Performance Evaluation of Combined Input Output Queued Switch with Finite Input and Output Buffers

Tsern-Huei Lee and Ying-Che Kuo

Institute of Communications Engineering
National Chiao-Tung University
Hsinchu, Taiwan 300, Republic of China
{thlee,mike}@atm.cm.nctu.edu.tw

Abstract. It has recently been shown that a combined input output queued (CIOQ) switch with a speedup factor of 2 can exactly emulate an output-queued (OQ) switch [1]-[6]. In particular, the maximal matching algorithm, named Least Cushion First/Most Urgent First (LCF/MUF) algorithm presented in [6], can be executed in parallel to achieve exact emulation. However, the buffer size at every input and output port was assumed to be of infinite size. This assumption is obviously unrealistic in practice. In this paper, we investigate via computer simulation the performance of the LCF/MUF algorithm with finite input and output buffers. We found that, under uniform traffic, a CIOQ switch behaves almost like an OQ switch if the buffer sizes at every input and output ports are 3 and 9 cells respectively. For correlated traffic, to achieve similar performance, the input and output buffer sizes have to be increased to about 7 and 11 times of the mean burst size, respectively.

1 Introduction

Over the years, many service scheduling algorithms have been proposed to provide quality of service (QoS) guarantees in an integrated services network [7]-[10]. Most of these algorithms were designed to be used at the output ports of an output queued (OQ) switch. The main problem of OQ switches is that the switching fabric of an $N \times N$ switch must run N times as fast as its line rate in the worst case. As such, OQ switches have serious scaling problem because the advancement in memory bandwidth is much slower than the advancement in transmission speed. Consequently, input queuing is unavoidable in building a large capacity switch.

On the other hand, input queued switches suffer from head-of-line blocking which limits the maximum throughput to about 0.586 (under uniform traffic assumption) [11]. It can be improved to approach 100% if cells are delivered from input ports to output ports based on maximum matching [12]. However, the high computational complexity of currently known algorithms prohibits maximum matching from being used in a high-speed switch. Although many maximal matching algorithms (e.g., PIM

I. Chong (Ed.): ICOIN 2002, LNCS 2343, pp. 203-214, 2002.

[13], LPF [14], and *i*SLIP [15]) have been proposed, none of these algorithms can provide QoS guarantee even though the complexity is lower than maximum matching. Another approach to improve the performance of an input queued (IQ) switch is to speedup the switching fabric. Because of speedup, an output port may receive cells faster than it can transmit. As a result, buffering at output is necessary and the switch becomes a combined input output queued (CIOQ) switch.

Obviously, to obtain good performance in a CIOQ switch, one has to wisely schedule the usage of switching fabric. Several algorithms had been proposed to achieve the goal. It was shown that a CIOQ switch with a speedup factor of 2 and well-designed matching algorithm can exactly emulate an OQ switch. In particular, a maximal matching algorithm named Least Cushion First/Most Urgent First (LCF/MUF) was proposed in [6]. The LCF/MUF algorithm can be executed with a parallel procedure of complexity $O(N)$, where N denotes the number of input and output ports. It was proved that, using the LCF/MUF algorithm, a CIOQ switch with a speedup factor of 2 can exactly emulate an OQ switch. However, to achieve exact emulation, the buffer size at each input and output port was assumed to be of infinite size. This assumption is obviously unrealistic in practice. In this paper, we investigate via computer simulations the performance of the LCF/MUF algorithm with finite input and output buffers. In our study, the input traffic is characterized by uniform or correlated model. With finite input and output buffers, the property of exact emulation is lost. However, simulation results show that, under uniform traffic, a CIOQ switch behaves almost like an OQ switch if the buffer size at every input port and every output port are 3 and 9 cells respectively. For correlated traffic, to achieve similar results, both input and output buffer sizes have to be increased to about 7 and 11 times of the mean burst size, respectively.

2 The LCF/MUF Matching Algorithm

In this section, we review the LCF/MUF matching algorithm proposed in [6]. Since the switching fabric is speeded up by a factor of 2, there are two scheduling phases in each slot, where a slot is defined as the time duration to transmit a cell. In each scheduling phase, the LCF/MUF scheduling algorithm matches each nonempty input with at most one output, and conversely, each output with at most one input. Cells are then delivered to output ports based on the matching outcome.

Before describing the LCF/MUF algorithm, we introduce some definitions. Let $x_{i,j}$ denote a cell at input port i destined to output port j.

Definition 1： The cushion of cell $x_{i,j}$ at input port i, denoted by $C(x_{i,j})$, equals the number of cells currently residing in output port j which will depart the emulated OQ switch earlier than cell $x_{i,j}$.

Definition 2： The cushion between input port i and output port j, denoted by $C(i,j)$, is the minimum of $C(x_{i,j})$ of all cells at input port i destined to same output port j. If there is no cell destined to output port j, then $C(i,j)$ is set to ∞.

Definition 3： The scheduling matrix of an $N\times N$ switch is an $N\times N$ square matrix whose $(i,j)^{th}$ entry equals $C(i,j)$.

The LCF/MUF matching algorithm is described below:

Step 1. Select the $(i,j)^{th}$ entry of the scheduling matrix which satisfies $C(i,j) = min_{k,l}\{C(k,l)\}$ (Least Cushion First). If the selected entry is ∞, then stop. If there are more than one entry with the least cushion residing in different columns, then choose the most urgent cell $x_{i,j}$ among those input ports which correspond to the selected entries (Most Urgent First).

