PAPER

Fault tolerance of the TBSF (Tandem Banyan Switching Fabrics) and PBSF (Piled Banyan Switching Fabrics).

Akira Funahashi[†], **Toshihiro Hanawa**[†], *Nonmembers and* **Hideharu Amano**[†], *Member*

SUMMARY 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 two of such networks (TBSF/PBSF) 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 diagnosis, and the network is available with some performance degradation. The bandwidth degradation under multiple faults on link/element is analyzed with both theoretical models and simulation.

Through the analysis, F-PBSF shows high fault tolerance under high traffic load and low reliability by using 3 or more banyan networks.

key words: Multistage Interconnection Network(MIN), MIN with multiple outlets, fault tolerance

1. Introduction

Multistage Interconnection Networks (MIN) with multiple outlets[14] 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) [9][10][8][7] 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 outlets is Expanded Banyan Switching Fabrics (EBSF) or Expanded Delta network which realizes multiple outlets by expanding the size of network[2][12]. 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[14], 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)[5][3] and Piled Banyan Switching Fabrics (PBSF) [14] support much better performance than those of the MBSF and EBSF[14]. 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 these two networks, and proposed Fault tolerant TBSF (F-TBSF) and Fault tolerant PBSF (F-PBSF) respectively. Then, the performance degradation when some elements of the F-TBSF/F-PBSF are damaged is analyzed both with probablistics model and simulation.

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 destination 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][15].

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.

[†]Dept. of Computer Science, Keio University, 3-14-1 Hiyoshi, Kohoku-ku, Yokohama, 223 Japan

2.2 The fault model

Like a common MIN, the MIN with multiple outlets consists 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 [13][16], 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, since 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 [17] is not treated. While the permanent damage is assumed in the link fault, intermittent faults can be treated in the element fault. Multiple faults are treated in our fault tolerant scheme. Increasing number of faults does not cause the system down but causes the performance degradation.

3. Fault Tolerant Tandem Banyan Switching Fabrics

3.1 Tandem Banyan Switching Fabrics

Tandem Banyan Switching Fabrics (TBSF)[5][6] is an advanced MIN with multiple outlets proposed for the ATM-based packet switching system †.

As shown in Figure 2, it consists of placing multiple banyan interconnection networks in tandem 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 of 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.

†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 [5].

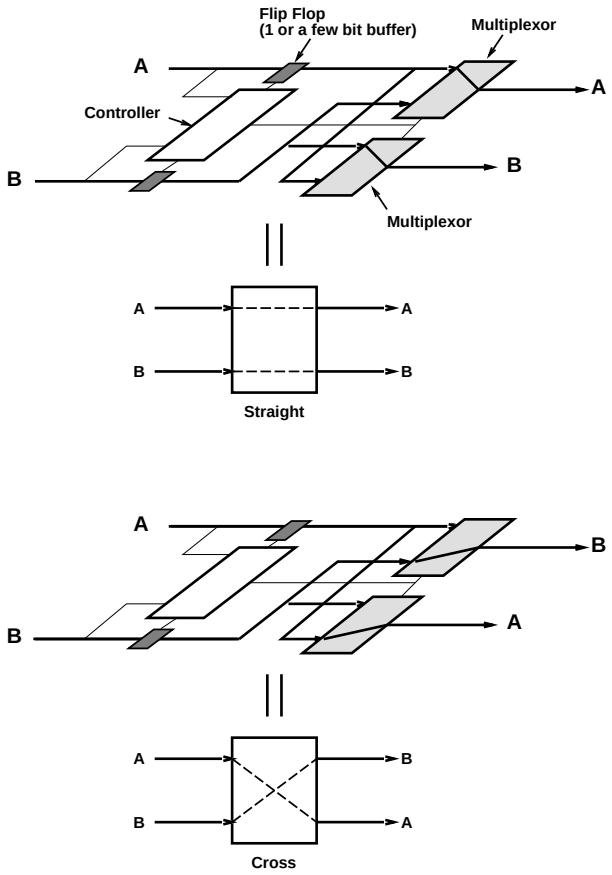

Fig. 1 Switching element

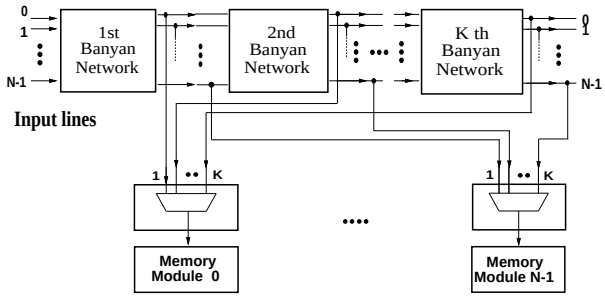

Fig. 2 Tandem Banyan Switching Fabrics

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[15].

