Fault tolerance of the MIN with multiple outlets

Akira Funahashi, Toshihiro Hanawa, Hideharu Amano
Dept. of Computer Science, Keio University
3-14-1, Hiyoshi Yokohama 223 JAPAN
+81-45-560-{1063,1064(fax)}
E-mail: {funa, hanawa, hunga}@aa.cs.keio.ac.jp

Abstract

Multistage Interconnection Networks (MIN) with multiple outlets are networks which can support higher bandwidth than those of nonblocking networks by passing multiple packets to the same destination.

Fault recovery mechanisms are proposed for one of such networks TBSF with the best use of its inherent fault tolerant capability. With these mechanisms, on-the-fly fault recovery is possible for multiple faults on switching elements. For the link fault, the networks are reconfigured after fault location, and the network is available with some performance degradation. The bandwidth degradation under multiple fault on link/element is analyzed with theoretical models and simulation.

1 Introduction

Multistage Interconnection Networks (MIN) with multiple outlets[12] are networks which can support higher bandwidth than those of nonblocking networks by passing multiple packets to the same destination.

The simplest MIN with multiple outlets is called Multi Banyan Switching Fabrics (MBSF) [8][9][7][6] which supports multiple independent banyan (omega) networks for traffic distribution. It has been well studied and load balancing algorithms were proposed as a circuit switching network or the ATM (Asynchronous Transfer Mode) packet switching network.

Another simple MIN with multiple network is Expanded Banyan Switching Fabrics (EBSF) or Expanded Delta network which realizes multiple outlets by expanding the size of network[2][10]. Conflict free access methods are proposed[2] and the efficiency as a processor-memory interconnection network of multiprocessors is demonstrated.

However, from the result of analysis[12], these two networks support poor bandwidth because of their simple structures. In the MBSF, multiple networks are only used independently. In the EBSF, the earlier stages do not contribute traffic distribution.

Two advanced MIN with multiple outlets called Tandem Banyan Switching Fabrics (TBSF)[4][3] and Piled Banyan Switching Fabrics (PBSF) [12] support much better performance than those of the MBSF and EBSF[12]. These networks can be efficiently used both for a processor-memory interconnection network in a multiprocessor and the ATM packet exchanger of a telecommunication switching system.

These two networks provide inherent fault tolerant capability, since they consist of combinations of multiple MINs. In this paper, a fault recovery mechanism is attached to the TBSF, and proposed Fault tolerant TBSF (F-TBSF). Then, the performance degradation when some elements of the F-TBSF are damaged is analyzed both with probablistics model and simulation. Simular method can be used for the PBSF with some modification.

2 The control model and fault model

2.1 The control model

The MIN with multiple outlets are proposed for a switching system with a simple structure and control. All packets are inserted into serially (in a few bits parallel) synchronized with a common frame clock from input packet buffers. Each switching element stores only one bit (or a few bits) of the packet, and the MIN behaves like a set of shift registers with the switching capability.

When a conflict occurs, one of the conflicting packets must be routed to the incorrect direction since there is no packets buffer in each switching element. When a packet is routed to the incorrect direction, the conflict bit in the routing tag is set. The packet whose conflict bit is set is treated as a dead packet, and never interferes the other packets. Since this control/structure enables to use high speed clock and high density implementation compared with the MIN providing packets buffers inside every element, it is commonly used in the ATM packet switching network, and sometimes used in a multiprocessor[1][13].

However, the packet conflicts inside the MIN will severely degrade performance because the conflicting packets must be inserted again in the next frame. In this case, the MIN with multiple outlets which can pass through multiple packets for the same destination are advantageous.

2.2 The fault model

Like a common MIN, the MIN with multiple outlets consist of simple 2×2 or 4×4 switching elements. As shown in Figure 1, the controller attached to the path checks the header of the packet, and decides the mode of the element 'straight' or 'cross' appropriately by setting the multiplexors. Here, like most of fault model of the MIN [11][14], following two types of faults:

1. a broken link within the multiplexors and between the switching units (link fault),
2. and the malfunction of the controller and multiplexor (element fault)

are considered. The latter type of fault causes the stuck of the switching element and misrouting of packets while the former type causes the loss of packet which to be routed to the faulty link. Usually, the area of the controller is larger than other part of the switching element, the possibility of the element fault is larger than that of the link fault. In the control model treated here, the packet is serially transferred. Thus, the partial damage of the packet [15] is not treated.

