Multistage Interconnection Networks with multiple outlets

Toshihiro Hanawa, Hideharu Amano†, Yoshifumi Fujikawa*
†Dept. of Computer Science
Keio University
3-14-1, Hiyoshi Yokohama 223 JAPAN
E-mail: {hanawa, hunga}@aa.cs.keio.ac.jp

Abstract: Multistage Interconnection Networks(MINs) with multiple outlets are networks which can support higher bandwidth than that of nonblocking networks by passing multiple packets to the same destination. A novel MIN topology with multiple outlets called Piled Banyan Switching Fabrics (PBSF) is proposed for the Simple Serial Synchronized (SSS)-MIN used in multiprocessors, and analyzed with other two types of MIN with multiple outlets called Multi-Banyan Switching Fabrics (MBSF) and Tandem Banyan Switching Fabrics (TBSF). The throughput of these MINs is evaluated and compared with both the theoretical model and simulation. The PBSF supports the best throughput and latency used for the SSS-MIN. Although the latency of the TBSF is large, the pass-through ratio is close to 1 if the number of connected banyan networks are more than 4. Therefore, the TBSF is useful for the ATM switching networks in which the relatively large latency is tolerable. The con ict-free access of these MINs is also analyzed, and it appears that rows, column, forward and backward diagonal of the matrix can be accessed without con ict.

1 Introduction

Multistage Interconnection Networks (MINs)[1] have been well researched as an interconnection mechanism between processors and memory modules of parallel machines. Most of them are blocking networks like the omega network[1], and packets are transferred in the store-and-forward manner between switching elements with bit-parallel (8-64bits) lines. In blocking networks, packets may collide with each other in the network even if their destinations are different. In this case, one of the con icting packets is routed for desired destination, another packet is stored in the packet buffer.

However, implementation of large scale MINs have been a cumbersome job because of their complicated structure and pin-limitation problem. For the store-and-forward operation, buffer storage and the control mechanism are required in each switching element. The pin-limitation problem caused by bit-parallel communication makes the high density implementation dif cult.

As an alternative approach, a novel Multistage Interconnection Network (MIN) architecture called Simple Serial Synchronized (SSS)-MIN[2] was proposed. All packets are transferred in the serial(a few parallel) and synchronized manner. Synchronized bit-serial communication simpli es the structure/control, and also solves the pin-limitation problem.

Although this architecture has simple structure, the internal packet con icts will severely degrade the performance in the SSS-MIN. Defeated packets by con ict must be inserted into the MIN again since there is no packet buffer inside the switching element. To cope with this problem, networks with large bandwidth are required for the SSS-MIN. The bandwidth of non-blocking networks[1] like the crossbar or Clos[3] is not enough since packets for the same destination still collide with each other inside the network.

Multistage Interconnection Networks with multiple outlets are networks which can support higher bandwidth than those of nonblocking networks by passing multiple packets to the same destination. In this paper, a novel MIN structure with multiple outlets called Piled Banyan Switching Fabrics (PBSF) is proposed and analyzed with other three types of MIN with multiple outlets called Multi Banyan Switching Fabrics (MBSF) [4][5] and Tandem Banyan Switching Fabrics (TBSF) [6][7] both of which were proposed for telecommunication switching systems, and Expanded Banyan Switching Fabrics(EBSF)[8].

2 The SSS-MIN

2.1 Basic operation of the SSS-MIN

The basic operation of the SSS-MIN is illustrated in Figure 1. Like the Asynchronous Transfer Mode(ATM)-based packet switching systems for telecommunication, all packets are inserted into the SSS-MIN serially (in a few bits parallel) synchronized with a common frame clock from input packet buffers prepared between processors and the MIN. Since each switching element stores only one bit (or a few bits) of the packet, the SSS-MIN behaves like a set of shift registers with the switching capability. After the delay for passing through all stages, head of packets come out at the outlets of the MIN.

When a con ict occurs, one of the con icting packets must be routed to the incorrect direction since the SSS-MIN provides no packets buffer in each switching element , and the con ict bit in the routing tag is set. The packet that set con ict bit is treated as a dead packet, and never interferes the other packets.

When the top of the address packet reaches at the outlet of the MIN, it can be judged whether the packet is correctly routed or not by referring to the con ict bit. An acknowledge signal (ACK: ACKnowledge or NAK: Not AcKnowledge) is returned to the processor which issues the access request.

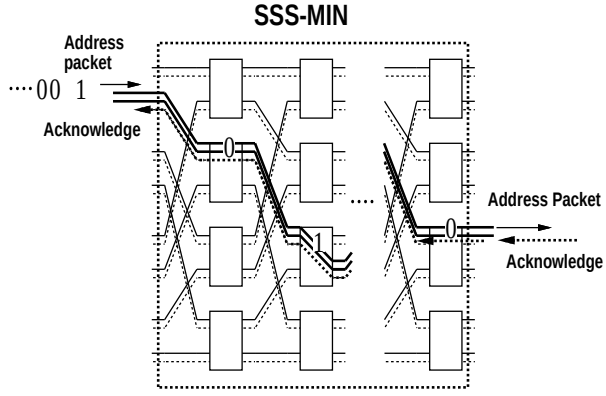

Figure 1: Structure of the SSS-MIN

All switching elements are set to be straight or exchange when an address packet reaches to the outlet. That is, a logical path is formed independently from the input to the output of the MIN. Here, this logical path is called the *trace*. The ACK/NAK signal is returned through this trace without being stored at each switching element,

*Microelectronics Products Development Lab., Hitachi, Ltd.

but only with the delay of wires and multiplexors. If the NAK signal is returned, the packet is inserted again from the input buffer synchronized with the next frame signal.

