Decomposed Crossbar Switches with Multiple Input and Output Buffers[†]

Seung Yeob Nam and Dan Keun Sung
Dept. of EECS, KAIST, 373-1, Kusong-dong, Yusong-gu, Taejon, 305-701, KOREA

Abstract— **Conventional input switches usually employ a single crossbar switch fabric to transfer cells from input buffers to output ports. This type of switches suffer from input and output cell contention problems which cause lower performance than for output buffer switches. However, dividing one crossbar fabric into several smaller crossbar fabrics, we can decrease the input and output contention probabilities. Based on this principle, we propose a new decomposed crossbar switch architecture. Since a decrease in input and output contention probabilities causes an increase in the grant probability for the cells at input buffers, the proposed decomposed crossbar switches yield better performance than conventional input switches. We derive the grant probability for a simple arbitration algorithm and evaluate the performance of the proposed switch architecture in terms of the average cell latency through simulation.**

I. Introduction

Recently, input queuing schemes have been adopted in many high-speed switching systems [1][2] because they can increase the throughput of input-buffered switches from 58.6% to 100% using a virtual output queuing (VOQ) scheme [3][4].

Even if input buffered switches have virtual output queues at each input port, there remains a problem that a head-of-line (HOL) cell at a VOQ may contend with the other HOL cells belonging to different VOQs because even though several HOL cells intend to pass through a switch fabric, only one cell from each input can pass the switch fabric during one cell time. This contention is called the input contention. In addition, even if the cells entering the switch fabric from different input ports may be destined to the same output port, only one cell is permitted to the output port during one cell time. Thus, these cell contentions called the output contention occur at the output port of the switch fabric.

In general, proper arbitration algorithms are required in order to solve the contentions at input and output ports and to increase the utilization of the switch. There have been a number of studies on arbitration algorithms: parallel iterative matching (PIM) [2], round-robin matching (RRM), iSLIP [5], FCFS in round-robin matching (FIRM) [6], wave front arbitration (WFA) [7], 2-dimensional round-robin (2DRR) [8], etc.

Although some efficient algorithms achieve 100% throughput [3][4][5], input buffered switches can not yield so good performance as output buffered switches due to input/output contentions which do not occur for output buffered switches. Since an incoming cell is promptly delivered to the destined output buffer for output buffer switches, the cell experiences queuing at the output buffer and thus, work-conserving service can be provided at each output port. However, for input buffered swtiches, an incoming cell may not be delivered to the output buffer because of contentions occuring at the ingress and egress points of the switch fabric. Therefore, all the input buffers can not always provide work-conserving service. This causes a degradation in the performance of input buffered switches, compared with output buffered switches.

In this paper we propose a new switch architecture which can decrease the contention probability. Conventional input buffered switches uaually employ one switch fabric. However, if we divide one crossbar switch fabric into several smaller crossbar fabrics, we can decrease the contention probability occuring at the ingress or egress points of the switch. Depending on how we divide one crossbar into a number of switching components, the input contention probability or the output contention probability, or both contention probabilities can be decreased. We derive the grant probability for a simple PIM arbitration algorithm and consider the effect of dividing the crossbar into a number of smaller crossbar fabrics on the grant probability.

This paper is organized as follows: In Section II, we propose a new switch architecture which consists of everal small crossbar fabrics and multiple input and output buffers. In Section III, we derive the grant probability for the proposed switch using the PIM algorithm and investigate the effect of dividing the swtich fabric into a number of smaller switch fabrics on the grant probability. In Section IV, we evaluate the performance of the proposed switch by simulation. Finally, we present conclusions in Section V.

II. Decomposed Crossbar Switch Architecture

Fig. 1 shows a conventional $N \times M$ input buffered switch architecture which adopts virtual output queueing (VOQ). Cells arriving at the input buffer are queued according to their destined output port number. $Q_{i,j}$ denotes the virtual output queue that stores the cells passing from the i-th input port to the j-th output port.

