HiMA: a hierarchical and modular ATM switch with partially shared output buffer

Z. Tsai
K. Yu
F. Lai

Indexing terms: B-ISDN, Switching systems, ATM

Abstract: The authors propose a hierarchical and modular ATM switch. To improve the queueing performance, they 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 nonuniform traffic patterns. Comparisons with the knockout switch, recursive switch, SCOQ, and Christmas-tree switch show that, in terms of complexity, crosspoint count, scalability and synchronisation, the proposed switch is superior.

1 Introduction

Broadband ISDN (integrated services digital network) offers versatile communication services such as voice, data, and video, and is expected to be incorporated into the future social infrastructure. The asynchronous transfer mode (ATM) has been widely accepted as the target solution for B-ISDN. A large number of switching fabrics have been proposed to implement an ATM switch.

ATM switches can be classified into two categories: time division and 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 [1, 2], or a shared memory [3, 4]. 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 cannot 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 contraints (e.g. device pinout, complexity*, synchronisation and crosspoint 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 [5–8]. A multiple-path network has a number of different paths available between any input and output pair [9, 10].

This paper proposes a high-performance, selfrouting, near-nonblocking ATM switch (HiMA), which has a hierarchical and modular architecture. The switch is of the space-division, single-path type. It has a hierarchical and modular architecture so that we can easily expand it to a very large size. To obtain an excellent queueing performance, we adopt the output-queueing concept, and group several output queues together into a shared buffer to save on total buffer space. The detailed switch architecture is illustrated in Section 2. Because the output queueing yields the best possible delay/throughput performance [11], in Section 3 we analyse the cell loss probability due to the knockout principle [5] under uniform and nonuniform traffic conditions, respectively. The numberical results in Section 4 show that HiMA has a high degree of endurance under a nonuniform traffic pattern and heavy traffic load. In addition, we find that the key parameters of HiMA can be arbitrarily adjusted so that its complexity measured in gate count, or the number of crosspoints of the interconnection wires, can be optimised, subject to a cell loss-rate constraint. In Section 5 therefore we compare the proposed switch with others and show the superiority of HiMA in terms of complexity, crosspoints, scalability and synchronisation.

2 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. For detailed design principles and operation of the knockout switch, readers are referred to Reference 5. In the knockout switch, 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

Paper 9720I (E7), first received 12th October 1992 and in revised form 6th July 1993

Z. Tsai and F. Lai are with the Department of Electrical Engineering, National Taiwan University, Taipei, Taiwan, Republic of China

K. Yu is with the Department of Product Development, Siemens Telecommunications Systems Limited, Taoyuan, Taiwan, Republic of China

* Throughout this paper, hardware complexity is measured by the gate count.

number of cell filters and the gate count of the concentrator is of 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 of the order of $O(N^2)$. To reduce the complexity of the original knockout switch, Chao [6] 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. However, Chao constructed each SM (switching module) with crossbar switching elements, which still makes the cost of the entire switching fabric high. Wang [8] proposed the interleaving of the filters and concentrators, so that a much lower complexity can be achieved under the same cell-loss requirement. But, as each level consists of only two SMs, the cost in building a large scale ATM switch will be too high. In this paper, we also construct each SM on the knockout principle, and interleave filters and concentrators as in Reference 8. However, the number of SMs in each level is allowed to be arbitrarily selected so that we can make the best choice according to complexity or crosspoint count. Fig. 1 shows the internal

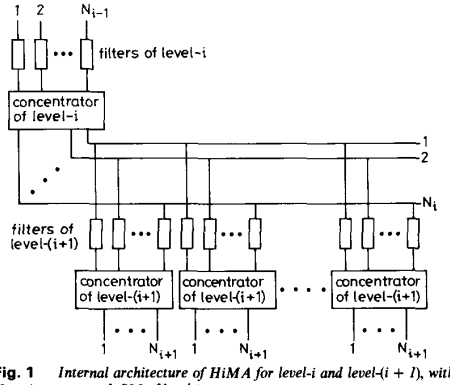

Fig. 1 *Internal architecture of HiMA for level-i and level-(i + 1), with N_{i-1} inputs to each SM of level-i*

architecture of HiMA. We can briefly say that the architecture of HiMA is like a tree. Fig. 2 is an example of the implementation of a 1024 × 1024 HiMA. The numbers