2.2 Pipelined circuit switching

The trace is utilized not only for the ACK signal but also for the transfer of data packets. The address transfer, ACK signal and data transfer can be performed in the overlapped manner. In the frame i, an address packet is transferred. When the head of the address packet arrives at the outlet of MIN, the trace is established and the ACK signal is transferred immediately. If the ACK signal is returned, the data packet is also transferred using the same trace (logical path) in the frame $i + 1$. If the pass through time of the MIN is large, the pipeline can be formed for these transfers. We call this operation overlapped or pipelined circuit switching.

In practice, the depth of the pipeline is only two since the packet passing through time is reduced by the stage hopping mechanism described later. In this case, memory reference & data transfer is overlapped with the next address transfer.

By sharing the trace, the structure of the SSS-MIN becomes simple unlike the conventional method which requires two independent MINs for bi-directional transfer.

To allow the overlapped/pipelined operation, address packet, data packet, and ACK/NAK signal have their own signal lines and multiplexors in a switching element. Since the ACK/NAK signals and data packet share logically the same trace which is set by the address packet, only one controller is required on the address path.

2.3 Other issues on the SSS-MIN

In order to enhance the performance of the SSS-MIN, the following two techniques are introduced:

- Bit serial message combining.
  Since packets are serially inserted in the synchronized manner, it is easy to judge whether two packets can be combined in the switching element. If two addresses are equal, the switching element is set to be the backward broadcast mode. After these address packets passed by, a tree-like trace remains in switching elements, and the returning data packets are multicast through the tree. This technique is called the *bit-serial message combining*[9], and implemented with only 20% increases of the hardware.

- Stage hopping.
  Once the trace is established, the rest of the address packet and the data packet pass through the same logical path. Therefore, the bit of a packet does not have to be stored in every switching element. That is, the ip- op provided in a switching element can be omitted in most stages, and packets is transferred without storing in these switching element. The latency is much reduced by this mechanism called the *stage hopping*.

Using these techniques, almost the same performance can be achieved with only 1/8 hardware or chip numbers compared with conventional MIN[2]. Now, a $1.0\mu m$ CMOS SOG LSI chip of the SSS-omega network with 16 inputs/outputs was developed[1], and works at 50MHz clock in a multiprocessor prototype called SNAIL[10].

3 MIN with multiple outlets

The SSS-omega network with 16 inputs/outputs supports suf cient performance for a relatively small scale multiprocessor with 16 processors and memory modules[10]. However, for a large size of the multiprocessor with hundreds of processors, the performance of a simple blocking network like the omega network is not enough. Since one of the con icting packets must be inserted again in the next frame, packet con icts inside the MIN will severely degrade performance of the SSS-MIN.

Even non-blocking networks like the crossbar, Clos[3], Batcher-banyan[11] and BDOC[12] can not support enough bandwidth because packets for the same destination collide each other inside the network. By this type of con ict, the pass-through ratio (bandwidth) of packets for random destinations is limited smaller than about 63.5%. In order to obtain large pass-through ratio, the MIN is required to provide multiple outlets for passing multiple packets for the same destination. Here, three MINs with this property are introduced.

3.1 Multi Banyan Switching Fabrics

Multi Banyan Switching Fabrics (MBSF) is the simplest MIN with multiple outlets. As shown in Figure 2, input traf c is distributed into independent banyan (omega) networks. It has been studied as a circuit switching network or the ATM packet switching networks because of its simple structure[13][14][4][5]. As a circuit switching network, the way of the load distribution sometimes in uences the performance [4][5]. However, as a packet switching network, input traf c is randomly distributed since the dynamic scheduling is impossible.

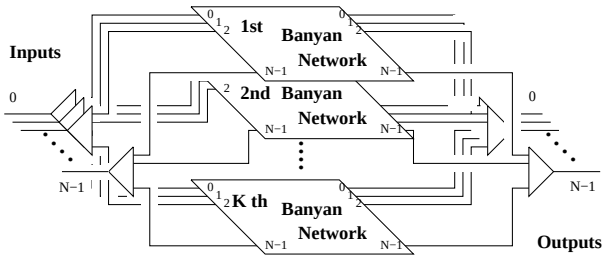

Figure 2: Multi Banyan Switching Fabrics

3.2 Expanded Banyan Switching Fabrics

Multiple outlets can be also achieved by expanding the size of the MIN[8]. If the size of the MIN is expanded three times as shown in Figure 3, three outlets are useful. We call this network the Expanded Banyan Switching Fabrics [2]. The con ict free access of the double sized EBSF is also analyzed by Lawrie[15].

3.3 Tandem Banyan Switching Fabrics

Tandem Banyan Switching Fabrics (TBSF)[6][16] is a another network with multiple outlets proposed for the ATM-based packet switching system [3].

As shown in Figure 4, 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 memory modules. All the misrouted packets, after resetting their con ict 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.

A problem of the TBSF used in the ATM packet switching is that the arriving time of packets is different depending on the number of the banyan networks from which the packet is fed. When it is used in the SSS-MIN, this problem changes to an advantage. Memory

[1] This chip was fabricated with Kawasaki Steel Co. Ltd.

[2] In [8], this network is called the Expanded Delta network.

[3] This network was proposed by Tobagi and Kwok rstly proposed in English 1990[7], but also was proposed independently by us and OKI Co. Ltd. in 1988 in Japanese paper [6].

