Joint Conference of 1996 International Computer Symposium
December 19~21, Kaohsiung, Taiwan, R.O.C.

Design and Implementation of a Fault Tolerant ATM Switch for B-ISDN

Kuochen Wang and Feng-Ming Lin

Department of Computer and Information Science
National Chiao Tung University
Hsinchu, Taiwan, R.O.C.
kwang@cis.nctu.edu.tw

Abstract

We present a new method to build a fault tolerant ATM switch. By this method, we can build an ATM switch which has two non-overlapping paths between each input/output pair. The key component in the proposed switch is a 2 × 2 FTSE (Fault Tolerant Switching Element) which can be the basic building block for high speed ATM switches. The FTSE is made with the ability of fault tolerance by adding a few spares to the traditional Switching Element (SE). We can construct an MIN which has two levels of fault tolerance ability and the redundant paths are in proportional to the network size. By mathematical analysis, we conclude that our ATM switch uses less SEs and has more redundant paths than other ATM switches. By VHDL simulation, we have verified the functionality of the switch. We also synthesize the ATM switch to evaluate its delay and area. The experimental results demonstrate that the reliability/cost ratio of the fault tolerant FTSE-based ATM switch is better than that of other switches.

1 Introduction

There are several MIN architectures suited for ATM switches such as Banyan, Baseline, Omega, and Benes [1]. Most of them are a variation of the Banyan switch. The configuration of a 16 × 16 Baseline switch is in Figure 1. Some MIN switches are called blocking switches, for example, Banyan, Baseline, and Omega. The method for resolving contention is to add queues to buffer the low priority cells in contention or to provide multiple paths between each input/output pair such as in the Benes switch [1]. There are some switches that are non-blocking. Batcher-Banyan [2] is one of them. It adds a switch named *Batcher* to sort the incoming cells according to a certain order in front of the Banyan switch in order to avoid internal blocking. Normally, the Batcher switch is more complicated and larger than the Banyan switch.

Several ATM switch architectures were presented in [3] [4] [5] [6] [7] [8] [9] [10] to satisfy the high speed switching requirement. We can classify them into two categories: with fault tolerance or not. The papers in

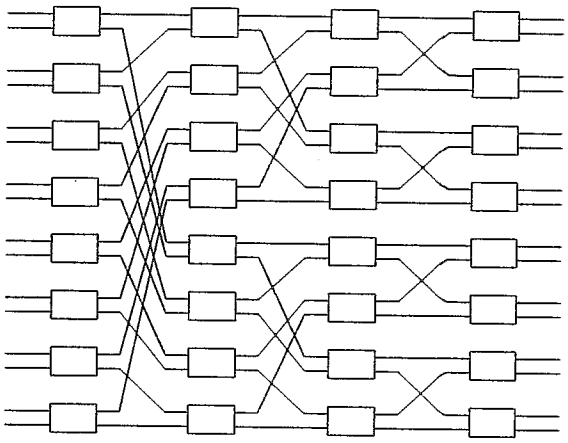


Figure 1: The configuration of a 16 × 16 Baseline switch.

[5] [6] [7] [8] deal with fault tolerant ATM switch designs . Their approaches are based on a 2 × 2 modified SE by adding extra SEs [7], links [5] or using multiple-pass [8] to provide fault tolerance. In [5], it is a modified Delta network for providing fault tolerance. One extra stage and double links are added to the original Delta network. Thus, a 2 × 2 SE is replaced by a 4 × 4 SE in the middle stage of the resulting network and a 4 × 2 SE in the last stage, respectively. In [7] it uses some *subswitches* to enhance the fault tolerance of the conventional multistage interconnection network by providing alternative paths between each input/output pair. The added SEs make the decision of selecting a path more complicated and the SEs in the switch have different sizes. A modified SE called PHOENIX in [6] is a 2 × 2 SE. The PHOENIX is based on a 2 × 2 crosspoint buffered switching element. The characteristics of PHOENIX make that it can be constructed as a multiplexer or demultiplexer. A larger ATM switch can be built by using the PHOENIX only. The PHOENIX employs several queues to buffer contention cells. The scheme in the PHOENIX is complex to initiate an adaptive routing or to trigger an alternative switch fabric when a fault

occurs. As to the multiple-pass method, the cell is rerouted by the faulty switch several times, if necessary, in order to reach the proper destination [8].