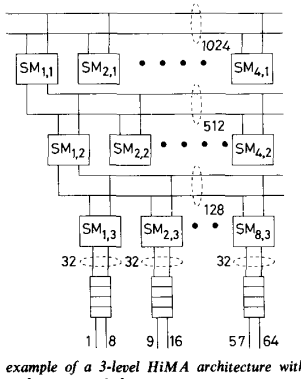

Fig. 2 *As example of a 3-level HiMA architecture with N = 1024, and 32 × 8 shared memory switches*

shown in this figure are derived using the evaluation procedure presented later.

As in Reference 12, the common buffer space requirement shared by several output ports can be decreased under the uncorrelated and correlated traffic conditions. Therefore in the last level of the HiMA, a fixed number of outputs are grouped together to share the common output buffer. Owing to the rapid development in shared memory switches in recent years [3, 4], 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, the filters belonging to it check the corresponding m address bits and rotate the address field of each cell m bit positions whenever this cell passes through the filter.

There are two ways of implementing the concentrator. One is to select the Batcher sorter as the concentrator [8]. We call a proposed switch of this type a Batcher HiMA. The Batcher sorter operates only on the activity bit of each entering cell, and separates those cells with activity bit equal to zero (empty) from cells with activity bit equal to one (active). The empty cells are then dropped because there are no connections between them and the next-level input ports. But the Batcher sorter has a severe drawback: the difference in length between the longest and the shortest wires is of 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 synchronise the cells entering the sorter in the same time slot. However, the Batcher sorter is less complex than other concentrator designs and is superior to other designs in a small-scale ATM switch.

The other way of implementing the concentrator construction is by means of the crossbar HiMA concentrator, as shown in Fig. 3. Each switching element in this

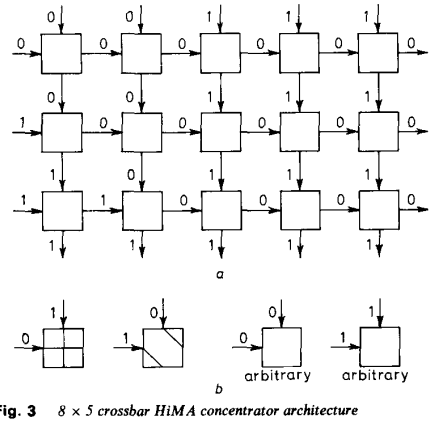

Fig. 3 *8 × 5 crossbar HiMA concentrator architecture*

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. 3*b*. In Fig. 3*a*, both the left and upper input phases are in skewed form, as in Fig. 4.

In Fig. 3, the N inputs of the $N \times L$ concentrator are divided into two groups equal to L and $N - L$, respec-

tively. The L inputs are placed in the vertical direction and the $N - L$ inputs are placed in the horizontal direction. The total number of switching elements of an $N \times L$

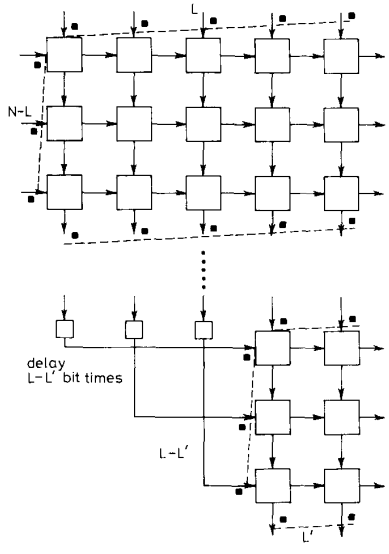

Fig. 4 *Phase diagram of an $N \times L$ concentrator which is connected to an $L \times L'$ concentrator of the crossbar HiMA architecture*

concentrator is $L(N - L)$, less than the proposed switch in Reference 6 (NL), and the difference is L^2. In Reference 6, the switching elements in the concentrator also contain the filtering function, and this will increase the total complexity of the concentrator design. If we separate the filtering from the concentration function, we can significantly decrease the total complexity.

