Paper

Methods for evaluation packet delay distribution of flows using Expedited Forwarding PHB

Sylwester Kaczmarek and Marcin Narloch

Abstract—The paper regards problem of providing statistical performance guarantees for real-time flows using Expedited Forwarding Per Hop Behavior (EF PHB) in IP Differentiated Services networks. Statistical approach to EF flows performance guarantees, based on calculation of probability that end-to-end packet delay is larger than certain value, allows larger network utilization than previously proposed deterministic approach. In the paper different methods of packet delay distribution evaluation are presented and compared. Considered cases comprise evaluation of delay distribution models for the core network and evaluation of end-to-end packet delay in the network consisted of edge node and chain of core nodes. Results obtained with aid of analytical models are compared with simulation results.

Keywords—packet delay distribution, Expedited Forwarding PHB, Differentiated Services, Service Level Specification, IP QoS.

1. Introduction

Rapidly increasing tendencies to provide services typical for traditional telecommunication networks in Internet rise new challenges for realizing services with guaranteed Quality of Service (QoS) in IP based networks. Particular field of interest is a problem of providing real-time services for streaming flows using Expedited Forwarding Per Hop Behavior (EF PHB) [6, 7, 17] in Differentiated Services (DiffServ) network [2, 24].

The paper is based on results of research effort presented in series of conference publications [18, 22, 23]. We present framework to evaluate statistical performance guarantees for flows using EF PHB and compare different methods to calculate the probability that packet delay is larger than certain value. We considered scenario with packet delay only in the network core nodes and scenario with edge node and core of the network (end-to-end delay including packet waiting in edge router). Among evaluated methods are Gaussian approximation, methods based on Large Deviation approach and approximation based on Erlang-n distribution. Despite discrepancies between presented methods, all provide possibility to evaluate statistical guarantees for packet delays of flows using EF PHB.

The paper is organized as follows. Section 2 presents previous work in the subject. In Section 3 model and methods for evaluation of packet delay distribution in the core are described. Section 4 regards influence of low priority traffic on EF packet delays in the node and presents appropriate numerical model. In Section 5 influence of packet waiting in edge node on end-to-end delay is presented together with respective analytical model. Section 6 presents configuration parameters of analysed and simulated networks together with obtained results. Section 7 concludes the paper.

2. Related work

There exist two distinct approaches to analysis of QoS for flows using EF PHB. First approach, derived from context of Integrated Services architecture [28] and represented by [1, 5], is based on deterministic bounds on performance guarantees with worst case assumptions for end-to-end delays. However, analysis in [5] led to very pessimistic bound on utilization for network with flow aggregation. The bound is order of $1/(n-1)$, where n is number of nodes the observed flow pass through. Such a small value indicates that deterministic approach cannot be applied in practice. The second approach, represented by [3], relies on statistical performance guarantees for flows using EF PHB. It allows larger level of utilization at the cost that DiffServ network assures certain packet loss ratio and guarantees that probability of packet transfer delay exceeding certain value (considered as a maximum) is smaller than certain level. That methodology is analogous to description of QoS for ATM CBR service. Also approach based on statistical performance guarantees appears in proposed standards of Service Level Specification (SLS) for DiffServ [14, 27]. An overview of the most current advances in Internet quality of service, including deterministic and statistical guarantees, can be found in [11] (see also references therein). Among other the most recent attempts to explore statistical performance guarantees is [29] where Large Deviations Theorems were applied to results obtained by the use of network calculus. However, we would like to point out that above result is limited to single node case. Thus suggested in [29] end-to-end delay calculation, which was obtained by summing bounds evaluated for single nodes in isolation, still seems to be conservative approach. In the paper we follow alternative approach to statistical guarantees based on the Better than Poisson—Negligible Jitter (NJ) conjecture, presented in [3]. That is the extension of the Negligible Cell Delay Variation notion, presented for ATM network in [4], to the case of IP environment with variable packet

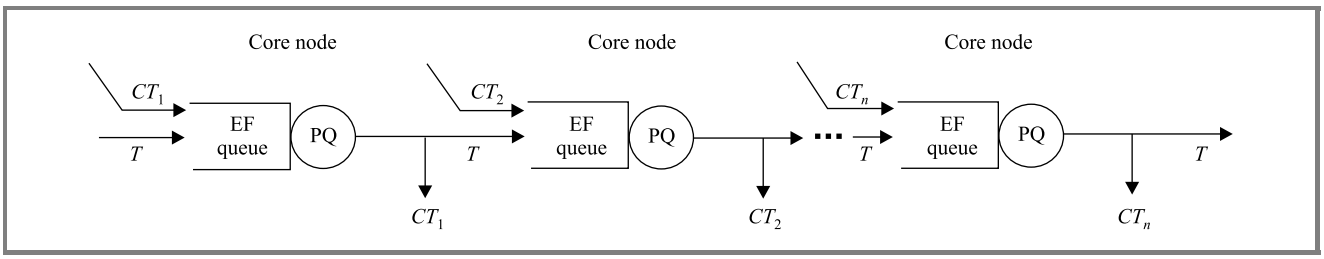


Fig. 1. Analysed network of tandem queues—core nodes case.

lengths. That approach allows radical simplification of the traffic management function, because worst case traffic inside a network can be modelled as Poisson stream of MTU size packets. Moreover, that approach is consistent with formulation of packet delay performance guarantees in the SLS specification, and allows realization of the real-time services with larger network utilization than methods based on worst case, deterministic bound on the delay. Thus we focus on statistical approaches and assume that NJ conjecture is valid.

