Design and Analysis of a Hierarchical and Modular Local ATM Switch

†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 hierarchical and modular local ATM switch. In order to improve the queueing performance, we adopt the output queueing technique and allow several outputs to partially share the same output buffer space. The performance of the proposed switch is evaluated under uniform and non-uniform traffic patterns. Comparisons with Knockout Switch[2], Recursive Switch[3], SCOQ[4], and Christmas Tree Switch[5] have shown that in terms of complexity, crosstalk, scalability and synchronization, the proposed switch is superior to the others.

1 Introduction

Local Area Networks (LANs) have completed two generations of development. A third generation LAN should provide real-time transport capabilities, scalable throughput, inter-working between LAN and wide area network. Among numerous approaches to meet the above criterias, the ATM (Asynchronous Transfer Mode) seems to be the most promising. ATM has been selected by international standard bodies as the basis for future B-ISDN (Broadband Integrated Services Digital Network). Thus, if we construct future LANs on ATM, we can benefit from common standards and protocols, as well as widely available hardware and software components. At the same time, the local ATM (LATM) networks can still easily connect to the future broadband networks.

A large number of switching fabrics have been proposed to implement an ATM switch, and they are all developed for B-ISDN. For LATM switch, there are some differences in service requirements. For examples, a LATM switch provides only below 1000 ports, due to the fact that it is located in the LAN environment. Delays in the switches themselves will dominate link delays, so the switching delay is much more important. Because the interconnections among hosts and switches are all point-to-point links, LATM switch must own both multicast and broadcast capabilities[1]. Beside the above differences, the design concept of ATM switch (B-ISDN) is the same as LATM's.

The ATM switches can be classified into two categories: (1) Time division, (2) Space division. In a switch fabric based upon time division, all cells flow across a single communication highway shared in common by all input and output ports. This communication highway may be either a shared medium such as a ring or a bus[6, 7], or a shared memory[8, 9]. The throughput of this single shared highway defines the capacity of the entire switching fabric and thus fixes an upper bound on the capacity for a particular implementation beyond which it can not grow.

Whereas in time division, a single communication highway is shared by all input and output ports, in space division, a plurality of paths is provided between the input and output ports. These paths operate concurrently so that many cells can be transmitted across the switching fabric at the same time. The upper bound of the total capacity of the switching fabric is therefore theoretically unlimited. In practice, however, it is restricted by physical implementation constraints (e.g., device pinout, complexity, synchronization and crosstalk considerations), which together limit the size of the switching fabric.

A space division switch is composed of a number of switching elements. Interconnection networks for a space division switching fabric can be classified into two basic categories: single-path and multiple-path networks. A single-path network has a unique path through the interconnection network between any given input and output pair[2, 3, 4, 5]. A multiple-path network has a number of different paths available between any input and output pair[10, 11].

This paper proposes a high performance, self-routing, near non-blocking LATM switch (*HiMA*), which has a hierarchical and modular architecture. This switch is of the space division, single path type. In addition, it also provides the desired features of a LATM switch, such as the multicast and broadcast capabilities, the low switching delay. In order to obtain an excellent queueing performance, we adopt the output queueing concept and group several output queues together into a shared buffer to save the total buffer space. The detailed switch architecture is illustrated in Sec. 2. Because the output queueing yields the best possible delay/throughput performance[12], we only analyze the cell loss probability due to ***knockout principle***[2] under uniform and non-uniform traffic respectively in Sec. 3. The numerical results in Sec. 4 show that HiMA has a high degree of endurance under a non-uniform traffic pattern and heavy traffic load. We compare the proposed switch with the others and show the superiority of HiMA in terms of complexi-

0742-1303/93 $03.00 © 1993 IEEE

ty, crosstalk, scalability and synchronization in Sec. 5. Finally, we conclude in Sec. 6.

2 The switch architecture

2.1 Basic concept

The original Knockout Switch takes advantage of the fact that an ATM switch with an output buffer scheme provides the best delay/throughput performance, and that the probability of more than L (e.g., 14) cells destined for any particular output in each cell time interval is very low (e.g., 10^{-12}). Under these conditions, the number of cell filters and the hardware complexity of the concentrator is on the order of $O(N)$, where N is the number of input ports or output ports of the ATM switch. However, the number of interconnection wires in the network is on the order of $O(N^2)$. In order to reduce the complexity of the original Knockout Switch, Chao[3] lumped a number of output ports into a group so that the vertical routing links belonging to the same group can be shared by the cells that are destined for any outputs in this group. But being that, Chao constructed each SM (switching module) with crossbar switching elements, this still makes the entire switching fabric's cost high. Weijia Wang[5] proposed the interleaving of the filters and concentrators, such that one can obtain a much lower complexity under the same cell loss requirement. But in [5], as each level consists of only two SMs, this will waste too much cost in building a large scale ATM switch. In this paper, we also construct each SM in knockout principle and interleave filters and concentrators as in [5]. However, the number of SMs in each level is dynamically allocated so that we can make the best choice according to complexity or crosstalk. Fig. 2 is the internal architecture of HiMA. We can briefly say that the whole architecture of HiMA is just like a tree. Fig. 3 is an example for implementing a 1024×1024 HiMA.

