Frame Occupancy-based Round-Robin Matching Scheme for Input-Queued Packet Switches

Roberto Rojas-Cessa and Chuan-bi Lin
Department of Computer and Electrical Engineering,
New Jersey Institute of Technology, University Heights,
Newark NJ 07102 USA.
Email: {rrojas, cl23}@njit.edu

***Abstract*— The use of virtual output queues (VOQs) in input-queued (IQ) switches can eliminate the head-of-line (HOL) blocking phenomena, which limits switching performance. An effective matching scheme for IQ switches with VOQs must provide high throughput under admissible traffic patterns while keeping the implementation feasible. This paper proposes a matching scheme for IQ switches that provides high throughput under uniform and a nonuniform traffic pattern, called unbalanced. The proposed matching scheme is primarily based on round-robin selection and the captured-frame concept. We show via simulation that this scheme delivers over 99% throughput under unbalanced traffic and retains the high performance under uniform traffic that round-robin matching schemes are known to offer.**

***Index Terms*— input-queued switch, round-robin matching, virtual output queue, captured frame, eligible frame**

I. Introduction

Input-queued (IQ) switches have been of research interest for several years as these switches work without the speedup requirement of an output-queued (OQ) switch. Because of the feasible implementability with current technologies, IQ switch architectures have been widely adopted by manufacturers of switches/routers. The use of virtual output queues (VOQs), where one queue per output port is allocated at an input port, is known to remove the head-of-line (HOL) blocking problem from IQ switches. HOL blocking causes idle outputs to remain so, even in the existence of traffic for them at an idle input [1].

It is common to find the following practices in packet switch design: 1) segmentation of incoming variable-size packets at the ingress side of a switch to perform internal switching with fixed-size packets, or cells, and re-assembling the packets at the egress side before they depart from the switch; 2) use of VOQs, to avoid head-of-line (HOL) blocking; and 3) use of crossbar fabrics for implementation of packet switches because of their non-blocking capability, simplicity, and market availability. This paper follows these practices.

One major requirement for an input-queued switch is the delivery of high throughput under different traffic conditions. We consider admissible traffic [2], with Bernoulli and bursty arrivals and uniform and nonuniform distributions. A single-stage IQ switch, based on a crossbar switch fabric and VOQs, has the throughput performance dependable mainly on the used matching scheme. In general, matching schemes are required to provide: a) low complexity, b) fast contention resolution, c) fairness, and, d) high matching efficiency.

Maximum weight matching (MWM) have been used to show that IQ switches with VOQs can provide 100% throughput under admissible traffic [2]. MWM schemes can make a switch deliver 100% throughput by using a quantitative differentiation among contending VOQs, based on queue occupancy or cell waiting time. However, MWM schemes have intrinsically high computation complexity that is translated into long resolution time and high hardware complexity. The complexity makes these schemes prohibitely expensive for a practical implementation with currently available technologies. An alternative is to use maximal-weight matching schemes, also based on quantitative differentiation of queues. However, the hardware and time complexity of these schemes can be considered still high for the ever increasing data rates. In addition, a large number of iterations may be needed to achieve satisfactory matching results, driving the complexity even higher. Moreover, weight-based schemes, based in queue occupancy, may starve queues with little traffic to provide more service to the congested ones, therefore, presenting unfairness [4].

Maximal-size matching schemes are an alternative to weight-based schemes [3]. Schemes based on round-robin matching that deliver 100% throughput under uniform traffic have been proposed. Some examples are iSLIP [5], DRRM [6], [7], and SRR [10], just to mention some. iSLIP showed that the desynchronization effect, where arbiters reach the point where each of them prefers to match with different input/outputs, is beneficial for switching under uniform traffic. Moreover, the round-robin policy has a low implementation complexity, which is attractive for developing high speed switches. Other schemes have employed further the advantage of the desynchronization effect [10], [11], [12].

However, to our knowledge, cell-based schemes based on round-robin selections have not been shown to provide nearly

This work was supported in part by New Jersey Institute of Technology under Grant 421070.

