Distributed Input and Deflection Routing Based Packet Switch Using Shuffle Pattern Network

Thai Thach Bao, Hiroaki Morino, Hitoshi Aida, and Tadao Saito

Department of Information and Communication Engineering, Faculty of Engineering,
University of Tokyo,
7-3-1 Hongo Bunkyo-ku Tokyo 113-8656 Japan

Abstract. In this paper, a scalable variable-length packet switch using multistage interconnection network is proposed. The architecture consists of multiple non-buffer switch elements, interconnected with perfect shuffle pattern and with ring topology. Each input port is connected to a switch element in a different stage, and packets applied from these ports are routed to their destined output ports based on principle of deflection routing. The proposed switch has following features; (1) To achieve certain packet loss rate, required number of stage to achieve certain packet loss rate is less than the same type of switches due to efficient use of switch elements. (2) Packet loss rate is almost unchanged regardless of traffic patterns. In performance evaluation, two types of non-uniform traffic are given and it is shown that the switch can achieve lower packet loss rate than existing deflection routing based switch.

1 Introduction

In recent years, demands of voice, data, video traffic increase explosively due to the success of the Internet. Corresponding to this situation, required capacity for backbone packet switch in the future will reach Tera bit/sec, one thousand times of today's packet switch capacity. Furthermore, traffic demand will be concentrated on large size file transfer and broadband video communication, and it is expected for the switching system to cope with these demands.

From this viewpoint, we have been investigating a new framework of multimedia network that is hybrid system of variable-length packet switch and TDM based switch in place of solution with ATM switch[1]. In this system, non real-time traffic such as data transfer traffic is transmitted via variable-length packet switch, while real-time traffics such as voice, video is transmitted via TDM based switch. Access network is constructed using integrated TDM technology, that provides unified interface to users. According to this framework, we have examined variable-length packet switch architecture having capability to handle traffic of Tera bit/sec. In order to design high capacity packet switch of this level, conventional shared bus architecture is insufficient while parallel switching using crosspoint switch is indispensable. Implementations of crosspoint switch are roughly classified into two categories; input buffering switch and output buffering switch.

G. Pujolle et al. (Eds.): NETWORKING 2000, LNCS 1815, pp. 74–84, 2000.
© Springer-Verlag Berlin Heidelberg 2000

Input buffering switch have advantage of simplicity of switch elements and buffer management and high throughput is achieved by combination of virtual output queuing and efficient scheduling[2]. However, maximum switch size will be limited by processing time of scheduling, and this problem will be more serious when port speed is increased.

For the purpose of implementation of large switch up to about 100×100, output buffering switch is noticeable due to its simple switching operation. The feature that it requires amount of wiring and pins is not a big problem with the recent progress of VLSI technology. Among several output queuing switches, authors have investigated deflection routing based switches, which are suitable for variable-length packet switch. For a deflection routing based switch in ATM, several methods have been propose d such as Tandem Banyan switch[3] or MS4[4] switch, but it is not optimized for variable-length packet switch.

In this paper, a deflection routing variable-length packet switch based on input port distribution is proposed. This switch consists of switching elements interconnected through shuffle pattern[5], and the switch exploits ring topology. Positions of input ports are distributed according to the rotation direction within the ring. Therefore, traffic applied to the switch is distributed over the switch, and utilization of switch elements is improved. As a result, required number of switch stages is less than conventional switches in order to achieve certain packet loss rate.

This paper is organized as follows. In Section 2, features of conventional type of deflection routing based switch are described, and further research issues are presented. In Section 3, architecture and operation of the proposed switch is described. In Section 4, results of performance evaluation are shown and it is indicated that performance of the switch does not change under both uniform and non-uniform traffic pattern . Finally, conclusion and future work are presented in Section 5.

