Multipath Banyan-based Multistage Networks for ATM Switches

Sheng-De Wang and Paul M. H. Dai
Department of Electrical Engineering
National Taiwan University, Taipei 106, TAIWAN
Email: sdwang@cc.ee.ntu.edu.tw

Abstract — Owing to the single path property of the banyan network, cell losses caused by the internal conflicts may be very pronounced. To overcome this problem, multipath ATM switches have been investigated. This paper proposes a class of ATM switch architectures that are constructed based on 3×3 switch elements (SEs). The proposed structures are constructed by adding lateral links to the banyan network to provide many paths between each input/output pair. As a result, the cell losses caused by internal conflicts are significantly reduced. And the proposed switch models are very modular, easy to expand, and use self-routing. The proposed switches use output queues to deal with the problem of cell sequence. From the analysis and simulation results, the proposed switches have the features of low cell loss probability and low cell delay.

Keyword: multipath switch, Banyan Network, ATM

I. Introduction

In the banyan networks, there exists one and only one path between each input port and each output port (single-path property). Owing to the single-path property, internal conflicts are happened usually. The conflict means that two cells in a switch element (SE) want to route to the same outlet of the SE, which results in the cell loss problem. Because there exists only one path from every given input port to any output port in a single-path network, the network can preserve the cell-sequence integrity. However, multiple-path MINs, which provide multiple paths between a given input port and any one of the output port, may have the cell out-of-sequence problem. This is the major disadvantage of a multipath MIN.

Almost all MINs have the modular architecture for system extension. In the banyan network, only $N\log_2 N$ SEs are used for an N×N switch network. The banyan network has the maximum throughput below 60% [1], which is a result of high cell loss probability due to internal contention. Many solutions are used to reduce the cell loss probability. Speed up the operation speed of the switch element is the simplest idea that was limited by the implementation technology. Although it is a straight way to improve the cell loss problem, the cost is very high. Another way to solve this problem is using the multi-path switch that provides multiple paths between each input port and each output port. Beside the above mentioned methods, adding the extra SEs in each stage to share the loads in the switch is another popular method [2]. Or, extra stages may be used in the switch [3]-[9]. To provide high conflict-tolerant ability, some additional buffers are provided in each SE. From another viewpoint, to prevent the conflict from occurring in the SE, the cell input order can be arranged by a sorting network before feeding cells into the switch. In ATM switches, conflict is a serious problem that may cause cell losses and decay of the throughput of the networks. A multipath augmented composite Banyan network is also studied in [10], whose basic building block is 4×4 SEs with logN stages. The augmented composite Banyan network is created by adding a link to each SE while the proposed approach in this paper is by adding some lateral links between SEs.

This paper is aimed to provide a scaleable method to design ATM switches. Based on the original banyan network, we change it into multipath switch by adding some lateral links, and the switch element is expanded to 3×3 SE. Then, some extra SEs in each stage are used to share the loads to decrease the cell loss. In the proposed models, each switch preserves high degree of parallelism and distributed control method. Without saying, all models use the *self-routing* method to route cells. According to the simulation result and the analysis, these models are multipath switches with high performance, low cell loss probability and low average delays.

II. The Proposed ATM Switches

The proposed ATM switches, are constructed by basic 3×3 switch elements (SEs). Each SE is with an identical hardware and with the same routing procedure.

A. The Specification of the Switch Element

Each 3×3 switch element has two formal inlets (IN_0, IN_1) , two formal outlets (OUT_0, OUT_1) , one redundant inlet (IN_2) and one redundant outlet(OUT_2) as shown in Fig. 1. The internal operation speed of each SE is the same with the external speed. In order to reduce the cell loss rate, the separate FIFO buffers are provided for each inlet of the SE. Besides the buffers, there are Central Controller (CC), Input Controllers (ICs), and Output Controllers (OCs) to perform the physical transmitting task of the SE.