Roberto Rojas-Cessa is the corresponding author. Email: rrojas@njit.edu

100% throughput under nonuniform traffic patterns without speedup or load-balancing stages [13]. The exhaustive round-robin matching (EDRRM) scheme [9] performs service exhaustion in an achieved match by keeping the match between a queue and an output port until the occupancy, produced by queued and arriving cells, of the matched queue is exhausted. This scheme has shown a throughput higher than iSLIP and DRRM under nonuniform traffic patterns at the cost of reduced performance under uniform traffic. Furthermore, frame-based matchings, where scheduling is performed on a set-of-cells basis instead of on cell basis, have been shown to have improved switching performance under different traffic scenarios [14], [16].

It is of interest to know if a round-robin based scheme can achieve throughput of nearly 100% under admissible traffic with nonuniform distributions, such as unbalanced traffic, using a single-stage switch, a single iteration, and no speedup.

This paper proposes a novel matching scheme for IQ switches, called frame occupancy-based round-robin matching (FORM), which is based on round-robin selection and on a capture frame concept. A frame is comprised by one or more cells that can be considered eligible for matching. The frame size associated with a VOQ is determined by the occupancy of the queue at the time when the previous frame has completed service. This paper shows that this arbitration scheme can achieve over 99% throughput, under a nonuniform traffic pattern, called unbalanced traffic [17], while keeping the high performance under uniform traffic that round-robin schemes are known to offer.

This paper is organized as follows. Section II presents the switch model under study and several definitions used in this paper. Section III introduces the proposed arbitration scheme. Section IV presents a simulation study of the throughput and delay performance of the resulting switch under uniform and nonuniform traffic patterns. Section V discusses the properties of the proposed matching scheme. Section VI presents the conclusions.

II. Switch Model and Preliminary Definitions

In this paper, we consider a single-stage IQ switch with N input and output ports. There are N VOQs at each input port. A VOQ at input port i that stores cells for output port j is denoted as $VOQ_{i,j}$. The following definitions are needed in the description of the proposed matching scheme.

Frame. A frame is related to a VOQ. A frame is the set of one or more cells in a VOQ that are eligible for matching. Only the HOL cell of the frame is eligible for matching at each time slot.

On-service status. A VOQ is said to be on-service status if the VOQ has a frame size of two or more cells and the first cell of the frame has been matched. An input is said to be on-service status if there is at least one on-service VOQ.

Off-service status. A VOQ is said to be off service if the last cell of the VOQ's frame has been matched (i.e., finished service) or no cell of the frame has been matched (i.e., not started service yet). Note that for a frame size of one cell, the associated VOQ is off-service during the matching of its one-cell frame. An input is said to be off-service if all VOQs are in off-service status.

Captured frame size. At the time t_c of matching the last cell of the frame associated to $VOQ_{i,j}$, the next frame is assigned a size equal to the minimum of the cell occupancy, denoted as $L_{i,j}(t_c)$, at $VOQ_{i,j}$ and a minimum limiting value f_m, where $1 \leq f_m \leq L_{i,j}(t_c)$. Cells arriving to $VOQ_{i,j}$ at time t_d, where $t_d > t_c$, are not considered for matching until the current frame is totally served and a new frame is captured.

In this paper, we consider two subsets of admissible traffic patterns: uniform and unbalanced traffic.

III. Frame Occupancy-based Round-robin Matching (FORM) Scheme

The proposed matching scheme is based on round-robin selection. For each output, there is an output arbiter a_j that selects a request among all received according to the policies described in the matching algorithm. For each input, there is an input arbiter a_i that accepts a grant among all received according to the policies described in the matching scheme. Each arbiter has a pointer that indicates the counter-part port with the highest priority position in a round-robin schedule.

For each VOQ there is a captured frame-size counter, $CF_{i,j}(t)$. We call this captured frame size as it is the equivalent of having a snapshot of the occupancy of a VOQ at a given time t, thus, the frame size is then equivalent to the occupancy at time t. The value of $CF_{i,j}(t)$, $|CF_{i,j}(t)|$, indicates the frame size; that is, the maximum number of cells that a $VOQ_{i,j}$ can have as candidates in the following and future time slots. $|CF_{i,j}(t)|$ takes a new frame-size value when the last cell of the current frame of $VOQ_{i,j}$ is matched. $|CF_{i,j}(t)|$ decreases its count by one each time a cell is matched, other than the last. VOQs are considered either on-service or off-service. All VOQs are initially considered with a frame size of one cell and in off-service status.

