A Performance Evaluation Procedure for a Class of Growable ATM Switches

†Zsehong Tsai, ‡Kangyei Yu, and †Feipei Lai
†Department of Electrical Engineering
National Taiwan University
Taipei, Taiwan, R.O.C.
‡Department of Product Development
Siemens Telecomm. System Limited
Taoyuan, Taiwan, R.O.C.

Abstract

In this paper, we propose a modular ATM switching network by modifying the original delta network. It is composed of several stages of switching modules, and all the switching modules are of the same type of basic building block. In order to improve the performance, the *channel group* connecting the switching modules in all stages are expanded to contain more than one link. The switching system is in fact a discrete-time queueing network, and each output of a particular switching module is modeled as a finite buffer, multi-server queue with multiple cell arrival streams. We obtain a general performance evaluation procedure for such a multi-stage, self-routing switching system.

1 Introduction

In recent years, the modularity problem in the ATM switch fabric design have become a major concern. In order to obtain the best delay/throughput performance, many researchers adopted the output queueing concept, but they also paid for such a design with the price of performance degradation or implementation difficulty. When this switching network expands larger and larger, i.e., the number of switch modules in the first stage grows to be very large, it is very much difficult to implement the path assignment function (especially in the high speed B-ISDN environment). In contrast, many manufacturers prefer the switching network to be composed of the same building blocks, such as shared memory or shared medium switches. By this method, they can construct the entire switching system based on one type of switch module.

For those modular architectures, the selection of appropriate system parameters such as buffer size and the number of intermediate switch module , etc, are essential to the switch performance such as cell loss rate and delay. Many general performance evaluation procedures such as [1, 2, 7] for ATM system can be applied to this task.

In this paper, we extend the delta network to a more general architecture and call it the *Extended Delta Network* (EDN). Such an architecture not only can grow in a modular fashion using only a single type of switch module, but also provides the self-routing capabilities to improve switching speed. The out-of-sequence problem can also be avoided. Next, we provide a general evaluation procedure to obtain the cell loss probability and delay performance experienced by a particular test stream in the switch. We find that this procedure is more useful to guarantee the Q.O.S. (quality of service) of a particular call. Furthermore, our approach can be used for the analysis of many other modular switches. This procedure is based on certain approximations but still provides an enough accuracy for the architecture design process.

2 The Switch Architecture

The EDN architecture is shown in Fig.1(a). In this figure, the $N \times N$ switching system can be split into two stages of switching modules ($M \times M$ and $L \times L$). The numbers M and L can be

0-8186-6507/94 $4.00

any power of 2 and $N = M \cdot L$. Both of these two types of switching modules can be constructed with smaller switching modules in the same way. (See also the 4-stage EDN example in Fig.1(b)) The number of internal links between the two stages can be expanded to more than one, in order to improve the performance. In fact, these switch modules may be any small scale, non-blocking switch. The only constraint is that all switching modules must be of the same type and the same scale, for the sake of simplicity. In order to operate easily in high speed, each switch module can self-route the incoming cells according to the corresponding fraction of their addresses (*local address*). If the dimension of the switch module in the first stage is sixteen, this switch module will route the cells according to the first four bits in their address fields. The EDN locates the buffers internally. In addition, the number of output ports belonging to the same local address in each switch module is expanded to be more than one to improve the performance, and these output ports are refered as a *channel group*. The production of channel group size and dimension of each switch module is required to be lower than the scale of the basic building block.

Since the out-of-sequence event is not allowed, we let the cells depart for the same module using the output ports of a particular channel group in a round-robin fashion. For instance, if the former cell leaves from the nth output port in its output channel group, the latter cell must be put on the $(n+1)$th output (or the first output when n is the group size). The dimensions of switch modules, sizes of channel groups and buffers of each stage can be different. So we can obtain an optimal combination under the constraint such as cell loss probability, total delay time, and admissible buffer space in each switch module. In the following sections, we modified the analysis in [1, 2] to analyse the EDN. With such an evaluation procedure, we can easily obtain the relations between performances and parameters of this switching network.

3 System State Analysis

In this section, we define the system states required for the general analysis procedure for a multi-stage, self-routing ATM switching network. Because the size of cell has been defined to be fixed, we have a discrete time queueing system. Although there have been many researches dedicated to the analysis of the ATM switch, they approximate the switching system to a continuous time queueing model. Other approaches such as [7] operate in accordance with a discrete time system, but restrict the usage in output queueing. For the proposed multi-stages switching system, we prefer to calculate in an iterative procedure, as in [1, 2].