The crossbar fabric of this $N \times M$ input buffered switch should manage cell switching from all the input ports to all output ports. Thus, in order to obtain a permission for the destined output port, the HOL cell of a non-empty VOQ should compete with other HOL cells of non-empty VOQs belonging to the same input port. Even if it acquires a priority at its input buffer, it should compete again with the cells from other input ports. Thus, the contention rate increases as the number of input or output ports increases regardless of the arbitration algorithm used.

We propose a new switch architecture which yields high performance by lowering the number of competitors and consequently, decreasing the contention probability. Fig. 2 shows a decomposed crossbar switch with multiple output buffers. The performance can be improved at the cost of adding output buffers, compared with conventional input buffered switches. However, output buffers used in the decomposed crossbar switch do not need to operate at the speed higher than the input link rate, and both the output and the input memory speeds are two times faster than the input link speed for read and write operations in this pa-

[†] This study was supported in part by the Ministry of Information and Communications.

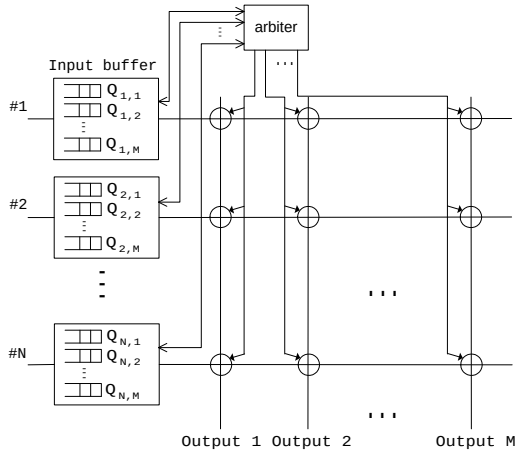

Fig. 1. An $N \times M$ input buffered crossbar switch

per. L seperate buffer memories are allocated for each output port. If L is equal to 1, then the switch architecture becomes a well-known combined input and output queued switch [9][10]. Let us assume that L is a divisor of N. Then, input ports with port numbers of $1, 2, \ldots, N/L$ only access the first buffer for every output port. In general, input ports with port numbers from $N(s-1)/L+1$ to Ns/L send their requests only to the s-th buffer of every output port. If we form a group from input ports $N(s-1)/L+1, N(s-1)/L+2, \ldots, Ns/L$ for each s and associate an identifier G_s with that group, all input ports belonging to group G_s access only the s-th buffer for every output port independent of other groups.

If R denotes the request matrix for an $N \times M$ input buffered switch, it can be expressed as

$$R = \begin{bmatrix} R_{1,1} & R_{1,2} & \cdots & R_{1,M} \\ R_{2,1} & R_{2,2} & \cdots & R_{2,M} \\ \vdots & & \ddots & \vdots \\ R_{N,1} & R_{N,2} & \cdots & R_{N,M} \end{bmatrix} \quad (1)$$

where $R_{i,j}$ is equal to 1 whenever $Q_{i,j}$ is non-empty and is equal to zero, otherwise.

On the other hand, the request matrix for the switch shown in Fig. 2 can be expressed as

$$R_{MOB} = \begin{bmatrix} R_{G1} \\ R_{G2} \\ \vdots \\ R_{GL} \end{bmatrix}, \quad (2)$$

where R_{Gi} is defined as

$$R_{Gi} = \begin{bmatrix} R_{N(i-1)/L+1,1} & R_{N(i-1)/L+1,2} & \cdots & R_{N(i-1)/L+1,M} \\ R_{N(i-1)/L+2,1} & R_{N(i-1)/L+2,2} & \cdots & R_{N(i-1)/L+2,M} \\ \vdots & & \ddots & \vdots \\ R_{Ni/L,1} & R_{Ni/L,2} & \cdots & R_{Ni/L,M} \end{bmatrix}. \quad (3)$$

Thus, introducing multiple output buffers for each output port, we can change one bipartite matching problem into