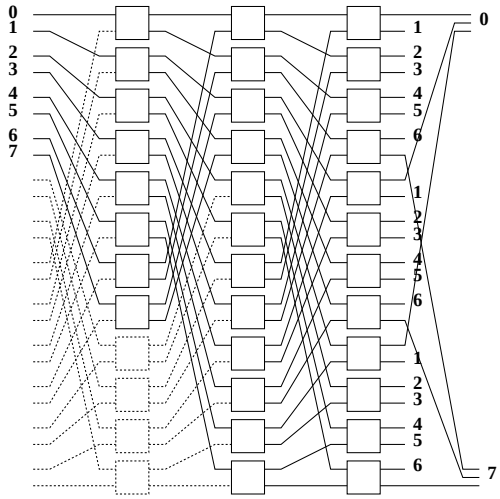

Figure 3: Expanded Banyan Switching Fabrics

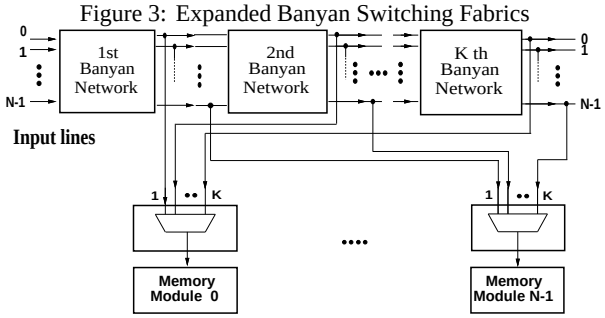

Figure 4: Tandem Banyan Switching Fabrics

module can be accessed as soon as a packet arrives at the module, and the memory utilization ratio is improved.

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.

3.4 Piled Banyan Switching Fabrics

The fourth network with multiple outlets is a novel network proposed here. Although the bandwidth of the TBSF is expected to be enough, the large latency will stretch the frame time of the SSS-MIN, thus, degrades the performance.

In order to reduce the latency with maintaining the high throughput, a new network topology called Piled Banyan Switching Fabrics (PBSF) is proposed. Unlike the TBSF which consists of tandem connection of networks, banyan networks are connected in the three dimensional direction in the PBSF (Figure 5). A switching element except in the highest and lowest layers provides four inputs/outputs(two for horizontal, and two for vertical direction.).

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. With several clocks delay corresponding to the stage, packets in the highest layer network reach to the outlet, then packets in the nth layer reach to the outlet with $n-1$ clock delay.

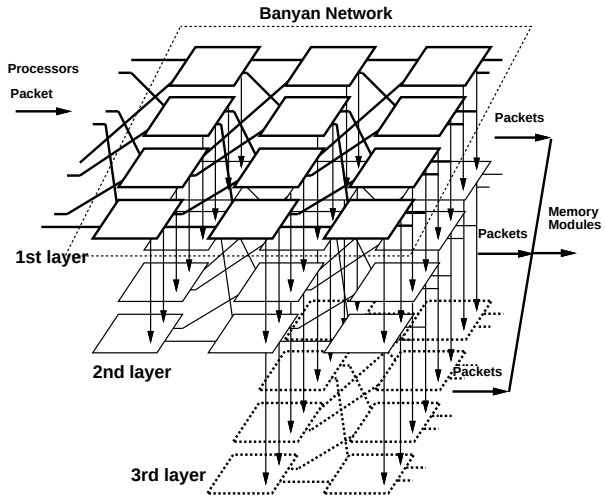

Figure 5: Piled Banyan Switching Fabrics

When three (one from the vertical, and two from the horizontal direction) packets for the same direction con ict 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". On the lowest layer, such a packet is discarded. Only the processor (or the interface units between processors) which issued the correctly routed packets receives the ACK signal through the trace. Other discarded packets are inserted again in the next frame from the input buffer. Like the TBSF, small packets buffer is provided for each outlets of the banyan network. When the buffer is full, the arrived packet is regarded as a discarded packet, and inserted again.

3.5 Latency and required hardware

Before quantitative analysis of the throughput, other factors (latency and required hardware) of three networks are compared.

Latency: The latency of the MBSF, EBSF, PBSF and TBSF are log_2N, log_2N, $K-1+log_2N$ and $K \times log_2N$, respectively where let N be the size of the network and K be the number of utilized banyan networks. For the large N and small K, the latency of the MBSF and PBSF are almost the $\frac{1}{K}$ as that of the TBSF.

Required hardware: Although the latency of the PBSF is better than those of the MBSF, EBSF and TBSF, the required hardware is increased since switching elements with 4 inputs and 4 outputs are required. However, the structure of the switching element is simpler than 4×4 crossbar because vertical wires are provided for the output of the switching element (Figure 6). A switching element used in the PBSF can be divided into two switching elements used in the MBSF/EBSF/TBSF and two multiplexors. In the SSS-MIN, the controller dominates the hardware of a switching element since the data path is serial line (or a few bit lines). The controller of the PBSF is not so complicated compared with others as the horizontal path is decided a half clock earlier than horizontal path. Thus, banyan network used in the PBSF requires less than twice hardware of that used in the MBSF/TBSF.