As in [13], the common buffer space requirement shared by several output ports can be decreased under the uncorrelated and correlated traffic conditions. So in the HiMA's last level, a fixed number of outputs have been grouped together to share the common output buffer. Due to the rapid development in shared memory switches in recent years[8, 9], we can choose an $M \times M$ shared memory switch (such as 32×32) as the common output buffer.

2.2 Filters and concentrators

In HiMA, there are several levels of SMs, and each SM consists of several filters and one concentrator. For a particular level, if it has 2^m SMs, then the filters belonging to it check the corresponding m address bits and rotate the address field of each cell m bit position whenever this cell passes through the filter.

There are two ways to implement the concentrator. One, as [5] did, selects the Batcher sorter as the concentrator. We call a proposed switch of this type as *Batcher HiMA*. The Batcher sorter operates only due to the activity bit (Fig. 1) of each entering cell, and separates those cells with activity bit equal to zero (empty) from cells with activity bit equal to one (active), then the empty cells are dropped by means of no connection between them and the next level's input ports. But the Batcher sorter has a severe drawback: the difference of length between the longest and the shortest wires is on the order of half the number of input ports. Whenever the number of input ports of the Batcher sorter is increased, this drawback will make it difficult to synchronize the cells entering this Batcher sorter in the same time slot. However, the Batcher sorter needs less complexity than other concentrator designs, so it is superior to the other concentrator designs in the small scale LATM switch.

The other way to implement the concentrator construction is called the *Crossbar HiMA* concentrator as shown in Fig. 4. Each switching element in this type of concentrator switches according to the activity bit of both entering cells on the upper and left input lag. The switching function is shown in Fig. 4(b). In Fig. 4(a), both the left and upper input phases are in skewed form as in Fig. 5.

In Fig. 4, the N inputs of $N \times L$ concentrator are divided into two groups with their sizes equal to L and $N - L$ respectively. The L inputs are placed on the vertical direction and $N - L$ inputs are placed on the horizontal direction. The total number of switching elements of an $N \times L$ concentrator is $L \cdot (N - L)$, less than the proposed switch in [3] $(N \cdot L)$ and the difference is L^2. In [3], the switching elements in the concentrator also contain the filtering function, this will increase the total complexity in the concentrator design. If we separate the filtering from concentration function, we can significantly decrease its total complexity.

Fig. 5 is the phase diagram of the proposed $N \times L$ concentrator of Crossbar HiMA. The output cells from the $N \times L$ concentrator are injected into the next level's $L \times L'$ concentrators, where the cells from the left hand side are delayed $L - L'$ bit times before entering this concentrator to match the time phase of the upper side. Now, we select the furthest left input of vertical direction of $N \times L$ concentrator as a view point. The largest possible difference of delay through the entire concentrator is $L - 1 - (L - L') = L' - 1$ bit times. Take the n levels' HiMA as an example, the largest difference of delay through the entire switching fabric is $L^{(n)} - 1$ bit times, where the $L^{(n)}$ is the number of output ports of the nth level's concentrator. Compare this to [3], which has the largest difference of delay of $L^{(1)} - 1$ switching unit times and each switching unit delay at least two bit times (due to the activity and address bits). Crossbar HiMA has no severe out-of-sequence problem. For example, if the number of SM in the first level is four, the largest difference of delay of [3] is at least $3 \times (L^{(1)} - 1)$ bit times. Whenever $L^{(1)} > 142$, the cell stream with the largest difference of delay is out-of-sequence. But Crossbar HiMA's each switching element operates due to only the activity bit, together with the fact that the largest difference of delay is $L^{(n)} - 1$ switching element time, it is sufficient to avoid out-of-sequence problem if $L^{(n)} - 1 < 424$. Because we adopt the 32×32 shared memory switch as the output buffer, this constraint is never a problem in the designing of Crossbar HiMA.

In order to avoid continuous losses under the bursty traffic, the switching element in Fig. 4(a) switches to an arbitrary state when its two input cells are both active. Under this simple switching function explained above, Yeh[2] estimated the complexity is 16 gates.