3.1 *System State Definitions*

As in [2], we consider three kinds of traffic streams to join the switching network. The first one is the test stream with batch arrivals, where the distributions of the inter-arrival time between arriving batches and the batch size are both general. We call it as T-stream (for *test*-stream). The other traffic streams are i) aggregated arrivals of cells which form a Bernoulli process with batch arrivals (M-stream), and ii) several traffic streams, each of which is modeled as a discrete time interrupted Poisson process (IPP)(B-stream). In the tagged stage, we select N_M inputs to carry the M-stream, N_B inputs to carry the B-stream, and μ inputs to bear the T-stream in the switch module where the test stream flow through. The dimension of switch module m is equal to the sum of N_M, N_B and μ. Other switch modules in the same stage carry only the B-stream and M-stream, and these two kinds of traffic streams occupy the inputs with the ratio of $N_B : N_M$. In this paper, we assume the output address is uniformly distributed, and all the cells belonging to the T-stream are destined for the same output port of this switching network. The size of the particular output channel group where the T-stream destined for is denoted as r. The switch modules can be all shared memory switches or of other type, so the capacity of each output queue may be variable. In order to simplify the analysis, we call the buffer area for cells waiting for transmission the *waiting buffer*, and assume the waiting buffer size of each channel group is fixed, and denote it as Q. Here we exclude the buffer for those cells in transmission and thus the waiting

buffer size is r less than the buffer size of the corresponding queue. For the discrete time queue where the T-stream flows through, we define the following random variables:

$\Phi^{(n)}_{i_k,k}$: Queue length of the tagged output in the current switch module at the end of kth slot after the nth T-stream batch arrival where i_k IPPs are ON.

$\Psi^{(n)}_{i_{k-1},k}$: System length of the tagged output (queue length plus number of servers) in the particular switch module at the beginning of the kth slot after the nth T-stream batch arrival on the condition that the number of B-streams in the ON state is i_{k-1} in the previous slot.

M: Number of cells from M-stream destined for the same output as the T-stream in the currently analyzing switch module.

B_i: Number of cells from B-stream destined for the same output as the test stream in the current stage while i IPPs are ON.

X: the batch size of the T-stream arriving at the current stage.

A: the inter-arrival time between batches from the T-stream in the analyzing switch module.

The probability density of M, denoted as $m(j)$, is easy to compute:

$$m(j) \triangleq Pr[M = j] = \binom{N_M}{j} \cdot \left(\frac{\rho_M}{m}\right)^j \cdot \left(1 - \frac{\rho_M}{m}\right)^{N_M - j}, \tag{1}$$

where ρ_M is the average traffic load of M-stream. The probability distribution of B_i is[2]

$$b_i(j) \triangleq Pr[B_i = j] = \binom{i}{j} \cdot \lambda^j \cdot (1-\lambda)^{i-j};\ 0 \le j \le i, \tag{2}$$

The symbol λ is defined to be the arriving probability of B-stream to the destination channel group of T-stream when the IPP is ON.

Similarly, the density functions $a(k) \triangleq Pr[A = k]$ and $x(j) \triangleq Pr[X = j]$ define the probability function of A and X respectively. The cells in an arriving batch from T-stream can only come from one input channel group, so $\sum_{j=1}^{\mu} x(j) = 1$. In order to reduce the computation time, we assume the inter-arrival time between batches from T-stream is finite, i.e., $\sum_{k=1}^{A_{max}} a(k) = 1$. We record the aggregated state of the arrival process as the number of IPPs' in ON phase, and denote $P^{(n)}(i_k)$, the probability that i_k IPPs from B-stream are ON at the end of kth slot after the nth T-stream batch arrival. To derive the these probability densities later, we employ the following operators as in [1]:

$$\pi^m(y(j)) \triangleq \begin{cases} y(j) & j < m, \\ \sum_{i=m}^{\infty} y(j) & j = m, \\ 0 & j > m. \end{cases} \tag{3}$$

$$\pi_n(y(j)) \triangleq \begin{cases} 0 & j < n, \\ \sum_{i=-\infty}^{n} y(i) & j = n, \\ y(j) & j > n. \end{cases} \tag{4}$$

