Fast Packet Switches with Shared Buffer Memory

Liljana Gavrilovska

Faculty of Electrical Engineering, Skopje, Macedonia

Modern networks (such as BISDN, gigabit networks, parallel computer networks, LANs, etc.) introduce fast packet switches as a new concept of a switching node. Fast packet switches which employ shared storage are able to utilize the buffer efficiently. Several analytical techniques to evaluate the performance of shared buffer switches have been proposed. They range from the simple convolution technique which is fast but inaccurate, to the approach proposed by Eckberg and Hou, which is accurate, but computationally slow. This paper proposes a new approach to performance analysis of shared buffer switches, called the ***reduced variance approximation (RVA)***. The new method appears to offer accurate results and efficient computation in comparison to other approaches. Implementation of this method provides reduction in the required shared buffer size.

Keywords: ATM switching; shared buffer; performance analysis; reduced variance approximation

1. Introduction

The ATM (Asynchronous Transfer Mode) has been selected as the multiplexing and switching technique for BISDN (CCITT,1989). Various ATM switching architectures have been proposed (Tobagi,1990). Most of them introduce some buffering strategies in order to avoid cell loss due to contention.

It is recognized that the output buffering offers the best delay and throughput characteristics (Karol,1987), (Yeh,1987). If the output buffer memories are organized so that they are completely shared by all the outputs, reduction in memory size can be achieved (Irland,1978). Also, the packet loss characteristics can be dramatically improved (Hluchyj,1988). Therefore, Shared Buffer Switches appear to be the most promising concept for fast packet switching architectures.

Attention has been focused on performance analysis of shared buffer switches. A new analytical method for performance analysis of shared buffer switches, called the ***reduced variance approximation (RVA)***, is proposed. In this approach, the variance of the input arrival process is reduced due to the *negative correlation* that exists between the streams destined to the different output ports. The main feature of this method is that it gives a comparable accuracy to the accurate method of Eckberg and Hou, but is much more efficient computationally. It also results in reduction of requested shared buffer size.

The brief overview of shared buffer architecture and it's characteristics is given in Section 2. A new approach is proposed in Section 3. The significance of the proposed approach is supported in Section 4, where some numerical results are presented and compared with simulation, convolution and Eckberg's methods. Conclusion follows in the last section.

2. Shared Buffer Architectures

The idea of the shared memory is conceptually realized in different manners inside the switching fabric (Garcia–Jaro,94).

The global concept of switching architecture is presented in Fig. 1. It consists of a *common memory* used for storage of the packets (or cells), until the packets arrived on the different *inputs*, and destined to the different *outputs*, are processed through the switch. The process of

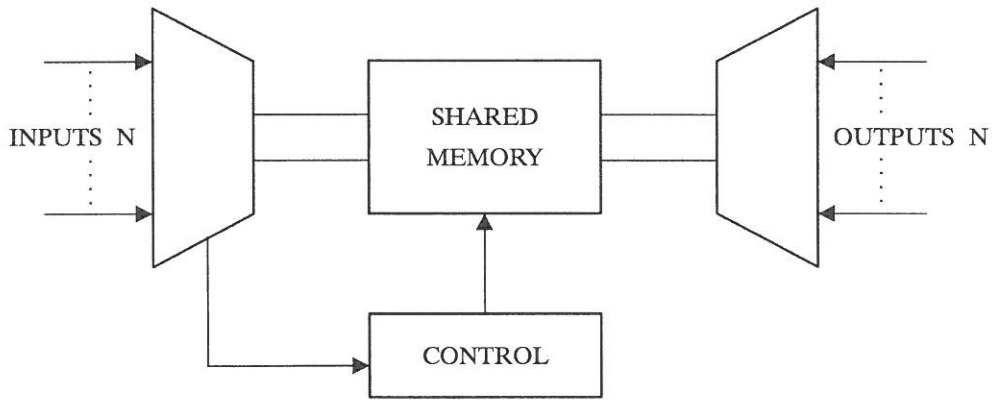

Fig.1. Shared Buffer Switch

write in and *read out of* the memory is controlled by the *control bloc*.