2 Background of Research

In Figure 1 and Figure 2, Tandem Banyan switch and Multi Single-Stage-Shuffling Switch(MS4) are presented as typical deflection-routing based switches. Though these are originally designed for fix-length ATM cell switch, they can be easily applied to variable-length packet switch. In these switches, when internal packet conflict occurs at some switch element, one packet is properly routed to destined port and others are misrouted to the rest of ports. In Tandem Banyan switch, misrouted packets will start routing again from a succeeding Banyan plain. On the other hand, they are allowed to restart routing from succeeding stage in MS4 switch. Comparing performance of two switches, MS4 switch will be better since it can utilize switch element more efficiently.

In both these switches, however, all applied packets enter the switch at the first stage. The utilization efficiency of switch elements in preceding stages is higher than that in latter stages resulting in a switch resources waste. As a result, probability of packet conflict in a switch element is high when the element

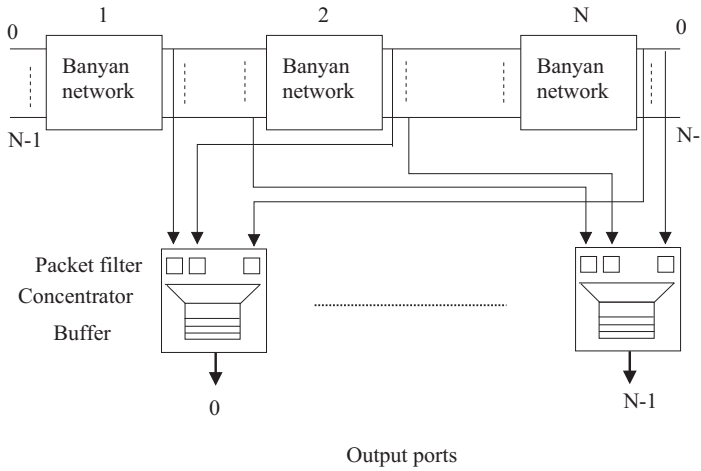

Fig. 1. Tandem-Banyan switch

is close to the first stage. If applied traffic load can be distributed equally over each stage of the switch, the number of stages required to achieve certain packet loss rate will decrea se.

As an approach to distributing traffic within switch , Ring Banyan switch[6] was proposed, which interconnects switch elements in ring topology. Ring Banyan is shown in Figure 3. The routing operation is the same as Tandem Banyan switch, and applied traffic is distributed uniformly within the network. It is shown that required number of stage is less than that of Tandem Banyan switch.

In Ring Banyan switch, however, misrouted packets waste switch elements until they enter next banyan plain. In variable-length packet switching, they may prevent other packets from being routed properly, which lead increase of packet loss. This type of loss can be reduced if misrouted packets can restart routing from the next stage.

3 Proposal of Distributed Input Shuffle Pattern Switch with Deflection Routing

In order to resolve problems in Ring Banyan switch, a new type of switch with ring topology is proposed, which optimizes the interconnection pattern of switch elements.

3.1 Basic Architecture and Operation

The proposed switch is shown in Figure 4, where switch elements are connected through shuffle pattern. An applied packet is routed based on principle of self