3.2 *Analysis of Random Variables*

Fig. 2 is the sample path for random variables defined above. At the beginning of the first slot following the nth T-stream batch arrival, there are $\Psi^{(n)}_{i_0,1}$ cells presented in the system. These

$\Psi^{(n)}_{i_0,1}$ cells are the sum of: i) the cells which have been located in the buffer immediately before the nth batch arrival from T-stream ($\Phi^{(n)}_{i_0,0}$); ii) the nth arrived batch from T-stream (X); iii) newly arrived M-stream cells (M); iv) newly arrived cells from B-stream (B_{i_0}). So

$$\Psi^{(n)}_{i_0,1} = \min(\Phi^{(n)}_{i_0,0} + M + B_{i_0} + X,\ Q + r);\ 0 \le i_0 \le N_B. \tag{5}$$

As in [2], we do not count the cell having been served in the first slot for computing $\Phi^{(n)}_{i_1,1}$, but do consider that the number of B-streams in the ON state is changed from i_0 to i_1. Prior to analyzing the queue length $\Phi^{(n)}_{i_1,1}$, we first compute the probability of the event that i_1 IPPs from B-stream are ON at the end of the first slot by

$$P^{(n)}(i_1) = \sum_{i_0=0}^{N_B} P^{(n)}(i_0) \cdot P_{i_0,i_1};\ 0 \le i_0,\ i_1 \le N_B. \tag{6}$$

Where P_{i_{k-1},i_k} is the probability that the number of B-streams in the ON state is changed from i_{k-1} to i_k at the end of the kth slot. We now can go on computing $\Phi^{(n)}_{i_1,1}$:

$$\Phi^{(n)}_{i_1,1} = \sum_{i_0=0}^{N_B} P_{i_0,i_1} \cdot \frac{P^{(n)}(i_0)}{P^{(n)}(i_1)} \cdot \max(\Psi^{(n)}_{i_0,1} - r,\ 0), \tag{7}$$

After the second slot, we have the following recurrence equations for $k = 2, 3, \ldots, A_{max}$ for the same reasons.

$$\Psi^{(n)}_{i_{k-1},k} = \min(\Phi^{(n)}_{i_{k-1},k-1} + M + B_{i_{k-1}},\ Q + r); \tag{8}$$

$$P^{(n)}(i_k) = \sum_{i_{k-1}=0}^{N_B} P^{(n)}(i_{k-1}) \cdot P_{i_{k-1},i_k}; \tag{9}$$

$$\Phi^{(n)}_{i_k,k} = \sum_{i_{k-1}=0}^{N_B} P_{i_{k-1},i_k} \cdot \frac{P^{(n)}(i_{k-1})}{P^{(n)}(i_k)} \cdot \max(\Psi^{(n)}_{i_{k-1},k} - r, 0); \tag{10}$$

3.3 *The Density Functions*

Next, we consider the density functions for the above random variables. Define $\phi^{(n)}_{i_k,k}(j)$ and $\psi^{(n)}_{i_{k-1},k}(j)$ to be the probability densities for $\Phi^{(n)}_{i_k,k}$ and $\Psi^{(n)}_{i_{k-1},k}$ respectively. That is, $\phi^{(n)}_{i_k,k}(j) \triangleq Pr[\Phi^{(n)}_{i_k,k} = j]$ ($k = 0, 1, \ldots, A_{max}$) and $\psi^{(n)}_{i_{k-1},k}(j) \triangleq Pr[\Psi^{(n)}_{i_{k-1},k} = j]$ ($k = 1, 2, \ldots, A_{max}$). Because the queue length of the tagged output; number of cells from M-stream, B-stream, and T-stream are all independent with each other (so do the system length of the tagged output), we employ the convolution operation on Eq. 5–Eq. 10 to obtain

$$\psi^{(n)}_{i_0,1}(j) = \pi^{Q+r}[\phi^{(n)}_{i_0,0}(j) * m(j) * b_{i_0}(j) * x(j)]; \tag{11}$$

$$\phi^{(n)}_{i_k,k}(j) = \sum_{i_{k-1}=0}^{N_B} P_{i_{k-1},i_k} \cdot \frac{P^{(n)}(i_{k-1})}{P^{(n)}(i_k)} \cdot \pi_r[\psi^{(n)}_{i_{k-1},k}(j + r)]; \tag{12}$$

$$\psi^{(n)}_{i_{k-1},k}(j) = \pi^{Q+r}[\phi^{(n)}_{i_{k-1},k-1}(j) * m(j) * b_{i_{k-1}}(j)]; \tag{13}$$