Employing the crossbar architecture in the concentrator construction leads to higher complexity than the Batcher sorter . But the crossbar concentrator has a much less serious synchronization problem, due to the fact that each connection wire between switching elements in the same concentrator has the same distance.

2.3 Multicast and broadcast capabilities

The multicast and broadcast capabilities are especially important in LATM, due to the fact that the interconnections among hosts and switches are all point-to-point links[1]. For HiMA, we can easily do these jobs.

Fig. 1 is the format of internal address which is appended to the cell's header by the switch's control logic. In Fig. 1, a is the ***activity bit***, b is the ***broadcast bit***, m_0–m_s are the ***multicast address***, and d_0–d_{n-1} are the output port address. In this example, we let the number of levels and number of output ports of the proposed switch be s and 2^n respectively. For activity bit, it is zero whenever the cell is empty, and set for non-empty cell. So when the filters detect this bit and find it is zero, they need not do anything.

For broadcast bit, it is set when the cell is decided to be broadcasted to all output ports. The filters of all levels must pass cells with broadcast bit being set, and each shared memory switch must send these cells to all its output ports. Since the architectures of all levels are of the common bus type (all outgoing lines connect to the same bus), it is very easy to do broadcast in HiMA.

Next, we consider the multicast capability. For level-i, the filters will first check the corresponding bit m_{i-1} of the multicast address. The event that m_{i-1} is set makes the filters of level-i unconditionally pass the cell. In opposite, the filters of level-i will check the corresponding output port address only after they have detected that m_{i-1} is zero. We ranked the outputs into several multicast groups, each with a particular multicast address. For the lth output of HiMA, it belongs to several multicast groups $\Phi_{l,i}$: $\Phi_{l,i}$=\{jth output$|j=(l-1) \bmod S_i+1+k\cdot S_i, 0\leq k\leq M_i-1$\}, $1\leq i\leq s$, $S_i=(2^n)/(M_1\cdots M_i)$=the number of possible output address for the cells passing through the level-i SM; where A mod B is the remainder of A/B, and M_i is the number of sons of level-(i-1)'s S-M (Sec. 3). $\Phi_{l,s+1}$=\{jth output$|j=l-(l-1) \bmod 8+k, 0\leq k\leq 7$\}. If we want to send data to $\Phi_{l,i}$, we must set m_{i-1} and a bits to be one, b to be zero, and d_0–d_{n-1} bits to correct values. For example, in Fig. 3, output port 1–output port 8 can be allocated to eight users within the same multicast group ($\Phi_{l,4}$, l=1–8, s=3). Whenever we want to send data to this multicast group, we set m_3 and a to be one, d_0–d_6, b and m_0–m_2 to be zero; and don't care d_7–d_9. So the cells will pass through the first SM of level-1–level-3 (because b and d_0–d_6 are all zero), and are then injected into the shared memory switch attached to $\text{SM}_{1,3}$. For all output buffers, we modify the control logic circuits of the shared memory switches so that they will send cells to all output ports whenever m_3 is set, in this case. If output ports $l=1+8\cdot k$ $(0\leq k\leq 7)$ belong to another multicast group ($\Phi_{l,3}$), the cells destined for this multicast group must be with the following address format: a and m_2 are one; d_0–d_3, d_7–d_9, m_0, m_1, m_3 and b are zero. Under the above arrangements, we can construct the whole HiMA with both broadcast and multicast capabilities. However, if arbitrary assignments of output ports to a particular multicast group is desired, a copy network may be implemented in front of this switching fabric.

3 Cell loss analysis

3.1 Uniform traffic

In this section, we first analyze the cell loss probability of HiMA under uniform traffic. We make the following assumptions: (1) the traffic loads on all the inputs of HiMA are the same, and denoted as ρ, (2) each entering cell has equal probability to be destined for any output. Now we construct the entire switching fabric as a tree structure. Let M_i denote the number of sons of level-$(i-1)$ SM if $i>1$. While M_1 represents the number of switching modules of level-1. Therefore we define the following variables:

- N_i = the number of output ports of each concentrator in level-i,
- N = the total number of input ports,
- K = the total number of levels in the entire switch.

The following random variables are employed:

- L_i = the number of lost cells of individual SM in level-i,
- O_i = the number of cells leaving individual SM in level-i,
- I_i = the number of cells entering individual SM in level-i.

To simplify the computation, we estimate the total lost cells of the proposed switch by

$$\sum_{i=1}^{K}(\prod_{j=1}^{i} M_j)\cdot E\{L_i\},$$

where $E\{\bullet\}$ is the expected value of $\bullet$.