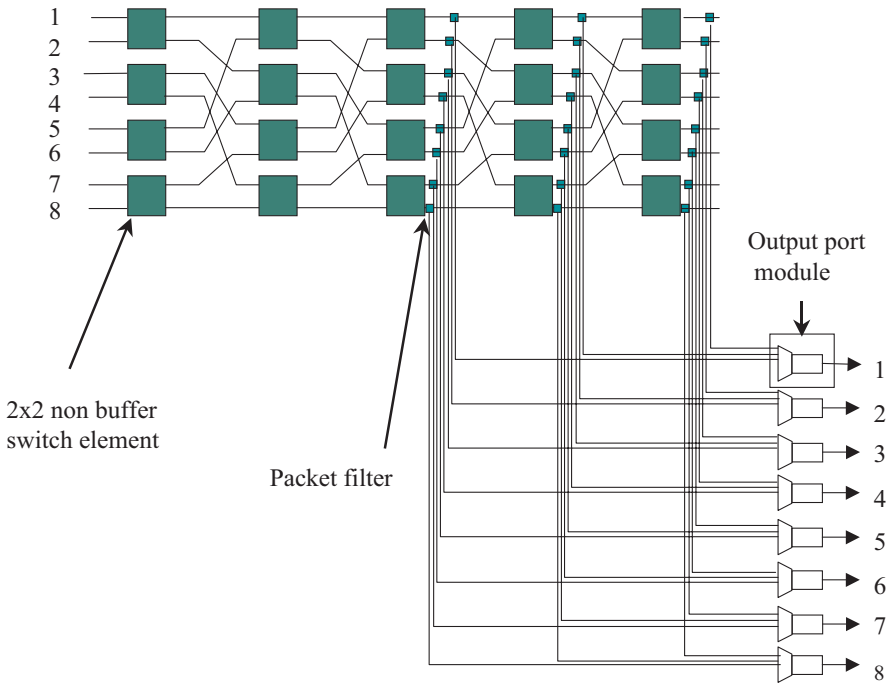

Fig. 2. Deflection routing based shuffle pattern switch(MS4 switch)

routing and rerouting. When a packet reaches a packet filter connected to the destination output port, it leaves the switch.

According to the ring topology, an applied packet can pass through arbitrary large number of stages. Therefore, a new packet may conflict with a packet coming from the previous stage. To prevent this problem, some buffer is provided at input port as shown in Figure 6. The new applied packet and a packet from the previous stage are stored in this buffer, which operates on first-comes-first-served basis. At each output module, a concentrator is provided to reduce the number of output buffers. Structure of concentrator and buffer management is the same as those in Knockout switch[7].

Before an operation of the switch is described in detail, several notations are given. N represents the number of input/output ports. $n = log_2 N$ represents the number of stages in an ordinary Banyan network. K is the number of stages in the switch. The local destination address of a packet will be represented in binary bits as $(d_1 d_2 \ldots d_n)$(where d_1 is the most significant bit). Proposed switch $(N \times N)$ is constructed using $K(\geq n)$ stages, each stage includes $N/2$ rows of 2×2 non-buffer switch elements.

Before a packet is applied to the network, the local header is generated and attached in front of each packet. The local header of each packet includes a des-

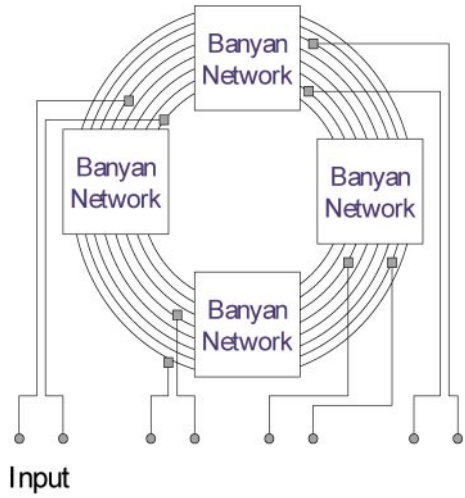

Fig. 3. Ring Banyan switch

tination address field, which contains the destination port represented $d_1 d_2 ... d_n$, and a counter field C, which is initially set to n. When a packet is applied to a switch element, output port of the element to which the packet is transmitted is decided according to the corresponding bit in the destination address field. The switch element refers the d_{n-C+1} bit of the field, and the packet is transmitted to the upper port if it is zero, and to the lower port if it is one.

When the packet is routed to the port successfully, the counter is decremented. Otherwise, when the port is occupied by another packet, the packet is routed to the other port, and the counter is set to $n = log_2 N$, and it continues routing from the next stage. If two packets are applied at the same time and they require the same output port of the switch element, the packet closest to its destination (smaller value of counter field) is served as the highest priority.

