Fast Cell Loss Rate Estimation of ATM Switches Using Importance Sampling

Junjie Wang*, K. Ben Letaief, M. Hamdi, and Xi-Ren Cao*
The Hong Kong University of Science and Technology
Hong Kong, P. R. China

Abstract

The performance evaluation of ATM switches is of paramount importance in designing an ATM network. In this paper, we focus on the evaluation of the cell loss rate (CLR) in nonblocking ATM switches using computer simulations. In particular, we investigate the potential of using importance sampling techniques as an "superfast" alternative to conventional Monte Carlo simulation in finding the CLR in nonblocking ATM switches. We propose a "split switch" method to decouple the input and output queue behaviors, along with the notion of regenerative cycles to achieve fast and accurate results. Numerical results will demonstrate that considerable computation cost can be saved using these proposed importance sampling techniques while maintaining a high degree of accuracy.

1 Introduction

The Asynchronous Transfer Mode (ATM) has widely been adopted as the transfer mode solution for future broadband ISDN. One of the most crucial design components of an ATM network is the ATM switch. Accordingly, a large number of architectures have been proposed for ATM switching systems. Among these architectures, nonblocking space-division switches have received the most attention because they present the best compromise in terms of hardware cost and switching performance [1, 2]. However, with this architecture come various challenges. For example, a cost-effective selection of speed-up parameters as well as a proper selection of buffer sizes should be carefully considered in these nonblocking ATM switches to guarantee a specified *Quality of service* (Qos) requirement for a given traffic (i.e., cell loss rate (CLR), maximum delay and delay jitter, etc.).

In this research, we focus on the performance of ATM switches with respect to CLR. Particularly, we consider the effects of speed-up factor and buffer sizes on the CLR of nonblocking ATM switches. These are clearly such important issues in the switch design that a sizable amount of work has been done on the CLR analysis [3-5]. Since ATM networks are intended for the transport of a variety of services with complex traffic conditions, it follows that closed form and/or tractable analytical solutions for ATM networks are quite difficult to obtain without resorting to some approximation [3].

Conventional Monte Carlo (MC) simulations offer us an attractive alternative for the estimation of the CLR of nonblocking ATM switches. Unfortunately, the required CLR for a typical ATM switch is smaller than 10^{-6} for most practical applications. Hence, a prohibitive number of simulation trials must be used to obtain accurate estimate of CLRs for a particular accuracy.

*Supported in part by Hong Kong UGC grant under HKUST 690/95E.

Importance Sampling (IS) [6-8] is one of the most promising techniques which can significantly reduce the simulation run time required to obtain accurate estimates. Specifically, the underlying probability distribution of the system inputs is replaced by another biased probability distribution which "favors" the occurrence of rare and important events such as cell loss. The IS estimator then weights the simulation data by an appropriate likelihood ratio in order to get an unbiased estimate.

In this paper, we illustrate the use of IS for the efficient and accurate estimation of the CLR in nonblocking ATM switches. It must be noted that the application of IS to the estimation of the CLR in ATM switches has been considered in [5, 7]. However, in these previous works, the switch model was a space-division ATM switch with only output queues. In our work, we consider the more practical case of ATM switches with both input and output queues, thereby, making our cell loss analysis more complicated than output-queued switches. This is because we have to confront the correlation between the HOL cells in different input queues, which remains the most obstacle in the analysis.

The rest of the paper is organized as follows. Section 2 gives a brief introduction of IS and regenerative simulation of rare events. Section 3 presents our proposed IS approaches for the simulation of ATM switches with both input and output queues using the notion of "*split switch*" model. Sample simulation results which illustrate the accuracy and efficiency of the proposed schemes are also included. Finally, we conclude in Section 4.