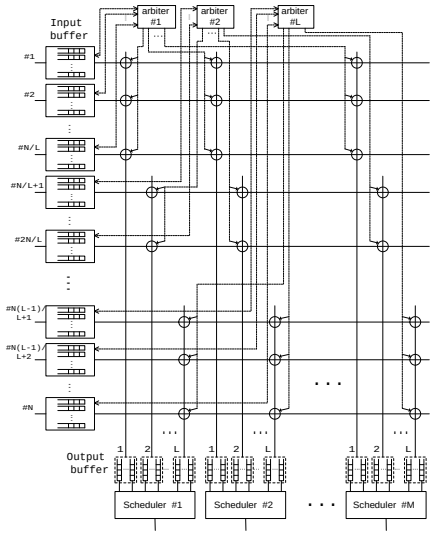

Fig. 2. Decomposed crossbar switch architecture with multiple output buffers

several independent, smaller matching problems. Independent matching problems can be solved simultaneously. A smaller matching problem is easier to solve and furthermore, increases the matching probability.

We discuss the effect of scaling down the matching problem in more detail. Thus far we have considered scaling down the matching problem by introducing multiple output buffers per each output port. However, there is another way to scale down the matching problem. In other words, we can observe a similar effect by dividing input buffers of input buffered switches into several seperate buffers. If we generalize this approach to scaling down the matching problem, the switch architecture shown in Fig. 3 can be derived. Fig. 3 shows a decomposed crossbar switch architecture with multiple input and output buffers. An $N \times M$ crossbar fabric is first divied into L $(N/L \times M)$ switches just as shown in Fig. 2, where L is the number of output buffer memories for each output port. Next, each $N/L \times M$ crossbar is divided into E $(N/L \times M/E)$ crossbar fabrics, where E is the number of input buffer memories for each input port. Consequently, the switching fabric consists of LE $(N/L \times M/E)$ crossbar fabrics. For this switch, the request matrix R_{MIOB} is written as

$$R_{MIOB} = \begin{bmatrix} R_{G11} & R_{G12} & \cdots & R_{G1E} \\ R_{G21} & R_{G22} & \cdots & R_{G2E} \\ \vdots & & \ddots & \vdots \\ R_{GL1} & R_{GL2} & \cdots & R_{GLE} \end{bmatrix}, \quad (4)$$

where R_{Gij} is defined as

$$R_{Gij} = \begin{bmatrix} R_{N(i-1)/L+1,M(j-1)/E+1} & \cdots & R_{N(i-1)/L+1,Mj/E} \\ R_{N(i-1)/L+2,M(j-1)/E+1} & \cdots & R_{N(i-1)/L+2,Mj/E} \\ \vdots & \ddots & \vdots \\ R_{Ni/L,M(j-1)/E+1} & \cdots & R_{Ni/L,Mj/E} \end{bmatrix}. \quad (5)$$

One seperate arbiter and one seperate crossbar unit are allocated for each R_{Gij}. If G_{ij} denotes the group of VOQs

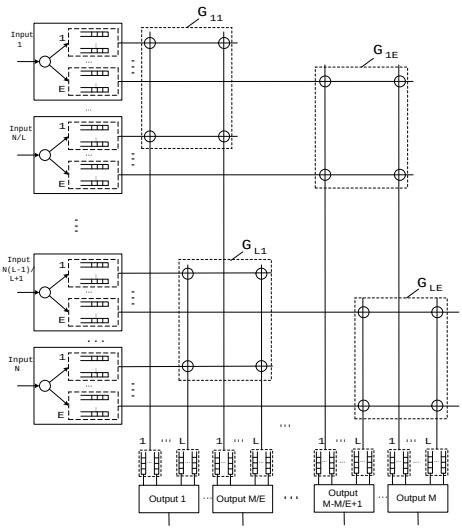

Fig. 3. Decomposed crossbar switch architecture with multiple input and output buffers

whose request belongs to R_{Gij}, all VOQs belonging to G_{ij} share the same arbiter and the same crossbar unit. Then, since each group uses one arbiter and one crossbar unit exclusively irrespective of other groups, arbitration can be performed for a smaller group of VOQs. Thus, introducing multiple input buffers for each input port, we can further scale down the size of a crossbar unit and the size of matching problem.