Fig. 4 is the phase diagram of the proposed $N \times L$ concentrator of the 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. In other words, the input cells to the next level have to be arranged in a skewed form, as shown in Fig. 4, so that each of these cells contends for the output only with other cells entering the switch at the same time slot. Now, we select the furthest left input in the vertical direction of the $N \times L$ concentrator as a viewpoint. The largest possible difference in the delay through the entire concentrator is $L - 1 - (L - L') = L' - 1$ bit times. Taking an HiMA with n levels as an example, the largest difference in the delay through the entire switching fabric is $L^{(n)} - 1$ bit times, where $L^{(n)}$ is the number of output ports of the nth level's concentrator. One can compare this to Reference 6, which has the largest difference in delay of $L^{(1)} - 1$ switching unit times, and each switching unit delay of at least two bit times (due to the activity and address bits). A crossbar HiMA also has no severe out-of-sequence problem. For example, if the number of SMs in the first level is four, the largest difference in the delay is at least $3 \times (L^{(1)} - 1)$ bit times [6]. Whenever $L^{(1)} > 142$, the cell stream with the largest difference in delay is out-of-sequence. Bus as each switching element operates only as a result of the activity bit, together with the fact that the largest difference in delay is $L^{(n)} - 1$ switching-element times, our design is sufficient to avoid the out-of-sequence problem if $L^{(n)} - 1 < 424$. Because we adpot the 32×32 shared memory switch as the output buffer, this constraint is never a problem in the design of the crossbar HiMA.

To avoid continuous losses with 'bursty' traffic, the switching element in Fig. 3*a* switches to an arbitrary state when its two input cells are both active. With this simple switching function, Yeh [5] estimated the complexity as 16 gates.

Employing the crossbar architecture in the concentrator construction leads to a higher complexity than that of the Batcher sorter. But the crossbar concentrator has a much less serious synchronisation problem, because each connection wire between switching elements in the same concentrator is the same length.

3 Cell loss analysis

In this Section, the cell-loss probability of HiMA SMs is analysed. The cell-loss probability due to output buffer contention is not included. Readers are referred to Reference 12 for related analysis.

3.1 Uniform traffic

We first analyse the cell loss probability due to concentration 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, and (2) each entering cell has an equal probability of being 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 $i > 1$. M_1 represents the number of switching modules of level-1. Therefore, we define the following variables

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

The following random variables are employed

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

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

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

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

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

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

Given the number of cell arrivals, we can obtain the probability of the number of cells entering an arbitrary SM in level-1 from the binomial distribution

$$\Pr\{I_1 = \beta_1 \mid O_0 = \alpha_0\} = \binom{\alpha_0}{\beta_1} \left(\frac{1}{M_1}\right)^{\beta_1} \left(1 - \frac{1}{M_1}\right)^{\alpha_0 - \beta_1} \tag{2}$$

Owing to the concentration, there are at most N_1 cells that can leave the arbitrary concentrator of level-1, so

that

$$E\{L_1 \mid O_0 = \alpha_0\} = \begin{cases} \sum_{\beta_1 = N_1+1}^{\alpha_0} (\beta_1 - N_1) \\ \Pr\{I_1 = \beta_1 \mid O_0 = \alpha_0\} & N_1 < \alpha_0 \leqslant N \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

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

$$\Pr\{O_1 = \alpha_1\} = \begin{cases} \Pr\{I_1 = \alpha_1\} & 0 \leqslant \alpha_1 < N_1 \\ \Pr\{I_1 \geqslant N_1\} & \alpha_1 = N_1 \end{cases} \quad (4)$$

With the same method, we can continue this process until all the mean numbers of lost cells of different levels are obtained, and evaluate the mean loss probability by dividing the total number of lost cells by $N\rho$. To obtain a set of parameters (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 (crosspoint or complexity).