B. The Routing Algorithm for the Switch Element

In the proposed switches, the same routing algorithm is adopted in each switch element. The routing algorithm can be divided into *the initiation phase* and *the transmitting phase*. In the initiation phase, each IC sends the self-routing bit and the priority number (according to

0-7803-7840-7/03/$17.00 ©2003 IEEE.

the age of the cell) of the *transmitting cell*.to the CC. And each OC sends to the CC the condition of the outlet after checking out each outlet is blocking or not blocking. Before entering the transmitting phase, each IC increases the age of all cells in the SE. In the transmitting phase, the CC allocates the usufruct of the outlet to the transmitting cells according to the above messages gathered from ICs and OCs. The transmitting cell with high priority (old age) have the highly usufruct of the outlet. If the desired outlet of the transmitting cell is blocked or used by the other transmitting cell with higher priority, the cell can be rerouted to the redundant outlet (OUT_2). However, the cell was stored in the queuing buffer in case of the redundant outlet is blocking or used by the cell with higher priority. The CC assigns the usufruct to each IC according to the priority number of the TCs. For inlet i, the *transmitting cell i* is defined as the cell that is transmitted from *inlet i.* The transmitting cell i may be from the FIFO buffer of inlet i or directly from the inlet i when the buffer is empty.

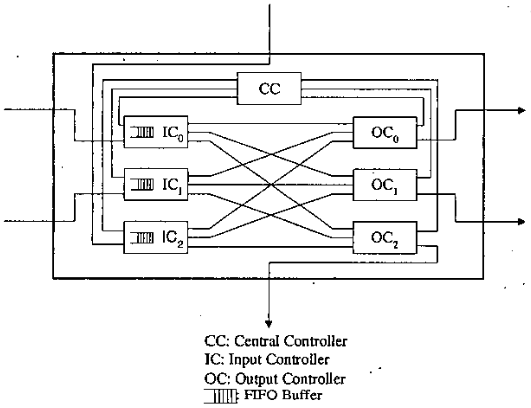


Fig. 1 The configuration of switching element

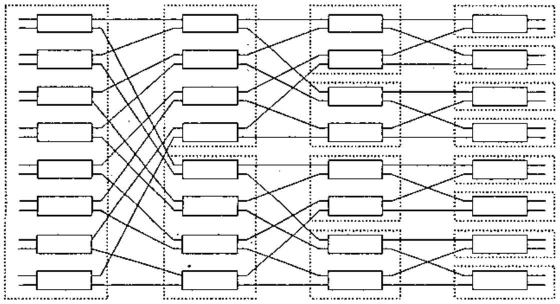


Fig. 2 The equality groups in the Banyan network

C. The Basic Model: Model 0

Before introducing the proposed models, a characteristic of the banyan network is illustrated. If the cell originally reaching the inport 0 of the input stage is sent to any other inport of the input stage instead, the cell can still arrive the required destination outport. By the above discovery, the term *equality group* is defined as following. In a given network stage, if the input cell arriving the switch elements in the stage can reach its destination, the group of switch elements is called *equality group.* The equality groups of the banyan network are shown in the Fig. 2.

By using the property of the equality group, the proposed model 0 is constructed by connecting each equality group with vertical cyclic links as shown in the Fig. 3. An NxN switch of model 0 was composed of logN stages. Each stage includes N/2 3×3 switch elements (SEs). At the final stage, the two outlets of each SE are connected to its dedicated output queues. Based on the normal banyan network, besides the two normal inlets and two normal outlets (4 normal links), there are one redundant inlet and one redundant outlet (two redundant links). While the normal links of each SE were connected as the banyan interconnection pattern, the redundant links were connected in a cyclic way. As shown in Fig. 3, these cycles are lateral in direction to the output.