3. Delay distribution in the core

In this section we present methods for analytical evaluation of delay experienced by packets from CBR flow in the presence of cross traffic in the core network. We considered a network (similar to presented in [3]) consisted of n FIFO queues arranged in tandem serving observed (tagged) CBR stream T passing through all queues, and interfering Poisson MTU-sized cross EF-traffic CT_i passing ith queue (Fig. 1). That type of cross traffic was chosen in order to obtain upper bound of delay distribution. We assumed that offered load ρ_T of observed traffic is relatively small in comparison with cross traffic offered load ρ_{CT}. We also assumed independence of queues in particular nodes. It should be noted that independence conjecture is reported in [3] as conservative, however it allows simplification of delay evaluation. For example, we do not have to consider the effects of distribution of queues with particular load within a chain. That assumptions are valid in the case of the core network (DiffServ network). We split packet delay into two components: deterministic service time equal to $n \cdot \tau_T$ (τ_T is observed CBR flow packet service time), and stochastic waiting time in queues modelled as a chain of n $\mathrm{M/D_{MTU}/1}$ queues. From practical point of view, we are interested in probability that waiting delay W exceeds certain value D of delay bound. Consequently probability that end-to-end delay exceeds value $D + n \cdot \tau$ (maximum packet delay stated in SLS) can be easily obtained. Thus we can write quality of service requirement as:

$$P(W > D) \leq L, \tag{1}$$

where L is a small number, i.e., $L \in \langle 10^{-2}, 10^{-6} \rangle$ [19]. In case of analysed network, core packet delay can be written as a sum of independent random variables W_i, denoting packet waiting time in ith queue of the core network:

$$W_{core} = \sum_{i=1}^{n} W_i . \tag{2}$$

If we assume that n is large, we can apply limit theorems. First approach is based on Gaussian approximation of delay distribution. It is simple extension of the model presented in [15], in the context of CBR service in ATM to the case of variable length IP packets. In that method, Gaussian distribution of packet delays has mean:

$$\mu = \sum_{i=1}^{n} \mu_i \tag{3}$$

and variance:

$$\sigma^2 = \sum_{i=1}^{n} \sigma_i^2 , \tag{4}$$

where μ_i and σ_i^2 are respectively mean and variance of waiting time in ith $\mathrm{M/D_{MTU}/1}$ queue. In next two approaches based on Theory of Large Deviations [8] we explore the fact that delay values for the probabilities of interest are largely deviated from the mean delay. Packet waiting distribution in the core network can be expressed with aid of approximation based on Chernoff theorem:

$$\log P(W_{core} \geq x) \leq -F(\theta^*) , \tag{5}$$

or refinement of the Chernoff-Cramer approximation based on Bahadur-Rao theorem (local limit theorem):

$$P(W_{core} \geq x) \approx \frac{e^{-F(\theta^*)}}{\sqrt{2\pi} \cdot \theta^* \cdot \sigma(\theta^*)}, \tag{6}$$

where large deviations rate function $F(\theta^*)$ is defined as:

$$F(\theta^*) = \sup_{\theta \geq 0} F(\theta),\ F(\theta) = \theta \cdot x - \sum_{i=1}^{n} \log M_i(\theta) , \tag{7}$$

and $\sigma^2(\theta)$ is second order derivate of large deviations rate function with respect to θ:

$$\sigma^2(\theta) = \sum_{i=1}^{n} \frac{M_i''(\theta) \cdot M_i(\theta) - \left(M_i'(\theta)\right)^2}{M_i^2(\theta)}. \tag{8}$$

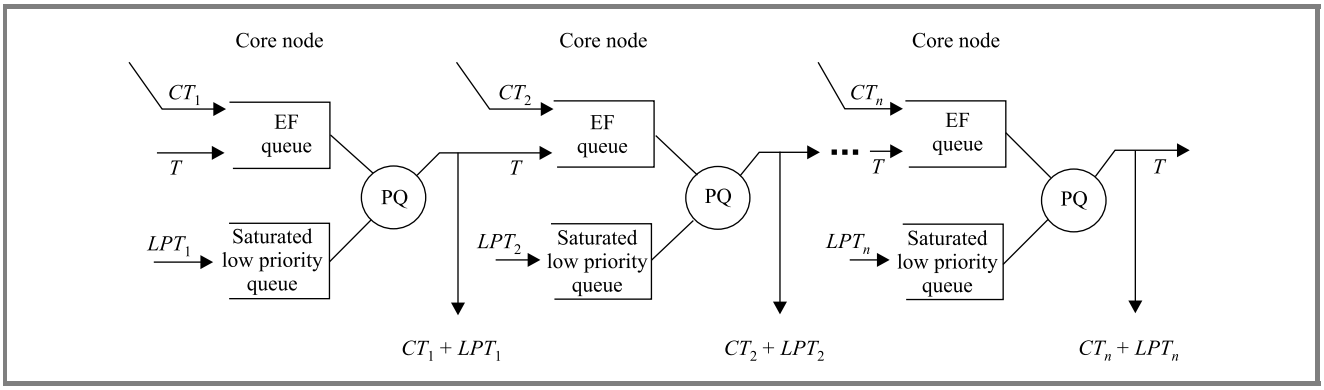


Fig. 2. Network of tandem queues with vacations—core nodes with low priority traffic.

$M_i(\theta)$ denotes moment generating function of packet waiting time in single queue for respective queuing model $\mathrm{M}/\mathrm{D}_{\mathrm{MTU}}/1$, etc. In order to compute desired probability we have to find θ^* for which supremum of $F(\theta)$ is attained, taking into consideration moment generating function $M(\theta)$ of packet waiting time for particular queuing model. Thus θ^* is positive root of the equation with derivative of $F(\theta^*)$:

$$F'(\theta^*) = 0, \text{ where } F'(\theta^*) = x - \sum_{j=1}^{J} n_j \frac{M_j'(\theta)}{M_j(\theta)}. \quad (9)$$

It is worth noting that in general case independent random variables W_i, denoting packet waiting time in ith queue, are not necessarily identically distributed. Last of evaluated methods of packet waiting distribution in the core is based on Erlang-n distribution, and for the sake of presentation clarity will be described in the next section.