III. The Effect of Switch Fabric Decomposition upon The Grant Probability

We derive the grant probability for the generic decomposed crossbar switches described at the previous section. We assmue that cells arrive at each input port according to a Bernoulli process with parameter λ, where λ is the probability that a cell arrives at an arbitrary time slot for each input port. We also assume cells are uniformly distributed for all the output ports. We consider a PIM arbitration algorithm for each crossbar unit for mathematical tractability. Under these conditions we analyticaly derive the probability that the HOL cell of a particular VOQ acquires a permission to be transfered to its destined output port.

In order to derive the grant probability as a function of the offered load λ, we first need to find the request probability for the offered load λ. For the switch shown in Fig. 3, we put $\hat{N} = N/L$ and $\hat{M} = M/E$. Because of the uniform traffic arrival condition, if the offered load for an input port is λ, the offered load for each VOQ belonging to that input port becomes λ/M. If the probability that the HOL cell of a non-empty VOQ acquires a permission for switching in the very time slot is assumed to be P_g in the steay state, then each VOQ can be modeled as a Geo/Geo/1 system. For that system, the cell arrival probability is λ/M and the service probability is P_g.

Let Π_n be the probability that the VOQ is occupied by n cells at an arbitrary time slot. Π_n can be solved using a discrete time Markov-chain (DTMC) as follows:

$$\begin{aligned}\Pi_0 &= 1 - \frac{\lambda}{MP_g}, \\ \Pi_i &= \frac{1-\lambda/(MP_g)}{1-P_g}\left\{\frac{(1-P_g)\lambda/M}{(1-\lambda/M)P_g}\right\}^i, \text{ if } i \geq 1 \end{aligned} \quad (6)$$

Since a VOQ sends a request to its corresponding arbiter whenever it has at least one cell to transmit, the request probability of each VOQ is $1 - \Pi_0$. We now derive the grant probability of the HOL cell of a non-empty VOQ as a function of the request probability $1 - \Pi_0$. Let us consider the grant probability of an arbitrary VOQ $Q_{a,b}$ belonging to a group G_{st}. Let V and W denote the number of other non-zero requests belonging to the same column as the request from $Q_{a,b}$ excluding $R_{a,b}$ and the number of other non-zero requests belonging to the same row as $R_{a,b}$ excluding $R_{a,b}$, respectively. The PIM algorithm first performs an arbitration for each column of R_{Gst} and randomly selects at most one winning request for each column. Then, the PIM algorithm performs an arbitration for each row of R_{Gst} among the winning requests and also randomly selects at most one winning request for each row. E_{col_sel} denotes the event that the request of interest $R_{a,b}$ is selected at the column arbitration stage. The grant probability P_g of the request $R_{a,b}$ can be expresses as:

$$P_g = P(grant|E_{col\_sel})P(E_{col\_sel}), \quad (7)$$

where

$$\begin{aligned} P(E_{col\_sel}) &= \sum_{i=0}^{\hat{N}-1} P(E_{col\_sel}|L=j)P(L=j) \\ &= \sum_{i=1}^{\hat{N}-1} \frac{1}{j+1}\binom{\hat{N}-1}{j}(1-\Pi_0)^j \Pi_0^{\hat{N}-1-j} \\ &= \frac{1}{\hat{N}}\frac{1-\Pi_0^{\hat{N}}}{1-\Pi_0}. \end{aligned} \quad (8)$$

$$P(grant|E_{col\_sel}) = \frac{1}{P(E_{col\_sel})}\frac{1-\{(1-\Pi_0)(1-P(E_{col\_sel}))+\Pi_0\}^{\hat{M}}}{\hat{M}(1-\Pi_0)}. \quad (9)$$