The arbitration process is as follows. This scheme follows request-grant-accept steps, as in the iSLIP algorithm [5]:

Step 1: Request. Non-empty on-service VOQs send a request to their destined outputs. Non-empty off-service VOQs send a request to their destined outputs only if the input is off-service.

Step 2: Grant. If an output arbiter a_j receives two or more requests, it chooses a request of an on-service VOQ (also called an on-service request) that appears next in a round-robin schedule, starting from the pointer position. If no on-service request exists, the output arbiter chooses an off-service request that appears next in a round-robin schedule, starting from its pointer position.

Step 3: Accept. If the input arbiter a_i receives two or more grants, it accepts one on-service grant that appears next in a round-robin schedule, starting from the pointer position. If none on-service grant exists, the arbiter chooses an off-service grant that appears next in round-robin schedule, starting from

its pointer position. The input pointers are updated to one position beyond the accepted ports. The output pointers are updated to one position beyond to the accepting port. In addition to the pointer update, the CF counter updates its value according to the following: If the input arbiter a_i accepts a grant from output arbiter a_j:

i) If $|CF_{i,j}(t)| > 1$: $|CF_{i,j}(t+1)| = |CF_{i,j}(t)| - 1$, and $VOQ_{i,j}$ is set as on-service.
ii) Otherwise (i.e., $|CF_{i,j}(t)| = 1$): $|CF_{i,j}(t+1)|$ is assigned the minimum of the occupancy of $VOQ_{i,j}$ and f_m, and $VOQ_{i,j}$ is set as off-service.

The variable f_m is a value to limit the captured frame size. Note that f_m may be equal to a constant or a variable value. In this paper, we use f_m as a constant. The frame size is used for determining the service status of a VOQ. Although the frame size is used to determine eligibility of a VOQ to participate in the matching process, matching is performed on time-slot basis. The value of f_m affects the performance of FORM in different traffic scenarios. We study the effects of using different f_m values in Section IV. Note that when $f_m = 1$, FORM becomes 1SLIP (iSLIP, with $i = 1$). The description above presents the matching procedure for a single iteration. FORM can consider multiple iterations. However, that is out of the scope of this paper.

IV. Performance Evaluation

We consider iSLIP (with one iteration, or 1SLIP) and EDRRM on this study for comparison purposes as our interest is to study the performance with a single iteration. The performance evaluations are produced through computer simulation. The traffic models considered have destinations with uniform and nonuniform distributions, the latter called unbalanced [17]. Both models use Bernoulli arrivals. The simulation does not consider the segmentation and re-assembly delays for variable size packets. Simulation results are obtained with a 95% confidence interval, not greater than 5% for the average cell delay. The VOQs are assumed of infinite capacity.

A. Uniform Traffic

Figure 1 shows simulation results of three 32×32 IQ switches, each one with a different matching scheme: 1SLIP, EDRRM, and FORM, all under uniform traffic with Bernoulli arrivals. In this figure, we use FORM, with $f_m = 2N$. This figure shows that FORM, as iSLIP, delivers 100% throughput under uniform traffic. FORM, with $f_m = 1$, is the equivalent of 1SLIP. Therefore, the average delay of FORM, with $F_m = 1$, is depicted by the $1SLIP$ curve. The desynchronization effect is also present in FORM under uniform traffic. This effect and the frame service policy allow FORM to deliver high throughput and low average cell delay under uniform traffic. The average cell delay of FORM is low as the frame consideration has an effect similar to having $f_m = 1$ and several iterations. After a frame starts being served, the VOQ in service will keep the match in a number of subsequent time slots equal to the frame size. This increases the number of matches by reducing the number of unmatched ports, resulting in a lower average delay than the other schemes.