4. Influence of low priority traffic

In order to model the influence of the lower priority, non-EF traffic on EF streams performance guarantees, we extended model of core nodes and considered non-preemptive static priority queues in core nodes (Fig. 2), with multiple and exhaustive vacations, with constant vacation time equal to MTU packet transmission time [9]. In that model arrivals and services have the same characteristics as in ordinary $\mathrm{M}/\mathrm{D}_{\mathrm{MTU}}/1$ queue, but in queue with vacations when high priority (EF) queue is empty, server takes vacation instead being idle waiting for EF packet to arrive for service. If queue server finds EF packets when returning from vacation, it serves them until EF queue becomes empty (exhaustive discipline), and than it takes next vacation. If there is no packet in high priority queue after returning from vacation, server takes another vacation (multiple vacation discipline). That allows modelling the real system with priority queuing and link transmitting non-EF, low priority packets, wherever there is no EF packets to transmit. This is also the worst case approach with respect to EF stream performance guarantees, because we assumed that link is saturated and there is always MTU sized low priority packet in node to send. In case of delay distribution approximation methods based on Large Deviations Theory, that extension of network node model results in application of appropriate moment generating function, regarding queuing model with vacation. Moment generating function $M(\theta)$ for $\mathrm{M}/\mathrm{D}_{\mathrm{MTU}}/1$ queue with vacation can be obtained by stochastic decomposition property described in [12, 13]. Stochastic decomposition property allows to consider the waiting time in the M/GI/1 queue with vacations, as the sum of two independent components: one distributed as the waiting time in the ordinary queue in the corresponding M/GI/1 queue without vacations, and the other as the equilibrium residual time of a vacation. Thus moment generating function $M(\theta)$ for $\mathrm{M}/\mathrm{D}_{\mathrm{MTU}}/1$ queue with vacation can be calculated as follows:

$$M(\theta) = \frac{U(\theta) - 1}{u\theta} M_{M/D/1}(\theta), \quad (10)$$

where $M_{M/D/1}(\theta)$ is moment generating function in ordinary $\mathrm{M}/\mathrm{D}_{\mathrm{MTU}}/1$ queue and $U(\theta)$ denotes moment generating function for the vacation time. In the considered case of constant vacation time equal to MTU packet transmission time $U(\theta) = \exp(\theta \cdot u)$, where $u = MTU/C$, C is link bandwidth. In case of presented in previous section Gaussian approximation of packet delay distribution in the core, appropriate formulas for μ_i and σ_i^2 can be obtained with the aid of respective derivatives of moment generating functions $M(\theta)$ of waiting time for queues with vacations. Last of the described packet waiting distribution approximation [3] can be expressed as a sum of $n \cdot x_{\min}$ and Erlang-n distribution of mean $(MTU \cdot n)/(r \cdot C)$, where r satisfies:

$$\rho_{CT} \cdot (e^r - 1) - r = 0, \quad (11)$$

and $x_{\min}$ is defined as:

$$x_{\min} = \frac{-MTU}{r \cdot C} \cdot \log \frac{1}{K}, \quad (12)$$

where

$$K = \frac{1 - \rho_{CT}}{\rho_{CT}^2 \cdot e^r - \rho_{CT}}. \quad (13)$$

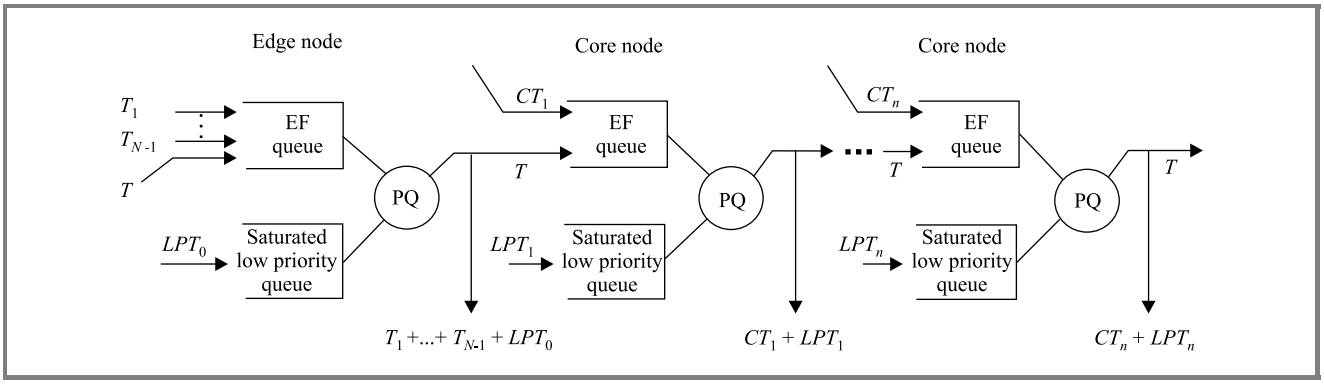


Fig. 3. Analysed network of tandem with vacations queues—edge and core nodes case.

That approach is based on presented in [25], in the context of ATM network approximation of queue size distribution:

$$P(Q > x) \approx K \cdot \exp(-r \cdot x), \tag{14}$$

with appropriate extensions to model influence of low priority traffic ($x_{\min}$ and different formula describing k). It is worth noting that Erlang-n approximation is limited to the homogenous case only and unfortunately that approach cannot be used directly to evaluate packet delay distribution for the heterogeneous case, which is typical for any practical network scenario.