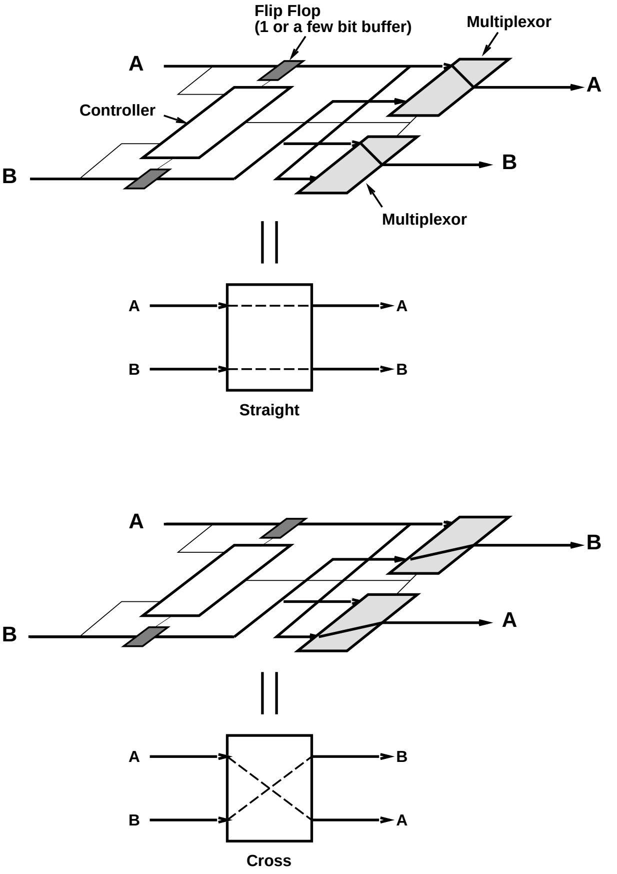

Figure 1: Switching element

3 Fault Tolerant Tandem Banyan Switching Fabrics

3.1 Tandem Banyan Switching Fabrics

Tandem Banyan Switching Fabrics (TBSF)[4][5] is an advanced MIN with multiple outlets proposed for the ATM-based packet switching system [1].

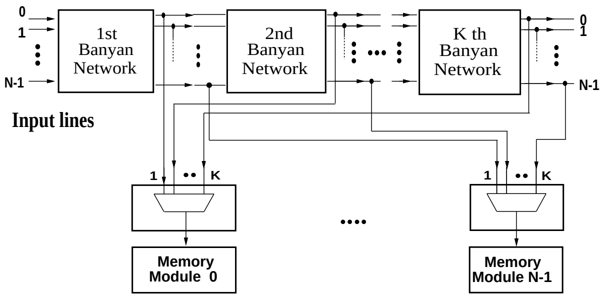

Figure 2: Tandem Banyan Switching Fabrics

As shown in Figure 2, it consists of placing multiple banyan interconnection networks in series such that, for each output of each banyan network, there is a connection feeding to the corresponding memory module, and a connection feeding the corresponding input of the following banyan network. At the end of a banyan network, all those packets which have succeeded in reaching their desired destination proceed to the output or memory modules. All the misrouted packets, after resetting their conflict bit, are fed to the next banyan network. The interface for the memory modules provides small packets buffer for each outlet of the banyan network.

Unlike the MBSF, only misrouted packets are rerouted in the next banyan network. That is, multiple banyan networks coordinately work for packet routing. The bandwidth is expected to be improved, while the latency is stretched. The 16-input/16-output TBSF is used in a real multiprocessor, and the performance is evaluated[13].