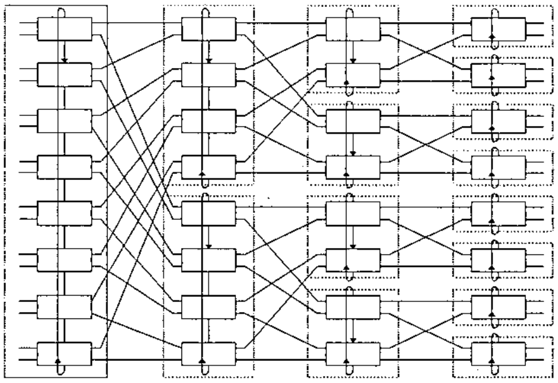


Fig. 3 The proposed switch architecture of the model 0.

D. The Enhanced Models: Model 1, Model 2, and Model 3

Model 0 is based on the concept of equality groups and thus can provide multipath and deflection ability. However, the other proposed models are based on the concept of adding redundant SEs in each stage to share the loads that can reduce the cell loss rate. The number of redundant SEs is increased in the sequence of Model 1, Model 2 and Model 3 as shown in the Fig. 4. The shadowed SEs are the redundant SEs and the white SEs are the original banyan SEs. The label in each redundant SE indicates which model the SE belongs to.

The SEs of the model 1 are the original banyan Ses (white SEs) and the shadowed SEs labeled (1). By observing of the model 1 in detail, we can find out that there are many inlets 1 are not used. In order to make full use of the SEs, we add some SEs to make use of these inlets. And the SEs belonged to the model 2 are the whole SEs of model 1 and the SEs labeled (2). It's easy to find out that some inlets of the SE labeled (2) is not used. With the same reason, to make full use of the switch, we add the SEs labeled (3). The union of the SEs of model 2 and the SEs labeled (3) forms the SEs of model 3.

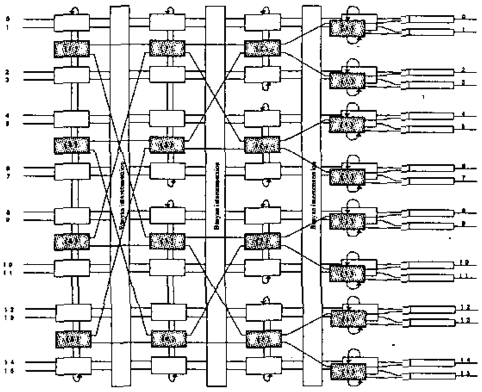
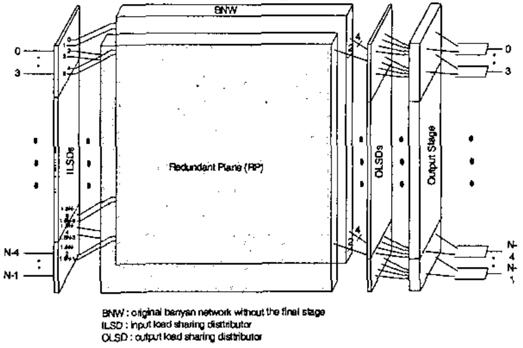


Fig. 4 The 16*16 switch architecture of the proposed model 1, model 2, and model 3.

The interconnections patterns between the redundant SEs are very simple. Each redundant SE is connected to the two nearest redundant SEs of the next stage that belong to different equality groups. In fact, the interconnection function of the proposed models is very flexible. The only requirement is that the two normal outputs of the SE are connected to two redundant SEs among different and adjacent equality groups. Owing to this flexibility, these models can easily change the interconnection patterns provided that it obeys the above mentioned requirement to connect the SEs. The interconnections of the redundant SEs can also be of the banyan type except the SEs in the final stage.

