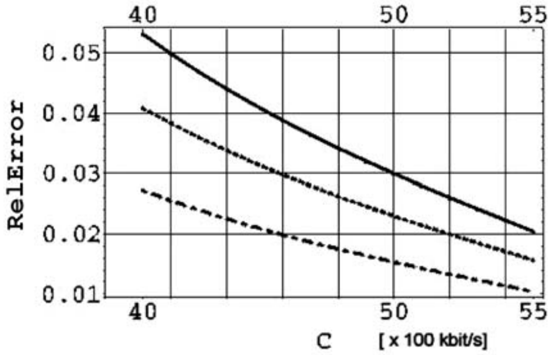


Figure 8.
The relative error of $\tilde{C}^{B\text{-}R}_{equ,WLR}$*, M1*

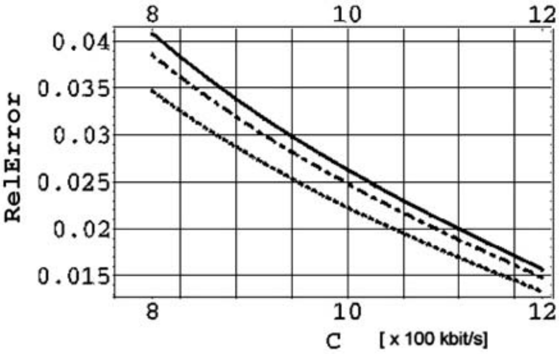


Figure 9.
The relative error of $\tilde{C}^{B\text{-}R}_{equ,WLR}$*, M2*

plotted with continuous line, while the dotted and dash-dotted lines refer to the equivalent capacity estimations for which $\tilde{\Lambda}_{X,ih}(s)$ or $\tilde{\Lambda}_{X,so}(s)$ was used.

The accuracy of the $\tilde{\Lambda}_{X,hoe}(s)$ based approximation is the worst in almost all cases, while the most accurate estimation is usually given by the one which uses the stochastic ordering based LMGF bound.

It can also be observed that the differences between the relative errors are bigger for smaller C values (i.e. for smaller γ values). It can also be seen that the investigated formulae give almost certainly a higher value than the exact one, if the $\tilde{\Lambda}_{X,ih}(s)$ or $\tilde{\Lambda}_{X,so}(s)$ bounds are used in the equivalent capacity estimator formulae. The absolute values of the relative errors decrease as γ increases (i.e. as the prescribed QoS level becomes more stringent).

6. Conclusions

In this article novel resource requirement estimator techniques were presented. With these methods the minimal required transmission capacity that should be provided for an aggregated network traffic flow in order to maintain a predefined QoS level can be computed. The most important advantage of our new formulae may be that they require very few input parameters: only the number of flows, their peak admission rates and the mean admission rate of the aggregated flow have to be known a priori.

For the computation of the presented equivalent capacity estimators the moment generating function of the aggregated traffic's rate distribution is needed. As it can not be determined exactly from the given parameters, three techniques capable to obtain an upper bound for the moment generating function was presented in Section 3. While the required parameters for these methods are the same, the performances of these bounds differ as we saw in Section 5.

Our numerical investigations also showed that for the best accuracy usually the new, stochastic ordering based bound should be applied in the equivalent capacity estimator formulae. However, if the computational simplicity is the most important factor, using the well-known Hoeffding bound may be the best idea.

With the aid of the presented resource requirement estimators efficient traffic load control mechanisms can be realized in packet based networks. The overload protection enables network operators to provide QoS guarantees for premium services, which in return ensures the satisfaction of their subscribers and encourages the evolution and spread of value-added services.

References

[1] C. Bouchat, S. van den Bosch, T. Pollet, "QoS in DSL Access", IEEE Communications Magazine, Vol. 41., Nr.9, pp.108–114, November 2003.

[2] W. Hoeffding, "Probability Inequalities for Sums of Bounded Random Variables", Journal of the American Statistical Association, 58:13–30, March 1963.

[3] Z. Heszberger, J. Zátonyi, J. Bíró, "Efficient Chernoff-based Resource Assessment Techniques in Multi-service Networks", Telecommunication Systems, 20(1):59–80, 2002.

[4] G. Mao, D. Habibi, "Loss Performance Analysis for Heterogeneous On-Off Sources with Application to Connection Admission Control", IEEE/ACM Transactions on Networking, 10(1):125–138, 2002.

[5] R. R. Bahadur, R. Rao, "On Deviations of the Sample Mean", Ann. Math. Statis., 31(27):1015–1027, 1960

[6] J. Y. Hui, "Resource Allocation for Broadband Networks", IEEE Journal on Selected Areas in Communications, 6(9):1598–1608, December 1988.

[7] M. Montgomery, G. de Veciana, "On the Relevance of Time Scales in Performance Oriented Traffic Characterizations", Proc. of the Conf. on Computer Communications, San Francisco, Vol. 2, pp.513–520, March 1996.

[8] J. Bíró, Z. Heszberger, F. Németh, M. Martinecz, "Bandwidth Requirement Estimators for Quality of Service Packet Networks", Proc. of the Intern. Network Optimization Conference, Evry, Paris, pp.95–100, October 2003.