An LSI implimentation of the Simple Serial Synchronized Multistage Interconnection Network

Takayuki KAMEI Masashi SASAHARA Hideharu AMANO

Dept. of Electrical Engineering
Keio University
3-14-1, Hiyoshi Yokohama, Japan
{kamei, sasahara, hunga}@aa.cs.keio.ac.jp

Simple Serial Synchronized (SSS) Multistage Interconnection Network (MIN) is a novel MIN architecture for connecting processors and memory modules in multiprocessors. Synchronized bit-serial communication simplifies the structure/control, and also solves the pin-limitation problem. The simple structure allows the use of a high frequency clock rate and a high throughput 3-dimensional multistage network called Piled Banyan Switching Fabrics (PBSF). Here, an LSI implementation of the SSS-PBSF chip is introduced. Using an 0.5 μ CMOS SOG technology, the PBSF consisting 3-dimensional connection of two omega networks (16 inputs/16 outputs) with the message combining function and priority control is implemented into a single chip (17356 gates).

I. Introduction

Multistage Interconnection Networks (MINs)[2] have been well researched as an interconnection mechanism between processors and memory modules of middle scale parallel machines. Most of them are blocking networks like the omega network, 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 each other in the network even if their destinations are different. In this case, one of the conflicting packets 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 difficult.

As an alternative approach, a novel Multistage Interconnection Network (MIN) architecture called Simple Serial Synchronized (SSS)-MIN[6] was proposed. All packets are transferred in the serial and synchronized manner. Synchronized bit-serial communication simplifies the structure/control, and also solves the pin-limitation problem. The simple structure allows the use of a high frequency clock rate and a highly integrated chip for MIN. A prototype multiprocessor called SNAIL was developed and is available now. In this system, 16 processors/memory modules are connected with SSS-omega chips. Through the evaluation for SNAIL, the performance of the SSS-omega network is satisfactory for the performance and the number of processors used in the multiprocessor [8].

However, for recent advanced high speed microprocessors, the large conflict ratio of the omega network is not tolerable. To cope with this problem, a multistage network with 3 dimensional structure called Piled Banyan Switching Fabrics (PBSF) is proposed[9]. Here, an LSI implementation of the SSS-PBSF chip is described. Using an 0.5 μ CMOS SOG technology, the PBSF consisting 3-dimensional connection of two omega networks (16 inputs/16 outputs) with the message combining function and priority control is implemented into a single chip (17356 gates).

II. The SSS-MIN and SNAIL

A. The SSS-MIN

A.1 Basic operations

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 from the outlets of the MIN.

When a conflict occurs, one of the conflicting packets must be routed to the incorrect direction since the SSS-

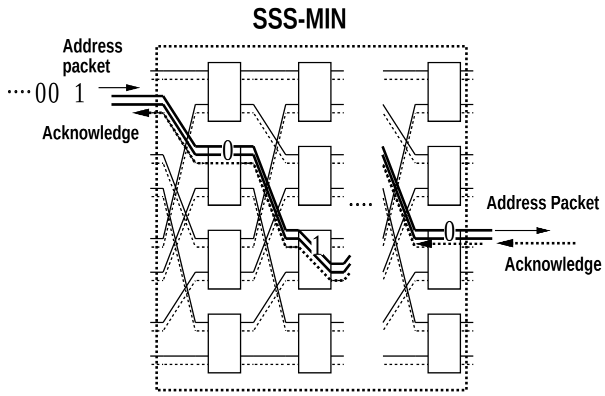


Fig. 1. Structure of the SSS-MIN

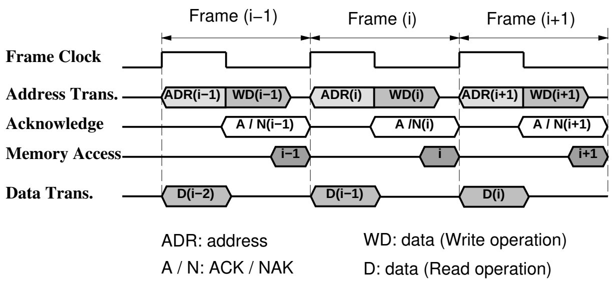


Fig. 2. Timing diagram of packets/signal transfers in the SSS-MIN

MIN provides no packet buffers in each switching element , and the conflict bit in the packet header is set. The packet that set conflict 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 only referring to the conflict bit. An acknowledge signal (ACK: ACKnowledge or NAK: Not AcKnowledge) is returned to the processor which issues the access request.