2 Importance Sampling

2.1 Monte Carlo Simulation

To illustrate the major problem of simulating a rare event by MC method, consider the following simplified example. Let X be a random variable with probability density function ($p.d.f.$) $f(x)$ and suppose that we are interested in estimating the probability, α, that X is in some set or event E. That is,

$$\alpha = \mathrm{E}[I_E(X)] = \int I_E(x) f(x)\, dx \qquad (1)$$

where the integral in the above expression can be replaced by a summation in the case of a discrete random variable. The variable $I_E(.)$ is the indicator random variable of the set E. That is, $I_E(x) = 1$ if $x \in E$. Otherwise, $I_E(x) = 0$.

Using MC simulation, the estimator for α is simply the sample mean estimator which generates L_{mc} i.i.d. (independent and identically distributed) random samples $X^{(1)}, ..., X^{(L_{mc})}$ from the density $f(.)$, and computes the

MC estimate of α as

$$\hat{\alpha} = \frac{1}{L_{mc}} \sum_{\ell=1}^{L_{mc}} I_E(X^{(\ell)}). \quad (2)$$

Note that if α is small, then we would not expect to "hit" the set E very often during the L_{mc} simulations. This then would require that L_{mc} must be very large to insure that $\hat{\alpha}$ is close to α. Normally, $100/\alpha$ samples are required to guarantee a 10% confidence interval.

2.2 Importance Sampling

Let $f^*(.)$ be a new probability density function such that $f^*(x) > 0$ whenever $f(x) > 0$. Importance sampling involves choosing $f^*(.)$ as the simulation density (instead of the true density $f(.)$) and observing that Eqn. (1) can be rewritten as

$$\alpha = \int I_E(x) w(x) f^*(x)\, dx \quad (3)$$

where $w(.)$ is the likelihood ratio of the true density $f(.)$ to the new density $f^*(.)$. That is,

$$w(x) \triangleq \frac{f(x)}{f^*(x)}. \quad (4)$$

Importance sampling can now be obtained by an "empirical evaluation" of the integral (3) instead of (1). For $\ell = 1, ..., L_{is}$, one generates a sequence of i.i.d. random numbers $X^{(1)}, ..., X^{(L_{is})}$ from the IS simulation density $f^*(.)$. The IS estimator for α is

$$\hat{\alpha}^* = \frac{1}{L_{is}} \sum_{\ell=1}^{L_{is}} I_E(X^{(\ell)}) w(X^{(\ell)}). \quad (5)$$

and it can be proved that $\hat{\alpha}^*$ is an unbiased estimator.

Typically, there are two commonly used figures of merit for IS schemes. One is *accuracy*, which is defined as the estimator standard deviation as a percentage of the true estimate. Generally, a 10% accuracy is often used as the performance measure. The other is the *efficiency*, which is defined as the ratio of sample size in MC to that in IS required to achieve the same accuracy.

2.3 Regenerative Simulations

The *regenerative method* is a simulation technique which is often used in the simulation of stochastic systems and in particular for the estimation of the steady state performance of such systems [9]. The basic principle is that there exists a sequence of random times, called regenerative times, such that at each of these times the random process starts anew according to the same probability structure.

The significance of the regenerative method is that observations or realizations obtained from different regenerative cycles are i.i.d.. Thereby, estimates can be made cycle by cycle, which makes the application of IS more manageable. In this paper, we will combine the concept of the regenerative method with IS to find CLR in a nonblocking ATM switch. The CLR, γ, will be defined as the ratio of average number of lost cells in a cycle to the average number of arriving cells in a cycle. That is,

$$\gamma = \frac{\mathbf{E}[\text{Number of cells lost in a cycle}]}{\mathbf{E}[\text{Offered cells in a cycle}]}. \quad (6)$$

3 Fast Estimation Of CLR

In this section, we consider the application of IS to the estimation of the CLR in input/output queuing nonblocking ATM switches. We denote N as the switch dimension, K, L as the input and output queue capacity, respectively, and m as the speed-up factor. We begin by considering ATM switches with only output queues. Some IS schemes developed in this case can be applied to the more complicated case of estimating the CLR of ATM switches when both input and output queues are present. In our work, the traffic is assumed randomly uniform with traffic load λ. In practice, the traffic characteristics rarely comply to this assumption. However, our major focus in this paper is to investigate the potential of using IS in the CLR estimation of ATM switches and develop some useful IS schemes, which yields much insight into the practical case.