In this way, the counter of a packet is decremented whenever the packet is successfully routed to the destined output port of a switch element, and the packet reaches packet filter connected to its destination output port of the switch when the counter is zero. In other words, a packet can reach its destination port of the switch fabric if and only if it successfully passes through n consecutive stages.

The function of packet filters at the output ports of each stage is to examine the content of the counter of each packet and to transmit packet to the following stage for continuing routing or to output port module.

4 Performance Evaluation

In this paper, it is assumed that variable-length packet switch of Tera bit/sec in the future has OC-12 (622.08 Mega bit/sec) interface and its size is about

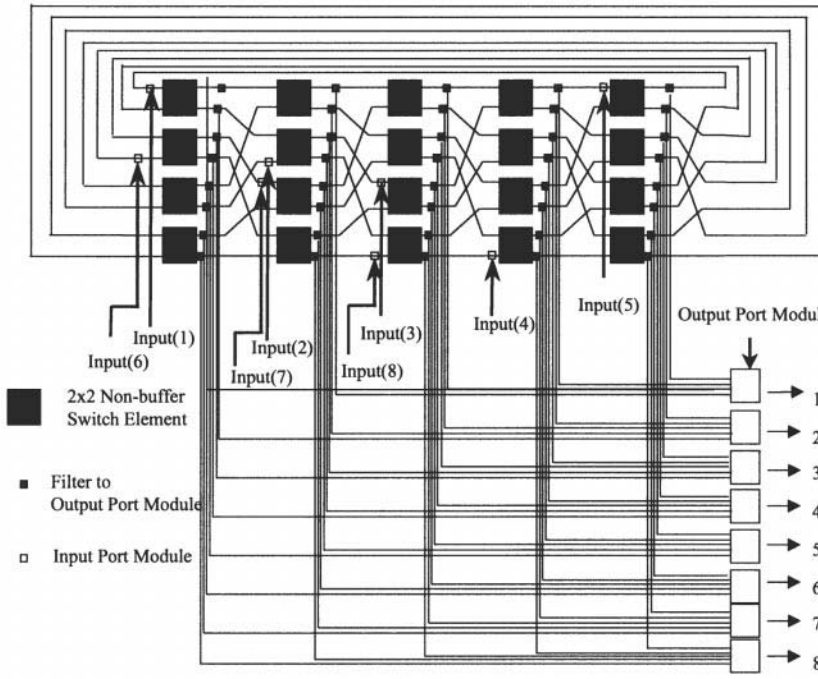

Fig. 4. Proposed switch

4000 x 4000. Considering today's VLSI technology, this size of switch will be constructed by two-stage interconnection of 64 x 64 switch implemented in a chip. On this assumption, size of the switch evaluated here is also set to 64 × 64.

In most cases, the switch performance strongly depends on patterns of applied traffic. In this section, two traffic conditions are prepared where destined output ports of applied packets are distributed in uniform and non-uniform pattern. Under these conditions, packet loss rate of the proposed switch is evaluated by computer simulation.

In our simulations, size of packets varies uniformly from 20 bytes to 1500 bytes according to the IP packet format, and time interval of packet arrival follows negative exponential distribution.

Packet loss shown in the following figures includes loss in the shuffle switch fabric and loss in the input modules, and it does not include loss in the output modules. It is because loss in output port module is approximately the same as other output buffering switches.