Under uniform traffic, the probability that there are α_0 cells entering the switch is

$$Pr\{O_0=\alpha_0\}=\binom{N}{\alpha_0}\cdot\rho^{\alpha_0}\cdot(1-\rho)^{N-\alpha_0}. \qquad (1)$$

Given the number of cell arrivals, we can get the probability of number of cells entering arbitrary SM in level-1 by binomial distribution:

$$Pr\{I_1=\beta_1 \mid O_0=\alpha_0\}=$$

$$\binom{\alpha_0}{\beta_1} \cdot \left(\frac{1}{M_1}\right)^{\beta_1} \cdot \left(1 - \frac{1}{M_1}\right)^{\alpha_0 - \beta_1}. \tag{2}$$

Due to the concentration, there are at most N_1 cells that can leave arbitrary concentrator of level-1, such that we get

$$E\{L_1 \mid O_0 = \alpha_0\} =$$

$$\begin{cases} \sum_{\beta_1 = N_1 + 1}^{\alpha_0} (\beta_1 - N_1) \cdot \\ \quad Pr\{I_1 = \beta_1 \mid O_0 = \alpha_0\} & N_1 < \alpha_0 \leq N, \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

In order to compute the mean number of lost cells of level-2, we must evaluate the probability of number of cells leaving the individual concentrator of level-1, which is

$$Pr\{O_1 = \alpha_1\} = \begin{cases} Pr\{I_1 = \alpha_1\} & 0 \leq \alpha_1 < N_1, \\ Pr\{I_1 \geq N_1\} & \alpha_1 = N_1. \end{cases} \tag{4}$$

With the same method, we can continue this process until all the mean numbers of lost cells of different level are obtained, and evaluate the mean loss probability by dividing the total number of lost cells with $N \cdot \rho$. To get a set of parameter (M_i, N_i) to meet the cell loss requirement, we set the cell loss probability requirement of each level to $1/K$ of the total cell loss requirement. Then we select the best parameter (M_i, N_i) separately for each level according to a particular objective (crosstalk or complexity).

3.2 Hot-spot traffic

As in [13], we define the Hot-spot traffic using a distribution matrix T_D, as shown in the follows.

$$T_D = \begin{bmatrix} h + \frac{1-h}{N} & \frac{1-h}{N} & \cdots & \frac{1-h}{N} \\ \vdots & \vdots & \ddots & \vdots \\ h + \frac{1-h}{N} & \frac{1-h}{N} & \cdots & \frac{1-h}{N} \end{bmatrix}. \tag{5}$$

The (i, j) entry of T_D, denoted as P_{ij}, gives the probability of a cell arriving at input-i and destined for output-j. In equation (5), h is the *concentration factor* such that h portion of input traffic is directed to the hot-spot destination output, while $(1 - h)$ fraction of the traffic are uniformly destined for all output ports. In this matrix, we select the output-1 as the hot-spot traffic destination output and suppose only one output carries this traffic in order to simplify the computation. This can be easily modified to any other conditions.

Three random variables are defined:

- X = the number of cells destined for output-1,
- Y = the number of cells destined for output-2–output-(N/M_1),
- Z_i = the number of cells destined for SM$_i$ of level-1.

Then,

$$Pr\{X = l \mid O_0 = \alpha_0\} =$$

$$\binom{\alpha_0}{l} \cdot P_{i,1}^{l} \cdot (1 - P_{i,1})^{\alpha_0 - l}. \tag{6}$$

$$Pr\{Y = m \mid X = l, O_0 = \alpha_0\} =$$

$$\binom{\alpha_0 - l}{m} \cdot \left(\frac{\frac{N}{M_1} - 1}{N - 1}\right)^{m} \cdot \left(1 - \frac{\frac{N}{M_1} - 1}{N - 1}\right)^{\alpha_0 - l - m}. \tag{7}$$

$$Pr\{Z_1 = k \mid O_0 = \alpha_0\} =$$

$$\sum_{l=0}^{k} Pr\{X = l, Y = k - l \mid O_0 = \alpha_0\}. \tag{8}$$

$$Pr\{Z_i = n \mid O_0 = \alpha_0, Z_1 = k\} =$$

$$\binom{\alpha_0 - k}{n} \cdot \left(\frac{1}{M_1 - 1}\right)^{n} \cdot \left(1 - \frac{1}{M_1 - 1}\right)^{\alpha_0 - k - n} ; i \neq 1. \tag{9}$$

Finally, we can get the probability of number of cells destined for SM$_1$ and SM$_i$ for i≠1, and estimate the mean number of lost cells of level-1 by $LOSS_1 + (M_1 - 1) \cdot LOSS_i$, where $LOSS_i$ is the mean number of lost cells of SM$_i$ of level-1, and is given by

$$LOSS_i = \sum_{k=N_1+1}^{N} Pr\{Z_i = k\} \cdot (k - N_1). \tag{10}$$

To simplify the analysis, we suppose there is no cell loss in the former levels when we analyze the mean number of losses of a particular level. We can then easily get the mean number of lost cells of level-i by changing M_1 to $\prod_{j=1}^{i} M_j$ and N_1 to N_i in equations (6)–(10). It is shown later that in this way we can obtain a worst case estimation.