3.2 Fault recovery mechanism for the TBSF

A large number of fault tolerant MINs[16] have been proposed and discussed. These MINs provide the extra stages or extra links for fault recovery. However, since the TBSF consists of multiple MINs, it provides multiple paths and stages. Although this redundancy is introduced for performance improvement, the structure of this MIN can be used for fault recovery.

Element fault For the element fault, the on-the-fly recovery can be realized with a simple additional hardware. In the original TBSF, the conflict bit of the packet is checked at the outlet of each banyan network, and if the bit is set, the packet is routed to the next banyan network.

For the fault recovery, the comparator is attached to the outlet of each banyan network, and the destination address of the packet is compared with the label of the output link. If the destination address is

[1] This network was proposed by Tobagi and Kwok firstly in English 1990[3], but also was proposed independently by us and OKI Co. Ltd. in 1988 in Japanese paper [4].

not matched to the output label, the packet is routed to the next banyan network even if the conflict bit is not set. Using this mechanism, misrouted packets with element faults are routed to the next banyan network and get an opportunity to be routed correctly. Of course, the comparator for fault recovery also may be faulty. To cope with this problem, a packet whose destination address is matched to the output label but the conflict bit is set is also routed to the next banyan network. By using this double check mechanism, the mis-judged packets with malfunction of the comparator can be saved.

Link fault For the link fault, the on-the-fly fault recovery is difficult since packets are lost inside the MIN. In this case, the switching system is stopped and after diagnosis, the banyan network including faulty link is bypassed as shown in Figure 3. For this purpose, bypassing paths are necessary for each banyan network. Although this method can avoid the loss of packets, the bandwidth of the network is much degraded compared with the case of element fault. Since the TBSF consists of common omega networks, the common diagnosis methods[11][14] can be used for the fault location.

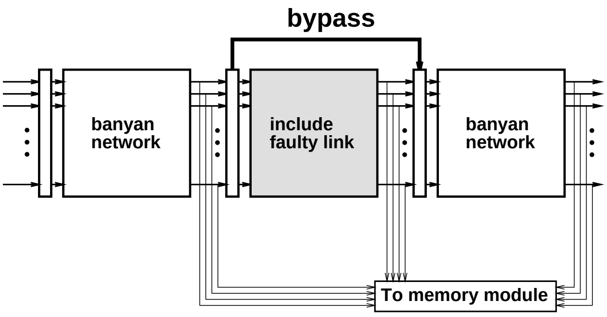

Figure 3: Bypassing path

We call the TBSF with both comparator and bypassing mechanism for fault recovery the Fault tolerant TBSF(F-TBSF).

4 Analysis of the throughput under the fault

Using the recovery hardware, the network can be available with single or multiple faults. However, in this case, the network throughput is degraded with faulty links or elements. Here, probabilistic models for analysis of the throughput (pass through ratio) of the F-TBSF with a single fault are proposed.

4.1 Analysis model of the F-TBSF

4.1.1 Analysis model of the fault-free TBSF

First, the state of each switching element is analyzed. Here, assume that a packet is inserted into an input of a switching element at a probability r. Then, the probability of the packets conflict is $0.25r^2$ because the conflict occurs only when packets are inserted from both inputs, and destinations of them are the same. When conflict occurs, a misrouted packet becomes a "dead packet", and never interfere other packets in the later stages. That is, these packets can be treated as being disappeared.

Therefore, when a packet is inserted into an input of a switching element in the i-th stage at a probability r_i, the probability for the next stage r_{i+1} is represented as follows:

$$\begin{aligned} r_{i+1} &= r_i - \frac{{r_i}^2}{4} \\ &= r_i\left(1 - \frac{r_i}{4}\right) \;=\; r_i f(r_i) \qquad (1) \end{aligned}$$

The probability that a packet arrives at the output of a single banyan network with n stages is represented as follows:

$$\begin{aligned} r_n &= r_0\left(1-\frac{r_0}{4}\right)\left(1-\frac{r_1}{4}\right)\left(1-\frac{r_2}{4}\right)\cdots\left(1-\frac{r_{n-1}}{4}\right) \\ &= r_0\prod_{j=0}^{n-1}\left(1-\frac{r_j}{4}\right) = r_0\prod_{j=0}^{n-1} f(r_j) \\ &= r_0 f^n(r_0) \qquad (2) \end{aligned}$$

where r_0 is the probability of the input packet for the MIN (That is, it is corresponding to the traffic load).