Combining (7), (8), and (9) yields the following equation:

$$P_g = \frac{1-\left\{1-\frac{1}{\hat{N}}(1-\Pi_0^{\hat{N}})\right\}^{\hat{M}}}{\hat{M}(1-\Pi_0)}. \quad (10)$$

By solving (6) and (10) simultaneously for a given offered load λ, we can obtain the grant probability P_g. However, it is necessary to note that (6) is valid only when P_g is larger than λ/M. When P_g is not larger than λ/M, the Geo/Geo/1 system becomes unstable. Thus, we can obtain P_g from (6) and (10) only when $P_g > \lambda/M$. However, it is possible to evaluate P_g in other case. When P_g is less than λ/M, the

queue occupancy increases indefinitely. Thus, Π_0 becomes zero in that case. Substituting zero for Π_0 in (10) yields the grant probability in the case that $P_g < \lambda/M$ as follows:

$$P_g = \frac{1}{\hat{M}} \left\{ 1 - \left(1 - \frac{1}{\hat{N}} \right)^{\hat{M}} \right\}. \qquad (11)$$

In order for P_g to be a continuous function with respect to λ, the value of P_g should also be defined as (11) for $P_g = \lambda/M$. Thus, we can finally summarize the result as follows: If λ is less than $\{1 - (1 - 1/\hat{N})^{\hat{M}}\}M/\hat{M}$,then the grant probability can be obtained from (6) and (10). Otherwise, the grant probability can be obtained from (11).

We now evaluate the effect of decomposing one crossbar fabric into several smaller crossbar units on the grant probability using the analytic results obtained previously. Fig. 4 shows the grant probabilities for a 16×16 decomposed switch with multiple input buffers per each input port. The grant probabilities are obtained from the numerical solution of (6) and (10) in the non-heavy load case and are obtained from (11) in the heavy load case. The grant probability increases as the number of input buffers per each input increases, and the grant probability improves significantly by dividing an input buffer into just two, compared with the conventional input buffered switches using the same arbitration algorithm. For the flat interval in Fig. 4, the system is unstable because the service probability is lower than the cell arrival probability. Thus, the offered load corresponding to the starting point of this flat region implies the maximum throughput of the switch.

Fig. 5 shows the grant probabilities for the decomposed switches with multiple output buffers per each output port. In this case we can observe the similar result that the grant probability improves as the number of output buffers per each output increases. Fig. 6 compares the grant probabilities of the decomposed switches with those of the switches with multiple input or output buffers. The grant probability improves much more if we divide the crossbar fabric both horizontally and vertically.

IV. Simulation Results

The performance of the proposed switch architecture is evaluated through simulation for an 8×8 switch. One traffic source is connected to each input port and the destined output ports of generated cells are randomly selected among 8 output ports.

Input traffic models for simulation include random and bursty traffic. For random traffic cell arrivals at each input port are generated according to a Bernoulli process with parameter λ, where λ is the offered load per each input port. Bursty traffic is modeled by an *on-off* arrival process where the *on* and *off* interval lengths are exponentially distributed with different parameters. The source alternately generates a burst of cells followed by an idle period of no cells. During the *on* period cells are generated at the link rate and the destined output ports of the cells belonging to the same *on* period are all identical. The average burst length is set to 16 cells.

The performance of the decomposed crossbar switches is compared with those of the input buffered switch, which uses the iSLIP and the wrapped wave-front arbitration

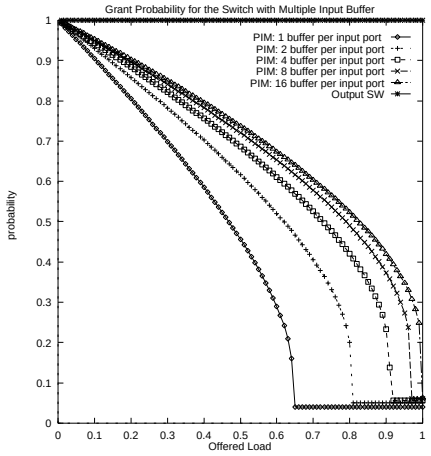

Fig. 4. Grant probabilities for the decomposed crossbar switches with multiple input buffers