3.3 Point-to-point traffic

We define the distribution matrix of the point-to-point traffic as follows:

$$T_D = \begin{bmatrix} q_{pp} & \frac{1-q_{pp}}{N-1} & \cdots & \frac{1-q_{pp}}{N-1} \\ \frac{1}{N} & \frac{1}{N} & \cdots & \frac{1}{N} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{1}{N} & \frac{1}{N} & \cdots & \frac{1}{N} \end{bmatrix}, \tag{11}$$

where $[T_D]_{i,j} = P_{i,j}$, $q_{pp} = \frac{1}{N} + \zeta$, and ζ is called the *non-uniform degree.* Using such a traffic matrix, we can easily get the probability of number of cells destined for SM$_1$ and SM$_i$ of level-1 for i≠1, and evaluate the mean number of lost cells of level-1. To simplify the computation, we also use the approximation method as in Sec. 3.2 and then get the worst case estimation of mean cell loss probability of the proposed switch.

4 Numerical results

In order to verify our analytical formulas, we consider the Crossbar HiMA with parameters: N=64, M_1=2, M_2=2, M_3=2, N_1=42, N_2=26, N_3=16. Fig. 6 illustrates the cell loss probability of the proposed switch under uniform traffic for various traffic loads. It indicates that the analysis technique presented in Sec. 3.1 yields very close numerical results to the simulation statistics. To show the effect of non-uniform traffic, we divide the traffic streams carried by the switching modules at the last level of the proposed switch into four groups: Group-0–Group-3, as shown in Table 1. The entries in Table 1 indicate whether the heaviest traffic stream shares switching modules at the indicated level with the considered group. The Group-0 is the most seriously affected and the Group-3 experiences the smallest impact. With no loss of generality, we select the output-1 as the hot-spot traffic destination output and the input-1 and output-1 as the point-to-point traffic source-destination pair. Fig. 7 shows the results under the hot-spot traffic. The 95% confidence interval is also provided. Fig. 8 presents the corresponding results in the point-to-point condition. These results convince us that the approximation method presented in Sec. 3.2 and Sec. 3.3 leads to a worst case estimation of the actual cell loss probability. Fig. 9–10 show the N=1024 case. These results verify again that HiMA can provide a very low cell loss probability even under the non-uniform traffic (if the buffer space is infinite).

Table 1: **Classification of traffic stream under non-uniform traffic.**

	level-1	level-2	level-3
Group-0	Y	Y	Y
Group-1	Y	Y	N
Group-2	Y	N	N
Group-3	N	N	N

5 Comparisons between HiMA and the other switches

In the practical implementation of an LATM switch, several factors can limit its development: total complexity in the whole switch, the number of crosstalks which constricts the switching speed, the delay through the entire switch which dominates the delay time of network, the scalability to expand the switch size to accommodate the future usage, and the synchronization problem. In fact, we can not make an accurate computation of the above parameters before practical chip design. To result in an approximate and fair comparison of the above limitative factors between the proposed and the other switches, we assume the following.

1. As in [2], we estimate the complexity of a switching element and a filter by 16 and 5 gates respectively, if each switching element's switching function depends only on the activity bit of input cell's header. When the switching function relies on more than one bit of the address field, there is no exact estimation of the complexity of each switching element. We assume that the actual complexity of this element is approximately equal to n times of 16 gates, where n is the number of bits for a switching element to work on.

2. In the estimation of crosstalk, it is impossible to obtain the real value in the chip layout level. We consider only the cross points of the wires connected between switching elements in the entire switching network .

3. To switch to the correct direction, the switching element must collect enough information about the switching function before doing its own task. Here we approximate the delay through a switching element working on n bits of information as n bit times long.

4. In the scalability comparison, we consider only those designs which can be easily expanded to an arbitrary size without changing the switching network's topology and disturbing existing connections to be truly scalable in Table 2. Also, we decide that only those architectures which have insignificant difference in the distance of connection wires between switching elements deserve a "Y" in the synchronization field of Table 2.

5. For the sake of simplicity, we only compare HiMA with some other switch proposals which have multicast and broadcast capabilities, due to the fact that these two capabilities are very important for LATM.