Thus, the pass-through ratio of this network is:

$$p_n = \frac{r_n}{r_0} = f^n(r_0). \qquad (3)$$

On the TBSF, correctly routed packets are sent for the destination, and traffic for the next banyan is reduced. Assuming that the input traffic for the k th banyan network is B_k, it is represented with the following equations:

$$B_k \;=\; B_{k-1} - B_{k-1}f^n(B_{k-1}) \qquad (4)$$

By solving these gradual equations, the pass-through ratio of the TBSF (PT_{TBSF}) can be represented as follows:

$$P_{TBSF} \;=\; \left(\sum_{k=1}^{l} B_k\right)/B_0 \qquad (5)$$

where let the number of whole banyan network is l[12].

4.1.2 Analysis model of the F-TBSF with faulty elements

A link fault causes the decrease of available banyan networks (thus, decreases l), and the damage on the throughput can be easily analyzed with the equation 5. Here, the throughput degradation caused by a faulty switching element is focused.

First, an element fault on a simple banyan network is considered. Assume that an element on the mth stage is faulty. Usually, the faulty element can not set the conflict packet even if the conflict occurs at the element. Thus, the number (thus the pass through ratio) of packets whose conflict bit is set is decreased by the faulty element. We refer this pretended pass through ratio as RA. Misrouted packets caused by the faulty element are detected at the outlets of the banyan network, and the real pass through ratio RF is degraded. The pass through ratio of the misrouted packet is referred as RM, and thus:

$$RF \quad = \quad RA - RM. \tag{6}$$

Calculation of the RA As shown in the Figure 4, the influence of a faulty element is propagated through the binary tree path.

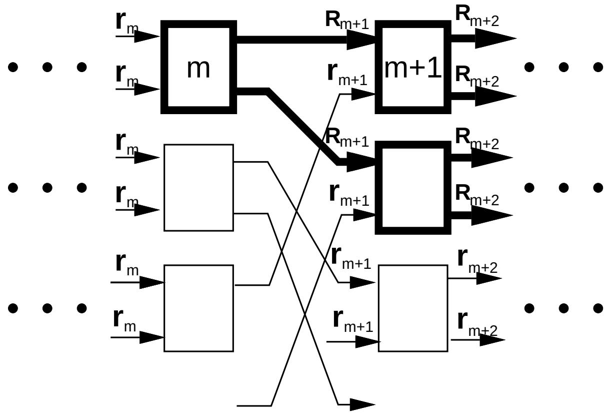

Figure 4: Influence of the faulty element

Here, the packet existing probabilities of the input link on the binary tree path are represented as follows:

- r_{i+1}: The packet existing probability for the input link on which no packet is passing through the fault-free element is inserted.
- R_{i+1}: The packet existing probability for the input link on which the packet is passing through the faulty element is inserted at $i + 1$ stage.

r_{i+1} is the same as the fault-free case calculated with the equation 1.

R_{i+1} is calculated with the input probability of r_i and R_i as shown in Figure 5.

Note that the faulty element on the m stage can not set the conflict bit, and the pretended pass through ratio on the stage m is 1, and thus, $R_{m+1} = r_m$.

The number of total switching element is 2^{n-1}, and the probability that an input of a switching element on the $m + 1$ stage may receive the packet which passed the faulty element is represented as $\frac{2^0}{2^{n-1}}$. Since the path on which packets are passing the faulty element forms the binary tree as shown in Figure 4, the probability that an input of a switching element receives the packet which passed the faulty element is represented as follows.