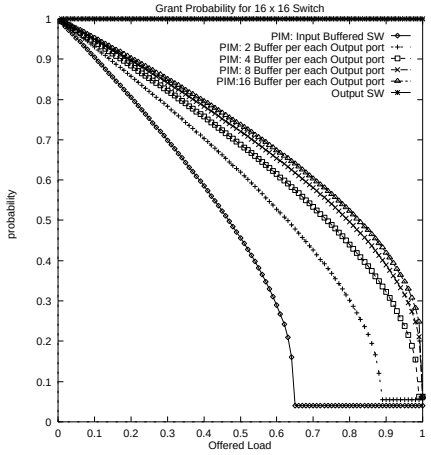

Fig. 5. Grant probabilities for the decomposed crossbar switches with multiple output buffers

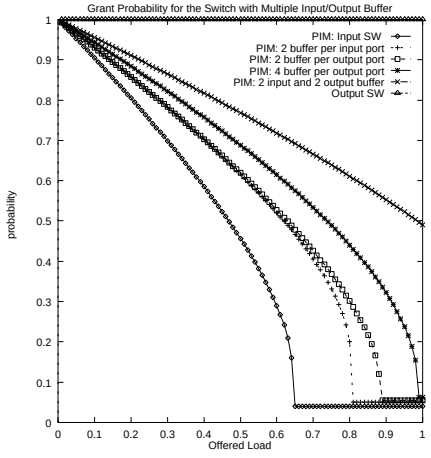

Fig. 6. Grant probabilities for the decomposed crossbar switches with multiple input/output buffers

(WWFA) arbitration algorithms, and an output buffered switch. We consider 3 types of decomposed crossbar switches. The first one uses two 8×4 crossbar units and allocates two input buffers for each input port. The second one uses two 4×8 crossbar units and allocates two output buffers for each output port. The last one uses four 4×4 crossbar units and allocates two input buffers for each input and two output buffers for each output. The WWFA algorithm is used for all types of decomposed crossbar switches [7].

Fig. 7 compares the average delay distributions for three different types of switches under a Bernoulli traffic load. It takes about $\log_2 N$ iterations for iSLIP to converge for an $N \times N$ switch [5]. We can observe that the WWFA algorithm yields slightly worse performance compared with the 3-SLIP algorithm (iSLIP with 3 iterations). However, dividing input buffers into two and a switch fabric into two crossbar units, the decomposed switch yields much better performance than the other input buffer switches. We can observe a similar performance improvement in the case of dividing a switch fabric into two crossbar units horizontally and putting two output buffers for each output port. More improvement is achieved by dividing one switch fabric both horizontally and vertically.

Fig. 8 compares the average delay for the same set of switches under a bursty traffic load. The average delay is longer than that under the random traffic because of the burstiness of the input traffic. However, we can observe a similar trend to Fig. 7. Dividing one crossbar into a few small crossbar units improves the switch performance in terms of the average delay.

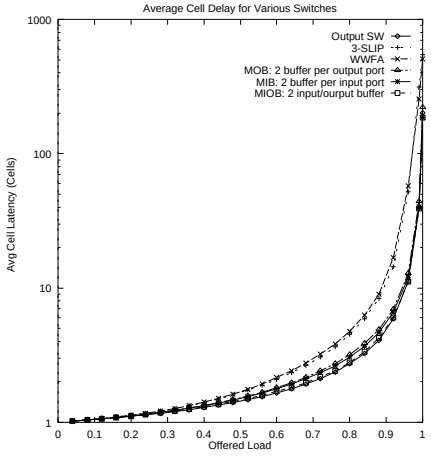

Fig. 7. Comparison of decomposed switches with the conventional input buffered and output buffered switches under random traffic load