E. The Final Model: Model 4

Noticing the models proposed foregoing, the interconnection and the organization are not very structural. As shown in the Fig. 4, all the inlets of the SEs labeled (3) are not used. For the sake of making full use of the SEs and making the switch more modular, the final structural model, model 4, with high modularity is proposed. For an NxN switch of the proposed model 4, there are (3/4)Nn+(1/2)N SEs, where $n = \log_2 N$. The switch is composed of the original banyan network with 3*3 SEs without the final stage (BNW), the redundant plane (RP) which contains all the redundant SEs, the input load sharing distributors (ILSDs), the output load sharing distributors (OLSDs), and the final output stage as shown in Fig. 5. All the components are connected with proper interconnection. As can be seen from Fig. 4, there are many unused inlets in the redundant SEs in the first stage and the redundant SEs in the final stage. In order to fully use the switch element, a simple architecture is proposed to distribute loads to these un-used inlets. The input load sharing distributors (ILSDs) are used to distribute the input loads. And the output load sharing distributors (OLSDs) are used to distribute the output loads. Using this approach of load sharing, the cell loss probability will decrease and the value of performance/cost of the switch will increase. For an NxN switch of model 4, there are $N/4$ ILSDs and $N/4$ OLSDs.



BNW : original banyan network without the final stage
ILSD : input load sharing disttributor
OLSD : output load sharing disttributor

Fig. 5 The N*N switch architecture of the model 4.

Each ILSD is connected to two adjacent SEs on the BNW and one SE on the RP. The ILSD is operated as a switch that changes the path of cell with 4 cycles. Here, the cycle time is the same as the internal time slot of the switch. In cycle *i*, the cell from inport *i* is re-directed to the inlet (*i mod 2*) of the SEs on the RP. That is to say the

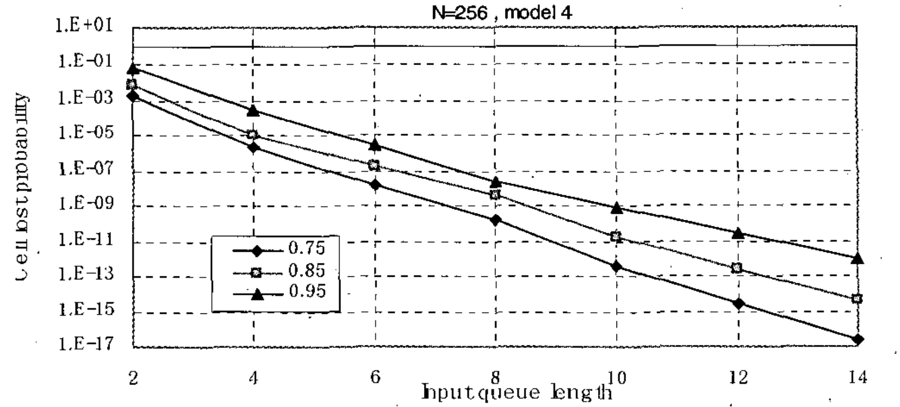


Fig. 6 Cell loss probabilities for the model 4, 256*256 switch with input queue length ranging from 2 to 14 and offered loads 0.75, 0.85,

original load of each inport is reduced to 3/4 load as compared to the original load.

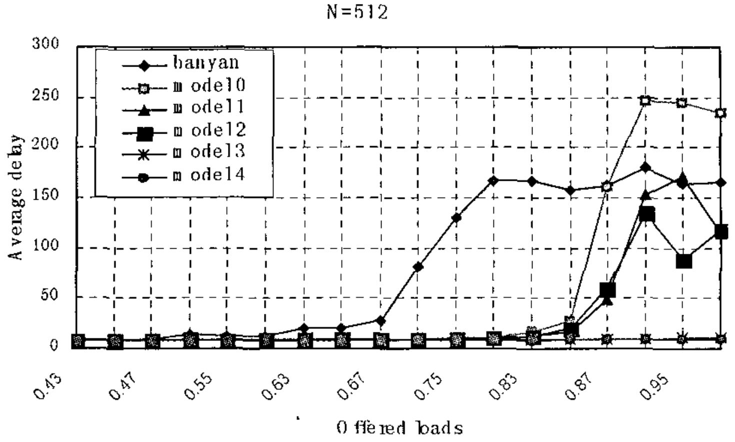


Fig. 7 The average delay of different models (Input queue lenth Q=40)