As in [2], when a new batch from the T-stream arrives in the kth slot following the last arrival, the new cells in this batch will find the queue length $\Phi^{(n)}_{i_k,k}$ upon their arrival. This event occurs

with the probability $a(k)$, so one can derive the queue length distribution immediately before the $(n+1)$th batch arrival from the T-stream by

$$\phi_{i_0,0}^{(n+1)}(j) = \sum_{k=1}^{A_{max}} a(k) \cdot \phi_{i_k(=i_0),k}^{(n)}(j); \tag{14}$$

For the same sake,

$$P^{(n+1)}(i_0) = \sum_{k=1}^{A_{max}} a(k) \cdot P^{(n)}(i_k = i_0), \tag{15}$$

Using Eq. 11–Eq. 15, we analyse the transient behavior of this system. The steady-state probabilities that there are j cells in the buffer of the tagged output immediately before a T-stream batch arrival, and the probability that the number of B-streams in the ON state at the end of the kth slot is i_k; are given by $\phi_{i_0,0}(j) = \lim_{n\to\infty} \phi_{i_0,0}^{(n)}(j)$, and $P(i_k) = \lim_{n\to\infty} P^{(n)}(i_k)$, respectively. Prior to an iterative calculation procedure, we assume the initial queue length to be zero by setting $\phi_{i_0,0}^{(1)}(0) = 1$ and $\phi_{i_0,0}^{(1)}(j) = 0$ for all $1 \le j \le Q$. The iterative calculation continues computing Eq. 11–Eq. 15 until the difference of the results between two iterations converges to be within an acceptable range.

4 Analysis of Module and Network Performance

First, we are concerned about the waiting time and loss probability of the test stream, T-stream within a single switching module. The analysis of the other streams can be derived by modifying the inter-arrival time distribution $a(k)$ to the M-stream and B-stream case.

We denote E_{i,i_0} as the event that the test cell from T-stream arrives in the ith position in the system (including the discarded cells) when the number of B-streams in ON state is i_0. Let $\hat{P}_{i,i_0}$ be the probability that event E_{i,i_0} occurs. Then

$$\hat{P}_{i,i_0} \triangleq Pr[E_{i,i_0}] = \sum_{l=0}^{\min(i-1,Q)} \phi_{i_0,0}(l) \cdot \varpi_{i_0}(i-l); \tag{16}$$

where $\varpi_{i_0}(j)$ is the probability that the test cell from the T-stream locates in the jth position among cells arriving in the same time slot when there are i_0 B-streams in the ON state. Then

$$\varpi_{i_0}(j) = \sum_{s=j}^{N_M+i_0+\mu} \frac{1}{s} \cdot [m(s) * b_{i_0}(s) * \chi(s)], \tag{17}$$

where $\chi(j) = j \cdot x(j)/E(X)$, is the probability that the test cell arrives in a T-stream batch with size equal to j.

To derive the waiting time, we introduce one more random variable. The random variable T_{i_0} represents the time interval that the test cell in the T-stream must wait for before its service under the condition that there are i_0 B-streams in the ON state. For the sake of analysis, we must choose the cell in position $1 \le i \le Q+r$ as a view point in waiting time computation. That means the analysis of waiting time must be conditional on this constraint and we don't care the $i > Q+r$ condition. We call this the *truncating effect*. Let $t_{i_0}(j)$ be the density function of T_{i_0} under this constraint, then

$$t_{i_0}(j) \triangleq Pr[T_{i_0} = j] = \frac{\sum_{i=j\cdot r+1}^{\min[(j+1)r,\, Q+r]} \hat{P}_{i,i_0}}{\sum_{i=1}^{Q+r} \hat{P}_{i,i_0}}; 0 \le j \le \left\lfloor \frac{Q-1}{r} \right\rfloor + 1, \tag{18}$$

where $\lfloor x \rfloor$ denotes the largest integer which is not more than x. It is easy to obtain the waiting time distribution $\overline{w}^{(i)}(j)$ of the particular stage i in steady-state by