3.1 Output-Queued ATM Switches

Under the uniform traffic assumption, the arrival process at each input port has the binomial distribution

$$\mathtt{P}_k = \mathtt{Pr}(A_n^i = k) = \binom{N}{k} (\frac{\lambda}{N})^k (1 - \frac{\lambda}{N})^{N-k}, k = 0, ..., N \quad (7)$$

where A_n^i denotes the number of cells destined for a particular output port i in the n-th slot. In this model, up to m cells are to be selected randomly if $A_n^i > m$. The rest of the cells are lost since there is no input queues. Furthermore, if the selected cells do not find enough space in the output queue, they will also be lost.

A close observation of Eqn. (7) reveals that the only randomness in the system is the arrival process. Hence, an intuitive way of applying IS in this application is to use a biasing model which increases the arrival of cells. Three different biasing schemes will be investigated in this section. A comparison of the efficiency of these schemes is also undertaken. Throughout Section 3, the *regenerative cycle* is defined as the time interval between two successive points when the tagged queue is emptied. The biased distribution is used at the beginning of each cycle, and then we switch to the original distribution when cell loss occurs. This is done to keep the system stable as well as to make the length of regenerative cycles finite.

A. Direct Bias

Intuitively, a larger traffic load can generate more cells into the switch, thereby, "forcing" more cells to be lost. To do so, one can bias the arrival process by using a new arrival rate $\lambda^* > \lambda$ so that the biased governing distribution under the IS model is given by [1]

$$\mathtt{P}_k^* = \binom{N}{k} (\frac{\lambda^*}{N})^k (1 - \frac{\lambda^*}{N})^{N-k}, \quad k = 0, 1, ..., N. \quad (8)$$

The bias parameter λ^* should be selected to achieve the smallest estimate variance.

B. Exponential Bias

In many previous efforts, the exponential twisting has been shown to be an asymptotically optimal biasing method among a large class of sampling distributions. Because of the optimality of the exponential twisting which involves an exponential change of measure, one can argue that a biasing approach which weights the probability $\mathtt{P_k}$ by an exponential bias can be employed. That is,

$$\mathtt{P}_k^* = \frac{\xi^k \mathtt{P}_k}{\sum_{i=0}^{N} \xi^i \mathtt{P}_i}, \quad k = 0, 1, ..., N \quad (9)$$

[1] Throughout the paper, the superscript $*$ indicates the distribution under the biased IS model

where ξ is the bias parameter. Similar to the case of the direct bias, the performance of this methodology will depend on the choice of ξ.

C. Uniform Bias

In the original binomial distribution, only 0 or 1 cell arrives at the tagged output queue in a slot most of the time, which essentially makes no contribution to the estimate of cell loss. The uniform bias ensures that the cell sources have an equal probability to generate $0, 1, ..., N$ cells in a slot. That is,

$$\mathtt{P}_k^* = \frac{1}{(N+1)}, \quad k = 0, 1, ..., N. \tag{10}$$

The key advantage of the uniform bias is that there is no need to find an optimal bias parameter.

D. Figure of Merit and Comparison

Fig. 1 shows some simulation results which illustrate the accuracy and potential of the proposed IS methods. It lists the IS estimates of the CLR of an ATM switch with $\lambda = 0.4$, $N = 8$ and $L = 10$ as a function of the speed-up factor m. These IS estimates were obtained using the above IS methods. For comparison, we also list the CLR of the above switch that were obtained using conventional MC simulation (using the original unbiased model). As can be easily seen, our IS results are in excellent agreement with the MC results, which are indicative of the exact and true system performance.