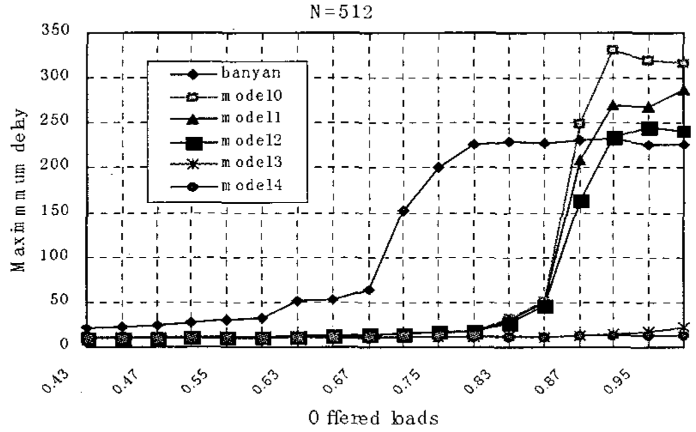


Fig. 8 The maximum delay of different models (Input queue lenth Q=40)

With the same concept, the load of the final stage of BNW is distributed to the inlet 1 of each SE on the redundant SEs of the output stage. The OLSD is also with 4 cycles. Each OLSD distributes the loads of two SEs on the BNW to two adjacent SEs on the redundant SEs of the output stage. Because there are four SEs, the OLSD distributes the cell from outlet 0 and outlet 1 of the upper SE to the upper inlet 1 of the redundant SE in cycle 0 and cycle 2, respectively. During the cycle 1 and cycle 3, the OLSD distributes the cell from outlet 0 and outlet 1 of the lower SE to the lower inlet 1 of the redundant SE, respectively.

The SEs on the RP are interconnected as banyan interconnections. For an NxN switch, the RP is composed of an (N/2) x (N/2) banyan network with 3*3 SEs. And the number of SEs on the output stage is N.

III.Performance Results

In this Section, some important performance results of the proposed switches are presented.

A. The Cell Loss Probability

In Fig. 6, the cell loss probabilities versus input queue lengths of a 1024x1024 switch of model 4 is plotted. And the relationships of the input queue length, offered loads, and the cell loss probability are illustrated. For a given offered loads, the cell loss probability decreased if the input queue length increase. For the given a fixed input queue length, the cell loss probability decreased if the offered loads decreased. As shown in Fig. 6, the cell loss probability is on the order of 10^{-9} with the input queue length greater than 12 even on a heavy load, 0.95.

B. The Average Delay and the Maximum Delay

The average delay and maximum delay is obtained from the simulation results. Each point in the result of simulation is running with 10^7 time slots. In Fig.7 and 8, the average delay and the maximum delay of a 512x512 switch with different models are plotted. Here the input queue lengths are all given 40 in the simulations to realize the real routing ability of the switch. The average delays of model 3 and model 4 are kept very low if the load is below 0.95. And the average delays of model 0, model 1, and model 2 are kept low if the load is below 0.85. However the average delay of the original banyan network is low only when the load is below 0.7. (The cell loss probability of the banyan is very high when the loads is greater than 0.5). Comparing Figs. 7 and 8, the differences of the maximum delay and the average delay between the proposed are very low. But the difference is very high for of the original banyan network. In the model 0, we can see that a small modification of the original banyan network can lead to a surprising outcome.

C. The Number of Nodes

As shown in the Table 1 and Table 2, the comparisons of the number of switch elements are listed. The proposed switches have the complexity of NlogN. The tables expressed that the proposed switches use the least number of switch elements. Furthermore, the proposed switches can afford a very high performance result.