5. Influence of waiting in edge node

In order to consider influence of queuing in edge node on EF packet delay, we extended evaluated network model. Model of the network core remained as previously described chain of $M/D_{MTU}/1$ queues with vacations, however we introduced edge node modelled as discrete time ND/D/1 queue with vacations. N denotes the number of CBR sources in the edge router. Each source generates packets with fixed length and constant, deterministic period P. Packet arrival instants from all N sources are independent and randomly spread with uniform distribution within the period P (assumption of random phases of the sources). We considered discrete time queuing model with time axis divided into slots. Slot duration is equal to CBR source packet transmission time τ_T in the link, between edge node and following core node. In our model one CBR stream plays the role of observed traffic T, which follows path through all nodes of considered network. The remaining $N-1$ CBR streams create background EF traffic in the ingress edge node and leave considered network path after edge node. That streams denoted in Fig. 3 as T_k, where $k \in <1, N-1>$, compete for resources in EF queue with the observed stream T only in the edge router. We also modelled influence of lower priority non-EF traffic on EF streams performance guarantees in non-preemptive priority queue of the edge node, by considering queue with multiple and exhaustive vacations, with constant vacation time equal to CBR packet transmission time. In case of analysed network model, evaluation of end-to-end packet delay distribution can be considered as a form of discrete convolution of delay in discrete ND/D/1 with vacations model, and delay distribution in chain of $M/D_{MTU}/1$ with vacations in the core:

$$P(W_{e2e} > x) = \\ = \sum_{k=1}^{N} P\left(W_{edge} = k\right) \cdot P\left(W_{e2e} > x \middle| W_{edge} = k\right), \tag{15}$$

where $P(W_{edge} = k)$ is probability that waiting delay in the edge router is equal to k slots, $P(W_{e2e} > x | W_{edge} = k)$ is conditional probability that end-to-end packet delay exceeds x conditioned on event that delay in the edge node equals k slots. Probability of packet waiting $P(W_{edge} = k)$ for discrete time ND/D/1 model is presented in [16]:

$$P\left(W_{edge} = k\right) = \frac{P}{N} P(Q = k), \tag{16}$$

where $k > 0$ and queue length distribution [16, 26] is given by:

$$P(Q > q) = \\ = \sum_{m=1}^{N-q} \frac{P - N + q}{P - m} \binom{N}{q+m} \left(\frac{m}{P}\right)^{q+m} \left(1 - \frac{m}{P}\right)^{N-q-m}, \tag{17}$$

where $q \geq 0$. Conditional probability $P(W_{e2e} > x | W_{edge} = k)$ can be determined regarding the assumption of queue independence in the network. Random variables denoting waiting time in node queues are independent and thus amount of packet delay encountered in the edge node does not influence value of delay in the chain of core nodes (queues). Because of that, we can express conditional probability as:

$$P\left(W_{e2e} > x \middle| W_{edge} = k\right) = P\left(W_{core} > x - k \cdot \tau_T\right), \tag{18}$$

and apply the approximations, describing packet delay distribution in the core of the network $P(W_{core} > x)$, presented in details in previous sections. In order to evaluate end-to-end packet delay distribution, we can also apply other method in which edge router is modelled as M/D/1 queue with vacations, where D denotes CBR source packet size (typically smaller than MTU) and core nodes are modelled as a chain of $n\,\mathrm{M}/\mathrm{D}_{\mathrm{MTU}}/1$ queues with vacations. In that method we can utilize approximations of delay distribution described in previous sections, but with respective modifications in formulas regarding presence of the edge node in the observed stream path. Hence, in formulas (5) and (6), W_{core} is replaced by W_{e2e} and, consequently, Large Deviations rate function $F(\theta^*)$ includes moment generating function of random variable, describing packet delay in the edge node (queue). Therefore, formula (7) should be rewritten as:

$$F(\theta^*) = \sup_{\theta \geq 0} F(\theta),\ F(\theta) =$$

$$= \theta \cdot x - \sum_{i=0}^{n} \log M_i(\theta), \tag{19}$$

where $i = 0$ regards edge node queue and $M_0(\theta)$ denotes moment generating function of waiting time in M/D/1 queue with vacations. Similarly, in case of formulas (8) and (9), range of index i should be extended to include edge node accordingly.

6. Numerical results

In order to verify accuracy of presented methods, we compared results obtained from theoretical derivations with simulation results. At first, we considered core network of 10 nodes with interconnecting links of 150 Mbit/s bandwidth and buffers for 20 packets from EF streams. We considered Poissonian cross EF-traffic with offered load ρ_{CT} of 0.1, 0.3 and 0.5 (three distinct cases), and MTU packet size equal to 1500 bytes. The observed traffic consisted of 1 CBR flow with rate 1.5 Mbit/s (offered load $\rho_T = 0.01$), and packet size equal to 100 bytes. We evaluated packet delay distribution for two scenarios. In the first scenario, a FIFO queue is dedicated to EF streams, which are the only traffic passing through the nodes. In the second scenario, we considered priority queue with vacations as described in Section 4. Figure 4 presents comparison of theoretical and simulation results for respective cases of offered load ρ_{CT} for both scenarios. We also considered influence of path length (the number of nodes the EF flow passes through). Corresponding results for ρ_{CT} of 0.3 in the network with 5 and 15 nodes are presented in the Fig. 5, respectively. In the figures one can observe that for delay distribution probabilities larger than 0.01, calculation based on Gaussian approximation provide very good results. However, Gaussian approximation provide results unacceptable from practical point of view for probabilities smaller than 0.01, i.e., calculation of delay probabilities becomes too optimistic, and comparing to simulation values packet delay distribution is significantly underestimated. For probability values smaller than 0.01 methods based on Large Deviations provide better calculation, particularly for tail probabilities. Only for delay values close to mean value, methods based on Large Deviations give moderate precision of approximation, overestimating packet delay probabilities. That comes from the fact that Large Deviations Theory is dedicated to describe rare events and tail probabilities. In all cases, method based on Bahadur-Rao (local limit theorem) approximation provides more precise results for the same queuing model than Chernoff-Cramer approximation, which should be considered as a conservative, upper bound of real delay distribution. Considering results obtained for case with different number of nodes the EF flow passes through, methods based on Large Deviations provide good approximation of packet delay distribution, even in case where the number of nodes is relatively small. That promising results was obtained regardless that limit theorems were used in formulation of proposed methods, i.e., demand for very large number of nodes. From practical point of view that feature is positive, particularly that bounds are relatively tight, regarding Bahadur-Rao approximation. However, Bahadur-Rao approximation provides delay distribution values close to simulation results for small probabilities and cannot be applied for probabilities larger than 10^{-2}. Approximation based on Erlang-n distribution is computationally very attractive and provides precise results, which are compared to results obtained by Bahadur-Rao approximation. Despite its simplicity and precision, Erlang-n approximation is limited to the homogenous case, which cannot be assured in practice for any typical network scenario.