The SE we present is a 2×2 FTSE (Fault Tolerant Switching Element) which can be the basic building block of an MIN switch for high speed broadband networks. The fault tolerance scheme in the FTSE is to employ one spare IC and two spare OCs to provide multiple paths between each IC/OC pair. We also proposes a method to build an MIN switch by using the FTSEs. The proposed method makes the switch has two non-overlapping paths between each input/output pair. We use the IEEE-1076 Standard, VHDL (VHSIC Hardware Description Language)[11] [12], to describe the ATM switch. VHDL simulation is used to verify the correctness of the functionality of the switch. VHDL synthesis is also used to analyze the *delay time* and *area* of the switch to make sure that it meets the speed requirement of the ATM switch with low overhead.

The organization of this paper is in the following manner. In the next section, we describe the FTSE and a method to build an FTSE-based MIN. In section 3, we will give an example to demonstrate how to route a cell in the resulting MIN. In section 4, we analyze the hardware complexity and the redundant paths of the proposed switch and give a comparison with other switches. In section 5, we evaluate and compare the reliability and cost effectiveness of our switch with other switches. In section 6, we will describe the proposed ATM switch by using VHDL. The VHDL simulation and synthesis results will help us validate the switch design and analyze the delay time encountered and the area of the proposed switch. Finally, we give some concluding remarks in section 7.

2 Design approach

2.1 The architecture of the FTSE and its functions

The configuration of the proposed 2×2 FTSE is shown in Figure 2. The FTSE is composed of the following basic parts : two ICs (Input Controllers), a spare IC, a Selector, two OCs (Output Controllers), two spare OCs, two MUXs (multiplexers), and a BFS_CTRL (backward fault signal controller). Now we describe each part of the FTSE as follows.

- *ICs (Input Controllers) and spare IC*
  The two ICs are the inlets of the FTSE. The use of a spare IC is for fault tolerance. When one of the normal IC_1 or IC_2 is faulty or can't accept a cell then the spare IC begins to work. There is a buffer of size one in the spare IC. If both IC_1 and IC_2 are broken, the low priority cell will be buffered in the buffer of the spare IC and the high priority cell will be passed to the selector via the spare IC.

- *Selector*
  The next stop of the incoming cells is a 3×2 selector. There are control signals coming from the ICs and the spare IC to activate the selector to make a proper selection among the three incoming cells.

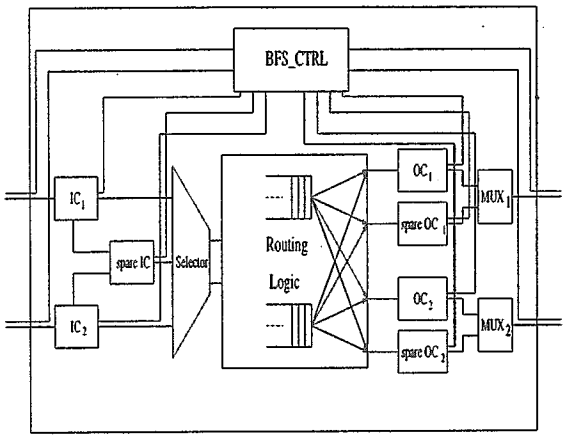


Figure 2: The configuration of the proposed 2×2 FTSE.

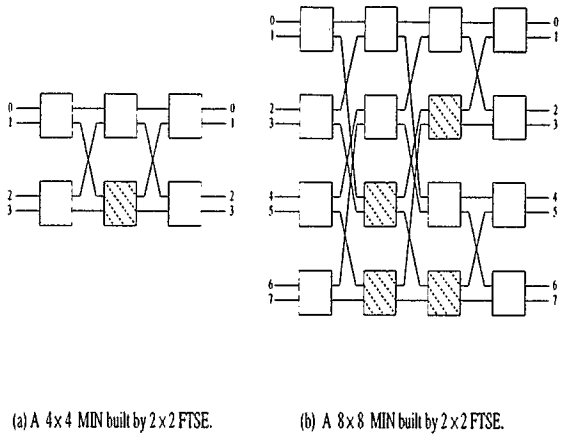
- *Routing Logic*
  The routing logic first checks whether there are cells in the two shared buffers or not. If there are cells in the shared buffers then the first cell in the corresponding shared buffer will be picked up and sent to the desired OC or spare OC. However, if the two shared buffers are empty, the incoming cell will be sent to its desired OC according to the i-th bit of the routing tag. Depending on the number of stages in the network, the routing tag will be inserted before each cell enters the switch.