3.2 Fault recovery mechanism for the TBSF

A large number of fault tolerant MINs [18] [21] [20] [22] [23] have been proposed and discussed. These MINs require the extra stages or links for fault recovery, and some of them provide independent multiple paths in the MIN[24][20] like MIN with multiple outlets treat-

ed here. The essential difference between the TBSF and such fault tolerant MINs is that the redundancy is introduced only for performance improvement in the TBSF. Therefore, the network topology of the TBSF is completely different from fault tolerant multipath networks. However, the redundancy in the TBSF can be used for fault recovery. Here we introduce simple and practical fault recovery methods by making the best use of the redundant structure of the TBSF.

(1) Element fault

For the element fault, the on-the-fly recovery can be realized with a simple additional hardware. In the original TBSF, only 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 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.

(2) 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 a fault 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[13][16] can be used for the fault location.

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

3.3 Piled Banyan Switching Fabrics

Another advanced MIN with multiple outlets is Piled Banyan Switching Fabrics (PBSF). Although the bandwidth of the TBSF is expected to be enough, the large latency will stretch the frame time of the MIN, thus, degrades the performance.

Unlike the TBSF which consists of tandem connection of networks, banyan networks are connected in

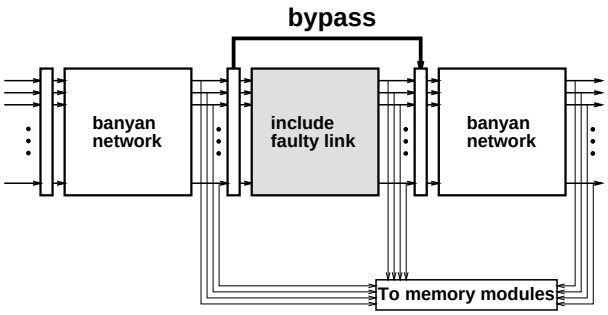

Fig. 3 Bypassing path

the three dimensional direction in the PBSF (Figure 4). A switching element except in the highest and lowest layers provides four inputs/outputs(two for horizontal, and two for vertical direction.).

The number of the layers in the PBSF can be changed considering the trade-off between the performance and hardware increase. The general structure of the PBSF is represented with $PBSF(n, l)$, where n is a number of the stage and l is the number of layers. The PBSF which has the same number of layers as the stages ($n = l$) is called the full-PBSF. Figure 4 shows the structure of full-PBSF with 8 inputs/outputs. Although this paper focuses on the full-PBSF, the extension to the non-full $PBSF(n, l) l < n$ is quite easy.

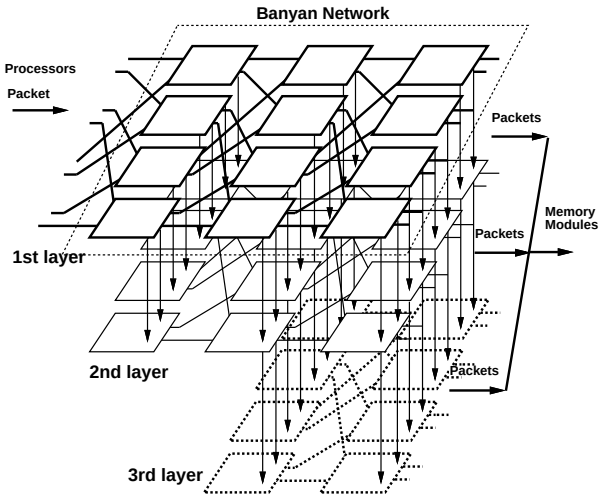

Fig. 4 Piled Banyan Switching Fabrics

Packets are inserted into the highest layer banyan network, and transferred to the horizontal direction. When two packets collide each other, a packet is fed to the corresponding switching element in the next lower layer banyan network with a clock delay. The vertically transferred packet may collide with both packets which are transferred to the horizontal direction also with a clock delay. In this case, one of horizontally transferred packets is selected and sent the next lower layer network

with a clock delay.

In the worst case, when three (one from the vertical, and two from the horizontal direction) packets for the same direction conflict in a switching element, a packet is routed to the correct direction, a packet is routed vertically, but the other packet cannot be routed to any direction. Like the TBSF, such a packet is routed to the incorrect direction, and treated as a "dead-packet".