In order to evaluate influence of delay in the edge node, we considered extended network with 1 ingress edge node and 5 core nodes. Edge node was connected to the core by 15 Mbit/s link. Links in the core had 150 Mbit/s bandwidth as in previous cases. Edge node served 6 homogenous, 1.5 Mbit/s CBR streams with packet size equal to 100 bytes, thus load of EF traffic in edge node was equal to 0.6 Erl. We considered lower priority, non-EF traffic in the edge node with packets of size 100 bytes. The observed traffic consisted of one CBR flow with rate 1.5 Mbit/s (offered load $\rho_T = 0.01$ in the core node link). In EF-queues in core nodes we considered heterogeneous scenario, regarding to offered load ρ_{CT} of Poissonian cross traffic with MTU-sized (1500 bytes) packets. The value of offered load in jth queue was equal to $0.1 \cdot j$, thus we cover the range of loads from 0.1 to 0.5 with step 0.1. Lower priority (non-EF) packets of size MTU = 1500 bytes filled remaining link capacity in every core node of the network. Simulation results for network with edge node were obtained with simulation method described in details in [18], which allows efficient evaluation of systems with ND/D/1 queues.

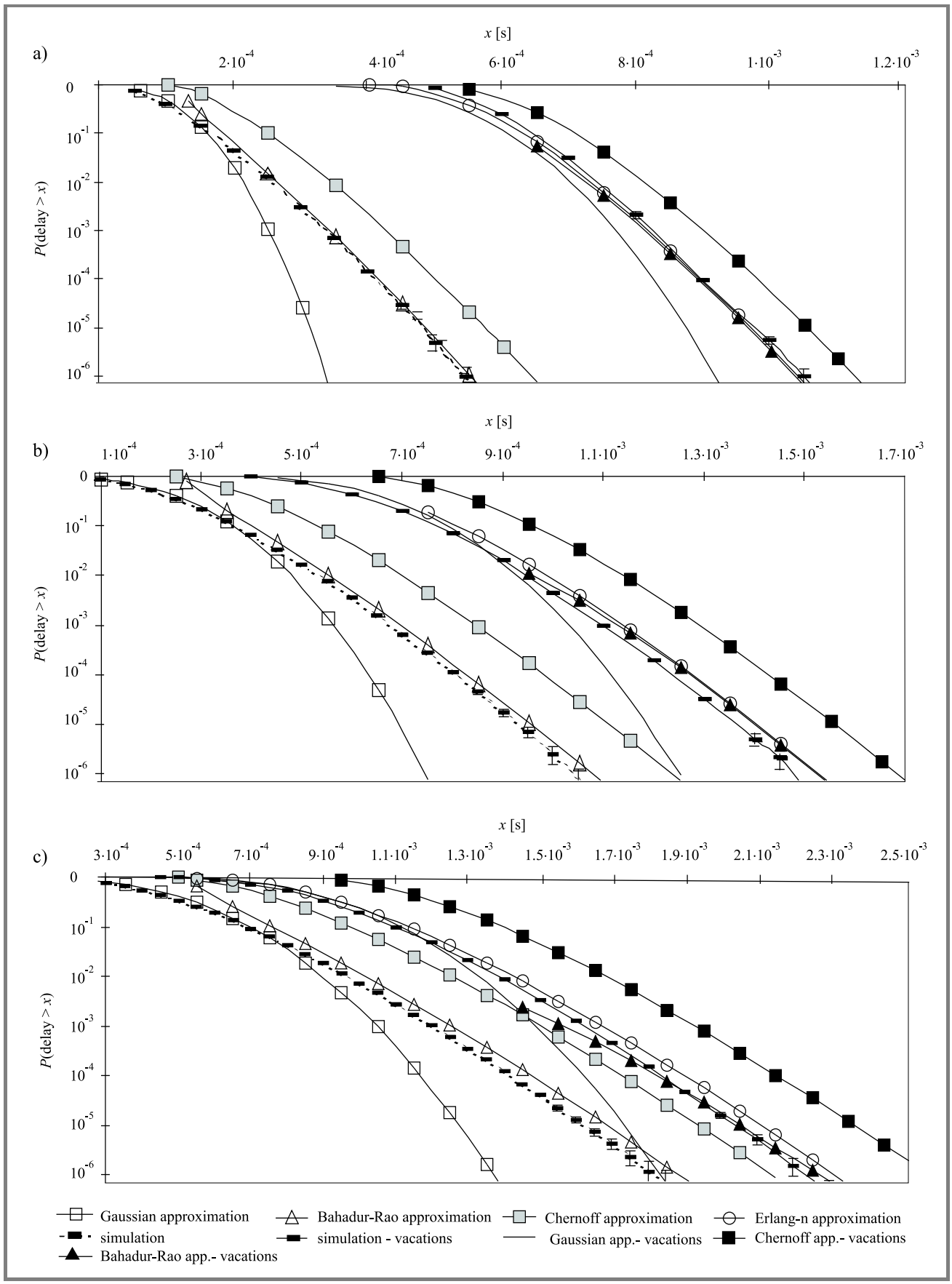


Fig. 4. Packet delay distribution of CBR flow in the core network: (a) $\rho_{CT} = 0.1$, $n = 10$; (b) $\rho_{CT} = 0.3$, $n = 10$; (c) $\rho_{CT} = 0.5$, $n = 10$.