- *OCs (Output Controllers) and spare OC*
  If a cell passes through the routing logic successfully, the next stop in the FTSE is OC_1, OC_2, spare OC_1, or spare OC_2. If faults occur in OC_1 or OC_2 (spare OC_1 or spare OC_2), the cell can only be sent to the spare OC_1 or spare OC_2 (OC_1 or OC_2).

- *MUX (Multiplexer)*
  When a cell reaches OC_1, OC_2, spare OC_1, or spare OC_2, the corresponding MUX (multiplexer MUX_1 or MUX_2) selects one of the cells from them. Since the high priority cell is in OC_1 or OC_2, the MUX will select the cell from OC_1 or OC_2 first.

- *BFS_CTRL (Backward Fault Signal Controller)*
  The function of the BFS_CTRL is to determine the state of the FTSE. It will receive the state of the FTSE in the next stage. If the state is faulty then the BFS signal is set to true. Otherwise, it will check the state of current FTSE and set the BFS signal to a proper value.

According to the design described above, we realize that there are two paths from an IC to the selector and two paths from the routing logic to an MUX.



(a) A 4x4 MIN built by 2x2 FTSE. (b) A 8x8 MIN built by 2x2 FTSE.

Figure 3: The construction of two MINs using FTSEs.

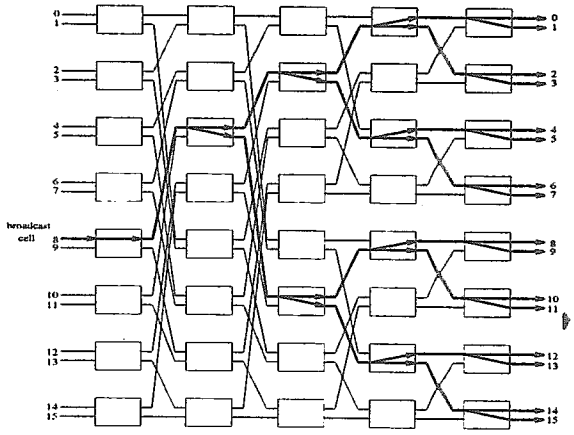


Figure 4: The broadcast routing in a 16 × 16 FTSE-based ATM switch.

Therefore, there are 4 paths between each IC/OC pair.

2.2 An ATM switch built by FTSEs

Here, we present a method to build an MIN by using the FTSE as a basic building block. An MIN built by FTSEs is called an *FTSE-based network*. Some MINs like Baseline, Banyan, and Omega, etc., have only one unique path between each input/output pair. Our fault tolerant MIN is based on the Baseline switch (see Figure 1). We add one extra stage to the Baseline switch. There are $N/2$ FTSEs in each stage of an $N \times N$ FTSE-based switch. The extra stage can be appended to the first or to last stage of the Baseline network. As depicted in [15], a general method is presented to insert e stages (to the back) and p planes to the $log_2 N$ network. In this paper, we on ly discuss the situation of appending the extra stage to the first stage of the Baseline network. To build a 4×4 MIN, we append 2 FTSEs to the first stage of the 4×4 Baseline network. As shown in Figure 3(a), the MIN has two independent halves: one is shown as a "plain" half and the other is shown as a "hatched" half. Thus, each input cell can reach its desired output by passing through either half of the MIN. For an 8×8 MIN, we add 4 FTSEs to the first stage of the 8×8 Baseline network. As shown in Figure 3(b), there is also two independent halves: one is a "plain" half and the other is a "hatched" half. In a similar fashion we can construct any size of the ATM switch.