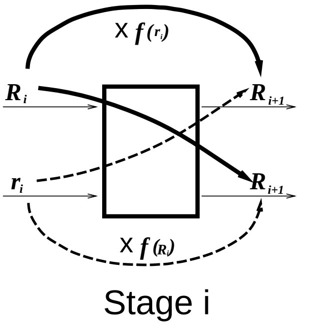

Figure 5: Consideration of R_{i+1}

$$\begin{aligned}
m + 1stage \quad &: \quad \frac{2^{n-1} - 2^0}{2^{n-1}} r_{m+1} + \frac{2^0}{2^{n-1}} R_{m+1} \\
m + 2stage \quad &: \quad \frac{2^{n-1} - 2^1}{2^{n-1}} r_{m+2} + \frac{2^1}{2^{n-1}} R_{m+2} \\
&\vdots
\end{aligned}$$

The pretended pass-through ratio RA is the probability at the output (stage $n+1$), and represented as follows.

$$\begin{aligned}
RA \quad &= \quad \frac{2^{n-1} - 2^{n-1-m}}{2^{n-1}} r_n + \frac{2^{n-1-m}}{2^{n-1}} R_n \\
&= \quad \left(1 - \frac{1}{2^m}\right) r_n + \frac{1}{2^m} R_n
\end{aligned} \tag{7}$$

Calculation of the RM Next, the probability that misrouted packets caused by the element fault are transferred output of the banyan network (RM) is calculated.

On the stage m, the probability of misrouting is as follows:

$$\frac{1}{2^{n-1}} r_m \times \frac{1}{2} \quad = \quad \frac{r_m}{2^n}$$

Since the path where misrouted packets are transferred forms the binary tree as shown in Figure 4, the probability that misrouted packets are existing on the input of the stage $m + 1$ is as follows:

$$\frac{r_m}{2^n} f(r_{m+1}) \times \frac{1}{2} \times 2 \quad = \quad \frac{r_m}{2^n} f(r_{m+1}).$$

Similarly, on the stage $m + 2$, the probability becomes:

$$\frac{r_m}{2^n} f(r_{m+1}) f(r_{m+2}) \times \frac{1}{2} \times 2 \quad = \quad \frac{r_m}{2^n} f(r_{m+1}) f(r_{m+2}).$$

RM is the probability on the output of the banyan network:

$$\begin{aligned} RM &= \frac{1}{2^n} r_m f(r_{m+1}) f(r_{m+2}) \cdots f(r_{n-2}) f(r_{n-1}) \\ &= \frac{1}{2^n} r_m \prod_{j=m+1}^{n-1} f(r_j). \qquad (8) \end{aligned}$$

From the equation 6, 7, and 8, the total pass through ratio of the banyan network in which an element on the stage m is faulty (RF) is represented as follows:

$$\begin{aligned} RF &= RA - RM \\ &= \left(1 - \frac{1}{2^m}\right) r_n + \frac{1}{2^m} R_n - \frac{1}{2^n} r_m \prod_{j=m+1}^{n-1} f(r_j) \end{aligned}$$

Like the pass through ratio of fault free TBSF (equation 5), the RF can be extended to the total banyan network just by replacing B_1 with RF.

$$F_{F-TBSF} = \left(RF + \sum_{k=2}^{l} B_k\right) / B_0 \qquad (9)$$

5 Evaluation of the throughput

Using proposed probabilistic model and computer simulation, the throughput (pass through ratio) of faulty F-TBSF/P-TBSF is analyzed.

5.1 F-TBSF

Table 1: Pass-through ratio vs. location of the fault on the TBSF (64 inputs, load:0.5, 2 banyan networks)

location of the fault	pass-through ratio	ratio (vs. no fault)
no fault	0.88050	1.00000
stage 0	0.87660	0.99557
stage 1	0.87666	0.99564
stage 2	0.87671	0.99570
stage 3	0.87675	0.99574
stage 4	0.87678	0.99578
stage 5	0.87681	0.99581

Single element fault Table 1 shows the relationship between the pass-through ratio and the location of the faulty element (64 × 64 single banyan network). The earlier the faulty element is located, the degradation of the path through ratio is large. However, the influence is not so large (under 1%). In the TBSF, the load of the first banyan network is maximum. Therefore, the influence of an element fault becomes maximum when the first-banyan stage-0 element is faulty.