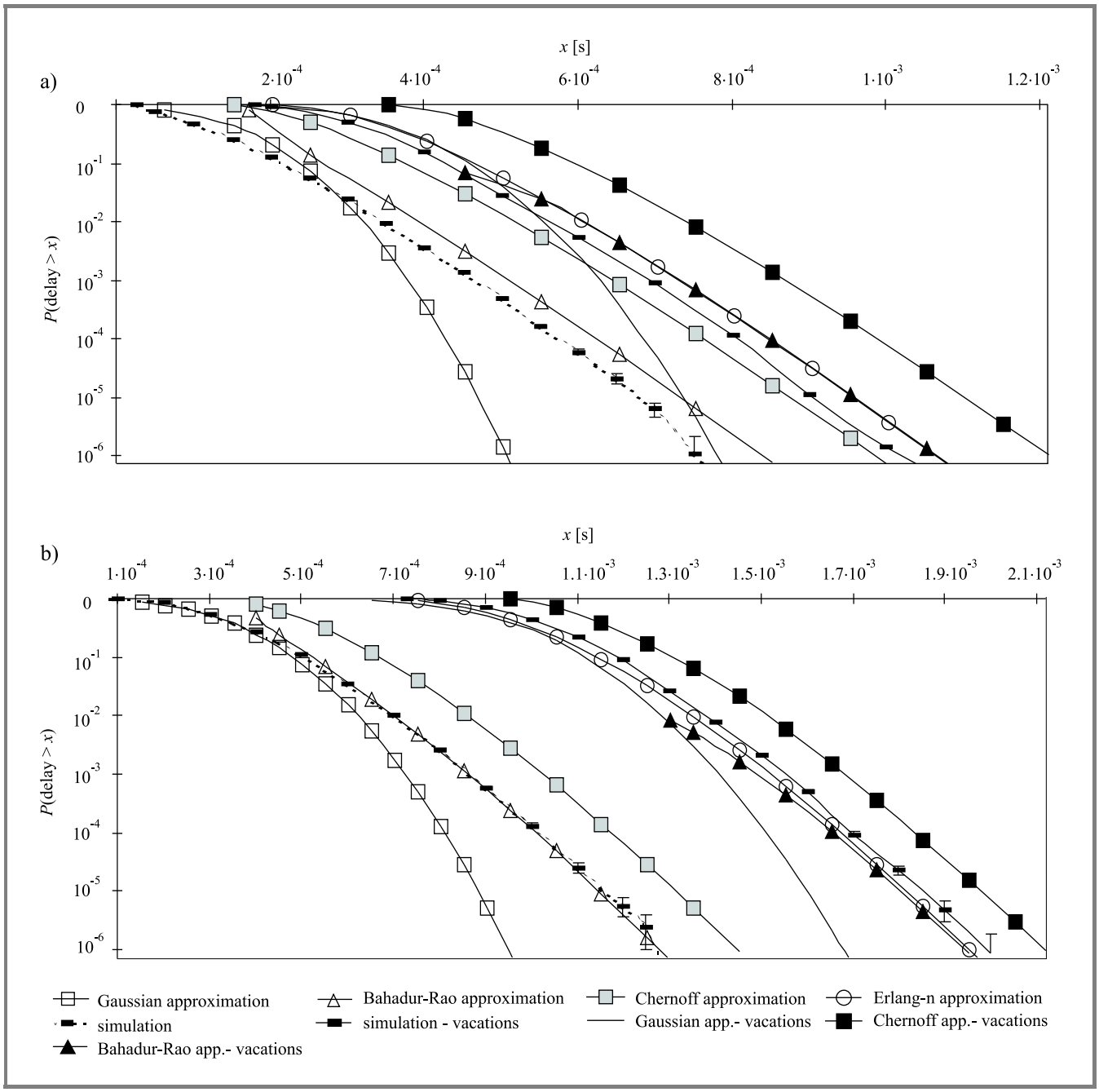


Fig. 5. Packet delay distribution of CBR flow in the core network: (a) $\rho_{CT} = 0.3$, $n = 5$; (b) $\rho_{CT} = 0.3$, $n = 15$.

In the Fig. 6 can be seen that method based on formula (14) and method in which edge router is modelled as M/D/1 queue with vacations and core nodes are modelled as a chain of n $\mathrm{M/D_{MTU}/1}$ queues with vacations provide almost similar results (under the condition of similar approximation application for calculation of delay distributions, for example, based on Bahadur-Rao theorem). However, the second approach to the calculation of end-to-end delay distribution seems to be simpler and more versatile than approximation based on formula (14), because approximation based on discrete time convolution (14) demands large number of calculations for large values of N. Alternatively, calculations with only few, significant values of $P(W_{edge} = i)$ can be applied as a form of formula (14) approximation. Moreover, in the Fig. 6 the accuracies of end-to-end packet delay distribution approximations based on different limit theorems can be compared. We would like to emphasise that precision of packet delay computation strongly depends on values provided by underlying approximation, and thus all remarks describing accuracy of approximation used to calculate packet delay distribution in the core regard calculations for end-to-end packet delay.