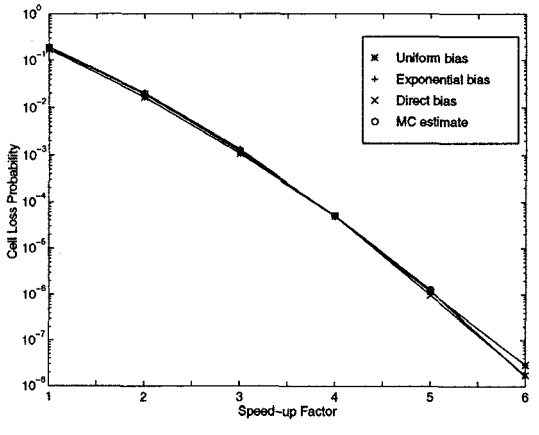

Figure 1: CLR estimation for output-queued switch, $\lambda = 0.4$, $N = 8$ and $L = 10$.

To further illustrate the efficiency of the proposed IS methods, we present in Table 1 the computational cost in terms of the number of cell slots generated during the simulation to achieve a 10% accuracy. In this table, the switch parameters are the same as those in Fig. 1 with the speed-up factor set to 5. The CLR obtained using MC method (denoted as γ in Table 1) was found to be 9.62×10^{-7} and the number of cell slots required to achieve a 10% accuracy was 7.5×10^8. The estimates obtained from the different IS methods (denoted as $\hat{\gamma}$) are also shown in Table 1 along with the number of cell slots required to achieve a 10% accuracy during the IS simulation trials. Using the number of slots one can then compute the amount of improvement that IS provides. A close observation of Table 1 indicates that all IS schemes result in large computational savings.

Table 1: The precision and improvement with IS ($\lambda = 0.4, N = 8, L = 10$ and $m = 5$)

Biasing Scheme	Precision $\frac{\hat{\gamma}-\gamma}{\gamma} \times 100\%$	Number of Slots	Efficiency
Direct Bias	5.4%	6.93×10^5	1.1×10^3
Exp. Bias	1.6%	1.81×10^6	4.1×10^2
Uniform Bias	4.7%	4.64×10^4	1.6×10^4

3.2 ATM Switches with Both Input and Output Queues

In this case, it may be lost at both input and output queues when the cell coming to the input and output queues find no space for it. We denote the total number of arriving cells at the switch as N_a and the number of lost cells due to input and output queue overflow as N_i and N_o. Thus, the CLR for the input queue is $\gamma_i = \frac{N_i}{N_a}$ and that for the output queue is $\gamma_o = \frac{N_o}{(N_a - N_i)}$. Since N_i is quite small compared with N_a, we can obtain the CLR for the whole switch as $\gamma = \frac{N_o + N_i}{N_a} \approx \gamma_i + \gamma_o$ without loss of accuracy.

One of the key contributions in this paper is the proposal of a "*split switch*" model, which estimates γ_i and γ_o, respectively, then combine them to get the overall result. However, γ_i (γ_o) is not the CLR of the conventional ATM switch with only input (output) queues. In our "*split switch*" model, two variants of input/output queuing schemes, named VIQ (Variant of Input Queuing) and VOQ (Variant of Output Queuing), as shown in Fig. 2, are developed to precisely describe the original switch model.

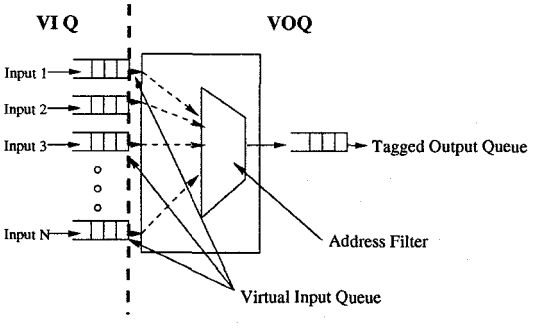

Figure 2: VIQ and VOQ schemes.