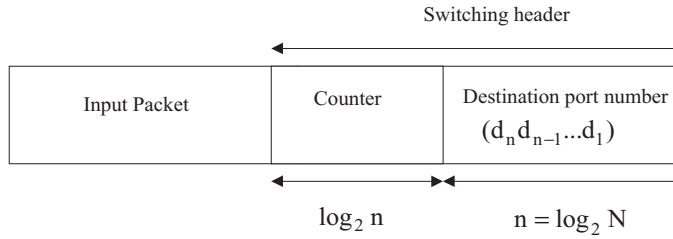

Fig. 5. Packet format

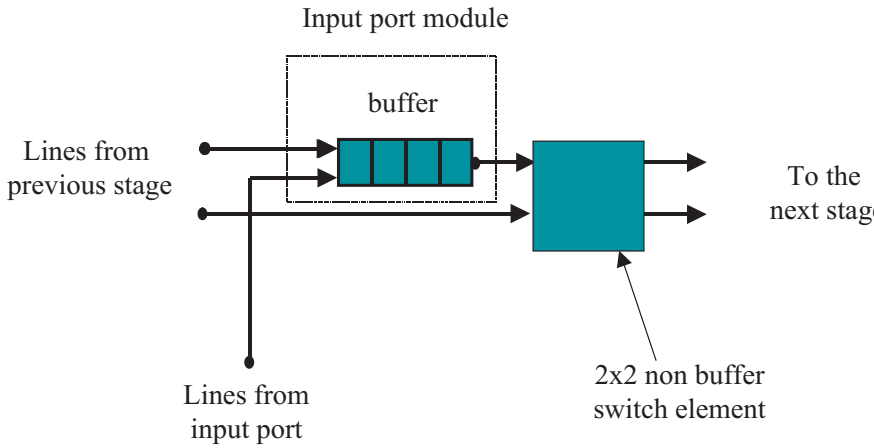

Fig. 6. Input module structure

4.1 Packet Loss Rate under Uniform Traffic Pattern

In this condition, the destination output port of a packet is uniformly distributed among all output ports.

Figure 7 plots the packet loss rate in the switch fabric as a function of K for $N = 64$ at load $p = 0.6$. The packet loss rate of the MS4 switch with the same traffic and switch conditions is also plotted for comparison. For example, a packet loss rate of 10^{-6} is achieved with $K = 24$ in the proposed switch while the MS4 switch needs 29 stages. Figure 8 illustrates probability distribution of required number stages by which a packets reaches the destination output p ort in the simulation. Under this condition, required number of stages is six in both switches when the packets do not suffer any conflict. The figure shows that probability that a packet arrives at the destination port by six stages in the proposed switch is 74%, while it is 38 % in MS4 switch. Therefore, it can be said that average transmission delay of packets in the proposed switch is less than