Every 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, 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.

A.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. Figure 2 shows a timing diagram of these transfers. 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. In the SSS-MIN, the state of an element (straight or exchange) is set by the address packet header. 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.

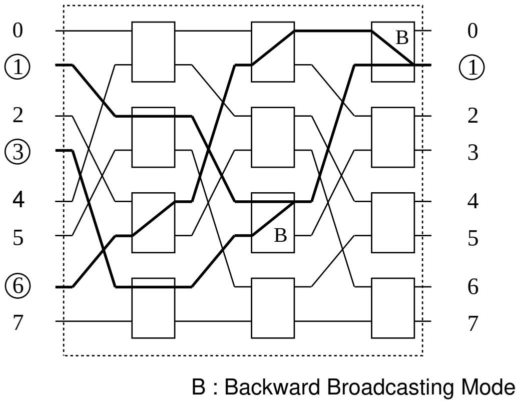
A.3 Other issues on the SSS-MIN

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

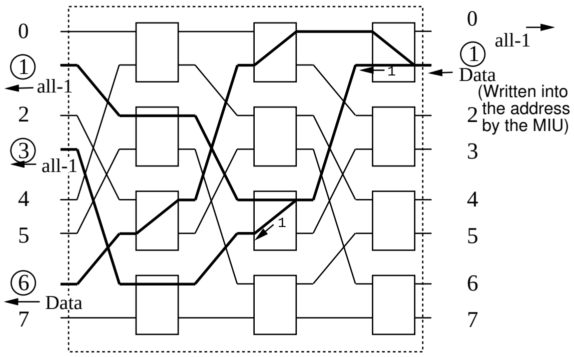
Bit serial message combining The message combining[3] is a technique which merges similar references into composite combined references at each stage of the network. Although a cost effective switch architecture with combining has been researched and implemented[7] for conventional MINs, it still requires a large amount of additional hardware.

In the SSS-MIN, the message combining is easily performed by a bit-by-bit comparison during the address packet transfer. If two addresses are equal, the switching element is set to be the backward broadcast mode B as shown in Figure 3a. In this example, three read references for the memory module 1 are combined together. After these address packets passed by, the tree-like trace remains in switching elements, and the returning data packets are multicast through the tree. This technique is also useful for the write-read combining.

To combine Test&Set operations, a simple additional hardware is required. The tree-like trace for data transfer is also formed like Figure 3a. However, the data is not multicast but one of the returned values is forced to '1' in each element as shown in Figure 3b. At the same time, the Memory Interface Unit (MIU) which is provided between the MIN and memory modules writes all-1 into the target address of the memory module. As a result, only one processor can read out the data in the target address, and other processors get all-1.



(a) Combining of read references



(b) Combining of Test&Set references

Fig. 3. Bit serial message combining

This technique is called the *bit-serial message combining*[5], 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 flip-flop provided in a switching element can be omitted in most stages, and packets are 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[6].

B. Prototype multiprocessor SNAIL

In order to evaluate the performance of the SSS-MIN using practical application programs, a multiprocessor prototype called SNAIL (SSS Network Architecture ImpLementation) has been developed. As shown in Figure 4, 16 processors with local memory are connected with the 16 shared memory modules using the SSS-MIN. a processor is connected with a host workstation through the Ethernet. The performance monitor and debugger system which works independently of the SNAIL are connected both sides of the MIN. All system modules share a single system clock (50MHz) and a frame clock (50/16=3.125MHz) for the synchronous operation.