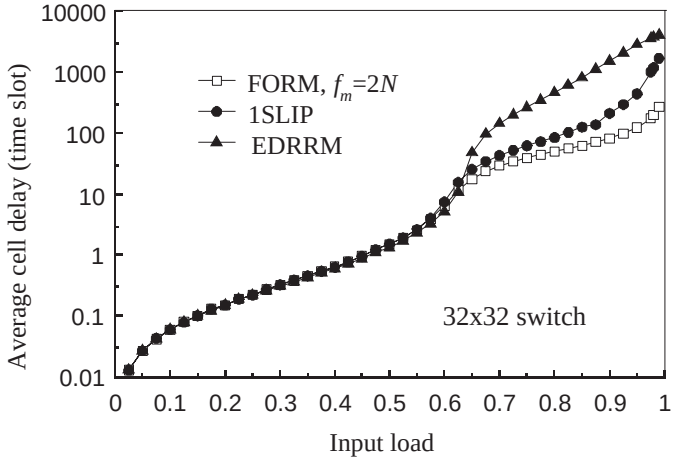

Fig. 1. Average delay of FORM scheme under Bernoulli uniform traffic

Figure 2 shows the average cell delay of switches of different sizes, all using FORM. In this case, we show the results for $f_m = 2N$. It can be seen that as the switch size increases, the average cell delay increases. However, in a load close to 1.0, small switches develop a long delay. Simulation experiments showed that small switches, $N = \{4, 8\}$, have higher performance when $f_m \leq N$, and larger switches are less sensitive to the f_m value. This figure shows that the performance of FORM with an intermediate f_m value, $f_m = 2N$, for both small and large switch sizes, is high in all cases.

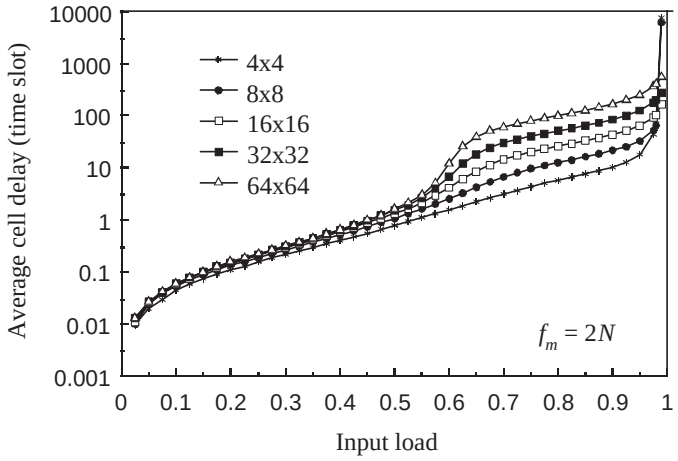

Fig. 2. Average delay of FORM in function of switch size, under Bernoulli uniform traffic

Figure 3 shows FORM with $f_m = 2N$ and an OQ switch under bursty traffic, modeled as on-off modulated Markov process, with average burst length l. The traffic has bursts with average lengths of 16 and 32 cells ($l = 16$ and $l = 32$), and Bernoulli traffic, $l = 1$. The simulation shows that the FORM scheme provides 100% throughput under uniform traffic under

Bernoulli and bursty arrivals. The curves for $l = 16$ and $l = 32$ show a constant delay of FORM over the OQ average cell delay. This constant delay is proportional to the burst length. Therefore, the frame concept used in FORM is not affected by bursty traffic.