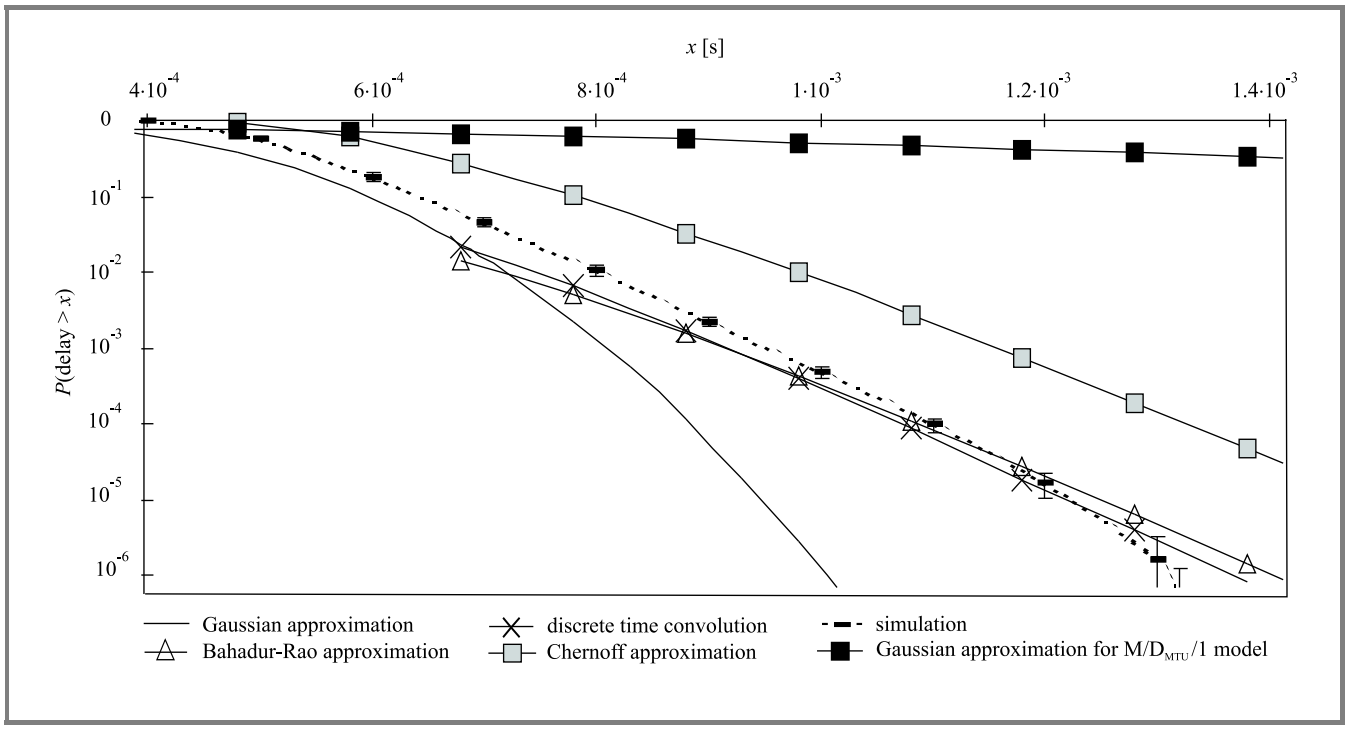


Fig. 6. End-to-end packet delay distribution of CBR flow in the network consisted of edge and core nodes.

7. Conclusions

Statistical performance guarantees allow to increase network utilization with sufficient margin of security, comparing to previously proposed deterministic approaches. Moreover, it is consistent with formulation of packet delay performance guarantees in Service Level Specification for IP QoS Differentiated Services.

Methods based on Large Deviations Theory are the most attractive from presented approaches to evaluate packet delay distribution. They provide bound on delay probabilities of packets from CBR flows, using Expedited Forwarding PHB in the region where exceeding maximum packet delay is allowed with certain, but very small probability. From practical point of view, application of that approximations allows realization of real-time services with statistical guarantees. We also extended core network to include edge node modelled by ND/D/1 queue, and applied it in evaluation of packet delay distribution. Obtained analytical and simulation results indicate that developed model allows evaluation of end-to-end packet delay distribution with accuracy which is satisfactory from practical point of view.

Knowledge of the packet delay distribution is very important in network dimensioning, network planning and traffic control algorithms. Methods presented above are used in calculation of *effective delay* [21] proposed as a new metrics for traffic control functions, for example a new EF flow admission control algorithm utilizing that notion. Consequently, precision of delay distribution calculation strongly influences accuracy of any traffic control function which relies on such approximations.

Future work in the subject should be directed toward extending model of edge router and considering more general low priority packet distribution. Also extension of presented simulation framework in order to evaluate more general EF and non-EF packet size distributions is intended.

References

[1] J. C. R. Bennett, K. Benson, A. Charny, W. F. Courtney, and J.-Y. Le Boudec, "Delay jitter bounds and packet scale rate guarantee for expedited forwarding", in *Proc. Infocom*, Anchorage, USA, 2001.

[2] S. Blake, D. Black, M. Carlson, E. Davies, Z. Wang, and W. Weiss, "An architecture for differentiated services", RFC 2475, Dec. 1998.

[3] T. Bonald, A. Proutiere, and J. W. Roberts, "Statistical performance guarantees for streaming flows using expedited forwarding", in *Proc. Infocom*, Anchorage, USA, 2001.

[4] F. Brichet, L. Massoulie, and J. W. Roberts, "Stochastic ordering and the notion of negligible CDV", in *Teletraffic Contributions for the Information Age. Proceedings of the 15th ITC*, V. Ramaswami and P. E. Wirth, Eds. Amsterdam [etc.]: Elsevier, 1997, pp. 1433–1444.

[5] A. Charny and J.-Y. Le Boudec, "Delay bounds in a network with aggregate scheduling", in *Quality of Future Internet Services. Proceedings of First COST 263 International Workshop, QofIS2000*, J. Crowcroft, J. W. Roberts, and M. Smirnov, Eds., *Lecture Notes in Computer Science*. Berlin [etc.]: Springer-Verlag, 2000, vol. 1922, pp. 1–13.

[6] A. Charny *et al.*, "Supplemental information for the new definition of the EF PHB (expedited forwarding per-hop behavior)", RFC 3247, 2002.