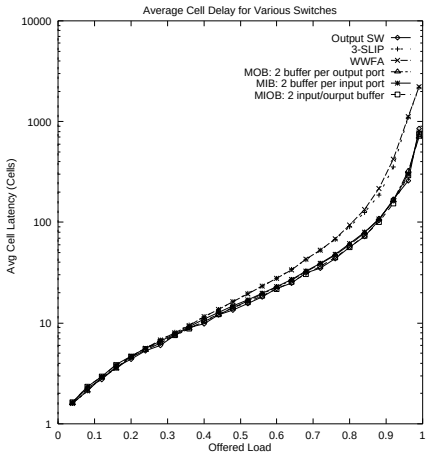

Fig. 8. Comparison of decomposed switches with the conventional input buffered and output buffered switches under bursty traffic load

V. Conclusions

In this paper we proposed a new switch architecture which can increase the grant probabilities of HOL cells of non-empty VOQs by dividing one switch fabric into several small independent crossbar units. For output buffered switches the grant probability is equal to one because there is no contention at the ingress and egress points of the switch fabric. Thus, an increase in the grant probability is directly related to the performance of the switch. We analyze the effect of decomposing a crossbar fabric into multiple smaller crossbar units on the grant probability. The simulation results show that various types of decomposed switches proposed in this paper yield better performance compared with the conventional input buffered switches and they yield the performance close to that of the output buffered switches.

Since the proposed switch architecture does not use memories faster than for the conventional input buffered switches, it can be applied for high speed switching systems. In addition, the proposed decomposition technique can be applied as a switch expansion mechanism.

References

[1] N. McKeown, M. Izzard, A. Mekkittikul, B. Ellersick, and M. Horowitz, "The tiny tera: A small high-bandwidth packet switch core," *IEEE Micro*, vol. 17, pp. 26-33, Jan.-Feb. 1997.

[2] T. Anderson, S. Owicki, J. Saxe, and C. Thacker, "High speed switch scheduling for local area networks," *ACM Trans. Comput. Syst.*, vol. 11, no. 4, pp. 319-352, Nov. 1993.

[3] N. McKeown, V. Anantharam, and J. Walrand, "Achieving 100% throughput in an input-queued switch," in *Proc. IEEE INFOCOM'96*, San Francisco, CA, pp. 296-302.

[4] A. Mekkittikul and N. McKeown, "A practical scheduling algorithm for achieving 100% throughput in input-queued switches," in *Proc. IEEE INFOCOM'98*, San Francisco, CA, vol. 2, pp. 792-799.

[5] Nick McKeown, "The iSLIP Scheduling Algorithm for Input-Queued Switches," *IEEE/ACM Trans. Networking*, vol. 7, no. 2, pp. 188-201, Apr. 1999.

[6] D. N. Serpanos and P. I. Antoniadis, "FIRM: A Class of Distributed Scheduling Algorithms for High-speed ATM Switches with Multiple Input Queues," in *Proc. IEEE INFOCOM 2000*, vol. 2, pp. 548-555, 2000

[7] Hsin-Chou Chi and Yuval Tamir, "Decomposed Arbiters for Large Crossbars with Multi-Queue Input Buffers," *IEEE International Conference on Computer Design: VLSI in Computers and Processors*, pp. 233-238, 1991.

[8] Richard O. LaMaire and N. Serpanos, "Two-Dimensional Round-Robin Schedulers for Packet Switches with Multiple Input Queues," *IEEE/ACM Trans. Networking*, vol. 2, no. 5, Oct. 1994.

[9] Shang-Tse Chuang, A. Goel, N. McKeown, B. Probhakar, "Matching output queueing with a combined input/output-queued switch," *IEEE J. Select. Areas Commun.*, vol. 17, pp. 1030-1039, June 1999.

[10] Cyriel Minkenberg and Ton Engbersen, "A Combined Input and Output Queued Packet-Switched System Based on PRIZMA Switch-on-a-Chip Technology," *IEEE Communication Magazine*, vol. 38, no. 12, pp. 70-77 , Dec. 2000.