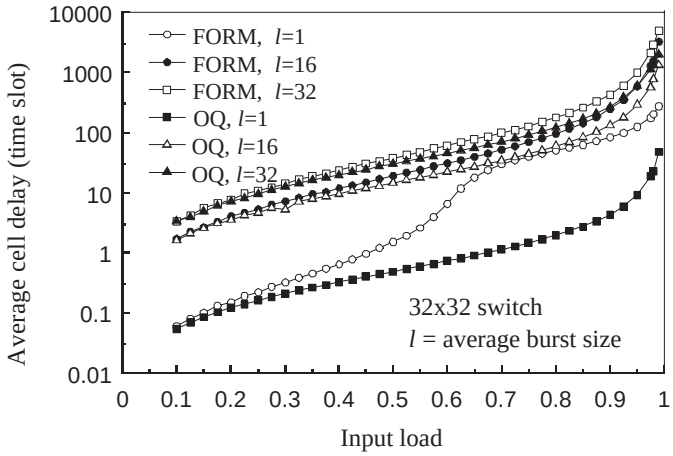

Fig. 3. Average delay of FORM, with $f_m = 2N$, under bursty uniform traffic

Under uniform traffic, the average frame size is small, as the uniform distribution of traffic among VOQs results in small average queue occupancies. Note that FORM does not suffer from VOQ starvation, even in the case when VOQ occupancy and f_m have large values, as the captured frame has a finite size, and the arrival of new cells does not affect the CF value arbitrarily.

B. Nonuniform Traffic

The study presented in this section uses a nonuniform traffic model, the unbalanced traffic [17]. The unbalanced traffic model uses a probability, w, as the fraction of input load directed to a single predetermined output, while the rest of the input load is directed to all outputs with uniform distribution. Let us consider input port s, output port d, and the offered input load for each input port ρ. The traffic load from input port s to output port d, $\rho_{s,d}$ is given by,

$$\rho_{s,d} = \begin{cases} \rho\left(w + \frac{1-w}{N}\right) & \text{if } s = d \\ \rho\frac{1-w}{N} & \text{otherwise.} \end{cases} \quad (1)$$

When $w = 0$, the offered traffic is uniform. On the other hand, when $w = 1$, it is completely directional, from input i to output j, where $i = j$. This means that all traffic of input port s is destined for only output port d, where $s = d$.

Three switches, of size 32, are considered. Each switch uses a different matching scheme: 1SLIP, EDRRM, and FORM. Figure 4 shows the throughput performance of 1SLIP, EDRRM, and FORM under unbalanced traffic. This figure shows that FORM ($f_m = 3N$) provides over 99% throughput under the complete range of w. This shows an improvement to the matching efficiency of FORM compared to the other schemes. The high throughput of FORM under this traffic

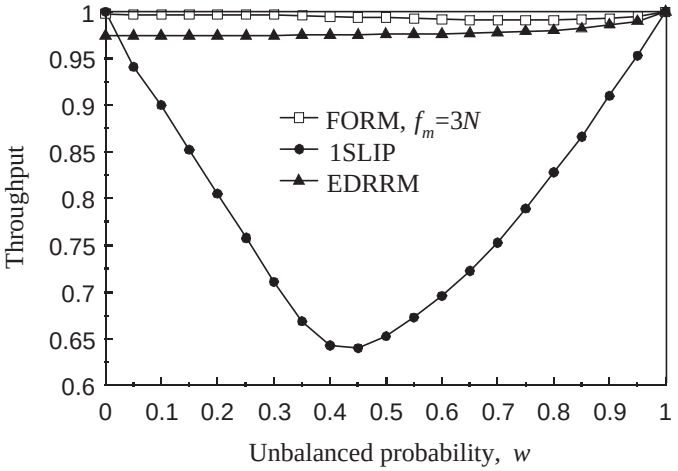

Fig. 4. Throughput performance of FORM under unbalanced traffic

model is the product of considering the VOQ occupancy. The occupancy of that queue can be expected to have a length in proportion to its received service and to the arrival rate. FORM ensures service to queues with high load by capturing a large frame size for each, and to the queues with low load by using round-robin selection.

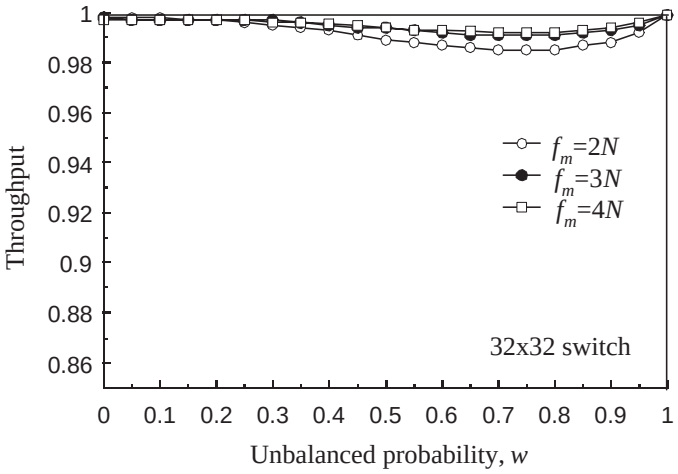

Fig. 5. Throughput performance of a 32x32 switch for different f_m values

Figure 5 shows a 32×32 switch with FORM under unbalanced traffic. This graph shows that for $f_m > 2N$, the throughput under unbalanced traffic is higher than 99%. Note that the lowest throughput value along the w range is the one considered.