4 Analysis of the throughput

Here, the throughput of three MINs with multiple outlets are analyzed by both probabilistic model and simulation. Since the SSS-MIN

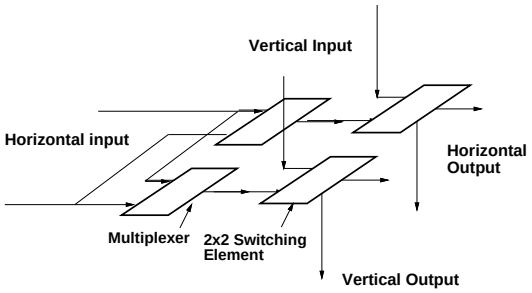

Figure 6: A structure of a switching element in the PBSF

does not provide the packet buffer inside the switching element, the throughput of the network is represented by the pass-through ratio of the network. Although the analysis is done for multiprocessors with the SSS-MIN, the models and analysis results are useful for the ATM packet switching networks.

The probabilistic model used here is similar to the analysis method for the circuit switched MIN by Wu[5] except with the following two points:

1. The target MIN provides multiple outlets while Wu's target has single.
2. All packets are inserted synchronized with the frame signal in the SSS-MIN.

Analysis is based on following assumptions:

- There are no buffers in the switching elements.
- The probability which packets arrive at each input of the MIN is the same.
- The destination of packets is uniformly distributed over all of the MIN outputs, that is, there are no 'hot-spot'.

4.1 Analysis model of the MBSF

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 con ict is $0.25r^2$ because the con ict occurs only when packets are inserted from both inputs, and destinations of them are the same. In the SSS-MIN, when con ict 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:

$$r_{i+1} = r_i - 0.25{r_i}^2 \equiv X(r_i) \tag{1}$$

The pass-through ratio (The probability that a packet arrives at the output) of a single banyan network with n stages is represented as follows:

$$r_n = X^n(r_0) \tag{2}$$

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

In the MBSF, the traf c is distributed into K banyan networks. Therefore, the pass-through ratio of the MBSF (PT_{MBSF}) is represented as follows:

$$PT_{MBSF} = KX^n(r_0/K) \tag{3}$$

4.2 Analysis model of the EBSF

Assume that the number of inputs of the EBSF is KN, and only N inputs are actually used. Thus, the size of the EBSF is K times expanded as the original banyan with N inputs.

In the rst stage, packets are inserted into an input port of N elements and so the packet existing probability is $r_0/2$, while the probability becomes 0 for other unused $(K-2)N/2$ elements. In the later stages, the packet existing probability is fairly distributed into all used inputs of the elements if k is an even number. The probability of a packet which passes through an element is represented with Equation (1).

However, if K is the odd number, the packet existing probability varies depending on the position of input links. In the second stage, there are N elements whose two input links are used, while other $(K-2)N/2$ elements use only a single input link (See Figure3). Here, we call former N elements the element group A, while the others are called the element group B. The packet existing probability of output links from the group A is twice as that from the group B.

Therefore, the packet existing probability on inputs links of elements in the third stage is determined the group in which the links is connected in the second stage. Fortunately, the combination of the input links never increase, and elements in the third stages are classi ed also into two groups: AA: both input links are connected to elements in the group A, AB: one is connected to elements in the group A, while the other is connected to elements in the group B.

Similarly, elements are classi ed into two groups in every stage, and the probability for each link is determined the stage by stage.

4.3 Analysis model of the TBSF

On the TBSF, correctly routed packets are sent for memory modules, and traf c for the next banyan is reduced. Assuming that the input traf c for the K th banyan network is R_K, it is represented with the following equations:

$$R_K = R_{K-1} - X^n(R_{K-1}) \tag{4}$$

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

$$PT_{TBSF} = R_0 - R_l \tag{5}$$

assumed that the number of whole banyan network is l.

4.4 Analysis model of the PBSF

In the rst (top) layer of the PBSF, the probability that a packet is forwarded to the horizontal direction is represented with the Equation 1. In the PBSF, a packet is forwarded to the next layer only when the con ict occurs. Therefore, the probability that a packet is inserted into a switching element in the 2nd layer i'th stage from vertical input ($d_{i,2}$) is represented as follows:

$$d_{(i,2)} = 0.25{r_i}^2 \tag{6}$$

For the layer lower than the 2, the probability that a packet is forwarded to the horizontal/vertical direction is represented in the Table 1.

4.5 Evaluation results

Figure 9,Figure 10, Figure 11 and Figure 12 show the pass-through ratio versus input traf c load (r_0) of the MBSF, EBSF, TBSF and PBSF, respectively. The size of the network is xed (256), while the number of connected banyan networks (K) varies. In order to certify analysis models, results from computer simulation based on the hardware of switching elements are also shown in Figure 9-12.

Table 1: Output rate of switching element(PBSF)

vertical input	holizontal input		generate probability	holizontal output	vertical output
$1-d$	$1-r$	r	$(1-d)r(1-r)$	0.5	0
$1-d$	r	$1-r$	$(1-d)r(1-r)$	0.5	0
$1-d$	r	r	$(1-d)r^2$	0.75	0.25
d	$1-r$	$1-r$	$d(1-r)^2$	1.0	0
d	$1-r$	r	$dr(1-r)$	1.0	0.5
d	r	$1-r$	$dr(1-r)$	1.0	0.5
d	r	r	dr^2	1.0	0.75

In the proposed probabilistic models, input load traf c for all banyan networks are assumed to be uniform. However, in the TBSF and the PBSF, the traf c fed to a banyan network may not be uniform since the previous banyan network may disturb the traf c uniformity. This is the reason why there are a little error between the probabilistic models and simulation. However, this error is only 7% in maximum, and this demonstrates that the results from the probabilistic models is useful in most cases.