3.2 Hot-spot traffic

As in Reference 12, we define the hot-spot traffic using a distribution matrix T_D as follows

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

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 eqn. 5, h is the concentration factor such that a fraction h of the input traffic is directed to the hot-spot destination output, while $(1 - h)$ of the traffic is uniformly destined for all output ports. In this matrix, we select the output-1 as the hot-spot traffic destination output and to simplify the computation we suppose that only one output carries this traffic. This can be easily modified for other conditions.

The probability of the number of cells destined for SM_1 and SM_i for $i \neq 1$ can be obtained as in the Appendix. We estimate the mean number of lost cells of level-1 by $LOSS_1 + (M_1 - 1)LOSS_i$, where $LOSS_i$ is the mean number of lost cells of SM_i of level-1, and is given by

$$LOSS_i = \sum_{k=N_1+1}^{N} \Pr\{Z_i = k\}(k - N_1) \quad (6)$$

where the random variable Z_i is defined to be the number of cells destined for SM_i of level-1 (see Appendix).

To simplify the analysis, we suppose that there is no cell loss in the former levels when we analyse the mean number of losses of a particular level. We can then easily obtain 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 eqns. 6 and 8–11. 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

$$T_D = \begin{bmatrix} q_{pp} & \frac{1-q_{pp}}{N-1} & \cdots & \frac{1-q_{pp}}{N-1} \\ \frac{1}{N} & \frac{1}{N} & \cdots & \frac{1}{N} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{1}{N} & \frac{1}{N} & \cdots & \frac{1}{N} \end{bmatrix} \quad (7)$$

where $[T_D]_{i,j} = P_{i,j}$, $q_{pp} = 1/N + \zeta$, and ζ is the nonuniform degree. Using such a traffic matrix, we can easily obtain the probability of the number of cells destined for SM_1 and SM_i of level-1 for $i \neq 1$ from the Appendix, and evaluate the mean number of lost cells of level-1. To simplify the computation, we also use the approximation method as in Section 3.2 and then obtain the worst-case estimation of the mean cell-loss probability of the proposed switch.

4 Numerical results

To verify our analytical formulae, we consider the 3-level crossbar HiMA with parameters: $N = 64$, $M_1 = 2$, $M_2 = 2$, $M_3 = 2$, $N_1 = 42$, $N_2 = 26$, and $N_3 = 16$. Fig. 5

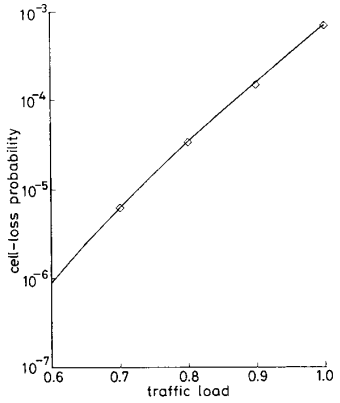

Fig. 5 *Cell-loss probability of crossbar HiMA against mean traffic load ρ under uniform traffic; number of input/output ports N = 64*

— analysis
◇ simulation

illustrates the cell-loss probability of the proposed switch under uniform traffic for various traffic loads. Here the simulation results are obtained via a simulation model consisting of the SMs, such that only the cell losses due to concentration are included in the statistics. No cell losses in the output buffers are taken into account. In addition, the cell-loss traffic patterns are generated exactly as described in Section 3.1. Fig. 5 indicates that the analysis technique presented in Section 3.1 yields numerical results very close to the simulation statistics. To show the effect of nonuniform traffic, we divide the traffic streams carried by the switching modules at the last level of the proposed switch into four groups: Group-0 to 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. Group-0 is the most seriously affected, and Group-3 experiences the smallest impact. With no loss of generality, we select the output-1 as the hot-spot

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*

traffic destination output and the input-1 and output-1 as the point-to-point traffic source–destination pair. Fig. 6 shows the results under hot-spot traffic. The 95% confidence interval is also provided. Fig. 7 presents the corresponding results in the point-to-point condition. These results convince us that the approximation method presented in Sections 3.2 and 3.3 leads to a worst-case estimation of the actual cell loss-probability. Figs. 8 and 9 show the $N = 1024$ case. These results verify again that HiMA can provide a very low cell-loss probability even under nonuniform traffic (if the buffer space is infinite).