[7] B. Davie *et al.*, "An expedited forwarding PHB (per-hop behavior)", RFC 3246, Apr. 2002.

[8] A. Dembo and O. Zeitouni, *Large Deviations Techniques and Applications*. Boston, London: Jones & Bartlet Publ., 1992.

[9] B. T. Doshi, "Queuing systems with vacations – a survey", *Queuing Syst. Theory Appl.*. vol. 1, pp. 29–66, 1986.

[10] "The ns Manual", K. Fall and K. Varadhan, Eds., Dec. 2003, http://www.isi.edu/nsnam/ns/ns-documentation.html

[11] V. Firoiu, J.-Y. Le Boudec, D. Towsley, and Z.-L. Zhang, "Theories and models for Internet quality of service", *Proc. IEEE*, vol. 90, no. 9, pp. 1565–1591, 2002.

[12] S. W. Fuhrmann, "A note on the M/G/1 queue with server vacation", *Oper. Res.*, vol. 32, pp. 1368–1373, 1984.

[13] S. W. Fuhrmann and R. B. Cooper, "Stochastic decompositions in a M/G/1 queue with generalized vacation", *Oper. Res.*, vol. 33, no. 5, pp. 1117–1129, 1985.

[14] D. Goderis *et al.*, "Service level specification semantics, parameters and negotiation requirements", Internet-Draft, work in progress, June 2001, draft-tequila-sls-01.txt

[15] M. Grossglauser and S. Keshav, "On CBR service" (extended version), in *Proc. Infocom*, San Francisco, USA, 1996.

[16] P. Humblet, A. Bhargava, and M. Hluchyj, "Ballot theorem applied to the transient analysis of nD/D/1 queues", *IEEE/ACM Trans. Netw.*, vol. 1, no. 1, 1993.

[17] V. Jacobson, K. Nichols, and K. Poduri, "An expedited forwarding PHB", RFC 2598, June 1999.

[18] S. Kaczmarek and M. Narloch, "End-to-end packet delay distribution of flows using EF PHB", in *Proc. 10th Polish Teletraffic Symp.*, Kraków, Poland, 2003.

[19] M. Karam and F. Tobagi, "Analysis of the delay and jitter of voice traffic over the Internet", in *Proc. Infocom*, Anchorage, USA, 2001.

[20] M. Mandjes, K. van der Wal, R. Kooij, and H. Bastiaansen, "End-to-end delay analysis for interactive services on a large-scale IP network", in *Proc. 7th IFIP Worksh. Perform. Model. Eval. ATM/IP Netw.*, Antwerp, Netherlands, 1999.

[21] M. Narloch and S. Kaczmarek, "Admission control method based on effective delay for flows using EF PHB", in *Architectures for Quality of Service in the Internet*, W. Burakowski, B. Koch, and A. Bęben, Eds., *Lecture Notes in Computer Science*. Springer-Verlag, 2003, vol. 2698.

[22] M. Narloch and S. Kaczmarek, "Methods for evaluation packet delay distribution of flows using expedited forwarding PHB", in *Proc. 2nd Polish-German Teletraffic Symp. PGTS*, Gdańsk, Poland, 2002, pp. 85–94.

[23] M. Narloch and S. Kaczmarek, "Quality of service problem in IP based network", in *Proc. First Inform. Technol. Conf.*, Gdańsk, Poland, 2003, vol. 1, pp. 301–314 (in Polish).

[24] K. Nichols *et al.*, "A two-bit differentiated services architecture for the Internet", RFC 2638, July 1999.

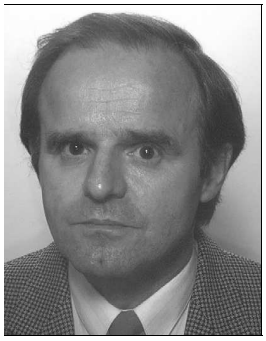
[25] *Broadband Network Teletraffic. Performance Evaluation and Design of Broadband Multiservice Networks. Final Report of COST 242*, J. W. Roberts, U. Mocci, and J. Virtamo, Eds. Heidelberg: Springer-Verlag, 1996.

[26] J. W. Roberts and J. T. Virtamo, "The superposition of periodic cell arrival streams in an ATM multiplexer", *IEEE Trans. Commun.*, vol. 39, no. 2, pp. 298–303, 1991.

[27] S. Salsano *et al.*, "Definition and usage of SLSs in the AQUILA consortium", Internet Draft, work in progress, November 2000, draft-salsano-aquila-sls-00.txt

[28] S. Shenker *et al.*, "Specification of guaranteed quality of service", RFC 2212, Sept. 1997.

[29] M. Vojnovic and J.-Y. Le Boudec, "Stochastic analysis of some expedited forwarding networks", in *Proc. Infocom*, New York, USA, 2002.



Sylwester Kaczmarek received his M.S./B.S. in electronics engineering, Ph.D. and D.Sc. in switching and teletraffic science from the Technical University of Gdańsk, Gdańsk, in 1972, 1981 and 1994, respectively. His research interests include: IP QoS and GMPLS networks, switching, routing, teletraffic and quality of service. He has published more than 125 papers.
e-mail: kasyl@eti.pg.gda.pl
Faculty of Electronics, Telecommunications and Informatics
Gdańsk University of Technology
G. Narutowicza st 11/12
80-952 Gdańsk, Poland

Marcin Narloch was born in Poland in 1974. He received his M.Sc. degree in telecommunications from Gdańsk University of Technology in 1998. Since 1998 he has kept assistant position at Gdańsk University of Technology, Faculty of Electronics, Telecommunications and Informatics. His research activities focus on traffic control in IP QoS and ATM networks.
e-mail: narloch@eti.pg.gda.pl
Faculty of Electronics, Telecommunications and Informatics
Gdańsk University of Technology
G. Narutowicza st 11/12
80-952 Gdańsk, Poland