The comparison results are shown in Table 2. The common system parameters are: number of I/O ports = 1024, traffic load = 1.0, cell loss probability $\leq 10^{-11}$. In order to emphasize the superiority of HiMA, we choose the largest dimension for a LATM switch, the lowest cell loss probability requirement in a high speed LAN environment. In the Batcher HiMA row, we select the following parameters: M_1=4, M_2=4, M_3=8, N_1=512, N_2=128, N_3=32. In the Crossbar HiMA row, we set M_1=8, M_2=4, M_3=4, N_1=203, N_2=74, N_3=32. By the knockout principle, it is easy to prove that eight output ports per common output buffer (32 x 32) can retain cell loss probability lower than 10^{-12} with infinite buffer space under arbitrary traffic pattern. From Table 2 we can see that the Batcher HiMA is better than Crossbar HiMA in cost and delay time. But considering the high speed constraint and the synchronization problem, the Crossbar HiMA seems more promising.

We then compare the Recursive Switch[3] with the Crossbar HiMA. It is evident that the proposed switch is superior to the Recursive Switch in complexity and delay columns. As described in Sec. 2.1, the Recursive Switch has more constraints than our proposed architecture in deciding the network parameters to

Table 2: **Comparisons between HiMA and the other switches.**

	crosstalk	complex.
Batcher HiMA	2×10^7	7.8×10^6
Crossbar HiMA	5.2×10^6	2.9×10^7
Knockout Switch [2]	2.1×10^9	2.0×10^8
Christmas Tree[5]	3.27×10^7	1.8×10^7
Recursive Switch[3]	2.84×10^6	1.0×10^8
SCOQ[4]	1.57×10^7	5.16×10^6

	delay	scal.	synch.
Batcher HiMA	0.318	Y	N

Crossbar HiMA	min = 2.74
	max = 2.82
		Y	Y

Knockout Switch	2.43	Y	N
Christmas Tree	0.776	Y	N
Recursive Switch	max = 10.7	Y	Y
SCOQ	1.325	N	N

avoid the out-of-sequence problem. In these comparisons, we select the best parameters to decrease the cost of the Recursive Switch, but the optimized Recursive Switch may not be able to avoid out-of-sequence at this moment. When one compares the Knockout Switch and the Christmas Tree Switch with the Batcher HiMA, the latter switch exhibits its excellent characteristic in all columns again.

From Table 2, we can make the comparisons between the proposed switch and SCOQ. The results show that the proposed switch's performance is close to SCOQ. But SCOQ adopted the Batcher-banyan concept, so it is not easy to deal with the synchronization problem, especially in a high speed environment. At the same time, when SCOQ is going to expand to a larger size, the connection wires between the Batcher sorter and the banyan networks must be removed and relocated again to fit the expansion requirement. These two drawbacks can become the potential difficulties for the practical application of SCOQ.

6 Conclusion

A high performance, space-division, near non-blocking LATM switch has been proposed. It employs partially shared output buffer to save the total buffer space. Its hierarchical and modular architecture can also be easily expanded to an arbitrary size. Its *common bus* topology makes it easy to implement both the multicast and broadcast functions. The cell loss analysis has been introduced and the numerical results show that it can bear extreme non-uniform traffic. Its performance comparisons with others exhibit its excellence in the future LATM switch implementation.

References

[1] *Network Compatible ATM for Local Network Applications*, ver 1.01, Oct. 19, 1992.

[2] Y. S. Yeh, M. G. Hluchyj, and A. S. Acampora, "The Knockout Switch: A Simple, Modular Architecture for High-Performance Packet Switching," *IEEE J. Select. Areas Commun.*, vol. SAC-5, pp. 1274–1283, Oct. 1987.

[3] J. Chao, "A Recursive Modular Terabit/Second ATM Switch," *IEEE J. Select. Areas Commun.*, vol. 9, pp. 1161–1172, Oct. 1991.

[4] D. X. Chen and J. W. Mark, "SCOQ: A Fast Packet Switch with Shared Concentration and Output Queueing," in *IEEE INFOCOM'91*, pp. 145–154.

[5] W. Wang and F. A. Tobagi, "The Christmas-Tree Switch: An Output Queueing Space-Division Fast Packet Switch Based on Interleaving Distribution and Concentration Functions," in *IEEE INFOCOM'91*, pp. 163–170.

[6] P. Barri and J. A. O. Goubert, "Implementation of a 16 to 16 Switching Element for ATM Exchanges," *IEEE J. Select. Areas Commun.*, vol. 9, pp. 751–757, June 1991.

[7] A. Itoh, *et al.*, "Practical Implementation and Packaging Technologies for a Large-Scale ATM Switching System," *IEEE J. Select. Areas Commun.*, vol. 9, pp. 1280–1288, Oct. 1991.