In the traditional input queuing scheme, the speed-up factor m is usually equal to one since the HOL blocking forms a bottleneck for the throughput of the output line. Therefore, no gain can be achieved through the use of multi-plane switch fabrics (i.e., having a speed-up factor $m > 1$). However, in order to be equivalent to the original switch model, the original speed-up factor m ($m > 1$) is kept in the VIQ scheme. Thus, up to m cells will be selected during an output contention. On the other hand, in the traditional output queuing schemes, some cells may be lost if not selected in the output contention. In our VOQ scheme, those cells will not be lost. Instead, they can stay

in the HOL of the *"virtual input queue"* which will be explained later. Specifically, the VOQ can be viewed as a combination of: (1) a group of virtual queues; (2) an address filter to solve output contention; and (3) a tagged output queue. Note that, the importance of our *"split switch"* model is not confined to CLR analysis. For example, the cell delay in ATM switch with both input and output queues can be obtained by taking the convolution of the cell delay in VIQ and VOQ schemes.

3.2.1 Cell Loss in VIQ Scheme

In the VIQ scheme, we can focus on a single input queue i, namely a tagged queue, to obtain the CLR estimate. Let I_n be the length of the tagged queue in the n-th slot. We also denote the number of arriving cells at the tagged queue as H_n and the number of departing cells from the tagged queue as G_n at n-th slot. Hence, the state transfer equation is given by

$$I_n = \max\{\min\{K, I_{n-1} + H_n - G_n\}, 0\} \tag{11}$$

where

$$\Pr(H_n = 1) = \lambda \tag{12}$$

$$\Pr(G_n = 1) = \min\{1, \frac{m}{D_n}\}. \tag{13}$$

In Eqn. (13), D_n is the number of HOL cells which have the same destination as the head of the tagged queue. If the cell coming to the tagged queue finds the queue full, i.e., $I_n = K$, then it will be lost. Note that although we focus on a single input queue, there exist correlations between the HOL cells in different queues. Hence, it is difficult to get a closed-form solution to such a problem without some independence approximations [3].

The IS scheme we developed in VIQ scheme is to bias the probability of a cell to be selected in the output contention. That is, we make the cell in the tagged queue less likely to be selected in the output contention. Thus, it is more likely to stay in the queue to hold back the arriving cells. In brief, this is done as follows: Suppose that the head of the tagged queue is destined for output j, the cells which are also destined for output j in all input queues except the tagged one, i.e., $D_n - 1$ cells, contend for $(m-1)$ winners. Then, only one chance of being selected exists. After that, all the failing cells in the above selection plus the tagged one contend for the last chance.

Using this procedure, we bias the departure process of the tagged queue as follows:

$$\Pr^*(G_n = 1) = \begin{cases} 1, & \text{when } D_n \le m \\ \frac{1}{D_n - (m-1)}, & \text{when } D_n > m \end{cases} \tag{14}$$

3.2.2 Cell Loss In VOQ Scheme

Let A_n^j be the number of cells destined for output port j which come to the HOL of input queues in the n-th slot. Next, we denote C_n^j as the number of cells destined for output port j in the n-th slot. Then, it follows that

$$C_n^j = \max\{C_{n-1}^j - m, 0\} + A_n^j. \tag{15}$$

Now, let's consider the evolution of the tagged output queue and suppose that O_n^j is the length of the output queue j during the n-th slot. Likewise, let S_n^j denote the number of cells which arrive at output j during the n-th slot. Then, the process of O_n^j is given by

$$O_n^j = \max\{O_{n-1}^j - 1, 0\} + S_n^j. \tag{16}$$

It is clear that S_n^j will never exceed the speed-up factor m and that $S_n^j = C_n^j$ if $C_n^j \le m$. Otherwise, m cells will be selected from these C_n^j cells. As a result, we have

$$\Pr(S_n^j = k) = \begin{cases} \Pr(C_n^j = k) & \text{when } k < m \\ \sum_{l=m}^{N} \Pr(C_n^j = l) & \text{when } k = m \\ 0 & \text{otherwise.} \end{cases} \tag{17}$$