IV. Conclusion

In this paper, several switch models associated with the performance results are presented. The switches are built up from 3×3 switch elements and using the simple self-routing technique to route cells. The proposed switches provide many redundant paths between each input/output pair. So, those switches can afford to provide high performance under the heavy input loads (which would lead to many conflicts). The complexity of the proposed switches is low, NlogN, for an N×N ATM switching system. This means that the proposed can use very few SEs to offer very high performance. There do not need any speed up in the internal switch elements. However, the output queue needs speed up factor 2 to eliminate the output cell loss. In the performance analysis, the uniform random traffic patterns are used. From the analysis results and the simulation results, the cell loss probability is kept low even small input queue buffers are used. And the average delay is very low as compared to the other ATM switches. The maximum delay does not deviate large from the average delay, which illustrates that the proposed switches are very stable under the highly loaded conditions.

Table 1 The formulas of the number of SEs for different type of switches.

Proposed□Model□4$= (3/4)Nn + (N/8)$
$B-tree = Nn(n+1)/4$
$Itoh = N(n-1)+1$
$Benes = N(n-1)/2$
$OLSS = Nk/2$

Table 2 The numerical results of the Table 1: comparison of the number of SEs with different switching architectures.

N	*4*	*8*	*16*	*32*	*64*	*128*	*256*	*512*	*1024*
n=log N	2	3	4	5	6	7	8	9	10
Proposed	6	19	50	124	296	688	1568	3520	7808
B-tree	6	24	80	240	672	1792	4608	11520	28160
Itoh	5	17	49	129	321	769	1793	4097	9217
Benes	6	20	56	144	352	832	1920	4352	9728
k	10	15	20	25	30	35	40	45	50
OLSS	20	60	160	400	960	2240	5120	11520	25600

References

[1] H. S. Kim and A. Leon-Garcia, "Performance of buffered banyan networks under nonuniform traffic patterns," *IEEE Transactions On Communications,* vol. 38, pp. 648-658, May. 1990.

[2] J.-J. Li and C.-M Weng, "B-tree: a high-performance fault-tolerant ATM switch," *IEE Proc.-Commun.,* vol. 141, pp. 20-28, Feb. 1994.

[3] Y. Mun and H. Y. Youn, "Performance analysis of finite buffered multistage interconnection networks," *IEEE Transactions On Computers,* vol. 43, pp. 153-162, Feb. 1994.

[4] I. Widjaja and A. Leon-Garcia, "The Helical switch: a multipath ATM switch which preserves cell sequence," *IEEE Transactions On Communications,* vol. 42, pp. 2618-2629, Aug. 1994.

[5] S. Bassi, M. Decina, P. Giacomazzi, and A. Pattavina, "Multistage shuffle networks with shortest path and deflection oruting for high performance ATM switching: The Open-Loop Shuffleout," *IEEE Transactions On Communications,* vol. 42, pp. 2881-2889, Oct. 1994.

[6] S. Bassi, M. Decina, P. Giacomazzi, and A. Pattavina, "Multistage shuffle networks with shortest path and deflection oruting for high performance ATM switching: The Closed-Loop Shuffleout," *IEEE Transactions On Communications,* vol. 42, pp. 3034-3044, Nov. 1994.

[7] J.-J. Li and C.-M Weng, "Self-routing multistage switch with repeated contention resolution algorithm for B-ISDN," *Computer communications,* vol. 17, pp. 788-798, Nov. 1994.

[8] S. Urushidani, "Rerouting network: a high performance self routing switch for B-ISDN," *IEEE Journal On Selected Areas In Communications,* vol. 9, pp. 1194-1204, Oct. 1991.

[9] A. Itoh, "A fault-tolerant switching network for B-ISDN," *IEEE Journal On Selected Areas In Communications,* vol. 9, pp. 1218-1226, Oct. 1991.

[10] H. -I. Lee, S. -W. Seo and T. -Y. Feng, "The Augmented Composite Banyan Network," *5th International Conference On High Performance Computing,* pp. 285 -292, Dec. 1998.