The main purpose of the SNAIL is evaluating the performance of the SSS-MIN. Particularly, the SNAIL is the first machine with more than 10 processors which provides the message combining. However, the SNAIL is also designed as a test-bed of the static scheduling with the best use of the predictability and schedulability of the SSS-MIN.

B.1 SSS-omega network

The heart of the SNAIL is the SSS-omega gate array chip which provides all functions including the bit-serial message combining [1] . An omega network [1] with 16-input/16-output is implemented into a single chip, and the Tandem Banyan Switching Fabrics[9] can be simply formed by the tandem connection of them.

The specification of this chip is shown in Table I

A single SSS-omega chip can connect 16 processors and 16 memory modules. However, in SNAIL, four chips are used in the bit-slice manner in order to reduce the latency. Thus, 8 chips (4×2) are used to form the TBSF with two banyan networks providing 12bits for address, and 8bits for data. The maximum bandwidth is 125Mbyte/sec for a processor/memory module, and 2Gbyte/sec in total.

From the empirical evaluation of SNAIL, it appears that the latency and synchronization overhead of the SSS-omega network are tolerable, and the bandwidth of the SSS-omega network is sufficient. However, when more than 12 processors are used, the ratio of the packet conflicts is up to 30%, and the performance degradation caused by the conflict cannot be ignored. If recent high performance microprocessors are used instead of the MC68040 used in SNAIL, the performance degradation of the SSS-omega network will bottleneck the system. The high bandwidth network structure called Tandem Banyan Switching Fabrics [9] (TBSF) can be easily realized by the tandem connection of the SSS-omega network. However, the latency is also stretched with the tandem connection, and it may degrade the performance multiprocessor even if the conflict is reduced. To cope with this problem, another network structure called Piled Banyan Switching Fabrics is proposed.

[1] This chip was designed and fabricated with the cooperation of Kawasaki Steel co. ltd. Japan.

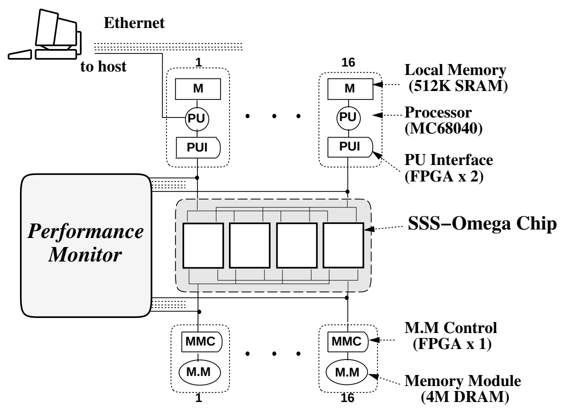


Fig. 4. Diagram of a multiprocessor SNAIL

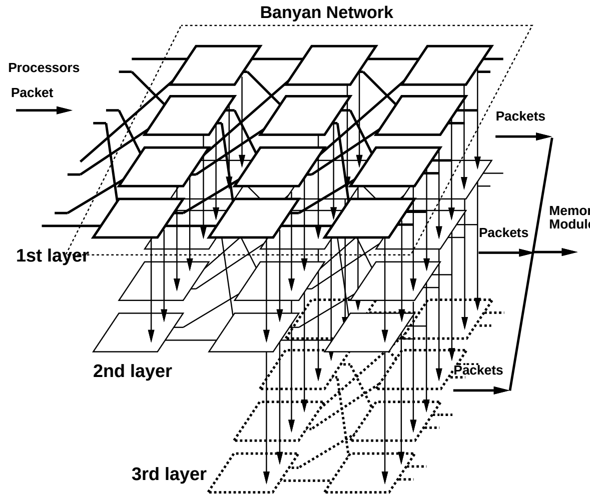


Fig. 5. Piled Banyan Switching Fabrics

TABLE I
The specification of the SSS-omega chip

Clock	50MHz
Bit width	3bit(address)2bit(data)
Size	16-input/output
Maximum bandwidth	$250 Mbits/sec \times 16$
Number of utilized cells	9838
Cell utilization ratio	49%
Number of signal pins	180
Technology	$1.0\mu m$ CMOS sea-of-gates

III. Piled Banyan Switching Fabrics

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 with 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. In this case, one of horizontally transferred packets is selected and sent the next lower layer network.