$$\overline{w}^{(i)}(j) = \sum_{i_0=0}^{N_B} t_{i_0}(j) \cdot P(i_0). \tag{19}$$

To analyze the cell loss probability, we use the random variable $L_{i_0,j}$ to represent the loss probability of the T-stream; under the condition that i_0 IPPs of B-stream are ON and the queue length immediately before the test cell arrival is j. And the average cell loss probability of T-stream in the current stage i is

$$L^{(i)} = \sum_{i_0=0}^{N_B} \sum_{j=0}^{Q} L_{i_0,j} \cdot \phi_{i_0,0}(j) \cdot P(i_0). \tag{20}$$

Based upon the above analysis, we can then proceed to obtain the inter-arrival time and batch size distribution to the next stage, using a procedure similar to [2]. For the B-stream and M-stream, we assume they have the same traffic characteristics for every physical output address. That is, when the dimension of the switch module grows to a given size, the number of inputs carrying the M-stream (B-stream) also increases with the same ratio. After obtaining the mean cell loss probabilities of all stages by Eq. 20, the total cell loss probability L can be easily derived by

$$L \cong 1 - (1 - L^{(1)}) \cdot (1 - L^{(2)}) \cdot \cdots \cdot (1 - L^{(n)}), \tag{21}$$

where we assume that the number of stages in the entire switching system is n.

For the total delay time distribution $d(j)$ of the switching system, we assume the waiting time of each stage is independent with each other. So

$$d(j) \cong \overline{w}^{(1)}(j-n) * \overline{w}^{(2)}(j-n) * \cdots * \overline{w}^{(n)}(j-n). \tag{22}$$

5 Numerical Results

In this section, we present numerical results to verify our evaluation procedure and show that the performance of EDN approaches to output queueing switch under appropriate arrangements of its system parameters. We employ the following parameters: 1) The dimension of the whole switching system is 64. The number of stage is 2. 2) The sizes of switching modules in stage-1 and stage-2 are both 8. 3) The output channel group size of the first stage's switching module is 4. 4) The total waiting waiting buffer size is 640; that means the buffer size of one output channel group of each switching module is 5. 5) The traffic load of the test stream (T-stream) is fixed at 0.2. The number of inputs carrying M-stream is 63, and there is no input carrying the B-stream.

Fig.3(a) shows the cell loss probability of T-stream versus the traffic load of M-stream for EDN and output queueing switch. For this figure, the *output-queueing-1* and *output-queueing-2* is the loss probabilities for single module 64×64 output queueing switches with total waiting buffer size equal to 640 and 320 respectively. That is, in the output-queueing-1 and output-queueing-2 switch, the waiting buffer size for an output port is given by 10 and 5 respectively. For the particular output where the test stream is destined for, the carried traffic load is approximately equal to the sum of loads of the test stream and the M-stream. For example, if the M-stream load is 0.75, the traffic load carried by the particular output is $0.2 + 0.75 \times \frac{63}{64} = 0.94$.

For the output-queueing-1 switch, the total waiting buffer size is equal to the EDN's but are located at the output ports of a single stage; thus its loss probability is lower than the latter by two order of magnitude under light traffic. For the output-queueing-2, each output experiences the same buffer size as an individual output channel group of each switching module of EDN. Hence, its loss probability curve is close to the latter. When the traffic load grows to heavy, these three curves are almost the same. The largest difference between simulation results and analysis results is never beyond 10%, and the major difference of this evaluation procedure is caused by the

computation of inter-departure time, where we assume the arrival and departure processes to be independent. Fig.3(b) also illustrates the curves of mean delay time of T-stream versus the traffic load of M-streams. The simulation and analysis results coincide better in medium loads. This is because we assume the waiting time of different stages to be independent with each other, and their correlations become significant when the traffic load is heavy. For the sake of multiple stages of buffers, the EDN delays more than the output-queueing-1 and output-queueing-2 switches. It is interesting to note that when the traffic load reaches a very heavy level, the delay time of output-queueing-1 becomes the largest among these three designs since it has the lowest cell loss rate.