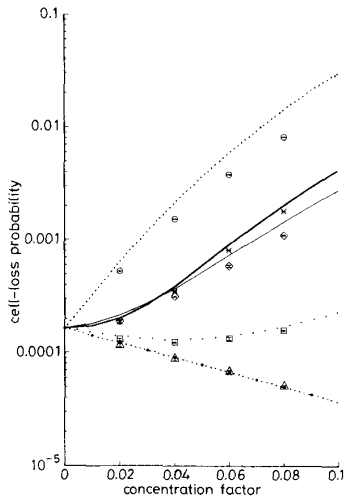

Fig. 6 *Cell-loss probability of crossbar HiMA against concentration factor h under hot-spot traffic for various groups; number of input/output ports N = 64, mean traffic load ρ = 0.9*

- analysis of group-0
- simulation of group-0
- analysis of group-1
- simulation of group-1
- analysis of group-2
- simulation of group-2
- analysis of group-3
- simulation of group-3
- analysis of total
- simulation of total

5 Comparisons between HiMA and other switches

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

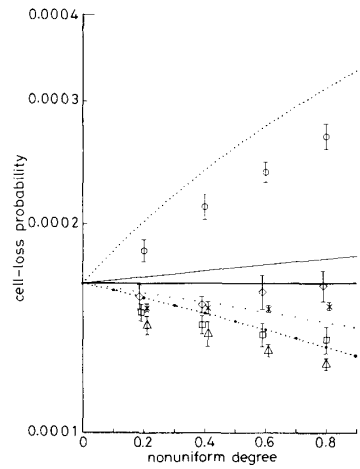

Fig. 7 *Cell-loss probability of crossbar HiMA against nonuniform degree ζ under point-to-point traffic for various groups; number of input/output ports N = 64, mean traffic load ρ = 0.9*

- analysis of group-0
- simulation of group-0
- analysis of group-1
- simulation of group-1
- analysis of group-2
- simulation of group-2
- analysis of group-3
- simulation of group-3
- analysis of total
- simulation of total

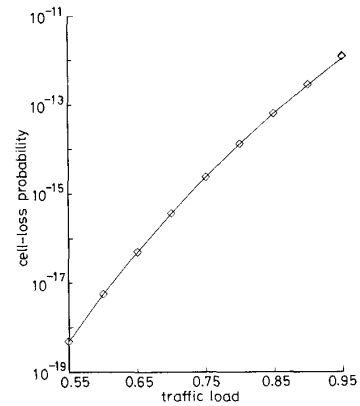

Fig. 8 *Mean cell-loss probability of HiMA against mean traffic load ρ under uniform traffic; number of input/output ports N = 1024*

- Batcher HiMA
- crossbar HiMA

(i) In this paper, we measure the complexity of a switch in terms of its gate count. As in Reference 5, we estimate the complexity of a switching element and a filter by 16 and five gates, respectively, if each switching element's switching function depends only on the activity bit of the

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.

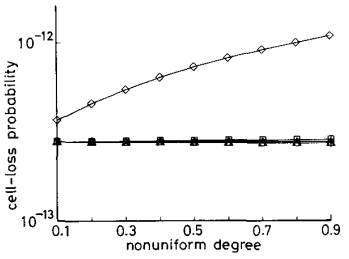

Fig. 9 *Mean cell-loss probability of crossbar HiMA against nonuniform degree ζ under point-to-point traffic for various groups; number of input/output ports N = 1024, mean traffic load ρ = 0.9*

× mean
◇ group-0
□ group-1
△ group-2
□ group-3

We assume that the actual complexity of this element is approximately equal to n times 16 gates, where n is the number of bits for a switching element to work on.