3.4 Fault recovery mechanism for the PBSF

(3) Element fault

The on-the-fly recovery can be also realized in the PBSF for the element fault. However, unlike the TBSF, the check mechanism is required for each switching element.

As shown in Figure 5, the vertical links are prepared to the output links and input links of the lower layer. The checking mechanism checks the conflict bit of the packet header and compares the appropriate bit with the state of the switching element. If conflict bit is set, the packet is routed to the output links of the lower layer just like the normal PBSF. If the header bit is not matched to the state of the switching element, the packet is considered to be misrouted. So it is routed to the input links of the lower layer to be routed at the same stage. However, there is the possibility that the misrouted packet will conflict with horizontal input packet at lower layer. To avoid this problem, all input links to the lower layer are connected to the input links of the lowest layer of the stage because the lowest layer of the stage doesn't have horizontal inputs(Figure6).

Although this mechanism increases the amount of the hardware compared with the TBSF, packets are not misrouted unless both the controller and check mechanism are faulty.

(4) Link fault

In the PBSF, the faulty link can be bypassed with vertical links between layers. However, for an element, it is impossible to recognize the fault on the connected output links. Therefore, diagnosis and fault location are required like the TBSF. If the faulty link are located, the previous element is forced to use the vertical links to bypass the faulty link. The dedicated hardware or control packet are required to set the specific element to this "bypassing mode".

We also call the PBSF with this recovery mechanism the Fault tolerant PBSF(F-PBSF).

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

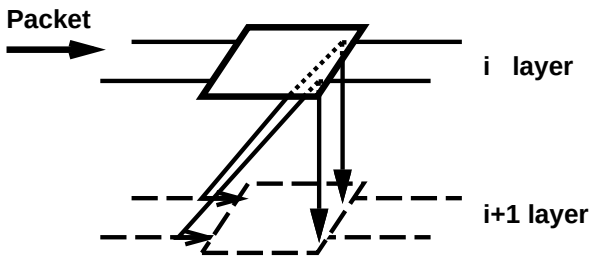

Fig. 5 F-PBSF switching element

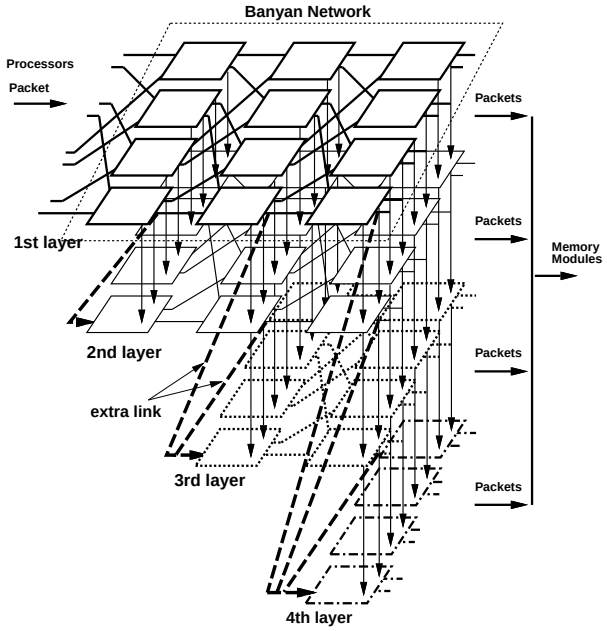

Fig. 6 F-PBSF

links or elements. Here, probabilistic models for analysis of the throughput (pass through ratio) of the F-TBSF with a single fault and F-PBSF with multiple faults 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 the network size is 2^n (thus, the number of stages is n), and r_i is the probability which packets are inserted at the i-th $(0 \leqq i \leqq n-1)$ stage input(Figure7). Figure 8 shows all the states of the switching element. To remark at the upper output, conflict occurs at one of four states, so the output probability is $r_i - \frac{{r_i}^2}{4}$. When conflict occurs, a misrouted packet becomes a "dead packet", and never interferes 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