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. 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. 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.

IV. Chip Design

The PBSF chip was developed by the pilot program of the chip design and fabrication service organization for LSI design education in Japan. In the pilot program, the type of chip is limited to be the $0.5\mu m$ CMOS SOG with 16k gates at maximum. The PGA package with 176 pins (116 pins for signals) can be used. Like the SSS-omega chip, the size of the network is set to be $16{\times}16$ which is the maximum one with this number of pins. The structure and function of the PBSF implemented in the chip is optimized so as to maximize the performance under the pin and gate limitation. For the optimization, the performance of the PBSF with various structures and functions is estimated with computer simulation.

A. Design choices on the structure

The number of layers The packet pass through ratio of the PBSF is improved as the number of layers becomes large. However, the number of the layers are limited both with the pin and gate limitation. In the SSS-MIN, if a packet cannot reach the destination, it is transferred again in the next frame. Figure 6 shows the average number of packet transfer trial required for a memory access versus the access probability of the shared memory. The number of trial is drastically improved with two layers, but not so improved with more layers. Considering the number of the gate and the performance improvement, we decided to use two layers.

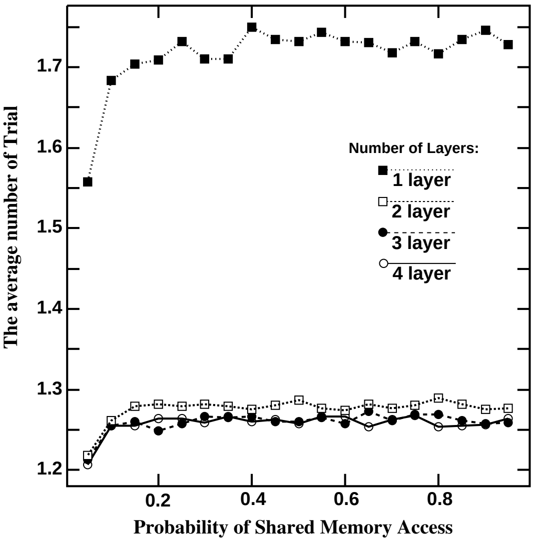


Fig. 6. The average number of trial for the number of layer

Multiplexed output For the pipelined circuit switching, at least 1 bit signal is required for address, data, and acknowledge respectively. However, address packet transfer with 1 bit signal takes larger time than that of the data packet transfer since the trace must be established in the address transfer. The difference between the address and data transfer introduces the difficulty to efficient control of the pipeline. In order to reduce the time for address packet transfer, 2 bits signal lines are provided. In this case, the sum of the total number of the required lines is beyond the pin limitation. To cope with this problem, we introduce a multiplexor at the output of the PBSF. The packet which can be issued is limited to be one. Although the performance of the PBSF without the multiplexor is 20% better than that with the multiplexor, enough packet pass through ratio (about 75%) is achieved even with the multiplexor.

As a result, we selected the two-layer PBSF with multiplexor shown in Figure 7.

B. Design choices on the functions

Message Combining Message combining is a powerful function for avoiding the hot spot contention, but requires about increase of 20% hardware according to our previous implementation (SSS-omega chip). This mechanism works effectively at the first layer since a lot of packets for the same direction conflict on the switching element.However, in the second layer, the opportunity of the packet combining is much smaller than that in the first layer. According to the result of the computer simulation, there is little performance improvement if the second layer provides the facilities for the combine. So, the message combining function is provided only in the first layer and the multiplexor. By omitting the facility in the second layer, 6-7% gates can be reduced.