The main characteristics of this concept are:

- The common memory can be accessed by any inlet and outlet on the shared manner;
- Shared buffer approach can be introduced on a switch level (Devault,1988) or on a basic switching element level (Kuwahara,1989);
- The control function, established to maintain management of the flow of the packets, can be centralized (Devault,1988) or distributed (Huang,1984). An example of distributed control is *self-routing* of the packets (Denzel,1992);
- The size of the packets that are switched through the switching fabric depend on the particular architecture. The packets can be ATM cells (Kuwahara,1989) or some internally specified blocks of smaller size (Henrion,1990);
- In order to maintain the speed requirements (speed-up due to multiple access to shared memory), internal parallelization is established. The parallelization can be performed on byte level, packet (cell) level, or in a form of bit-slice configurations of identical shared buffer memories (Kozaki,1991).

The main problem in hardware realization of the shared buffer switches is centralized control of large architectures. That makes them more adequate for smaller switching fabrics (for example, 16 × 16). Processing time through the switch can also be critical. It can be overcome by very fast write in and read out memories and control blocs. The size of the memory block and its price is another important parameter.

The main feature of these architectures is that they utilize the buffers most efficiently. They are insensitive to the unbalanced traffic and burstiness. The lowest loss probabilities can be achieved with minimum number of buffers.

Due to the stricter requirements from the future applications, the fast packet switches should:

- support multicast and broadcast functions;
- be capable of building growable architectures (up to 1024×1024 inlets/outlets) (Garcia–Jaro,1994)
- become independent of external environment (Henrion,1991)
- work with very fast input/output lines, from 150 Mbit/s and 600 Mbit/s up to Gbit/s (Denzel,1992).

Several different arcitectures, based on output shared buffering strategy were proposed in the last few years: Prelude Switch (Devault,1988), Hitachi Switch (Kuwahara,1989), Alcatel's Multipath Self Routing Switch (Henrion,1990), (Srodi,1992), (Henrion,1993), IBM's switch called PRIZMA (Denzel, 1992). Some benefits of shared buffering, in a more general sense, are used in well known archutectures such as Starlite (Huang,1984), Knockout (Yeh,1987), and others. In some architectures, as in Starburst (Widjaja,1992), the principle of shared buffer memory is combined with the dedicated buffering to optimize and improve the performances of the switch.

3. Analysis of Shared Buffer Switches

The ***reduced variance approximation (RVA)*** method is presented in the following Section. The results of the performance analysis are compared with the convolution method, simulation results and Eckberg's method. The convolution approach is computationally efficient but tends to produce quite conservative results, whereas Eckberg's approach is more accurate but is much less efficient. The reduced variance approximation compares favorably with the convolution method in terms of computation speed, but with a much greater accuracy.

We assume that the packet arrival processes on each of the N input ports are independent Bernoulli trials, with parameter p, and each packet is equally likely destined for each output port. Our approach is to first assume that the shared buffer is of infinite capacity. Then we compute the *overflow probability* by truncating the tail of probability distribution beyond the size of the shared buffer.

If we focus on the packets that are destined for a tagged output, then the probability that k packets arrive in the time slot is given by the binomial probability distribution

(1)
$$a(k) \triangleq P[A=k] = \binom{N}{k} \left(\frac{p}{N}\right)^k \left(1-\frac{p}{N}\right)^{N-k}, \quad k = 0, 1, \ldots, N.$$

The probability generating function (*PGF*) of the random variable A is (Kleinrock,1976):

$$(2) \quad G_A(z) = \sum_{k=0}^{\infty} P[A=k] z^k = \left(1 - \frac{p}{N} + z\frac{p}{N}\right)^N.$$

Let the random variable Q denote the steady-state number of packets in the tagged output buffer. The *PGF* of the random variable Q can be easily obtained (Karol,1987):

$$(3) \qquad G_Q(z) = \frac{(1-p)(1-z)}{G_A(z) - z}$$

Let random variable S denote the total number of packets in the shared buffer. In the convolution method, the random variables Qj's (the number of packets in output buffer j) are assumed independent. Thus, the distribution of S is simply N-fold convolution of the distributions of Q's. In the z-transform domain, the *PGF* of the random variable S can be expressed as $G_{s(z)} = [G_{q(z)}]^N$. Thus, the distribution of S can be obtained by inverting its *PGF* using, for example, the Fast Fourier Transform (*FFT*) algorithm.