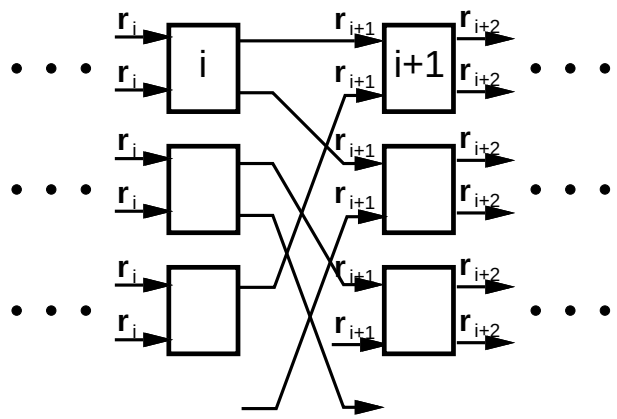

Fig. 7 Packet existing probability

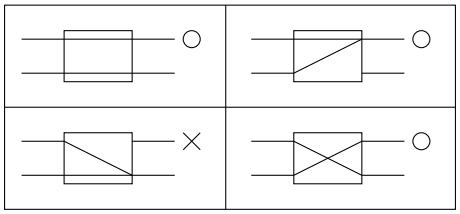

Fig. 8 States of the switching element

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:

$$r_{i+1} = r_i - \frac{{r_i}^2}{4} = r_i\left(1 - \frac{r_i}{4}\right) = r_i f(r_i) \quad (1)$$

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

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

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:

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

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 (thus, $B_0 = r_0$), it is represented with the following equations:

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

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

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

where let the number of whole banyan network be l[14].

4.1.2 Analysis model of the F-TBSF with faulty elements

A link fault causes the decrease of the number 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:

$$RF = RA - RM. \quad (6)$$

(1) Calculation of the RA

As shown in the Figure 9, the influence of a faulty element is propagated through the binary tree path.

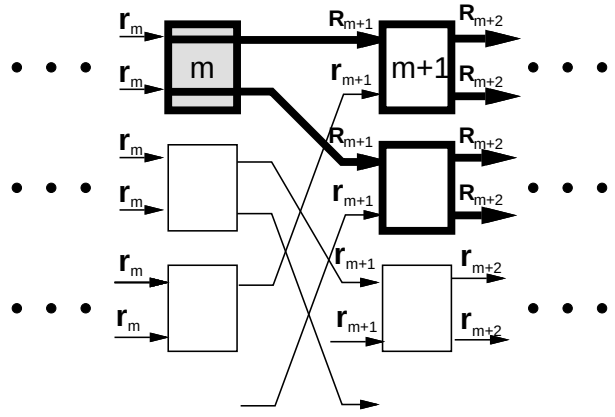

Fig. 9 Influence of the faulty element

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 at the $i+1$ stage's input link on which the packet is passing through the fault-free element.
- R_{i+1}: The packet existing probability at the $i+1$ stage's input link on which the packet is passing through the faulty element.

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 10, thus:

$$R_{i+1} = \frac{r_i f(R_i) + R_i f(r_i)}{2}. \tag{7}$$

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

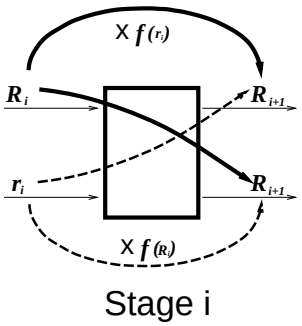

Fig. 10 Consideration of R_{i+1}

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 through 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 9, the probability that an input of a switching element receives the packet which passed the faulty element is represented as follows.

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

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

$$\begin{aligned} RA &= \frac{2^{n-1}-2^{n-1-m}}{2^{n-1}} r_n + \frac{2^{n-1-m}}{2^{n-1}} R_n \\ &= \left(1 - \frac{1}{2^m}\right) r_n + \frac{1}{2^m} R_n \end{aligned} \tag{8}$$

(2) 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:

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

Since the path where misrouted packets are transferred forms the binary tree as shown in Figure 9, the probability that misrouted packets are existing on the input of the stage $m+1$ is as follows:

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

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

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

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

$$\begin{aligned} RM &= \frac{1}{2^n} r_m f(r_{m+1}) f(r_{m+2}) \cdots f(r_{n-2}) f(r_{n-1}) \\ &= \frac{1}{2^n} r_m \prod_{j=m+1}^{n-1} f(r_j). \end{aligned} \tag{9}$$

From the equation 6, 8, and 9, 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.

$$P_{F-TBSF} = \left(RF + \sum_{k=2}^{l} B_k\right) / B_0 \tag{10}$$

4.2 Analysis model of the F-PBSF

4.2.1 Analysis model of the fault-free PBSF

In the first (top) layer of the PBSF, the probability that a packet is forwarded to the horizontal direction is the same as TBSF (because there is no vertical input packet), so it is represented with the equation 2.