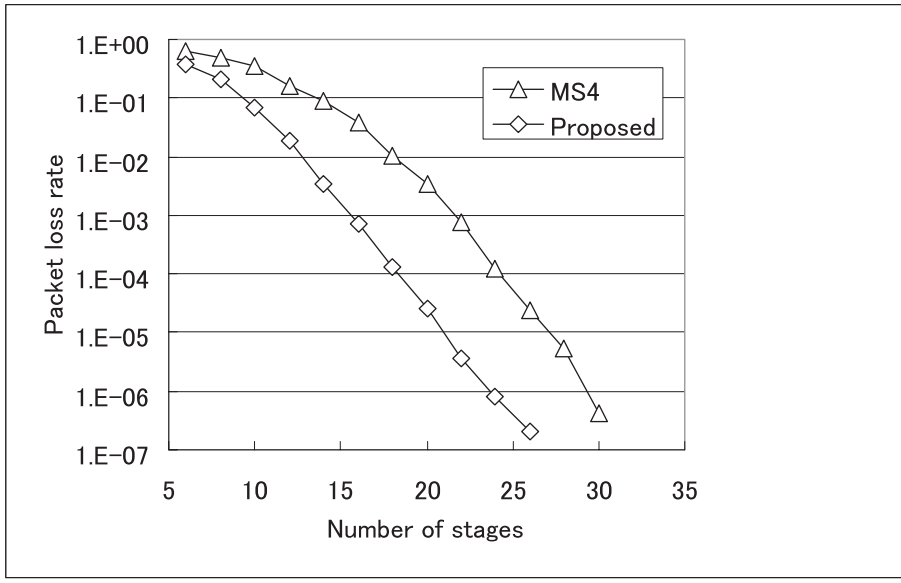

Fig. 7. Packet loss rate in the switch fabric under the uniform traffic

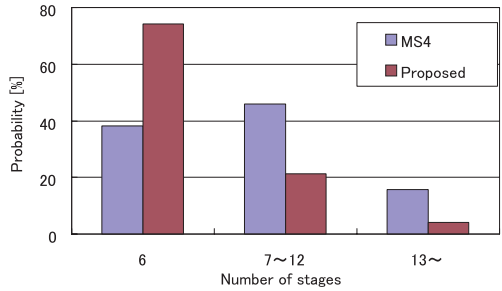

Fig. 8. Probability distribution of required number of stages by which a packet reaches its destination port

that of MS4 switch. This result shows that technique of distributing input port has great effectiveness in reducing probability of conflict within the network.

4.2 Packet Loss Rate under Non-uniform Traffic Pattern

All the non-uniform traffic patterns studied here are considered to satisfy properties shown below[8]:

– packets arrive at all input ports with the same probability p

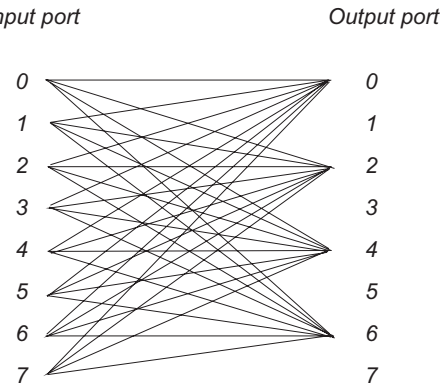

Fig. 9. Rectangular non-uniform traffic pattern (An example for 8 x 8 switch)

– arriving packets request a destination port according to a non-uniform distribution.

For the destined output port distribution, rectangular traffic pattern and hot-spot traffic pattern are provided.

Rectangular Traffic Pattern In this traffic pattern, all arriving packets are addressed to only a subset of all output ports. We consider the case that number of the ports is $L = N/2$ and the distance between the ports is the furthest. An example of the pattern for 8 x 8 switch is shown in Figure 9. Figure 10 shows the packet loss rate at load $p = 0.5$ as a function of K , where each of rectangular traffic pattern and uniform traffic pattern (UT) is offered. Though larger number of sta ges K required in rectangular pattern to achieve the same loss rate as UT, the difference can be ignored.

Hot Spot Traffic Next, the performance under hot spot traffic is studied. A hot-spot traffic is defined as a traffic pattern in which a single hot-spot of high access percentage is superimposed on a background of UT[9]. Supposed that h is the rate of packets addressed at the hot-spot, we consider the case of $h = 10\%$ and eimulation results are plotted in Figure 11. The figure shows that effect of hot spot traffic on increase of loss rate is quite small.

The above results attained by simulations under various non-uniform traffic patterns show that the distributed input shuffle switch not only achieves high performance under uniform traffic, but also it is robust under non-uniform traffic pattern which may rise in many real applications.