Figure 6 shows the pass-through ratio versus input traffic load (r_0) with a single banyan network (64×64)

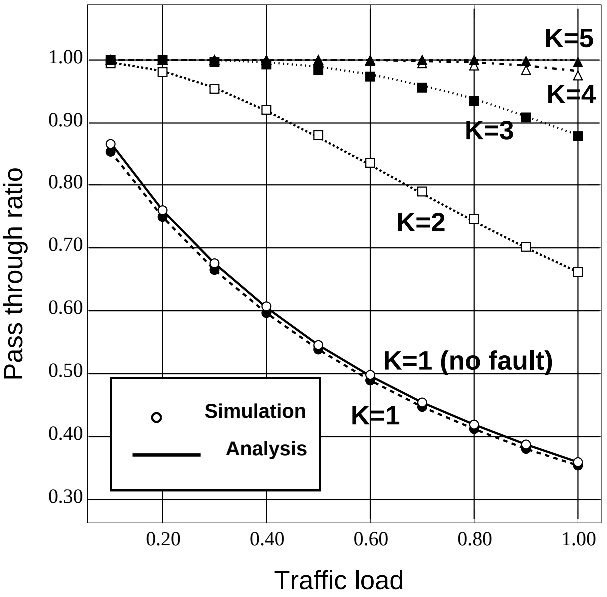

Figure 6: Pass-through ratio vs. traffic load on the F-TBSF (64 inputs , location of fault: 0 stage)

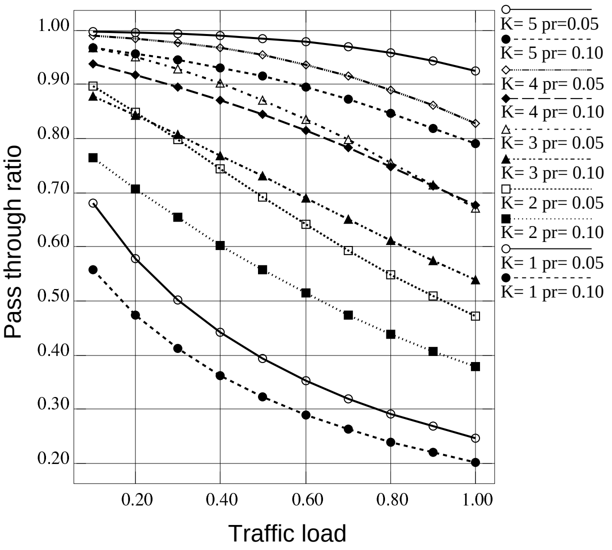

Figure 7: Pass-through ratio vs. traffic load on the F-TBSF (256 inputs ,fault probability: 0.05, 0.1)

when there is a faulty element at the first stage of the first banyan network. In this figure, the pass-through ratio of the banyan with the faulty element is only a few % lower than that of without fault even if the connected banyan network is one. From this figure, It also appeared that the difference of the faulty banyan and the fault free banyan does not changed when the traffic load is changed. Figure 6 also shows the result of computer simulation. The results from computer simulation is almost equal to those from analysis results.

Multiple elements fault Figure 7 shows the pass-through ratio versus input traffic load under the fixed network size(256×256) when the a switching element of the F-TBSF is faulty with a certain probability (0.05 and 0.1). Since the theoretical analysis is difficult under this assumption, this result comes from computer simulations. In this figure, the pass-through ratio of the banyan is strongly influenced by the difference of fault probability. But even in the severe situation such as the fault probability is 0.1 and traffic load is 1.0, the 80% packets can be saved with 5 banyan networks.

Link fault If there is a link fault, the banyan network must be bypassed. Thus, the number of connected banyan network (K) is decreased. As shown in Figure 6, the degradation of the pass through ratio is large especially with small K. For maintaining enough pass through ratio under the link fault, four or five banyan networks are required.