In the PBSF, a packet is forwarded to the next layer only when the conflict occurs. Therefore, the probability that a packet is inserted into a switching element in the 2nd layer i-th stage from the vertical input ($d_{i,2}$) is represented as follows:

$$d_{i,2} = \frac{{r_i}^2}{4} \tag{11}$$

For the layer lower than the 2, there are horizontal input packet and vertical input packet. All possible combinations of input/output packets are shown in Table 1. In this table, 0/1 shows that the destination of the input packets in the switching element are upper output (0) or lower output (1) respectively. If a packet is actually reaches to the output link, O is marked. Otherwise, X is marked. From this table, the probability that a packet is forwarded to the horizontal/vertical direction is represented in the Table 2.

From Table 2, the packet existing probability at a horizontal output link (r_i) and vertical output link ($d_{i,j}$) of a jth layer ith stage element are calculated as follows for each combinations of input packets.

$$\Sigma(\text{input probability}) \times (\text{output probability})$$

So, r_i and $d_{i,j}$ are represented as follows respectively.

$$r_i = r_{i-1} - \frac{1}{4}r_{i-1}^2 + \frac{1}{2}d_{i,j-1}\left(1 - r_{i-1} + \frac{1}{4}r_{i-1}^2\right) \tag{12}$$

$$d_{i,j} = d_{i,j-1}r_{i-1} + (1 - d_{i,j-1})\, r_{i-1}^2 \tag{13}$$

4.2.2 Analysis model of the faulty F-PBSF

Unlike the TBSF, the packet is transferred to the lower layer both with the link fault and element fault. If there is a fault in an element/link, the combination of output packets are modified as shown in Table 1. In this case, the probability that a packet is forwarded to the horizontal/vertical direction is represented in the Table 2.

From Table 2, the packet existing probability at a horizontal output link (r_i), vertical output link ($d_{i,j}$) and vertical output link to lowest layer's input link ($s_{i,j}$) of a jth layer ith stage element on condition that each switching element or link will be faulty at probability f, are modified as follows respectively.

$$\begin{aligned} r_i = {} & \frac{1}{2}f(1 - d_{i,j-1})r_{i-1}(1 - r_{i-1}) \\ & + \frac{1}{2}f(1 - d_{i,j-1})r_{i-1}^2 \\ & + \frac{1}{2}f d_{i,j-1}(1 - r_{i-1})^2 + \frac{3}{4}f d_{i,j-1}r_{i-1}^2 \\ & + \frac{5}{4}f d_{i,j-1}r_{i-1}(1 - r_{i-1}) \\ & + (1 - f)r_{i-1} - \frac{1}{4}(1 - f)r_{i-1}^2 \\ & + \frac{1}{2}(1 - f)d_{i,j-1}\left(1 - r_{i-1} + \frac{1}{4}r_{i-1}^2\right) \end{aligned} \tag{14}$$

$$\begin{aligned} d_{i,j} = {} & \frac{1}{2}f d_{i,j-1}r_{i-1}(1 - r_{i-1}) + f d_{i,j-1}r_{i-1}^2 \\ & + \frac{1}{2}(1 - f)d_{i,j-1}r_{i-1}^2 \\ & + \frac{1}{2}(1 - f)(1 - d_{i,j-1})r_{i-1}(1 - r_{i-1}) \\ & + \frac{1}{2}(1 - f)\,(1 - d_{i,j-1})\,r_{i-1}^2 \end{aligned} \tag{15}$$

$$s_{i,j} = \frac{1}{2}f r_{i-1}(1 - r_{i-1}) + \frac{1}{2}f r_{i-1}^2 \tag{16}$$

Applying these equations from the first stage of the first layer to every output, the packet existing probability of the F-PBSF can be analyzed.

5. Evaluation of the throughput

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

5.1 F-TBSF

Table 3 Pass-through ratio vs. location of the fault on the F-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

(1) Single element fault

Table 3 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 larger is the degradation of the path through ratio. 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 11 shows the pass-through ratio versus input traffic load (r_0) with a single banyan network (64×64) 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

Table 1 All input/output combination in a switching element (F-PBSF)