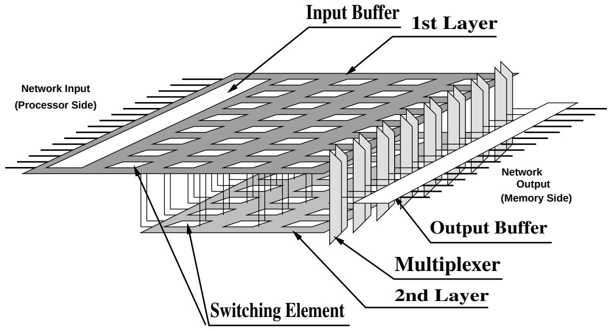


Fig. 7. The implemented structure

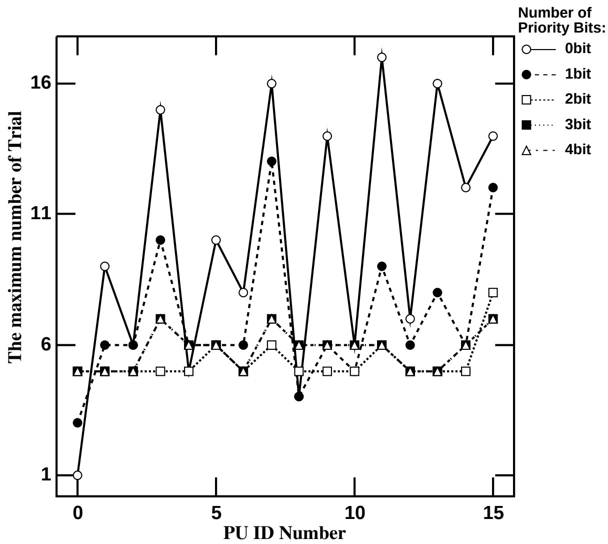


Fig. 8. The maximum number of trial versus the number of priority level

Priority Control In order to keep the fairness when packets collide at the switching element, the priority tag is attached to the address packet header. Although a large number of tag bits enable the precise fairness control, they stretch the latency of trace setup and increase the hardware. Figure 8 shows the maximum number of trials which is required for a packet from a processor arriving at the shared memory. In this simulation, the priority of a failed packet is incremented in the next frame. If there is no priority control, the difference between processors is large, thus, some processors take a long time to access the shared memory compared with others. According to Figure 8, the difference is sufficiently reduced with 2 bits (4 level priority control) priority tag.

V. Implementation

A. Detail of the structure

As shown in Figure 7, the implemented PBSF consists of the input interface, PBSF core, and the output interface including the multiplexor.

Input interface As shown in Figure 9, the input interface holds a small buffer to store the header of the address packet while the trace is established. Since the packet is sent to the stage where the trace was set up in the previous clock, the head of the packet reaches the output of the MIN just a clock after the trace setting up. Once the trace is established, each flit hops all stages, and reaches the output interface with a clock delay. Timing signal for controller of each switching element is also generated in the input interface.

PBSF core Figure 10 shows the structure of a switching element in the first layer.

By making the best use of the control strategy in the SSS-MIN, the structure and control of the switching element is quite simple. It consists of the address switch with control block, acknowledge switch, and data switch. In order to reduce the latency for setting up the trace with priority control, 2 bits extra wires are provided. Thus, the number of wires for the address switch is expanded to be 4. As the address switch is forward direction, there is paths to the second layer for bypassing the conflicting packet.

In the SSS-MIN, the state of an element (straight or exchange) is set by the address packet header. Since the acknowledge signal and data packet share the same trace which is set by the address packet, only one controller is required. The acknowledge switch and data switch consists of simple multiplexors controlled by the controller. Both the acknowledge and data switch are reverse direction (The data switch only passes the data packets for read operations. Data packets for write operations are sent through the address switch), and the data packet and acknowledge signal are joined in the switching element. The message combining is implemented in the switching element of the first layer using the backward broadcasting mode[4].

Output interface Output interface selects one of packets from two layers with the multiplexor. If two address packets are entirely the same, these packets are combined in the multiplexor.

B. Specification

The chip is designed using the RTL(Register Transfer Level) description of Verilog-XL. The design was done only with three months by two students.