A close observation of (15)-(17) indicates that the only randomness in the system is A_n^j, which is in the form of a binomial distribution that is given by

$$\Pr(A_n^j = k) = \binom{F_n}{k} (\frac{1}{N})^k (1 - \frac{1}{N})^{F_n - k} \tag{18}$$

where F_n is the total number of cells coming to the HOL of all input queues in the n-th slot, i.e., $F_n = \sum_{j=1}^{N} A_n^j$. Intuitively, we can derive the probability mass function of A_n^j, then we can use some IS schemes to bias it to improve the estimation efficiency. However, note that F_n is not constant but depends on all cell sources, which makes it difficult to derive an explicit probability mass function of F_n. Thereby, making the application of IS not straightforward.

Now, let's turn to the VOQ scheme. Recall that in this case, there is only output queuing. The key problem is to generate the cells with the same characteristics as A_n^j in the absence of input queues. First, we define the concept of *"virtual input queue"*, which is a logical queue in contrast to the physical input queue in the original switch model, but acts just like the physical input queues. The name "virtual" is used because in the VOQ scheme of the *"split switch"* model, we assume no queuing at the switch input line. The only goal of the *"virtual input queue"* is to generate the cells with the same statistical properties of A_n^i as stated in (18). That is, it is simply a cell generator. Two biasing schemes are developed for the VOQ scheme as follows.

A. Accurate Biasing Scheme

In the original switch model, An incoming cell is equal likely to be destined for any output port. That is,

$$\Pr(Destination = j) = \frac{1}{N}, \quad j = 1, 2, ..., N. \tag{19}$$

To apply IS, we bias the routing probability such that the incoming cells are more likely to be destined for the tagged output queue (for simplicity, the tagged output queue is labeled as 1). That is,

$$\Pr^*(Destination = j) = \frac{M}{N}, \quad j = 1, 2, ..., \frac{N}{M} \tag{20}$$

where M is defined as the *routing weight*. As a result, the distribution of A_n^j is biased as

$$\Pr^*(A_n^j = k) = \binom{F_n}{k} (\frac{M}{N})^k (1 - \frac{M}{N})^{F_n - k}. \tag{21}$$

The scheme is called "accurate" since no approximation is made here (in contrast to the other biasing scheme we formulate below).

B. Approximate biasing scheme

In [3], it has been demonstrated that when the dimension of ATM switches, N, goes to infinity, A_n^j is subject to a Poisson distribution with rate λ. That is, we have

$$\Pr(A_n^j = k) = \frac{\lambda^k e^{-\lambda}}{k!}. \quad (22)$$

Recall that, A_n^j is the arrival process for a VOQ scheme. Therefore, we can use the IS schemes similar to those we developed in section 3.1 to bias the "*virtual input queues*". Such an approximation is reasonable when a large-scale ATM switch is considered and/or if only a rough estimate is required.

After obtaining the CLR in VIQ and VOQ schemes respectively, we can simply add them up, to get the overall CLR of the ATM switches as shown in Fig. 3. The simulation is done in the situation where the total buffer size of input and output queue is fixed at 32 so that different CLRs are observed in different allocation approaches. It seems that more buffers should be allocated to the output queue than to the input queue in order to achieve the lowest CLR. Again, we point out that the IS estimates are highly consistent with MC estimates.

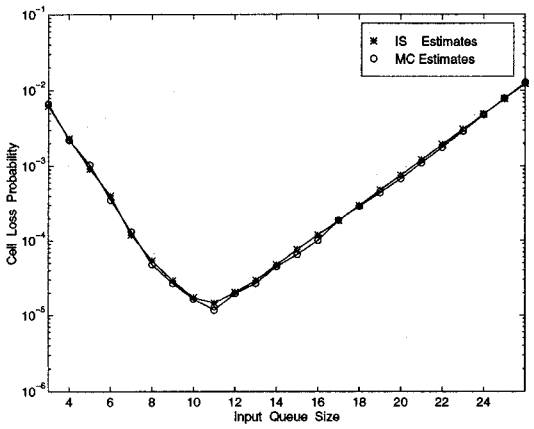

Figure 3: CLR of the I/O queued ATM Switch. $\lambda = 0.8$, $N = 16$, $m = 2$.