(ii) In the estimation of crosspoint count, it is impossible to obtain the real value in the chip layout level. We consider only the crosspoints of the wires connected between switching elements in the entire switching network.

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

(iv) 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 (Table 2). Also, we decide that only those architectures which have insignificant differences in the lengths of connection wires between switching elements deserve a 'Y' in the synchronisation field of Table 2.

The comparison results are shown in Table 2. During the computation of the crosspoint count and complexity of other switches, we first select appropriate parameters for them such that their cell-loss constraint is satisfied, using the procedures described in the original work whenever possible (such as References 5 and 7), or using the same procedure as for the HiMA switch. The common system parameters are: number of I/O ports = 1024, traffic load = 1.0. We have fixed the cell-loss probability of all switches in this comparison to be within the range of $(10^{-12}$–$10^{-11})$ for fairness (under uniform traffic). To emphasize the superiority of HiMA, we choose a large enough dimension for an ATM switch and the lowest cell-loss probability requirement in a high-speed B-ISDN environment. In the Batcher HiMA row, we select the following parameters: $M_1 = 4$, $M_2 = 4$, $M_3 = 8$, $N_1 = 512$, $N_2 = 128$, and $N_3 = 32$. In the crossbar HiMA row, we set $M_1 = 8$, $M_2 = 4$, $M_3 = 4$, $N_1 = 203$, $N_2 = 74$, $N_3 = 32$. These parameters are obtained via an exhaustive research procedure. By the knockout principle, it is easy to prove that eight output ports per common output buffer (32 × 32) can retain cell-loss probability lower than 10^{-12} with infinite buffer space under an arbitrary traffic pattern.

From Table 2, we can see that the Batcher HiMA is better than that crossbar HiMA in complexity and delay time. But considering the high speed constraint and the synchronisation problem, the crossbar HiMA seems more promising.

We then compared the recursive switch [6] with the crossbar HiMA. It is evident that the proposed switch is superior to the recursive switch in complexity and delay time. As described in Section 2.1, the recursive switch has more constraints than our proposed architecture in deciding the network parameters to avoid the out-of-sequence problem. In these comparisons, we selected the best parameters to decrease the cost of the recursive switch, but the optimised recursive switch may not be able to avoid the out-of-sequence problem at this time.

When one compares the knockout switch and the Christmas-tree switch with the Batcher HiMA, the latter exhibits its excellent characteristics in all columns again.

Table 2 shows also that the performance of the proposed switch is close to that of the SCOQ. But the SCOQ is based on the Batcher-banyan concept, so it is not easy to deal with the synchronisation problem, especially in a high-speed environment. At the same time, when the SCOQ is to be expanded 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 limit the practical application of the SCOQ.

6 Conclusions

A high-performance, space-division, near-nonblocking ATM switch has been proposed. It employs partially shared output buffer to save on total buffer space. Its hierarchical and modular architecture can also be easily expanded to an arbitary size. Cell-loss analysis has been introduced and the numerical results shown that the switch can bear extremely nonuniform traffic. Its performance compared with other switches demonstrates its excellence for future ATM switch implementation.

7 References

1 BARRI, P., and GOUBERT, J.A.O.: 'Implementation of a 16 to 16 switching element for ATM exchanges', *IEEE J. Sel. Areas Commun.*, 1991, **9**, pp. 751–757

Table 2: Comparisons between HiMA and other switches

	Crosspoints	Complexity (gate)	Delay (cell time)	Scalability	Synchronisation
Batcher HiMA	2×10^7	7.8×10^8	0.318	*Y*	*N*
Crossbar HiMA	5.2×10^6	2.9×10^7	min = 2.74 max = 2.82	*Y*	*Y*
Knockout switch [5]	2.1×10^9	2.0×10^8	2.43	*Y*	*N*
Christmas tree [8]	3.27×10^7	1.8×10^7	0.776	*Y*	*N*
Recursive switch [6]	2.84×10^6	1.0×10^8	max = 10.7	*Y*	*N*
SCOQ [7]	1.57×10^7	5.16×10^8	1.325	*N*	*N*