[8] 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.

[9] T. Kozaki, *et al.*, "32 × 32 Shared Buffer Type ATM Switch VLSI's for B-ISDN's," *IEEE J. Select. Areas Commun.*, vol. 9, pp. 1239–1247, Oct. 1991.

[10] J. N. Giacopell, *et al.*, "Sunshine: A High-Performance Self-Routing Broadband Packet Switch Architecture," *IEEE J. Select. Areas Commun.*, vol. 9, pp. 1289–1298, Oct. 1991.

[11] F. A. Tobagi, T. Kwok, and F. M. Chiussi, "Architecture, Performance, and Implementation of the Tandem Banyan Fast Packet Switch," *IEEE J. Select. Areas Commun.*, vol. 9, pp. 1173–1192, Oct. 1991.

[12] M. G. Hluchyj and M. J. Karol, "Queueing in High-Performance Packet Switching," *IEEE J. Select. Areas Commun.*, vol. 6, pp. 1587–1597, Dec. 1988.

[13] D. X. Chen and J. W. Mark, "A Buffer Management Scheme for the SCOQ Switch Under Non-uniform Traffic Loading," in *IEEE INFOCOM'92*, pp. 132–140.

d_{n-1}	....	d_0	m_s	....	m_0	b	a

Figure 1: Format of address field in the cell header (a: activity bit, b: broadcast bit, m_i: multicast bit for level-$(i-1)$, d_i: output address bit).

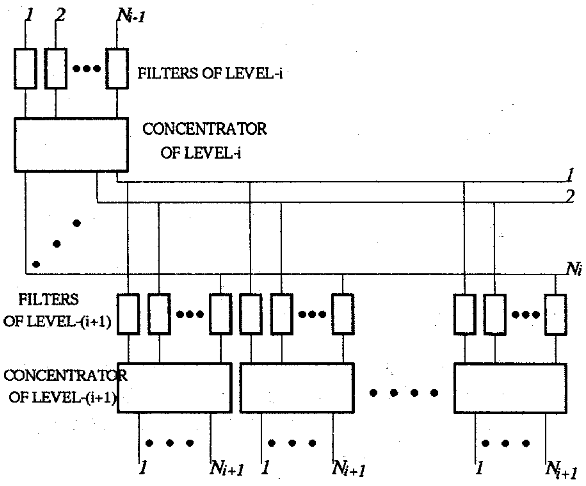

Figure 2: The internal architecture of HiMA.

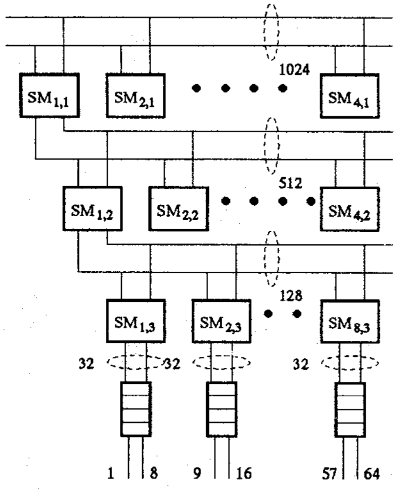

Figure 3: An example of HiMA architecture for N=1024.

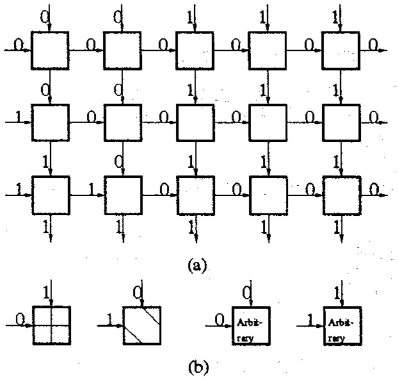

Figure 4: The 8×5 Crossbar HiMA concentrator architecture.

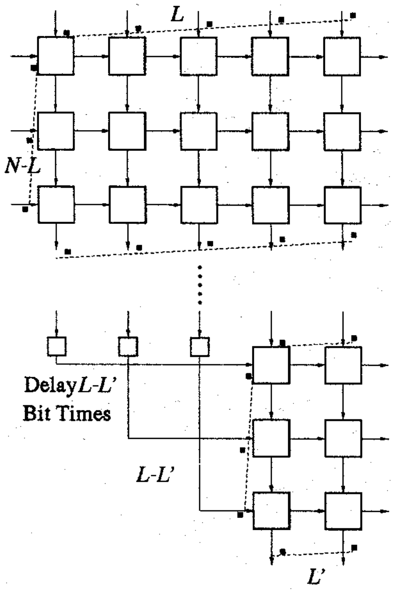

Figure 5: The phase diagram of $N \times L$ concentrator of Crossbar HiMA.