combination			no fault			fault				
vertical input	horizontal input		horizontal output		vertical output	horizontal output		vertical output	lowest layer output	
d	r_0	r_1	0	1		0	1		0	1
-	-	-	×	×	×	×	×	×	×	×
-	0	-	○	×	×	○	×	×	×	×
-	1	-	×	○	×	×	×	×	○	×
-	-	0	○	×	×	×	×	×	×	○
-	-	1	×	○	×	×	○	×	×	×
-	0	0	○	×	○	○	×	×	×	○
-	0	1	○	○	×	○	○	×	×	×
-	1	0	○	○	×	×	×	×	○	○
-	1	1	×	○	○	×	○	×	○	×
0	-	-	○	×	×	○	×	×	×	×
0	0	-	○	×	○	○	×	○	×	×
0	1	-	○	○	×	○	×	×	○	×
0	-	0	○	×	○	○	×	×	×	○
0	-	1	○	○	×	○	○	×	×	×
0	0	0	○	×	○	○	×	○	×	○
0	0	1	○	○	○	○	○	○	×	×
0	1	0	○	○	○	○	×	×	○	○
0	1	1	○	○	○	○	○	×	○	×
1	-	-	×	○	×	×	○	×	×	×
1	0	-	○	○	×	○	○	×	×	×
1	1	-	×	○	○	×	○	×	○	×
1	-	0	○	○	×	×	○	×	×	○
1	-	1	×	○	○	×	○	○	×	×
1	0	0	○	○	○	○	○	×	×	○
1	0	1	○	○	○	○	○	○	×	×
1	1	0	○	○	○	×	○	×	○	○
1	1	1	×	○	○	×	○	○	○	×

Table 2 Packet output rate of a switching element (F-PBSF)

combination			packet existing probability at input link	no fault		fault		
vertical input	horizontal input			horizontal output	vertical output	horizontal output	vertical output	lowest layer output
$1-d$	$1-r$	$1-r$	$(1-d)(1-r)^2$	0	0	0	0	0
$1-d$	$1-r$	r	$(1-d)r(1-r)$	0.5	0	0.25	0	0.25
$1-d$	r	$1-r$	$(1-d)r(1-r)$	0.5	0	0.25	0	0.25
$1-d$	r	r	$(1-d)r^2$	0.75	0.5	0.5	0	0.5
d	$1-r$	$1-r$	$d(1-r)^2$	0.5	0	0.5	0	0
d	$1-r$	r	$dr(1-r)$	0.75	0.5	0.625	0.25	0.25
d	r	$1-r$	$dr(1-r)$	0.75	0.5	0.625	0.25	0.25
d	r	r	dr^2	0.875	1.0	0.75	0.5	0.5

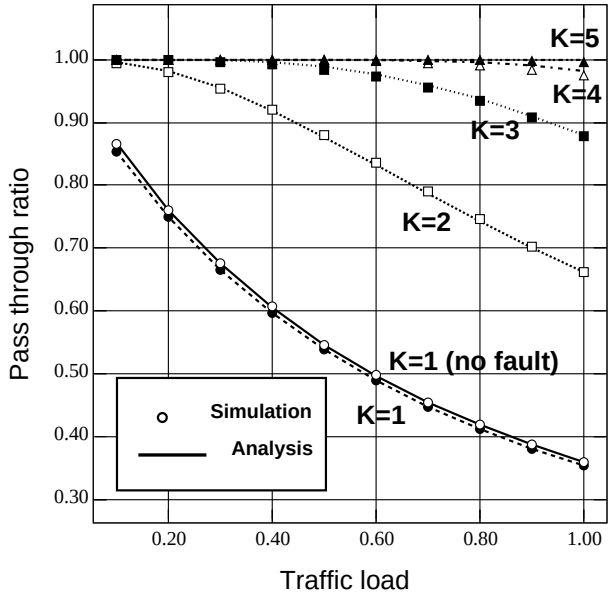

Fig. 11 Pass-through ratio vs. traffic load on the F-TBSF (64 inputs , location of fault: 0 stage)

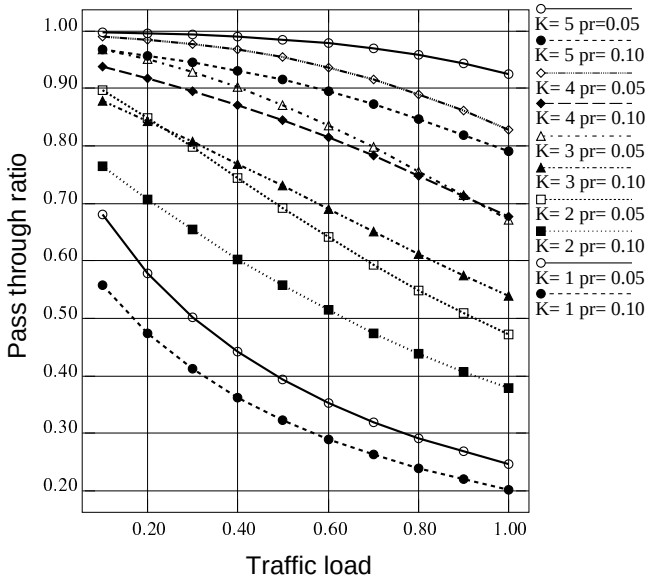

Fig. 12 Pass-through ratio vs. traffic load on the F-TBSF (256 inputs ,fault probability: 0.05, 0.1)

figure, it also appeared that the difference of the faulty banyan and the fault free banyan does not change when the traffic load is changed. Figure 11 also shows the result of computer simulation. The results from computer simulation are almost equal to those from analysis results.