3 Routing in the ATM switch

3.1 Singlecast routing

Since our fault tolerant ATM switch has two-level of redundant paths, a routing path is selected based on two levels: one is in the FTSE and the other is in the resulting MIN. The routing scheme for the proposed FTSE-based ATM switch is similar to that of the Baseline switch, except that our switch needs extra control for the first stage. There is a *Backward Fault Signal* (BFS) associated with each FTSE. The function of the BFS is similar to the Backward Availability Signal (BAS) proposed in [13]. It is to tell the previous stage the state of the current stage.

3.2 Broadcast and multicast routing

Broadcast Routing If a cell has to be broadcast, at each stage, except for the first stage, the cell is sent to the upper and lower OCs within each FTSE. Figure 4 shows the fashion of broadcast routing in a 16×16 FTSE-based ATM switch.

Multicast Routing If a cell has to be multicast, at each stage, a cell is selectively sent to the upper, lower, or both OCs in an FTSE depending on the destination addresses. Let the stage number of an $N \times N$ FTSE-based switch be from 0 to $log_2 N$. Under normal operations, assume that a cell from inlet 8 has to be multicast to outlets 4, 6, 9, 14, 15 in a 16×16 FTSE-based ATM switch as shown in Figure 5. The binary representation of destination addresses 4, 6, 9, 14, 15 is "00100", "00110", "01001", "01100", "01111". "0" is appended to the leftmost of each destination address for normal operations. At each stage, the FTSE first checks the multicast bit then the destination addresses. At stage 0, since the first bits of all the destination addresses are all "0", the cell is sent to the upper OC. At stage 1, the second bits of the destination addresses contain "0" and "1", so the cell is copied and sent to the upper OC and lower OC, respectively. At the subsequent stages, each FTSE repeats the same procedure described above until cells reach the last stage.

4 Comparison with other fault tolerant ATM switches

The comparison criteria are based on hardware complexity (expressed in terms of the number of SEs) and fault tolerance ability (expressed in terms of the number of redundant paths).

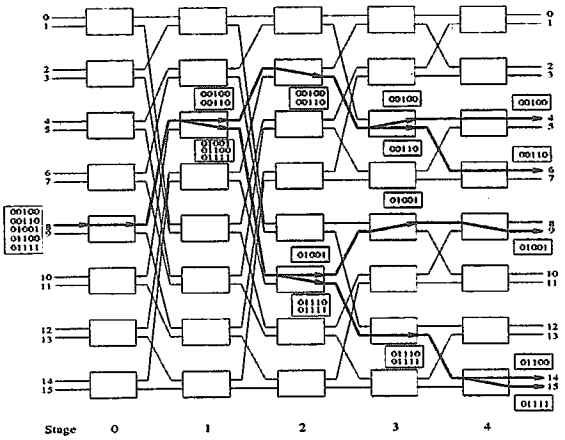


Figure 5: The multicast routing in a 16 × 16 FTSE-based ATM network.

4.1 Hardware complexity

Here, we compare the FTSE-based switch with the MD-Omega network [14] built by the 2×2 PHOENIX and the Itoh's network [7]. We know there are $n+1$ stages in our FTSE-based ATM switch. Therefore, we can calculate the number of FTSEs needed in an $N \times N$ MIN. Since there are $N/2$ FTSEs in each stage, the total number of FTSEs required is

$$\begin{aligned} H_{FTSE-based} &= N/2\ (n+1) \\ &= N/2\ (log_2 N + 1) \\ &= N/2 log_2 N + N/2 \end{aligned} \quad (1)$$

Figure 6 shows the number of SEs used in each network in variance of network size. By inspecting Figure 6, we conclude that to build the same size of an $N \times N$ fault tolerant MIN, our FTSE-based switch uses the least SEs.