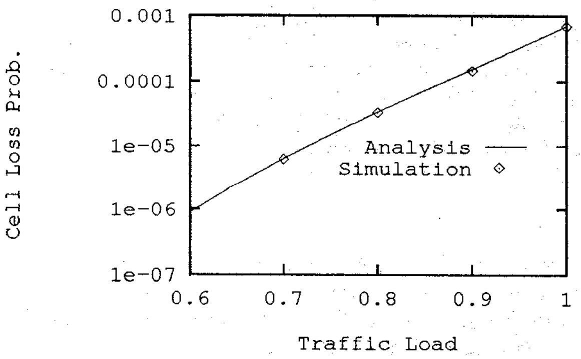

Figure 6: Cell loss probability of Crossbar HiMA versus mean traffic load ρ under uniform traffic; number of input/output ports N = 64.

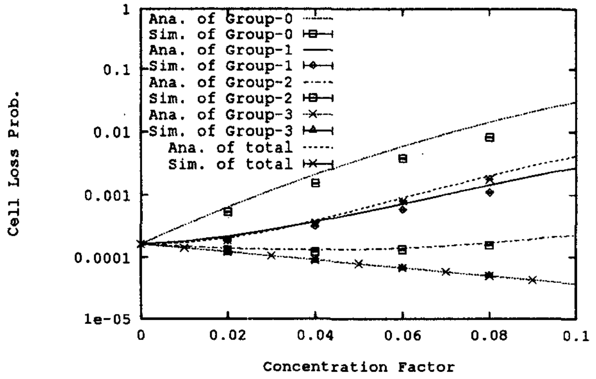

Figure 7: Cell loss probability of Crossbar HiMA versus concentration factor h under hot-spot traffic for various groups; number of input/output ports N = 64, mean traffic load ρ = 0.9.

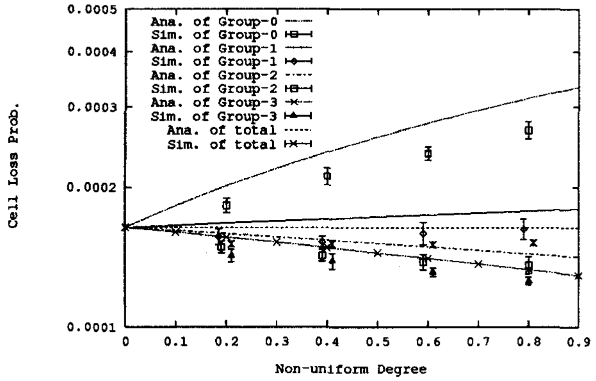

Figure 8: Cell loss probability of Crossbar HiMA versus non-uniform degree ζ under point-to-point traffic for various groups; number of input/output ports N = 64, mean traffic load ρ = 0.9.

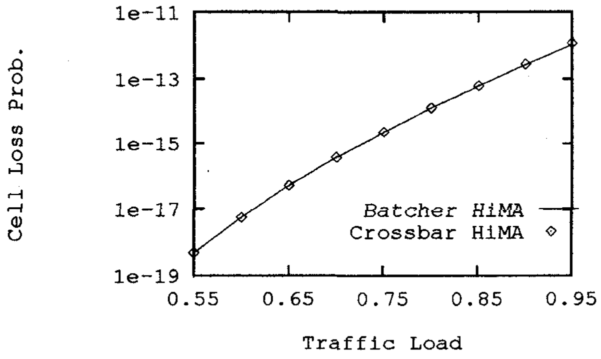

Figure 9: Mean cell loss probability of HiMA versus mean traffic load ρ under uniform traffic; number of input/output ports N = 1024.

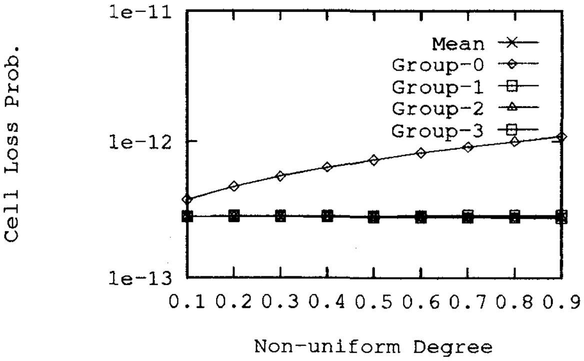

Figure 10: Mean cell loss probability of Crossbar HiMA versus non-uniform degree ζ under point-to-point traffic for various groups; number of input/output ports N = 1024, mean traffic load ρ = 0.9.