Eckberg and Hou were the first to note that, during this computation, large errors can occur, when the switch size is small and the offered load is high (Eckberg,1988). In particular, it is noted that the total number of packet arrivals at each time slot destined for individual output ports is not independent, and, in fact, they are *negatively correlated*. It means that if all the packets are destined to one particular output, then no packets are addressed to the other outputs. Because of the negative correlation, the random variable S would be stochastically smaller than the sum of the random variables Q's, assuming the output queues are independent. In fact the variance of S is given by

(4)
$$\mathrm{Var}(S) = N\,\mathrm{Var}(Q_i) + N(N-1)\,\mathrm{Cov}(Q_i, Q_j)$$

where $\mathrm{Cov}(Q_i, Q_j)$ is the covariance between any parts of individual buffers. The mean of S is still given by the sum of means of individual buffers, which is:

$$(5) \qquad E[S] = NE[Q].$$

Based on the information of this two moments, (Eckberg,1988) proposes to use the Gamma distribution to approximate the distribution of S. Unfortunately, the computation of $\mathrm{Cov}(Q_i, Q_j)$ in (4) requires the computation of the joint distribution of any pair of two queues. Moreover, it is found that the joint generating function of the two queues can not be determined analytically. They propose an iterative numerical method for solving the joint generating function.

Recall that the number of arriving packets to the tagged output port at each time slot forms a Binomial distribution. Thus the expected number of packet arrivals is $E[A] = p$ and the variance is $\mathrm{Var}(A) = p(1 - p/N)$. The approximation of the shared buffer switch, assuming independence among the output queues, would give a conservative result. This is because of the existence of the negative correlation among the packet streams destined for different out-

put ports. To be able to catch this correlation we need to look at the arriving process to the entire shared buffer system. A simple computation shows that the variance of the arriving process to the entire shared buffer system is less than the sum of the variances of the individual arriving processes to the output queues, assuming independence among them. Thus, it appears natural to reduce the variance of the packet arrivals to a given output, to improve the approximation, while keeping the mean unchanged. Through extensive experiments, we have found that choosing the variance of the arrival process to be $p(1-p/\sqrt{N})$ would virtually transfer the effect of the negative correlation to the input process and tends to give good estimates on the performance. Given the mean and the reduced variance, our next task is to find the associated new packet arrival distribution to an output queue.

At this point, we employ the maximum *entropy method (MEM)* to find the distribution satisfying the constrains of the mean and the variance (Garcia,1994), (Kouvatos,1989). Specifically, we consider the following problem:

$$\max\Big\{-\sum_{i=0}^{N}\hat{a}_i \ln \hat{a}_i\Big\} \tag{6}$$

subject to:

$$\begin{aligned}\sum_{i=0}^{N}\hat{a}_i &= 1\\ \sum_{i=0}^{N} i\hat{a}_i &= \overline{A}\\ \sum_{i=0}^{N} i^2\hat{a}_i &= \overline{A^2}\end{aligned} \tag{7}$$

where $\hat{a}_i$ is the probability that there are i packet arrivals destined to a given output port in a time slot, $\overline{A} = p$ (the means), and $\overline{A^2} = (1 - p/\sqrt{N})p + p^2$ (the second moment). If the constraints are absent from the problem, the distribution that maximizes the entropy is known to have a uniform distribution. In our case, where the mean and the second moment (or the variance) are predetermined, the optimization can be found through the Lagrange multiplier approach. Applying this approach the distribution can be obtained by maximizing the expression

$$\begin{aligned}&-\sum_{i=0}^{N}\hat{a}_i \ln \hat{a}_i + \lambda_1\Big(\sum_{i=0}^{N}\hat{a}_i - 1\Big)\\ &+ \lambda_2\Big(\sum_{i=0}^{N} i\hat{a}_i - \overline{A}\Big) + \lambda_3\Big(\sum_{i=0}^{N} i^2\hat{a}_i - \overline{A^2}\Big)\end{aligned} \tag{8}$$

by differentiation with respect to $\hat{a}_k$. The constants $\lambda_1, \lambda_2, \lambda_3$ are the Lagrange multipliers which need to be determined. After differentiating the expression (8), setting the derivative to zero, and solving for $\hat{a}_k$, we have

$$\hat{a}_k = e^{(-1+\lambda_1+k\lambda_2+k^2\lambda_3)}. \tag{9}$$

Applying the constrains in (7) to the solution in (9), we obtain the following system of nonlinear equations:

$$\begin{aligned}\sum_{k=0}^{N} e^{(k\lambda_2+k^2\lambda_3)} &= e^{-(\lambda_1-1)}\\ \sum_{i=0}^{N} k e^{(k\lambda_2+k^2\lambda_3)} &= \overline{A}e^{-(\lambda_1-1)}\\ \sum_{i=0}^{N} k^2 e^{(k\lambda_2+k^2\lambda_3)} &= \overline{A^2}e^{-(\lambda_1-1)}\end{aligned} \tag{10}$$

The problem of finding the constraints λ's in (10) translates to the problem of finding the root of a multidimensional function. This system of nonlinear equations must be solved numerically. Fortunately, there exists an algorithm based on a multidimensional Newton–Raphson method that can solve the constraints very efficiently (Press,1990).