2 ITOH, A., TAKAHASHI, W., NAGANO, H., KURISAKA, M., and IWASAKI, S.: 'Practical implementation and packaging technologies for a large-scale ATM switching system', *IEEE J. Sel. Areas Commun.*, 1991, **9**, pp. 1280–1288
3 BANWELL, T.C., ESTES, R.C., HABIBY, S.F., HAYWARD, G.A., and HELSTERN, T.K.: 'Physical design issues for very large ATM switching system', *IEEE J. Sel. Areas Commun.*, 1991, **9**, pp. 1227–1238
4 KOZAKI, T., ENDO, N., SAKURAI, Y., MATSUBARA, O., MIZUKAMI, M., and ASANO, K.: '32 × 32 shared buffer type ATM switch VLSIs for B-ISDNs', *IEEE J. Sel. Areas Commun.*, 1991, **9**, pp. 1239–1247
5 YEH, Y.S., HLUCHYJ, M.G., and ACAMPORA, A.S.: 'The knockout switch: a simple, modular architecture for high-performance packet switching', *IEEE J. Sel. Areas Commun.*, 1987, **SAC-5**, pp. 1274–1283
6 CHAO, J.: 'A recursive modular terabit/second ATM switch', *IEEE J. Sel. Areas Commun.*, 1991, **9**, pp. 1161–1172
7 CHEN, D.X., and MARK, J.W.: 'SCOQ: a fast packet switch with shared concentration and output queueing', *IEEE/ACM Trans. Netw.*, 1993, **1**, (1), pp. 142–151
8 WANG, W., and TOBAGI, F.A.: 'The Christmas-tree switch: an output queueing space-division fast packet switch based on interleaving distribution and concentration functions'. IEEE INFOCOM'91, pp. 163–170
9 GIACOPELL, J.N., HICKEY, J.J., MARCUS, W.S., SINCOSKIE, W.D., and LITTLEWOOD, M.: 'Sunshine: a high-performance self-routing broadband packet switch architecture', *IEEE J. Sel. Areas Commun.*, 1991, **9**, pp. 1289–1298
10 TOBAGI, F.A., KWOK, T., and CHIUSSI, F.M.: 'Architecture, performance, and implementation of the tandem banyan fast packet switch', *IEEE J. Sel. Areas Commun.*, 1991, **9**, pp. 1173–1192
11 HLUCHYJ, M.G., and KAROL, M.J.: 'Queueing in high-performance packet switching', *IEEE J. Sel. Areas Commun.*, 1988, **6**, pp. 1587–1597
12 CHEN, D.X., and MARK, J.W.: 'A buffer management scheme for the SCOQ switch under non-uniform traffic loading'. IEEE INFOCOM'92, pp. 132–140

8 Appendix

8.1 Probability of number of cells destined for tagged SM for hot-spot traffic

For hot-spot traffic, three random variables are defined

X = number of cells destined for output-1
Y = number of cells destined for output-2–output-(N/M_1)
Z_i = the number of cells destined for SM_i of level-1

Given that α_0 cells arrive at the input ports of HiMA, the probability of l cells destined for output-1 is

$$\Pr\{X = l \mid O_0 = \alpha_0\} = \binom{\alpha_0}{l} P_{i,1}^{l}(1 - P_{i,1})^{\alpha_0 - l} \tag{8}$$

where the definitions of O_0, M_i and N_i are the same as in Section 3.1. As above, given α_0 cell-arrivals at the input ports of HiMA and l cells destined for output-1, the probability of m cells destined for output-2–output-(N/M_1) is

$$\Pr\{Y = m \mid X = l, O_0 = \alpha_0\} = \binom{\alpha_0 - l}{m}\left(\frac{\frac{N}{M_1} - 1}{N - 1}\right)^{m}\left(1 - \frac{\frac{N}{M_1} - 1}{N - 1}\right)^{\alpha_0 - l - m} \tag{9}$$

It is then easy to obtain the probability $\Pr\{X, Y \mid O_0\}$.