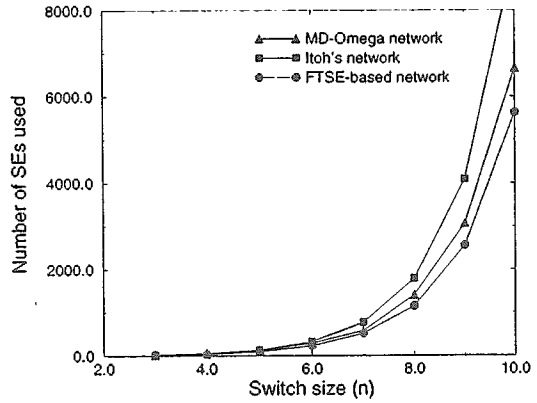


Figure 6: Number of SEs used in each $2^n \times 2^n$ MIN.

4.2 Redundant paths

If a switch network can provide more redundant paths, it will be able to survive under more faults. We give a comparison of redundant paths among the MD-Omega network, the Itoh's network and FTSE-based network. The redundant paths of the MD-Omega network built by PHOENIX is a fixed 2 regardless of the network size. For each node in the Itoh's network, there are $M(i,j)$ paths to the last node, where (i,j) is the node position. The redundant paths of the Itoh's network ($R_{Itoh's}$) [7] is expressed as:

$$\begin{aligned} &M(n,0) = M(n-1,0) + M(n,1) \\ &M(n,k) = M(n-1,k-1) + M(n,k+1) \\ &M(n,n-1) = 1 \end{aligned} \quad (2)$$

As mentioned before, there are 4 redundant paths between each IC/OC pair in the FTSE. Therefore, for an $N \times N$ FTSE-based switch, the redundant paths for each input/output pair is

$$\begin{aligned} R_{FTSE\_based} &= (2^2)^{(log_2 N+1)} \times 2 \\ &= 2^{2(log_2 N+1)+1} \end{aligned} \quad (3)$$

Figure 7 shows that the number of redundant paths of the FTSE-based network is far larger than those of the other two networks of the same size.

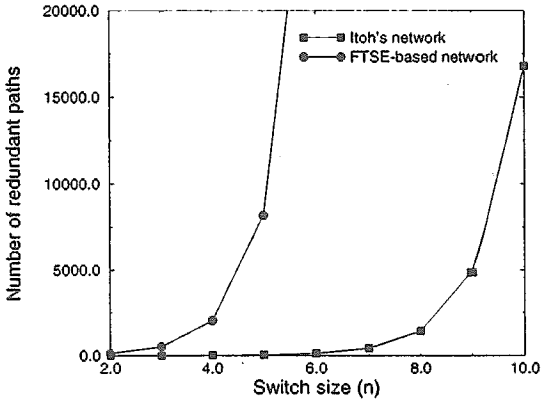


Figure 7: Number of redundant paths in a $2^n \times 2^n$ MIN.

5 Reliability analysis

5.1 Fault model

The fault model used in this section is similar to that proposed in [7]. Faults can occur in ICs, spare IC, OCs , spare OCs and links that connect two adjacent SEs. In our method, the functionality of a spare IC is similar to an IC except that it can accept two cells concurrently. Hence, we treat a spare IC as two independent ICs. Similar to [17], the two ICs that connect to the same OC and spare OC in the previous stage is called an *input module*. The OC and the

corresponding spare OC is called an *output module*. An *element* is formed by an input module in the current stage, an output module in the previous stage, and the links between them. If there are faults in the input module, output module, or links, an element is considered as faulty. When an element is marked as faulty, then it remains faulty permanently. Our analysis is based on the following assumptions [7]:

- Events in which an element becomes faulty are independent and occur randomly.
- When there are faults that prevent the connection of an arbitrary input/output pair, a network is considered to be failed.

Similar to [17], the analysis only concentrates on the elements between the first stage and the last stage of an network.

5.2 The reliability of the FTSE-based ATM switch

The FTSE-based switch is considered to be faulty if the two non-overlapping paths are both broken. Thus, for each output port in stage i, two elements can be selected to connect to stage $i+1$ (due to the two nonoverlapping paths). The m FTSEs within a stage can be divided into $m/2$ subsets of 2 elements. A switching network is faulty if both elements in the same subset are faulty. Thus, in an $N \times N$ FTSE-based switching network, there are $N/2 \times log_2 N$ subsets. Here, we can calculate the average number of faults, K, that causes the network failure. K is expressed as follows [16]:

$$K = \sum_{i=2}^{L} i\, P(i) \tag{4}$$

where

$P(i)$ = *Prob* {the network fails due to the i-th fault}

L = $N/2 \times log_2 N$, the number of subsets in an $N \times N$ FTSE-based network