To illustrate the dependency of N, Figure 6 shows the throughput of FORM for different switch sizes, $N = \{4, 8, 16, 32, 64\}$, where $f_m = 2N$ for switches of sizes $N = \{4, 8, 16\}$ and $f_m = 4N$ for switches of sizes $N = \{32, 64\}$. The figure shows that the smaller switches offer high performance (nearly 99% throughput) when $f_m = 2N$, while larger size switches offer higher performance with rather larger values of f_m. In this case, a 32×32 switch offers a throughput above 99% under this traffic model with $f_m > 2N$. As the

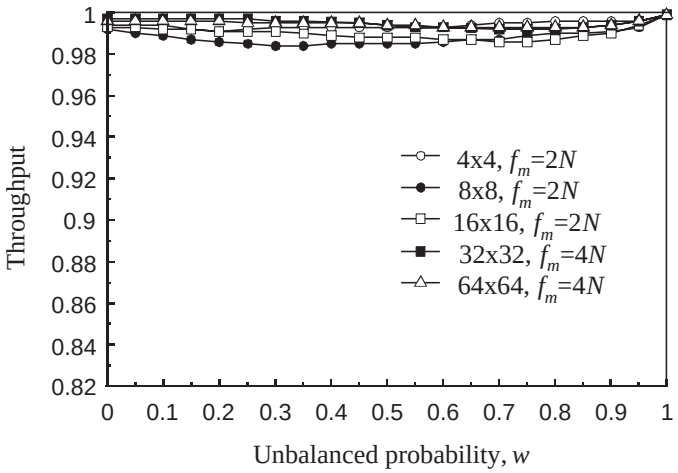

Fig. 6. Throughput performance of FORM for different switch sizes under unbalanced traffic

switch size increases, FORM is less sensitive to the f_m value for delivering high throughput. The decreased dependency on f_m with the increase of the switch size was observed not only under unbalanced traffic, but also under uniform traffic.

V. Properties of FORM

The use of a captured frame size and the service concepts used here make FORM deliver high performance under uniform and unbalanced traffic patterns. FORM with $f_m = 1$ is the particular case of 1SLIP. When f_m is large, the selected frame size is then the VOQ occupancy. Note that in the case where a VOQ has no cells at the capturing time, the VOQ can still participate in a matching after a new cell arrives and as long as the input is off-service.

When a VOQ changes its status to on-service, that VOQ has higher priority than the others to continue sending its request in subsequent time slots. Therefore, the pair $i-j$ matched are latched for the duration of the frame. When an input is off-service, all nonempty VOQs (independently of their CF value) send a request to their respective outputs.

Under uniform traffic, the captured frame sizes are not expected to reach large values because of the cell distribution among all queues. Therefore, most queues may remain in off-service status while completing service for one-cell frames. The performance is mainly determined by the round-robin policy. Under unbalanced traffic, some queues are expected to have heavier loads than others. The queues with large occupancies have a higher service than the queues with lower occupancy. The difference on frame sizes results in more service for queues with a larger number of arrivals than those for queues with a small number of arrivals. Moreover, the round-robin policy ensures that all queues receive service.

The implementation complexity of FORM is low because of the following reasons: the arbitration scheme is round-robin based, FORM performs no comparisons among different queues, and the hardware additions are the CF counters to each VOQ and a service flag.