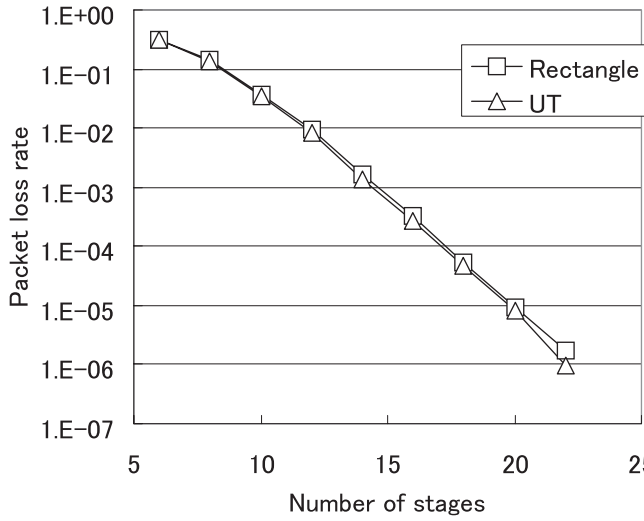

Fig. 10. Packet loss rate in the switch fabric under the rectangular traffic

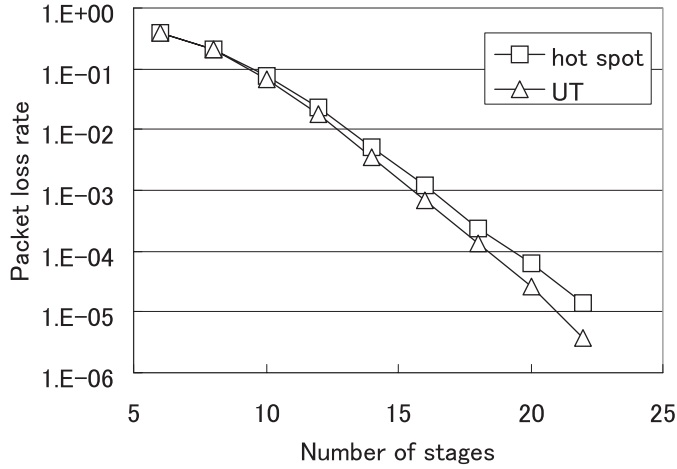

Fig. 11. Packet loss rate in the switch fabric where hot-spot traffic is superimposed

5 Conclusion

In this paper, we have introduced a deflection routing based packet switch using shuffle pattern network and technique of input ports distribution. By distributing

input ports over all stages of the switch, probability of internal packet conflict is reduced, and low packet loss rate is achieved. Moreover, computer simulations show that the switch is robust under non-uniform traffic pattern.

In a future work, several technical requirements for actual implementation of the switch will be investigated.

Acknowledgement

The authors would like to acknowledge the support of JSPS(JSPS-RFTF96P00601).

References

1. T.Aoki, "Post-ATM Network Architecture", Ph.D. Thesis, The University of Tokyo, Dec. 1997.
2. Nick McKeown, Martin Izzard, Adisak Mekkittikul, Bill Ellersick and Mark Horowitz, "The Tiny Tera: A Packet Switch Core" , IEEE Micro Jan/Feb 1997, pp. 26-33.
3. F.A.Tobagi, T.Kwok and F.M.Chiussi, "Architecture, Performance, and Implementation of the Tandem Banyan Fast Packet Switch", IEEE Journal on Selected Areas in Communications, Vol.9, No.8, Oct. 1991, pp.1173-1193.
4. Ra'ed Y Awdeh and H T Mouftah, "MS4-a high performance output buffering ATM switch", Computer communications, Vol.18, No.9, Sep. 1995, pp.631-644.
5. H.S.Stone, "Parallel processing with the Perfect Shuffle", IEEE Transactions on Computers, Vol.C-20, No.2, Feb. 1971, pp.153-161.
6. T.Aoki, H.Aida, T.Saito, "A multi stage interconnection switch for high capacity IP switching", Proc. of The 55th IPSJ conference Sep. 1997. (In Japanese).
7. Y.Yeh, M.G.Hluchyj, and A.S.Acampora, "The knockout switch: A simple, modular architecture for high-performance packet switching", IEEE Journal on Selected Areas in Communications, Vol.SAC-5, No.8, Oct. 1987, pp.1274-1283.
8. S. Bassi, M. Decina, A. Pattavina, "Performance analysis of the ATM Shuffleout switching architecture under non-uniform traffic patterns", *Proc. of IEEE INFOCOM*, Florence, Italy 1992, pp.734-742.
9. G.F.Pfister and V.A.Norton, "Hot spot contention and combining in multistage interconnection networks", IEEE Transactions on Computers, Vol.C-34, No.10, Oct. 1985, pp.943-948.