6 Conclusion

A performance evaluation procedure for a particular cell stream is then presented. The difference between analysis and simulation results is never beyond 10%. This analysis procedure may be applied to another multi-stage switching system such as those in [5, 6] by modifying the address distribution. It is also applicable to other multiple-stage output queueing switch. The only additional modification work is adjusting the dimensions of all switching modules and setting the buffer spaces of the first two stages to be zero. We also find that the analysis in [7] is just a special case of our analysis. The cell loss probability of EDN is found to be higher than that of the single-module output queueing switch by two order with the same total buffer size. However, this should not be consider as a drawback of EDN, and be treated as the cost associated with modularity. Considering the importance of modularity for large size ATM switches, the EDN should still be more promising than the single module output queueing switches.

References

[1] M. Murata, *et al.*, "Analysis of a Discrete-Time Single-Server Queue with Bursty Inputs for Traffic Control in ATM Networks," *IEEE J. Select. Areas Commun.*, vol. 8, no. 3, pp.447–458, April 1990.

[2] Y. Ohba, M. Murata, and H. Miyahara, "Analysis of Interdeparture Processes for Bursty Traffic in ATM Networks," *IEEE J. Select. Areas Commun.*, vol. 9, no. 3, pp. 468–476, April 1991.

[3] K. Y. Eng, M. J. Karol, and Y. Yeh, "A Growable Packet (ATM) Switch Architecture: Design Principles and Applications," *IEEE Trans. Commun.*, vol. 40, no. 2, pp. 423–430, Feb. 1992.

[4] S. C. Liew and K. W. Lu, "A 3-Stage Interconnection Structure for Very Large Packet Switches," *ICC'90*, pp. 771–777.

[5] Y. Sakurai, *et al.*, "Large-Scale ATM Multistage Switching Network with Shared Buffer Memory Switches," *IEEE Commun. Magz.*, pp. 90–96, Jan. 1991.

[6] T. C. Banwell, *et al.*, "Physical Design Issues for Very Large ATM Switching System," *IEEE J. Select. Areas Commun.*, vol. 9, pp. 1227–1238, Oct. 1991.

[7] Y. Oie, *et al.*, "Performance Analysis of Internally Unbuffered Large Scale ATM Switch with Bursty Traffic," *IEEE Infocom'93*, pp. 1270–1279.

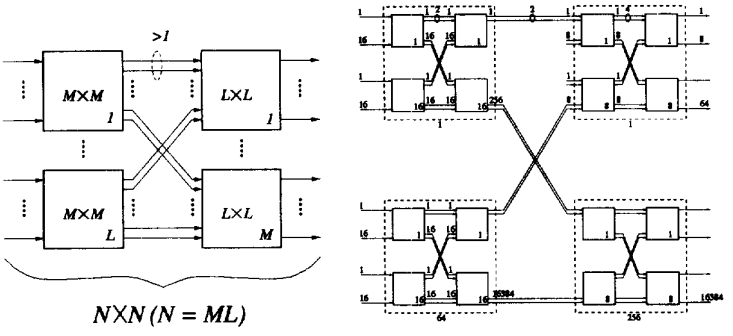

Figure 1: (a) General architecture of the Extended Delta Network (b) a 4-stage example.

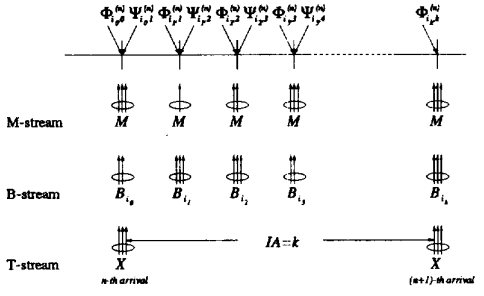

Figure 2: Sample path for random variables.

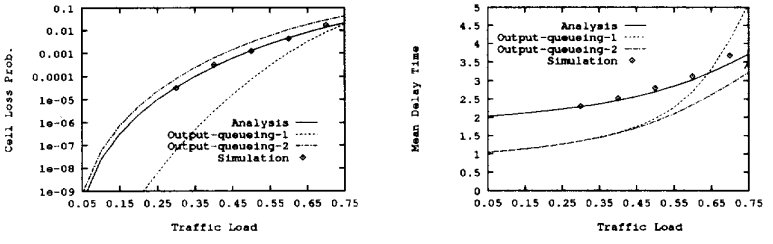

Figure 3: (a) The cell loss probability of the proposed switch versus the traffic load of the M-stream.(b) The delay of the proposed switch versus the traffic load of the M-stream.