The probability of k cells destined for SM_1 of level-1 given that α_0 cells arrive at the input ports is

$$\Pr\{Z_1 = k \mid O_0 = \alpha_0\} = \sum_{l=0}^{k} \Pr\{X = l, Y = k - l \mid O_0 = \alpha_0\} \tag{10}$$

The probability of n cells destined for SM_i ($i \neq 1$) given that α_0 cells arrive at the inputs and k cells destined for SM_1 is

$$\Pr\{Z_i = n \mid O_0 = \alpha_0, Z_1 = k\} = \binom{\alpha_0 - k}{n}\left(\frac{1}{M_1 - 1}\right)^{n}\left(1 - \frac{1}{M_1 - 1}\right)^{\alpha_0 - k - n} \quad i \neq 1 \tag{11}$$

Then we can obtain $\Pr\{Z_i \mid O_0\}$.

8.2 Probability of number of cells destined for tagged SM for point-to-point traffic

To separate the effect of cells belonging to point-to-point traffic, we define a Bernoulli random variable C_{ij} to indicate that a cell arrives at input-i and destines to output-j, and C_1 to indicate that there is a cell arrival at input-1. It is straightforward to derive

$$\Pr\{C_{i,j} = 1\} = \frac{\rho}{N} \quad i \neq 1 \tag{12}$$

$$\Pr\{C_{1,j} = 1\} = \begin{cases} \rho q_{pp} & j = 1 \\ \dfrac{\rho(1 - q_{pp})}{N - 1} & j \neq 1 \end{cases} \tag{13}$$

$$\Pr\{C_1 = 1\} = \rho \tag{14}$$

The probability of a cell arrival at input-1 given that α_0 cells arrfive at the inputs is

$$\Pr\{C_1 = 1 \mid O_0 = \alpha_0\} = \frac{\alpha_0}{N} \tag{15}$$

Given that input-1 has a cell arrival

$$\Pr\{C_{1,j} = 1 \mid C_1 = 1, O_0 = \alpha_0\} = \begin{cases} \dfrac{1 - q_{pp}}{N - 1} & \alpha_0 > 0 \quad j \neq 1 \\ q_{pp} & \alpha_0 > 0 \quad j = 1 \\ 0 & \text{otherwise} \end{cases} \tag{16}$$

We next compute the conditional probability distribution for Z_i

$$\Pr\{Z_1 = k \mid O_0 = \alpha_0, C_{1,j} = 1)\} = \begin{cases} \binom{\alpha_0 - 1}{k - 1}\left(\frac{1}{M_1}\right)^{k-1}\left(1 - \frac{1}{M_1}\right)^{\alpha_0 - k} & 1 \leqslant j \leqslant \frac{N}{M_1} \\ \binom{\alpha_0 - 1}{k}\left(\frac{1}{M_1}\right)^{k}\left(1 - \frac{1}{M_1}\right)^{\alpha_0 - k - 1} & \frac{N}{M_1} < j \leqslant N \end{cases} \tag{17}$$

$$\Pr\{Z_i = k \mid O_0 = \alpha_0, C_1 = 0\} = \binom{\alpha_0}{k}\left(\frac{1}{M_1}\right)^{k}\left(1 - \frac{1}{M_1}\right)^{\alpha_0 - k} \tag{18}$$

$$\Pr\{Z_i = k \mid O_0 = \alpha_0, C_{1,j} = 1\} = \begin{cases} \binom{\alpha_0 - 1}{k - 1}\left(\frac{1}{M_1}\right)^{k-1}\left(1 - \frac{1}{M_1}\right)^{\alpha_0 - k} \\ \quad (i - 1)M_1 < j \leqslant iM_1 \\ \binom{\alpha_0 - 1}{k}\left(\frac{1}{M_1}\right)^{k}\left(1 - \frac{1}{M_1}\right)^{\alpha_0 - k - 1} \\ \quad 1 \leqslant j \leqslant (i - 1)M_1 \text{ or } iM_1 < j \leqslant N, 1 < i \leqslant M_1 \end{cases} \tag{19}$$