From these gures, the pass-through ratio of the MBSF is smaller than those of the TBSF and PBSF re ecting that multiple networks are used independently in the MBSF. Since elements in the earlier stages do not contribute traf c distribution in the EBSF, the pass through ratio is not much improved compared with the MBSF. The pass-through ratio of the TBSF and PBSF is kept high in the heavy input traf c load. In the TBSF and PBSF with 5 banyan networks, almost packets are passed even if packets are inserted to each input in every frame.

Figure 13 shows the pass-through ratio versus the number of connected banyan network under the xed (100%) input load. In this gure, the pass-through ratio of the non-blocking network like the crossbar is also shown. The TBSF and PBSF requires only two banyan networks to achieve higher pass-through ratio than that of non-blocking network. Even the MBSF, three banyan networks are enough to overcome non-blocking networks. This demonstrates the advantages of providing multiple outlets.

Figure 11 and Figure 12 shows that the pass-through ratio of the PBSF is up to 7% larger than that of the TBSF if the connected banyan networks are small (2 or 3). In the PBSF, the routing before con ict is effective since the con icting packet is fed to the element in the same position of the next layer. On the other hand, the con icting packet must be routed from the rst stage in the TBSF. This is the reason why the pass-through ratio of the PBSF is better than that of the TBSF.

However, in the PBSF, if three packets are concentrated into a switching element, a packet is misrouted horizontal direction, and this packet is never saved. On the other hand, in the TBSF, misrouted packet can be always saved in the next banyan network. Therefore, if the number of the connected banyan is large and the pass- through ratio is close to 100%, the pass-through ratio of the TBSF becomes better than that of the PBSF. The pass-through ratio of the PBSF is saturated at 98% if a large number of banyan networks are used. On the other hand, in the TBSF, a pass-through ratio more than 99.9% can be achieved by connecting a large number of banyan networks.

These results suggest that the PBSF is suitable for the multiprocessor with the SSS-MIN because of its small latency and large pass-through ratio with the small number of connected banyan networks, while the TBSF is suitable for the ATM switching network which allows relatively a large latency but requires extremely high pass-through ratio.

5 Con ict-free access using the MINs with multiple outlets

Con ict-free access of the memory module is critical for vector/matrix operations frequently used in scienti c calculations. Because of their multiple outlets, the MBSF, EBSF, TBSF and PBSF have powerful permutation facility compared with other blocking networks. By making the best use of this facility, con ict-free access to the various data slices of vector/matrix is possible under an appropriate data alignment.

Lawrie proposed the data alignment method for the con ict-free access with double size omega network and double size memory modules[15]. The similar data alignment can be applied for the con ict-free access in the SSS-MIN with MBSF, TBSF and PBSF since their multiple (two) outlets support the similar permutation facility to the double size omega network.

5.1 Assumptions

First, let $N = n^2$ be the number of processors, and thus, let $M = N$ be the number of memory modules (n is a power of two).

Here, the detail of MINs are assumed as follows.

- Network topology : The inverse omega network[15] is assumed as the banyan network used in the MBSF/TBSF/PBSF. Other types of blocking networks like the omega and baseline can be used similarly.
- The number of utilized banyan network (K) : 2

Here, con ict-free access to common slices of an array: rows, columns, forward diagonals, and backward diagonals are considered.

First of all, data alignment which solves the access con ict to the above data slices on a memory module must be considered. A data skewing scheme which allows such accesses was proposed by Lawrie[15]. In this scheme, an element a_{jk} in $N \times N$-matrix is stored in the $j \cdot (N^{\frac{1}{2}} + 1) + k \cdot 2$ memory module (Figure 7). Note that the size of both memory modules and the inverse omega network is $2N$.

Since the MBSF/TBSF/PBSF with N inputs is assumed here, two memory modules are squeezed into one module as an element a_{jk} is stored in the memory $\lfloor (j \cdot (N^{\frac{1}{2}} + 1) + k \cdot 2)/2 \rfloor$ (Figure 8). In this

MEMORY

0	1	2	3	4	5	6	7
a_{00}	a_{01}	a_{02}	a_{03}				
a_{13}	a_{10}	a_{11}	a_{12}				
a_{21}	a_{22}	a_{23}	a_{20}				
a_{30}	a_{31}	a_{32}	a_{33}				

Figure 7: Skewing scheme by Lawrie

MEMORY

0	1	2	3
a_{00}	a_{01}	a_{02}	a_{03}
a_{13}	a_{10}	a_{11}	a_{12}
a_{21}	a_{22}	a_{23}	a_{20}
a_{30}	a_{31}	a_{32}	a_{33}

Figure 8: Skewing scheme on this paper

paper, only the TBSF and PBSF are discussed because the proofs for the double sized omega networks shown in [15] are almost directly applied for the MBSF.

In the TBSF/PBSF, packets are routed in the distributed control according to the header. To realize the con ict-free access, the priority of the con icting packets must be controlled. A priority bit is added to the packet header, and the packet with the high priority overcomes that with the low priority. Here, packets issued from even input lines have the high priority while ones from odd input lines have the low priority.