(2) Multiple elements fault

Figure 12 shows the pass-through ratio versus input traffic load under the fixed network size(256×256) when the 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.

(3) Link fault

If there is a link fault, the banyan network must be bypassed. Thus, the number of connected banyan networks (K) is decreased. As shown in Figure 11, 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.

5.2 F-PBSF

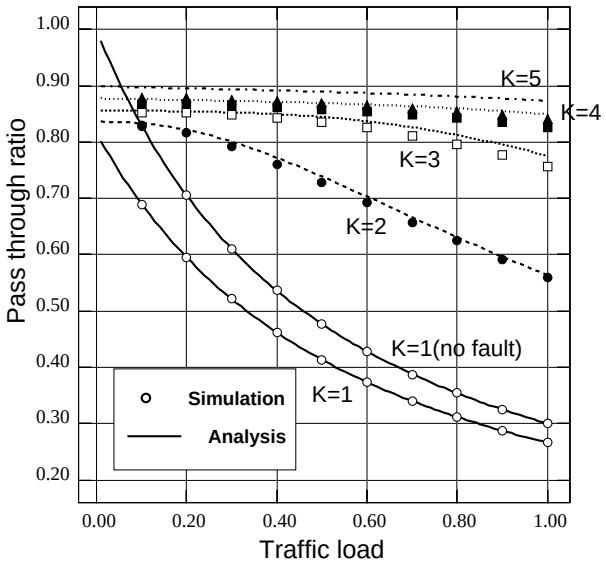

Fig. 13 Pass-through ratio vs. traffic load on the F-PBSF (256 inputs , fault probability: 0.05)

Unlike the F-TBSF, the fault recovery mechanism is attached to each switching element of the F-PBSF. Thus, packets are transferred to the lower layer both with the case of link fault and element fault. Only difference is that the on-the-fly fault recovery is possible for the element fault while the diagnosis and setting bypass mode is required for the link fault. Therefore, the throughput degradation of both cases is the same.

Figure 13 shows the pass-through ratio versus input traffic load under the fixed network size(256×256)

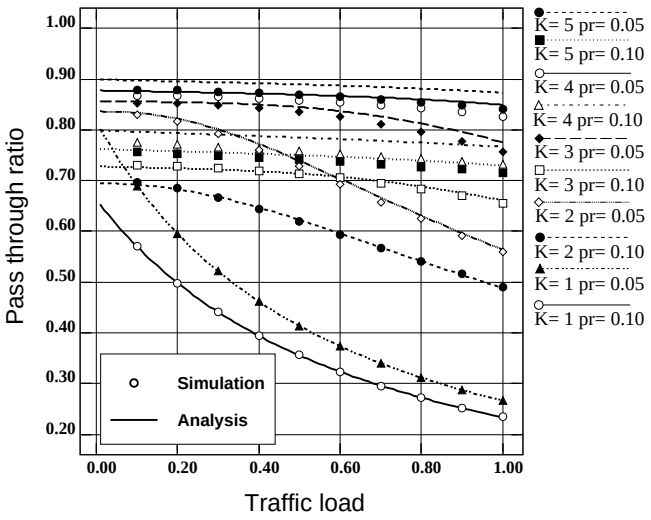

Fig. 14 Pass-through ratio vs. traffic load on the F-PBSF (256 inputs , fault probability: 0.05, 0.10)

when the fault probability is 0.05. If the number of layers is larger than 3, there are a little error between the results of the probabilistic models and those of the simulation. In the proposed probabilistic models, input load traffic for all banyan networks are assumed to be uniform. However, in the PBSF, the traffic fed to a banyan network may not be uniform since the previous banyan network may disturb the traffic uniformity. This is the reason why there are a little error between the probabilistic models and simulation. However, this error is 5% in maximum, and this demonstrates that the results from the probabilistic models are useful in most cases.

From this figure, the pass-through ratio of first banyan network includes 5% faulty element is 4 to 18% lower than that of fault free banyan network. And if the number of layers K is increased, the pass-through ratio will increase from 28% to 55% when K=2, and to 76% when K=3. This is higher than that of the F-TBSF.