Table 2 contains the computation cost of IS using the *"split switch"* model compared with the MC method applied to the original switch model. In order to illustrate the potential of IS, the estimates are obtained in the case where the total buffer budget is fixed at 48, which makes the cell loss more rare compared with the results of Fig. 3. All the IS estimates are within the 10% accuracy. It is clear that the computation gains increase when the CLR decreases.

4 Conclusion

In this paper, we considered the application of IS to the efficient and accurate estimation of the cell loss rate of non-blocking ATM switches. In particular, we presented various IS methods which were developed using our knowledge of the ATM switch operation. We developed some new IS schemes for ATM switches with only output queues. Then, we extended these methods and used the notion of "split switch" model to estimate the CLR in the more complicated case of ATM switches with both input and output queues. The IS estimates obtained by the proposed methodologies were shown to be in excellent agreement with MC simulations. In addition, it has been demonstrated that a considerable computation cost can be saved using our proposed IS schemes. Finally, we plan to extend these results in the future to include more realistic traffic models, and to investigate the potential of using IS techniques as a realtime method for estimating other *Quality of service* parameters in ATM networks.

Table 2: The computation gains with IS ($\lambda = 0.8, N = 16, m = 2$ and $K + L = 48$)

(K,L)	$\hat{\gamma}$	Number of Slots(IS)	Number of Slots(MC)	Efficiency
$(15,33)$	8.14×10^{-8}	1.9×10^6	1.5×10^9	794
$(16,32)$	1.27×10^{-7}	1.8×10^6	9.8×10^8	553
$(17,31)$	1.98×10^{-7}	1.6×10^6	6.3×10^8	407

References

[1] R. Y. Awedeh and H. T. Mouftah, "Survey of ATM Switch Architectures," *Computer Networks and ISDN Systems 27*, pp. 1567-1613, 1995.

[2] E. W. Zegyra, "Architecture for ATM switching systems," *IEEE Commun. Mag.*, vol. 31, pp. 28-37, Feb. 1993.

[3] M. J. Karol, M. G. Hluchuj and S. P. Morgan, "Input and output queuing on a space-division packet switch," *IEEE Trans. on Commun.* , vol. 35, pp. 1347-1356, Dec. 1987.

[4] M. J. Lee and David S. Ahn, "Cell loss analysis and design trade-offs of nonblocking ATM switches with nonuniform traffic," *IEEE/ACM Trans. on Networking* , vol. 3, No.2, pp. 199-209, Apr. 1995.

[5] Q. L. Wang and V. S. Frost, "Efficient estimate of cell loss blocking probability for ATM systems," *IEEE/ACM Trans. on Networking* , vol. 1, No. 2, pp. 230-235, Apr. 1993.

[6] K. Ben. Letaief and K. Muhammad, "An efficient new technique for accurate bit error probability estimation of ZJ decoders," *IEEE Trans. Commun.*, vol. 43, June 1995.

[7] J. A. Freebersyser and J. K. Townsend, "Efficient simulation of CLR using standardized ATM connection traffic descriptors," *Proc. IEEE Int. Conf. Commun., ICC '95*, pp. 298-303, June 1995.

[8] K. Ben. Letaief, "Performance analysis of digital lightwave systems using efficient computer simulation techniques," *IEEE Trans. Commun.*, vol. COM-43, No. 2, pp. 240-251, Feb, 1995.

[9] G. S. Shedler, "Regenerative stochastic simulation," *Academic Press, INC.* , 1993.