Having found the *modified* distribution of the packet arrivals to an output port $\hat{a}_k = \hat{a}(k) = P[\hat{A} = k]$, we compute the distribution of packets in an output buffer of infinite size, $q(k)$, as

(Karol,1987)

(11)

$$q(0) = \frac{(1-p)}{\hat{a}(0)}$$

(12)

$$q(1) = \frac{(1-\hat{a}(0)-\hat{a}(1))}{\hat{a}(0)} q(0)$$

(13)

$$q(k) = \frac{(1-\hat{a}(1)}{\hat{a}(0)} q(k-1) - \sum_{i=2}^{k} \frac{\hat{a}(i)}{\hat{a}(0)} q(k-1)$$

The distribution of packets in the shared buffer, $P[S = k]$, can be obtained by performing N-fold convolutions of $q(k)$'s. Finally, the overflow probability can be computed by truncating the tail of the probability distribution beyond the size of the shared buffer.

4. Performance Comparison

In this section we investigate the accuracy of the proposed reduced variance approximation against other methods. For the purpose of comparison, we also include the convolution method, the Eckberg's method, and the simulation results. Fig. 2 plots the overflow probability versus the shared buffer size for a switch of size 16×16. The shared buffer size is normalized with respect to the switch size N. The figure provides two sets of curves at offered loads $p = 0.8$ and $p = 0.9$. The simulation results are provided with 95% confidence interval. It is obvious from the figure that the convolution method gives very conservative results, which in general can be orders of magnitude larger than the simulation results. Particularly, at overflow probability 10^{-6} and offered loads from $p = 0.8$ to $p = 0.9$, the overestimation from 20% to 25% of the required buffer size appears. This value grows up at lower overflow probabilities. The Eckberg's approach gives more accurate results then the reduced variance approximation at high overflow probability (> 0.01). However, at low overflow probability ($< 10^{-6}$) which is the natural operating region, the estimate given by the reduced variance approximation is competitive with that given by Eckberg's approach. In terms of the computation time, the reduced variance approximation is comparable to the convolution method and is much more efficient than the Eckberg's approach.

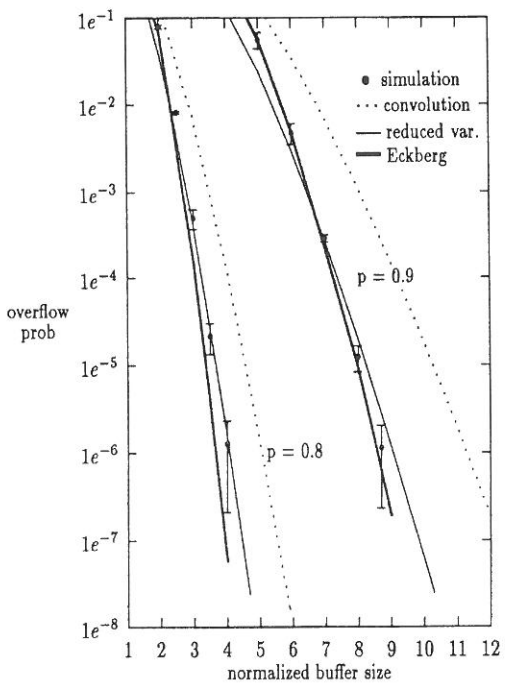

Fig.2. Overflow probability versus normalized buffer size for a switch of size 16×16

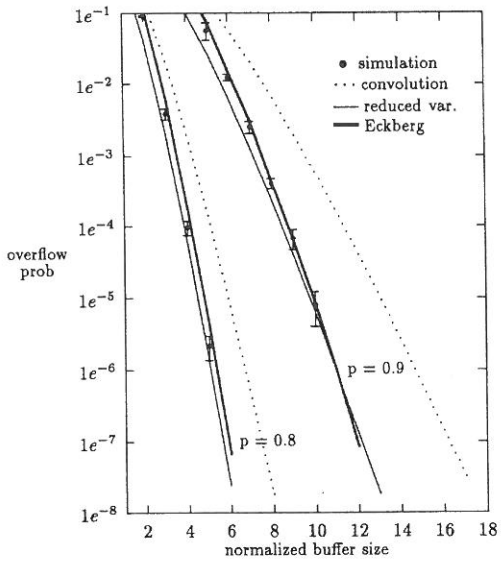

Fig.3. Overflow probability versus normalized buffer size for a switch of size 8×8