VI. Conclusions

This paper introduced a novel matching scheme, FORM, for input-queued packet switches. This scheme is based on round-robin selection and uses the concept of captured frame size, where the frame size depends on VOQ occupancy at complete-service time. In this paper, we presented a study when the maximum frame size is limited to several constant values. As the switch size increases, FORM shows above 99% throughput under unbalanced traffic models, while retaining the high performance of round-robin based schemes under uniform traffic. This matching scheme does not need to compare the status of different VOQs as it is based on simple round-robin. The hardware and timing complexity of FORM is low. This makes FORM an efficient and implementable scheme.

References

[1] M. Karol, M. Hluchyj, "Queuing in High-performance Packet-switching," *IEEE J. Select. Area Commun.*, vol. 6, pp. 1587-1597, December 1988.

[2] N. McKeown, A. Mekkittikul, V. Anantharam, J. Walrand, "Achieving 100% Throughput in an Input-queued Switch," *IEEE Trans. Commun.*, vol. 47, no. 8, pp. 1260-1267, August 1999.

[3] T.E. Anderson, S.S. Owicki, J.B. Saxe, and C.P. Tacker, "High-speed Switch Scheduling for Local Area Networks," ACM Trans. on Computer Systems, vol. 11, no. 4, pp. 319-352, November 1993.

[4] N. McKeown, "Scheduling algorithms for input-queued cell switches," Ph.D. dissertation, Dept. Elect. Eng. Comput. Sci., Univ. California at Berkeley, Berkeley, CA, 1995.

[5] N. McKeown, "The iSLIP scheduling algorithm for Input-queued Switches," IEEE/ACM Trans. Networking, vol. 7, no. 4, pp. 188-201, April 1999.

[6] H.J. Chao, J-S. Park, "Centralized Contention Resolution Schemes for a large-capacity Optical ATM Switch," IEEE ATM Workshop 1998, pp. 11-16, May 1998.

[7] E. Oki, R. Rojas-Cessa, and H. J. Chao, "PMM: A Pipelined Maximal-Sized Matching Scheduling Approach for Input-Buffered Switches," IEEE Globecom 2001, pp. 35-39, November 2001.

[8] G. Nong, M. Hamdi, J.K. Muppala, "Performance evaluation of multiple input-queued ATM switches with PIM scheduling under bursty traffic," IEEE Trans. on Commun., Vol. 49, Issue 8, pp. 1329-1333, Aug. 2001.

[9] Y. Li, S. Panwar, H.J. Chao, "The Dual Round-robin Matching Switch with Exhaustive Service," IEEE HPSR 2002, pp. 58-63, May 2002.

[10] Y. Jiang and M. Hamdi, "A fully Desynchronized Round-robin Matching Scheduler for a VOQ Packet Switch Architecture," IEEE HPSR 2001, pp. 407-411, May 2001.

[11] Y. Jiang and M. Hamdi, "A 2-stage Matching Scheduler for a VOQ Packet Switch Architecture," IEEE ICC 2002, vol. 4, pp. 2105-2110, May 2002.

[12] D. N. Serpanos and P. I. Antoniadis, "FIRM: A Class of Distributed Scheduling Algorithm for High-speed ATM Switch with Multiple Input Queues," INFOCOM 2000, vol. 2, pp. 548-555, March 2000.

[13] C-S. Chang, D-S. Lee, and Y-S. Jou, "Load Balanced Birkhoff-von Newman Switches," IEEE HPSR 2001, pp. 276-280, May 2001.

[14] A. Bianco, M. Franceschinis, S. Ghisolfi, A.M. Hill, E. Leonardi, F. Neri, R. Webb, "Frame-based Matching Algorithms for Input-queued Switches," IEEE HPSR 2002, pp. 69-76, May 2002.

[15] H.J. Chao, S.Y. Liew, and Z. Jing, "A dual-level Matching algorithm for 3-stage Clos-network Packet Switches," 11th Symposium on High Performance Interconnects, pp. 38-44, August 2003.

[16] S. Li and N. Ansari, "Input-queuing Switching with QOS Guarantees," IEEE INFOCOM 1999, vol.3, pp. 1152-1159, March 1999.

[17] R. Rojas-Cessa, E. Oki, Z. Jing, and H. J. Chao, "CIXB-1: Combined Input-One-cell-crosspoint Buffered Switch," IEEE HPSR 2001, pp. 324-329, May 2001.