$P(i)$ can be further expressed as:

$$P(i) = Q(i-1) \times R(i) \tag{5}$$

where

$Q(i-1)$ = *Prob* {there are $i-1$ faults in the network and the network does not fail }

$R(i)$ = *Prob* {a fault that makes the network fail | $i-1$ faults have already existed in the network }

We have [16],

$$Q(i-1) = 2^{i-1} \times \binom{L}{i-1} / \binom{2L}{i-1} \tag{6}$$

$$R(i) = (i-1)/(2L-i+1) \tag{7}$$

We evaluate the cost effectiveness of an ATM switching network in terms of K/T, where T is the total number of elements in the middle stages of the network. The cost effectiveness of the FTSE-based network and the Itoh's network, respectively, is shown in Figure 8. In Figure 8, we conclude that the cost effectiveness of our FTSE-based network is better than that of the Itoh's network.

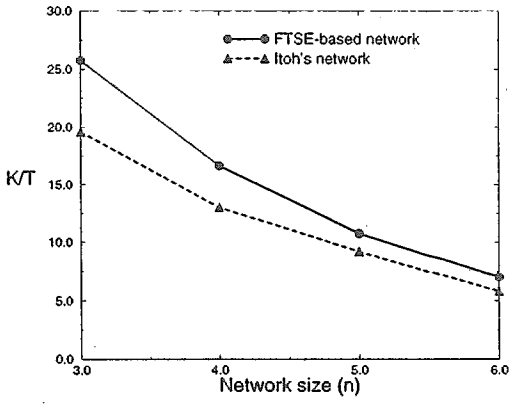


Figure 8: The cost effectiveness comparison between the FTSE-based network and the Itoh's network.

6 Simulation and synthesis

6.1 VHDL simulation

According to the design described in Section 2, an FTSE-based ATM switch is described with VHDL. Each component of the FTSE is described independently. Then, we group each component of the FTSE to form a complete 2×2 FTSE. The VHDL simulation result verifies the functionality of the FTSE. Finally, we use the FTSE as a building block to construct an example 4×4 FTSE-based ATM switch.

6.2 VHDL synthesis

The ATM switch described with VHDL can be further synthesized into the gate level representation in a chosen technology. According to the synthesis result, the overall delay for a fault-free FTSE is 5.12 ns. However, the overall delay becomes 5.47 ns when there is a fault in an IC and the spare IC is used. The total area of the FTSE is 58617 (unit area). Based on the FTSE, we can expand it to a 4×4 FTSE-based ATM switch. The overall delay for the 4×4 FTSE-based ATM switch is 15.36

ns when fault-free and 16.41 ns when a fault occurs in an IC and the spare IC is used. The total area of the 4×4 FTSE-based ATM switch is 351702 (unit area). We need two clock cycles to transfer a cell from *IC* to *Routing Logic* and a delay time to transfer it from *OC* to *MUX*. Thus, the maximum throughput of a $2^n \times 2^n$ FTSE-based ATM switch is $\lfloor \frac{8 \times 53}{(2 \times clock\_cycle\_time + delay(OC+MUX)) \times (n+1)} \rfloor$.

7 Conclusions

An ATM switch built by an MIN is inherently a parallel switch, since it accepts and processes incoming cells concurrently. Our proposed ATM switch is also an MIN. The basic component of our FTSE-based ATM switch is a 2×2 FTSE. The FTSE is very flexible and can be the basic building block of any kind of MINs. The FTSE can not only process incoming cells quickly, but also have the ability of fault tolerance. It offers multiple paths between each IC/OC pair and makes the ICs and OCs fault tolerant.. The problem of cells contention for the same output is resolved by using shared buffers to store the low priority contention cells. Thus, we offer a guarantee service to high priority cells. The method of constructing an MIN with FTSEs will provide each input/output pair with two non-overlapping paths. Mathematical analysis shows that our switch is better than other fault tolerant ATM switches in terms of the numbers of SE used and redundant paths, and the reliability/cost ratio. The VHDL simulation results have verified the functionalities of the switch and the VHDL synthesis results also support that the switch throughput can meet the ATM network requirement.