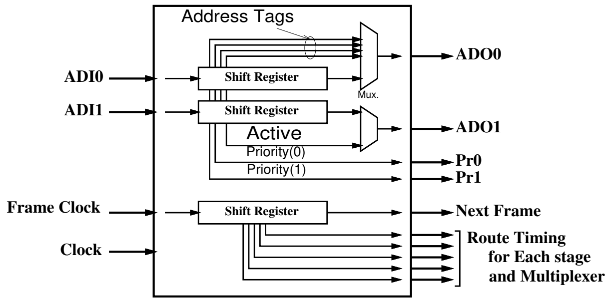


Fig. 9. The structure of an input buffer

TABLE II
THE SPECIFICATION OF THE PBSF CHIP

Clock	100MHz
Bit width	2bit(address) 1bit(data)
Size	16-in/out
Max bandwidth	$300 Mbits/sec \times 16$
the number of cell	17356
signal pin	116pin
Technology	$0.5\mu m$ CMOS sea-of-gates

Table II shows the specification of the implemented PBSF chip. Compared with the simple SSS-omega chip shown in Table I, the maximum clock frequency used in the PBSF chip is twice. On the contrary, the number of the lines is smaller than that of the SSS-omega chip because of the pin limitation of the package. As a result, the maximum bandwidth of the PBSF chip is about 1.2 times as that of the simple omega network. However, since the pass through ratio of the PBSF is almost twice as that of the simple omega network, the average bandwidth of the implemented PBSF chip is almost 2.4 times of the omega network chip. Although the PBSF chip requires 1.76 times number of gates (17356/9838), the increase of the hardware is reasonable considering the performance improvement.

C. Chip test environment

The sample chips were tested on a simple breadboard shown in Figure 11. A packet is generated in the packet generator and sent to the PBSF chip through a shift register. The destination of the packet is decided with a dip switch. Exchanged packets with the PBSF chip are received by shift registers and displayed on the LEDs. High speed discrete CMOS ACT/FCT gates are used to cope with a high speed clock.

Now, the PBSF chip is working on the board with 90MHz clock. Figure 12 shows the input and output waveforms of the PBSF chip. The upper wave (2V range) shows the input packet and the lower (2V range) shows