Figure 14 shows the pass-through ratio versus input traffic load like Figure 12 when the fault probability is 0.05 and 0.10. In this figure, the pass-through ratio of the banyan can keep more than 65% in both reliabilities even traffic load is 1.0 by using 3 banyan networks. This figure also shows that the influence of reliability will decrease by using more than 3 banyan networks.

Figure 12 and Figure 14 shows that the pass-through ratio of the F-PBSF is up to 10% larger than that of the F-TBSF if the connected banyan networks are small (2 or 3). In the PBSF, the routing before conflict is effective since the conflicting packet is fed to the element in the same position of the next layer. On the other hand, the conflicting packet must be routed from the first stage in the TBSF. This is the reason why the pass-through ratio of the F-PBSF is better than that of the F-TBSF.

6. Conclusion

Fault recovery mechanisms are proposed for MINs with multiple outlets TBSF/PBSF 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 faults on link/element is analyzed with theoretical models and simulation. Through the analysis, the F-PBSF shows high fault tolerance under high traffic load and low reliability by using 3 or more banyan networks.

In this paper, only uniform traffic is assumed, that is, the destination of all packets is assumed to be distributed uniformly to all memory module. However, irregular traffic including hot spots should be considered for more practical evaluation. The analysis model treated here can be extended for traffic including a hot spot[4][11]. However, since the existence of faults also disturbs the traffic, the theoretical analysis under the irregular traffic is difficult. The extensive simulation study including hot spots and other irregular traffic is the future work.

The TBSF chip has been implemented and now utilized in a real multiprocessor [15], and the implementation of the PBSF chip was also finished[19]. Through these implementations, the number of gates or areas of each switching element can be calculated. We will structure precise fault models based on these real hardwares.

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] K. Gaye, T. Hanawa, H. Amano, "An Analysis of the Hot Spot Contention and Message Combining on the SSS-MIN" IEICE Trans. inf. & syst. Vol.J77-D-I, No.5 pp.354-364, May. 1994.

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

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

[7] 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.

[8] C.T. Lea, "Multi-$Log_2 N$ networks and their applications in high-speed electronic and photonic switching systems,"

IEEE. Trans. Comm. Vol. 38, No. 10 pp.1740-1749, Oct. 1990.

[9] 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.

[10] 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.

[11] H. Hanawa, J. Terada, H. Yasukawa, T. Kamei, H. Amano, "An Analysis of Message Combining on the SSS-MIN", Technical Report of IEICE. CPSY 95-49, Aug. 1995. (in Japanese)

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

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

[14] 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).

[15] 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).

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

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

[18] 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).

[19] T. Kamei, M. Sasahara, H. Amano, "An LSI Implementation of a high speed multi stage interconnection network," Proc. of the 5th workshop of Synthesis And System Integration of MIxed Technologies (SASIMI95), pp.199-206, (Aug. 1995).

[20] F. T. Chong, T. F. Knight,Jr, "Design and Performance of Multipath MIN Architectures," SPAA '92. 4th Annual ACM Symposium on Parallel Algorithms and Architectures, pp.286-295, 1992.

[21] J. T. Blake, K. S. Trivedi, "Multistage interconnection network reliability," IEEE Trans. Computers, Vol.38, pp.1600-1604, (Nov. 1989).

[22] J. H. Park, H. K. Lee, J. H. Cho, "Ring-Banyan Network: a fault tolerant multistage interconnection network and its fault diagnosis," Dependable Computing - EDCC-1. First European Dependable Computing Conference Proceedings, pp.511-528, 1994.

[23] P. K. Bansal, K. Singh, R. C. Joshi "Routing and path length algorithm for a cost-effective four-tree multistage interconnection network," INT. J. Electronics, Vol. 73, No. 1, pp.107-115, 1994.

[24] S. M. Reddy, V. P. Kumar, "On Fault Tolerant Multistage Interconnection Networks," 1984 Int'l Conf. Parallel Processing, Computer Society Press, Silver Spring, Md., 1984, pp. 155-164.

Akira Funahashi received the B.E. from Keio University, Japan, in 1995. He is a Master course student in the Department of Computer Science, Keio University, Japan. His research interests include interconnection network and fault tolerant network.

Toshihiro Hanawa received the B.E., M.E. from Keio University, Japan, in 1993, 1995. He is a Ph.D. candidate in the Department of Computer Science, Keio University, Japan. His research interests include analysis of interconnection network.

Hideharu Amano received the B.E., M.E., and Ph.D. degrees from Keio University, Japan, in 1981, 1983, and 1986, respectively. He is now an associate professor in the Department of Electrical Engineering, Keio University. His research interests include the area of parallel processing.