For the access to the above data slices under the skewing scheme shown in Figure 8, the mapping from memory module M to processor N , $P = \{(x, y) | 0 \leq x, y < N\}$ becomes as follows (b is constant) :

rows:$P = \{(x, x+b) | 0 \le x < N\}$ (7)

columns:

$$\begin{cases} P = \left\{ \left(x, \frac{x}{2}(n+1)+b\right) \;\middle|\; 0 \le x < N \right\} & (x = \text{even}) \\ P = \left\{ \left(x, \frac{x}{2}(n+1)-\frac{1}{2}+b\right) \;\middle|\; 0 \le x < N \right\} & (x = \text{odd}) \end{cases} \quad (8)$$

forward diagonals:

$$\begin{cases} P = \left\{ \left(x, \frac{x}{2}(n+3)+b\right) \;\middle|\; 0 \le x < N \right\} & (x = \text{even}) \\ P = \left\{ \left(x, \frac{x}{2}(n+3)-\frac{1}{2}+b\right) \;\middle|\; 0 \le x < N \right\} & (x = \text{odd}) \end{cases} \quad (9)$$

backward diagonals:

$$\begin{cases} P = \left\{ \left(x, \frac{x}{2}(n-1)+b\right) \;\middle|\; 0 \le x < N \right\} & (x = \text{even}) \\ P = \left\{ \left(x, \frac{x}{2}(n-1)-\frac{1}{2}+b\right) \;\middle|\; 0 \le x < N \right\} & (x = \text{odd}) \end{cases} \quad (10)$$

In order to discuss on the con ict-free access using the TBSF/PBSF, the following de nitions and a theorem are required.

De nition *Let Ω_N be the inverse omega network with N inputs, where $N = 2^l$. If Ω_N can produce the mapping speci ed by P_N then we say Ω_N passes P_N or $\Omega_N \uparrow P_N$. Otherwise, we say $\Omega_N \not\uparrow P_N$.*

De nition *Let TBSF $_{N,K}$ be the TBSF consisting of K inverse-omega networks with N inputs, where $N = 2^l$. If TBSF$_{N,K}$ can produce the mapping speci ed by P_N then we say TBSF$_{N,K}$ passes P_N or TBSF$_{N,K} \uparrow P_N$. Otherwise, we say TBSF$_{N,K} \not\uparrow P_N$.*

De nition *Let PBSF $_{N,K}$ be the PBSF consisting of K inverse-omega networks with N inputs, where $N = 2^l$. If PBSF$_{N,K}$ can produce the mapping speci ed by P_N then we say PBSF$_{N,K}$ passes P_N or PBSF$_{N,K} \uparrow P_N$. Otherwise, we say PBSF$_{N,K} \not\uparrow P_N$.*

Theorem 1 *Let*

$$P_N(a, b, c, d, \xi) = \{(ax+b, cx+d) | 0 \le x < \xi\} \quad (11)$$

where $(ax+b, cx+d)$ represents a connection from input $ax+b \pmod N$ to output $cx+d \pmod N$. If $\gcd(a, N) \ge \gcd(c, N)$ and $\xi \le N/\gcd(a, N)$ then $\Omega \uparrow P_N(a, b, c, d, \xi)$.

This theorem was proved by Lawrie[15].

Here, de ne $\alpha = \gcd(a, N), \gamma = \gcd(c, N), a' = a/\alpha$ and $c' = c/\gamma$.

5.2 Con ict free access with the PBSF

Theorem 2 *PBSF$_{N,2} \uparrow P$(rows) .*

Proof By theorem 1 where α and $\gamma = 1$ so $\alpha \ge \gamma$ and $\xi = N/2$ so $\xi \le N/\alpha$. Thus *PBSF$_{N,2} \uparrow P$*(rows) .

Theorem 3 *PBSF$_{N,2} \uparrow P$(columns) .*

Proof This proof is divided into two parts: for $x = 2m$ (even input), and for $x = 2m+1$ (odd input) (Equation 8).

$$\begin{aligned} \text{even: } P &= \left\{ \left(2m, m(n+1)+b\right) \;\middle|\; 0 \le m < \tfrac{N}{2} \right\} \\ \text{odd: } P &= \left\{ \left(2m+1, m(n+1)+\tfrac{n}{2}+b\right) \;\middle|\; 0 \le m < \tfrac{N}{2} \right\} \end{aligned} \quad (12)$$

(I) Even : Since the priority of a packet from the even input is higher than that from odd, the routing of the packet never disturb with packets from odd inputs. $\gamma = 1$ since $n+1$ is prime to 2 . $\alpha = 2$ and so $\alpha \ge \gamma$. $\xi = N/2$ and so $\xi \le N/\alpha$. Thus, the condition is satis ed.

(II) Odd : Since the priority is low, a packet from the odd input is transferred to the second layer. However, since the routing on the rst layer is effective before the con ict, the packet can be treat so that it is routed in the single inverse omega network. Like the case of packets from the even input, since $\alpha = 2$ and $\gamma = 1$, $\alpha \ge \gamma$ and $\xi \le N/\alpha$. Thus, the condition is satis ed.

From (I) and (II), PBSF$_{N,2} \uparrow P$(columns) .

Similarly, the PBSF passes forward diagonals and backward diagonals.

5.3 Con ict free access with the TBSF

Theorem 4 *A connection from input $2m$ or $2m+1 \pmod N$ to destination $cm+d$ or $cm+d' \pmod N$ is represented as follows:*

$$\begin{aligned} \text{even: } P &= \left\{ (2m, cm+d) \;\middle|\; 0 \le m < \tfrac{N}{2} \right\} \\ \text{odd: } P &= \left\{ (2m+1, cm+d') \;\middle|\; 0 \le m < \tfrac{N}{2} \right\} \end{aligned} \quad (13)$$

where c is odd. Assume that all of packets from the odd line can not reach the destination in the rst banyan network. Then, these packets are routed to the following output lines at the end of the rst banyan network:

$$P(\textit{miss}) = \left\{ \left(2m+1, (N-c)m+d\right) \;\middle|\; 0 \le m < \frac{N}{2} \right\} \textit{(odd)} \quad (14)$$

Proof Since $N-c$ is prime to 2 , we have $\gamma = 1$ and $\alpha = 2$. Thus, $\alpha \ge \gamma$ and $\xi = N/2$, and so $\xi \le N/\alpha$. Then, by Theorem 1, TBSF passes P(miss) .