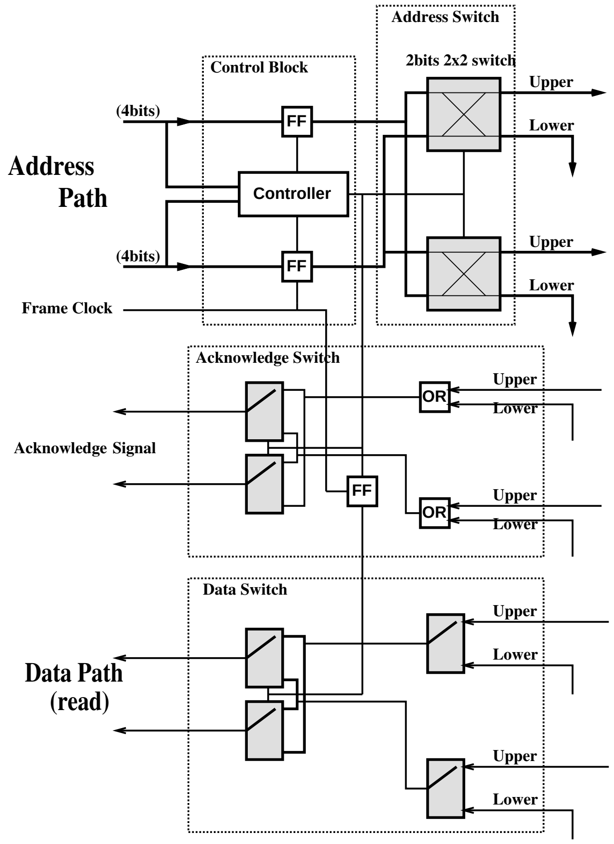


Fig. 10. The structure of an upper switching element

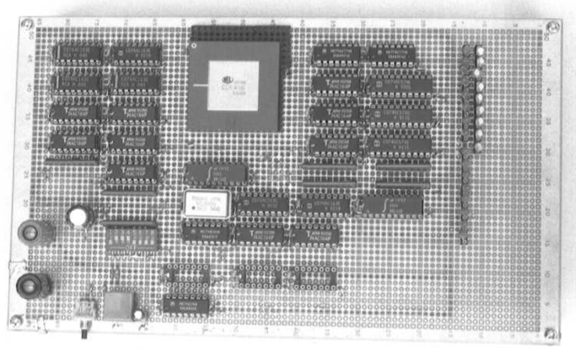


Fig. 11. A tested breadboard

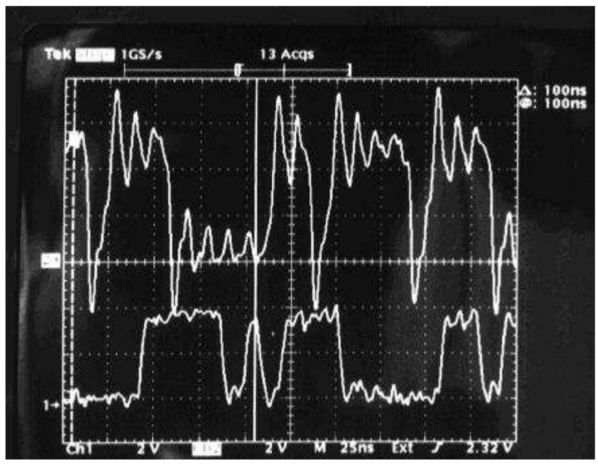


Fig. 12. Input and output waveforms (with 90MHz clock)

the output after switching. Unlike the output wave from the PBSF chip, input wave includes a large overshooting and undershooting. This comes from the insufficient electrical condition of the breadboard, and this is the reason why the chip is not stable at 100MHz clock.

VI. Conclusion

An LSI implementation of the SSS-PBSF chip is introduced. Using an 0.5 μ CMOS SOG technology, the PBSF consisting 3-dimensional connection of two omega networks (16 inputs/16 outputs) with the message combining function and priority control is implemented into a single chip (17356 gates). The simple architecture and Verilog-XL based design reduce the cost and time for development. The design was done only with three months by two students. Now, the chip is confirmed to work with at least 90MHz clock, and the total throughput of the chip is 540Mbyte/sec(270Mbits/sec). The average throughput of the SSS-PBSF is almost 2.4 times as that of simple SSS-omega chip.

We are now designing another multiprocessor called SNAIL-2 based on the PBSF chips. The real performance of the chip will be evaluated on this multiprocessor using parallel application programs.

Acknowlegement

The PBSF chip was developed by the pilot program of the chip design and fabrication service organization for LSI design education in Japan, and fabricated by NTT Electronics Technology Co. Ltd. The authors would appreciate with the staff of the pilot program and NTT Electronics Technology Co. Ltd. for their sincere advice for the chip design. The authors also thank Toshihiro Hanawa, Hideki Yasukawa, Akira Funahashi for their comments that improved the presentation of the paper.

References

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

1975.

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

[3] 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 Trns. on Comput.*, vol.c-32, No. 2, 1983

[4] G. Kalidou, H. Amano, "Design and Implimentation of the Batcher Double Omega Network with Combining," *Proc. of ISMM Int. Workshop on Parallel Computing*, pp.121-124, Sep. 1991

[5] H. Amano, G. Kalidou, "A Batcher Double Omega network with Combining," *Proc. of International Conference on Parallel Processing*, volume I, pp.718-719, 1991

[6] 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.1, pp.571-577, Sep. 1992

[7] Dickey, S.R. and Percus, O.E. , "Performance differences among combining switch architectures," *Proc. of the International Conference on Parallel Processing*, pp/I110-117, 1992

[8] M. Sasahara, et al, "SNAIL: a multiprocessor based on the Simple Serial Synchronized multistage interconnection network architecture," *Proc. of the International Conference on Parallel Processing*, pp.I117-120, Aug. 1994

[9] T. Hanawa, H.Amano, Y.Fujikawa, "Multistage Interconnection Networks eith multiple outlets," *Proc. of the International Conference on Parallel Processing*, pp.I1-8, Aug. 1994