Fig. 3 shows the corresponding curves for a switch of size 8×8. Note that both, the reduced variance approximation and the Eckberg's approach are very accurate. On the other hand, the convolution results become poor for small

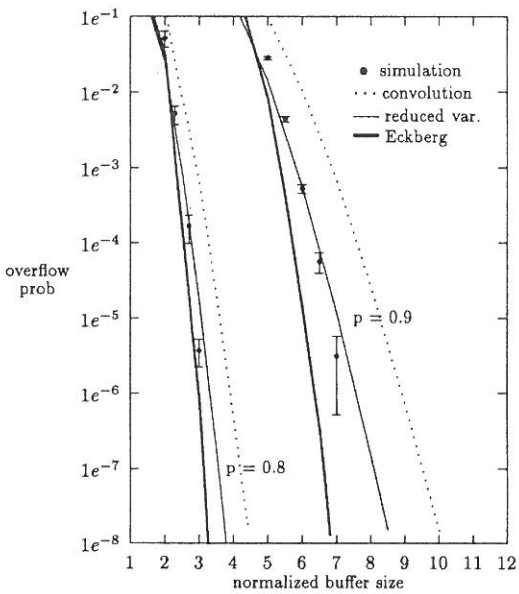

Fig.4. Overflow probability versus normalized buffer size for a switch of size 32 × 32

switch size. The reduction in the required buffer size for overflow probability 10^{-6}, and offered loads $p = 0.8$ and $p = 0.9$, varies from 20% to 36%.

Fig.4 shows the corresponding curves for a switch of size 32 × 32. We note that Eckberg's approach loses its accuracy at high loads while the reduced variance approximation remains highly accurate over a wide range of traffic loads. We also note that the convolution method improves as the switch size increases. This is because the negative correlation among the arrival processes to different output queues diminishes as the number of inputs/outputs increases. The reduction in the required buffer size for the overflow probability 10^{-6}, in this case, goes from 18% to 20%.

5. Conclusion

In this paper, we have developed a simple and accurate method, called the ***reduced variance approximation***, for computing the performance of shared buffer ATM switches. The basic idea of the method is that the negative correlation existing among the packet streams destined to different outputs can be used to reduce the variance of packet arrivals destined to a given output. The modified distribution of packet arrivals, given the mean and the reduced variance, is then found through the maximum entropy method, and the shared buffer distribution is found through N-fold convolution of individual output queues. We have investigated the performance of shared buffer switches under random traffic. Introducing the negative correlation, for overflow probability 10^{-6} and different switch size, we acquire reduction in the required buffer memory from 18%–36%. It is of current interest to extend the analytical method to the case of bursty traffic.

Acknowledgments

The author would like to thank Prof. A.Leon–Garcia, from the Department of Electrical Engineering, University of Toronto and I.Widjaja, from the Teletraffic Research Center, University of Adelaide, for leadership and helpful suggestion during this work.

References

H. BRUNEEL, B. STEYAERT (1988), "Tail Distribution of Shared Buffer Queue Contents", Technical Repoprt CNET–123–030–CD–CC, CNET, France CCITT, (1989), CCITT Recomendations I.121, "Broadband Aspects of BISDN", Blue Book, Geneva, Switchzerland

W. E. DENZEL, A. P. J. ENGBERSEN, I. ILIADIS, G. KALSSON (1992), "A Highly Modular Packet Switch for Gb/s Rates," Proceedings of ISS, Yokohama, Japan, vol.2;

M. DEVAULT , J. COCHENNEC, M. SERVEL (1988), "The Prelude ATD Experiment: assessments and future prospects", IEEE Journal on Selected Areas in Communications, vol. SAC–6, pp. 1528–1537;

A. ECKBERG , T. HOU (1988), "Effects of Output Buffer Sharing on Buffer Requirements in an ATM Packet Switch" in Proceedings of INFOCOM, New Orleans, Lousiana, pp. 459–466;

J. GARCIA–JARO , A. JAJZCZYK (1994), "ATM Shared Memory-Switching Architectures," IEEE Network Magazine, vol. 8, No. 4, pp. 18–26;