Theorem 5 *The packets from odd input lines, which can not be routed to their destination at the end of the rst banyan network, pass the second banyan network if the condition is satis ed at Equation 13.*

Proof The packets are routed at the end of the rst banyan network as shown in Equation 14. Then, these packets are transferred to their original destinations shown in Equation 13 in the second banyan network shown as follows.

$$\begin{aligned} &\text{TBSF}_{N,2} \uparrow P(\text{miss} \to \text{right}) \\ &= \left\{ \left((N-c)m+d, cm+d'\right) \;\middle|\; 0 \le m < \frac{N}{2} \right\} \text{(odd)} \end{aligned} \quad (15)$$

$\gamma = 1$ since c is odd. $N-c$ is prime to 2, and so $\alpha = 1$. Thus, $\alpha \ge \gamma$, $\xi = N/2$, and $\xi \le N/\alpha$. By Theorem 1, TBSF$_{N,2} \uparrow P$(miss $\to$ right) .

Theorem 6 *TBSF$_{N,2} \uparrow P$(rows) .*

Proof It is similar to the case of the rows on the PBSF.

Theorem 7 *TBSF$_{N,2} \uparrow P$(columns) .*

Proof This proof is divided into two parts: for $x = 2m$ (even input), and for $x = 2m+1$ (odd input) (Equation 8).

$$\begin{aligned} \text{even: } P &= \left\{ \left(2m, m(n+1)+b\right) \;\middle|\; 0 \le m < \tfrac{N}{2} \right\} \\ \text{odd: } P &= \left\{ \left(2m+1, m(n+1)+\tfrac{n}{2}+b\right) \;\middle|\; 0 \le m < \tfrac{N}{2} \right\} \end{aligned} \quad (16)$$

Since the priority of a packet from the even input is higher than that from the odd input, all of these packets can pass on the rst network. Next, odd parts are transferred to the second network, and pass the second one.

(I) Even : Since $n+1$ is prime to 2 , we have $\gamma = 1$. Since $\alpha = 2$, $\alpha \geq \gamma$. $\xi = N/2$ and so $\xi \leq N/\alpha$. Thus, by Theorem 1, all packets from even inputs pass on the rst banyan network.

(II) Odd : Since packets are routed as shown in the Equation 14 , we get

$$P(\text{right}) = \left\{ \left(2m+1, m(n+1) + \tfrac{n}{2} + b\right) \;\middle|\; 0 \leq m < \tfrac{N}{2} \right\}$$
$$\Downarrow$$
$$P(\text{miss}) = \left\{ \left(2m+1, m\big(N-(n+1)\big) + N - (n+1) + b\right) \;\middle|\; 0 \leq m < \tfrac{N}{2} \right\}. \quad (17)$$

Next, in the second network, $\gamma = 1$ since $n+1$ is prime to 2. Then, by Theorem 5, the packets from odd line can reach the correct destination on the second banyan network.

Thus $TBSF_{N,2} \uparrow P(\text{ columns })$.

Similarly, the TBSF passes forward diagonals and backward diagonals.

6 Conclusion

A MIN with multiple outlets called the MBSF, EBSF, TBSF and PBSF are analyzed and compared. By passing packets for the same destination using multiple outlet, the throughput of these three networks can overcome those of the non-blocking networks. Through the evaluation, the PBSF, which is proposed in this paper can support the best throughput and latency used in the SSS-MIN. Although the latency of the TBSF is large, the pass-through ratio becomes maximum if the number of connected banyan networks are more than 4. Therefore, the TBSF is useful for the ATM switching networks in which the relatively large latency is tolerable.

Con ict-free access to the common data slices of the matrix (rows, column, diagonal, and backward diagonal) is also possible with these three networks.

References

[1] G.Broomell, and J.R.Heath, \Classi cation Categories and Historical Development of Circuit Switching Topologies," ACM Computing Surveys, Vol.15, No.2, Jun. 1983.

[2] 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.

[3] C.Clos, \A study of non-blocking switching networks," Bell System Tech. J. 32, 2 Mar. 1953.

[4] 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.

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

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

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

[8] R. Awdeh, and H. Mouftah, \The Expanded Delta Fast Packet Switch," IEEE Int. Conf. Commun. , 1994.

[9] H.Amano, K.Gaye, \A Batcher Double Omega network with Combining," Proc. of the International Conference on Parallel Processing, Vol I, pp.I-718-I719 Aug. 1991.

[10] 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 the International Conference on Parallel Processing, Aug, 1994.

[11] M.J.Narasimha, \The Batcher-banyan self-routing network: Universality and Simpli cation," IEEE Trans. Commun.,vol.36, No 10, pp.1175-1178, Oct.1988.

[12] K.Gaye, H.Amano, \A Batcher Double Omega network with Combining," IEICE Trans. inf. & syst. Vol.E75-D, No.3 pp.307-314, May. 1992.

[13] 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.

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

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

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

[17] A. Gottlieb, R. Grishman, C. P. Kruskal, K.P. Mcauliffe, L. Rudolf, and M. Snir, \The NYU Ultracomputer - Designing an MIMD Shared Memory Parallel Computer", IEEE Trans.on Comput. vol. c-32, No.2, Feb.1983.