6 Conclusion

Fault recovery mechanisms are proposed for MINs with multiple outlets TBSF with the best use of their inherent fault tolerant capability. With these mechanisms, on-the-fly fault recovery is possible for multiple faults on switching elements. For the link fault, the networks are reconfigured after fault location, and the network is available with some performance degradation.

The bandwidth degradation under multiple fault on link/element is analyzed with theoretical models and simulation. Through the analysis, F-TBSF with 5 banyan networks can save 80% packets under 100% traffic load even when the fault probanility is 0.1.

The similar fault recovery and analysis method can be useful for another high bandwidth network PBSF. In this case, a hihger bandwidth can be maintained although a larger amount of recovery hardware is required.

References

[1] H.Amano, L.Zhou, K.Gaye, "SSS(Simple Serial Synchronized)-MIN: a novel multi stage interconnection architecture for multiprocessors," Proc. of the IFIP 12th World Computer Congress, Vol.I, pp.571-577, Sept. 1992.

[2] D.H.Lawrie, "Access and Alignment of Data in an Array Processor," IEEE Trans. on Comput. vol. c-24, No.12, Dec. 1975.

[3] F.A.Tobagi, "Fast Packet Switch Architectures For Broadband Integrated Services Digital Networks," Proceedings of the IEEE Vol.78, No.1 Jan. 1990.

[4] H.Sakamoto, T.Masaki, H.Inoue, H.Amano, "Configuration and evaluation of self routing switches," ISSE88-30 No.8, 1988, (in Japanese).

[5] F.A.Tobagi and T.Kwok, "The Tandem Banyan Switching Fabric: a Simple High-Performance Fast Packet Switch," Proc. INFOCOM91, Apr. 1991.

[6] C.L.Wu, M.Lee, "Performance Analysis of Multistage Interconnection Network Configurations and Operations," IEEE. Trans. Comput., Vol. 41, No.1 pp.18-27, Jan. 1992.

[7] C.T. Lea, "Multi-Log_2N networks and their applications in high-speed electronic and photonic switching systems," IEEE. Trans. Comm. Vol. 38, No. 10 pp.1740-1749, Oct. 1990.

[8] C.P. Kruskal, M. Snir, "The performance of multistage interconnection networks for multiprocessors," IEEE Trans. Comput. Vol.C-32, No.12, pp.1091-1098, Dec. 1983.

[9] M. Kumar, and J.R. Jump, "Performance of unbuffered shuffle-exchange networks," IEEE Trans. Comput. Vol.C-35, No.6, pp.573-577, Jun. 1986.

[10] R. Awdeh , H. Mouftah, "The Expanded Delta Fast Packet Switch ", IEEE International Conference Commun, (ICC) 1994.

[11] Tse-Yun Feng, "Fault Diagnosis for a Class of Multistage Interconnection Networks", IEEE Trans. on Computer C-30, 10, pp.351-366 (Oct. 1981).

[12] T. Hanawa, H.Amano, Y.Fujikawa, "Multistage Interconnection Networks with multiple outlets," Proc. of International Conference on Parallel Processing, Vol.I pp.1-8 (Aug. 1994).

[13] M.Sasahara, J.Terada, L.Zhou, K.Gaye, J.Yamato, S.Ogura, H.Amano, "SNAIL: a multiprocessor based on the Simple Serial Synchronized multistage interconnection network architecture," Proc. of International Conference on Parallel Processing, Vol.I pp.76-80 (Aug. 1994).

[14] N.J.Davis IV, W.T.Hsu, H.J.Siegel, "Fault location techiniques for Distributed Control Interconnection Networks," IEEE Trans. on Computer C-34, 10, pp.902-910 (Oct. 1985).

[15] A. Jajszczyk J,Tyszer, "Fault Diagnosis of Digital Switching Networks," IEEE Trans. on Communication, COM-34, 7, pp.732-739, (July 1989).

[16] G.B.Adams III, D.P.Agrawal, H.J.Siegel, "A Survey and COmparison of Fault Tolerant Multistage Interconnection Networks," IEEE Computer Vol.20, pp.14-27, (Jun. 1987).