Acknowledgement

This research was supported in part by National Science Council, ROC under Grant NSC85-2622-E009-0102.

References

[1] V. E. Benes, *Mathematical Theory of Connecting Networks and Telephone Traffic*, Academic Press, New York, 1965.

[2] J. Y. Hui and E. Arthurs, "A Broadband Packet Switch for Integrated Transport," *IEEE Journal on Selected Areas in Communications*, vol. SAC-5, pp. 1264-1273, Oct. 1987.

[3] J. N. Giacopelli, J. J. Hickey, W. S. Marcus, W. D. Sincoskie, and M. Littlewood, "Sunshine: A High Performance Self-routing Broadband Packet Switch Architecture," *IEEE J. Sel. Areas in Commun.*, vol. 9, no. 8, pp. 1289-1298, Oct. 1991.

[4] Y. S. Yeh, M. G. Hluchyj, and A. S. Acampora, "The Knockout Switch: A Simple, Modular Architecture for High Performance Packet Switching," *IEEE J. Sel. Areas in Commun.*, vol. 5, no. 8, pp. 1274-1283, Oct. 1987.

[5] Wen-Shyen E. Chen, Young Man Kim, Yow-Wei Yao, and Ming T. Liu, "FDB: A High-Performance Fault-Tolerant Switching Fabric for ATM Switching Systems," *IEEE International Pheonix Conference on Computers and Communications*, pp. 703-709, Mar. 1991.

[6] V. P. Kumar, J. G. Kneuer, D. Pal, and B. Brunner, "PHEONIX: A Building Block for Fault Tolerant Broadband Packet Switches," *IEEE GLOBECOM'92*, pp. 228-233, Dec. 1992.

[7] Arata Itoh, "A Fault Tolerant Switching Net work for B-ISDN," *IEEE Journal on Selected Areas in Communications*, pp. 1218-1226, Oct. 1991.

[8] T. H. Lee and J. J. Chou, "Fault Tolerance of Banyan Using Multiple-Pass," *IEEE INFOCOM'92*, pp. 867-875, May 1992.

[9] H. Ahmadi, W. E. Denzel, C. A. Murphy, and E. Prot, "A High-Performance Switch Fabric for Integrated Circuit and Packet Switching," *IEEE INFOCOM'88*, pp. 9-18, Mar. 1988.

[10] H. S. Kim, "Design and Performance of Multinet Switch: A Multistage ATM Switch Architecture with Partially Shared Buffers," *IEEE/ACM Transactions on Networking*, vol. 2, no. 6, Dec. 1994.

[11] Z. Navabi, *VHDL Analysis and Modeling of Digital Systems*, McGRAW-HILL Inc., 1993.

[12] Roland Airiau, Jean-Michel Berge, and Vincent Olive, *Circuit Synthesis with VHDL*, Kluwer Academic Publishers, 1994.

[13] M. A. Henrion, G. J. Eilenberger, G. H. Petit, and P. H. Parmentier, "A Multipath Self-Routing Switch," *IEEE Communications Magazine*, pp. 46-52, Apr. 1993.

[14] V. P. Kumar and S. J. Wang, "Reliability Enhancement of Multistage Interconnection Networks by Space and Time Redundancy," *IEEE Transactions on Reliability*, vol. 40, no. 4, pp. 461-473, Oct. 1991.

[15] D. J. Shyy and C. T. Lea, "Log_2(N, m, p) Strictly nonblocking Networks," *IEEE Transactions on Communications*, vol. 39, no. 10, pp. 1502-1510, Oct. 1991.

[16] L. Ciminiera and A. Serra, "A Fault Tolerant Connecting Network for Multiprocessor Systems," *International Conference on Parallel Processing*, pp. 113-122, Aug. 1982.

[17] N. Tzeng, P. Yew, and C. Zhu, "A Fault-Tolerant Scheme for Multistage Interconnection Networks," *International Symposium on Computer Architecture*, pp. 368-375, Jun. 1985.