M. HENRION , G. EILENBERGER, G. PETTIT, P. PARMENTIER (1990), " A Multipath Self-Routing Switch", IEEE Communications Magazine, Vol. 31, No. 4, pp. 46–52;

M. HENRION, K. SCHRODI, D. BOETTLE, M. DE SOMER, M. DIUEDONNE (1990), "Switching Network Architecture for ATM Based Broadband Communications", in Proceedings of ISS, Stochkholm, Sweden, vol. V, pp. 1–7;

M. HLUCHYJ , M. KAROL (1988), "Queueing in High Performance Packet Switching", IEEE Journal on Selected. Areas in Communication, vol. 6, No. 9, pp. 1587–1597;

A. HUANG, S. KNAUER (1984), " Starlite: A Wideband Digital Switch", in Proceedings of GLOBECOM, Atlanta, GA, pp. 121–125;

M. IRLAND (1978), "Buffer Management in a Packet Switch", IEEE Transactions on Communications, vol. 26, No. 3, pp. 328–337;

M. KAROL, M. HLUCHYJ, S. MORGAN (1987), "Input versus Output Queueing on a Space-Division Packet Switch", IEEE Transactions on Communications, vol. 36, No. 12, pp. 1347–1356;

L. KLEINROCK (1976), Queueing Systems, VolumeI: Theory, New York: John Wiley & Sons;

D. KOUVATOS, N. XENIOUS (1989), "MEM for Arbitrary Queueing Networks with Multiple General Servers and Repetitive-service Blocking", Performance Evaluation, No. 10, pp. 169–195.

T. KOZAKI, N. ENDO, O. SAKURAI, M. MATSUBARA, K. ASANO (1991), "32 × 32 Shared Buffer Type ATM Switch VLSI's for B-ISDN's", IEEE Journal on Selected Areas in Communications, vol. 9, No. 8, pp. 1239–1247;

H. KUWAHARA, N. ENDO, M. OGINO, T. KOZAKI (1989), "A Shared Buffer Memory Switch for an ATM Exchange", in Proceedings of ICC, Boston, MA, pp. 4.4.1–4.1.5;

A. LEON–GARCIA (1994), Probability and Random Processes for Electrical Engineering, Addison–Welsley, Second Ed., 1994;

G. PETTIT, E. DESMET (1990), "Performance Evaluation of Shared Buffer Multiserver Output Queue Switces Used in ATM", in Proceedings of 7–th ITC Seminar, Morristown, New Jersey;

W. PRESS , B. FLANNERY, S. TEUKOLOSKY, W. VETTERLING (1990), Numerical Recipes in C, Cambridge University Press, Cambridge;

K. SCHRODI , B. PLEIFFER, J. DELMAS, M. DE SOMER (1992), "Multicast Handling in a Self-Routing Switch Architecture", in Proceedings of ISS, Yokohama, vol. 2, pp. 156–160;

F. TOBAGI, (1990), "Fast Packet Switch Architectures for Broadband Integrate Services Digital Network", in Proceedings of IEEE, vol. 78, No. 1, pp. 133–167;

I. WIDJAJA , A. LEON–GARCIA (1992), "Starburst: A Flexible Output-Buffered ATM Switch with $N \log^2 N$ Complexity", in Proceedings of ISS, Yokohama, Japan, vol. 2, pp. 226–230;

Y. YEH, M. HLUCHYJ, A. ACAMPORA (1987), "The Knockout Switch: A Simple Modular Architecture for High-Performance Packet Switching", IEEE Journal on Selected Areas in Communications, vol. 5, No. 8, pp. 1274–1283;

Received: October, 1994
Accepted: January, 1995

Contact address:
Liljana Gavrilovska
Faculty of Electrical Engineering
Orce Nikolov bb
91 000 Skopje
Republic of Macedonia

LILJANA GAVRILOVSKA was born in Moscow, 1951. She received her B.Sc. degree from the University of Skopje and M.S. degree from the University of Belgrade, both in electrical engineering.

She is currently an Assistant Professor on the Department of Telecommunication, Faculty of Electrical Engineering of Skopje, Republic of Macedonia.

She spent a year (1992–1993) with the communication group at the University of Toronto, Canada. As a part of her Ph.D. thesis she was working on fast packet switching analysis.

Her research interest include ATM, queueing theory, fast packet switches architecture, teletraffic analysis.