[18] F.Chiussi, H.Amano, F.A.Tobagi, \A 0.8μm BiCMOS Sea-Of-Gates implementation of the Tandem Banyan Fast Packet Switch," Proc. of IEEE Custom Integrated Circuits Conference May. 1991.

[19] S.R.Dickey, R.Kenner, \Combining Switches for the NYU Ultracomputer," Proc on IEEE Frontiers to Massively Parallel Processing, pp.521-523, Nov. 1992.

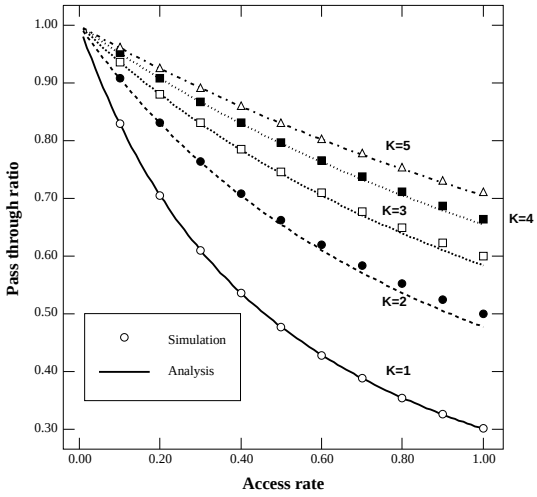

Figure 9: Relation between load and pass-through ratio on MBSF

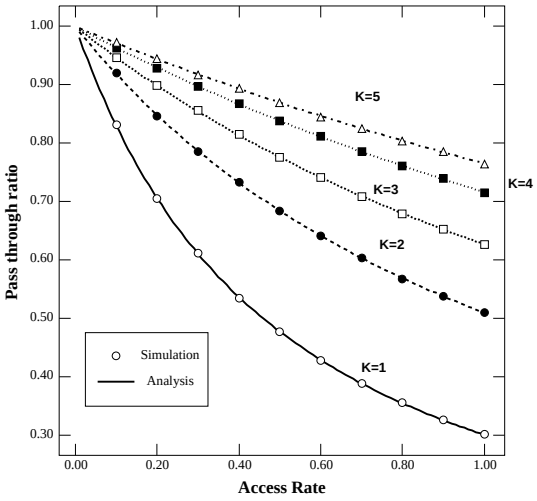

Figure 10: Relation between load and pass-through ratio on EBSF

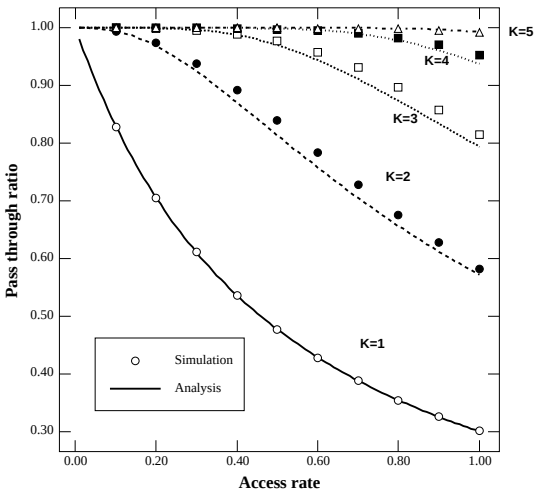

Figure 11: Relation between load and pass-through ratio on TBSF

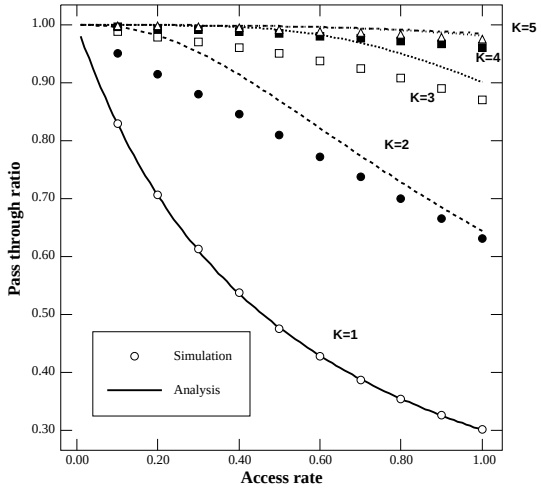

Figure 12: Relation between load and pass-through ratio on PBSF

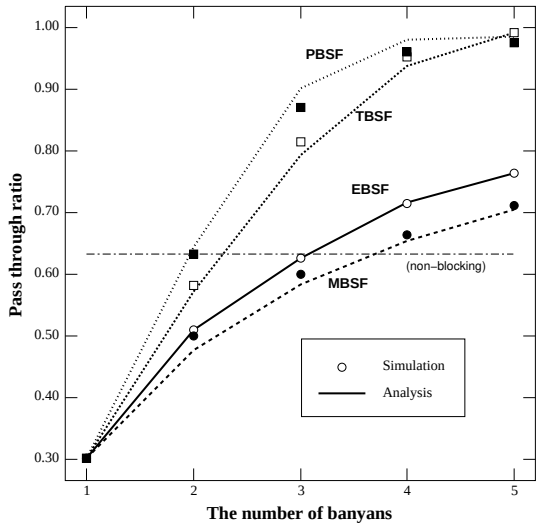

Figure 13: Relation between the number of connected banyan network and pass-through ratio at 100% input load