Step 2. Eliminate the i^{th} row and the j^{th} column (i.e., match output port j to input port i) of the scheduling matrix. If the reduced matrix becomes null, then stop. Otherwise, use the reduced matrix and go to *Step 1*.

3 System Model

As a benchmark with which the CIOQ switch we studied is compared, we assume there exists a shadow OQ switch. The traffic arrived at the CIOQ switch is identical to that arrived at the shadow switch. Our goal is to arrange for each cell to depart from the CIOQ switch at exactly the same time it departs from the shadow OQ switch, i.e., to exactly emulate the OQ switch with the CIOQ switch.

3.1 The Shadow OQ Switch Which Adopts Strict Priority Service Scheme

For ease of simulation and feasibility in implementation, we select the strict priority scheme as the service discipline of the shadow OQ switch. We assume that every output port j maintains K FIFO queues named as $Q_{j,k}$ with priority *1* to priority K respectively (priority *1* has the highest priority). At each output port j, the strict priority service algorithm is described below.

```
for k=1 to K {
        if queue Q_{j,k} is non-empty  then
                the head-of-line cell of Q_{j,k} is served and break for-loop.
}
```

3.2 The CIOQ Switch We Studied

Fig. 1 illustrates the conceptual model of an $N\times N$ CIOQ switch with finite buffers. Every input port maintains per output port, per priority queues. In other words,

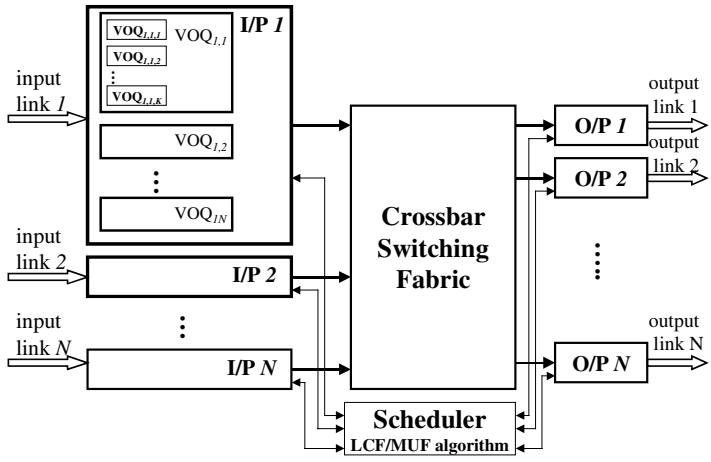

Fig. 1. The CIOQ switch scheduled by LCF/MUF algorithm.

each input port i ($1<=i<=N$) maintains a separate set of FIFO queues for each output port j, named as $VOQ_{i,j}$ for $1<=j<=N$. Therefore, there are N sets of $VOQ_{i,j}$ queues at each input port and each set of $VOQ_{i,j}$ consists of K priority FIFO queues, named as $VOQ_{i,j,k}$ for $1<=k<=K$. As a result, there are $N\times K$ queues to be managed at each input port. A new arrival cell is placed at the tail of its appropriate queue.

Since the service discipline of the shadow OQ switch is assumed to be the strict priority scheme, one can easily determine the service order of cells in the set of $VOQ_{i,j}$ queue. In other words, the entries of the scheduling matrix for the LCF/MUF algorithm can be easily determined.

4 Input Traffic Model

To evaluate switch performance quantitatively, we describe practical input traffic models that will be sent simultaneously to both the shadow OQ switch and the CIOQ switch in each simulation experiment. The following introduces two input traffic models — uniform and correlated models.

4.1 Uniform Input Traffic

The uniform input traffic model is used to characterize the interactive behavior of data arrivals in computer networks [11]; there is no time correlation between arrivals. Cells arrival on the N input links are governed by independent and identical Bernoulli processes. Specially, in any given time slot, the probability that a cell will arrival on a

particular input port is *r*. Each cell has equal probability *1/N* of being addressed to any given output port, and successive packets are independent.

Consider a particular output port (the "tagged" port) and define random variable *A* as the number of cell arrival at the tagged port during a given time slot. It follows that *A* has the binomial distribution as shown in Equation 1 that can be used to estimate the utility of output buffers.

$$\Pr[A=i]=\binom{N}{i}(r/N)^i(1-r/N)^{N-i}, \quad i=0,1,\ldots,N \tag{1}$$

4.2 Correlated Input Traffic

The correlated traffic model we studied is characterized by a 2-state Markov process alternating between active and idle states with probabilities *p* and *q*, respectively [16][17] (see Fig. 2). This model is often used to describe IP packets fragmented into ATM cells. Based on this model, the traffic source will generate a cell every slot when it is in the active state. Note that there is at least one cell in a burst. Define random variable *B* as the number of time slots that the active period (burst) lasts. The probability which the active period lasts for a duration of *i* time slots (consists of *i* cells) is

$$\Pr[B=i] = p(1-p)^{i-1}, \qquad i \geq 1 \tag{2}$$

The mean burst length is given by

$$E[B] = \sum_{i=1}^{\infty} i \cdot \Pr(B=i) = 1/p \tag{3}$$

Similarly, define random variable *I* as the number of time slots which the idle period lasts. The probability that an idle period lasts for *j* time slots is

$$\Pr[I=j] = q(1-q)^j, \qquad j \geq 0 \tag{4}$$

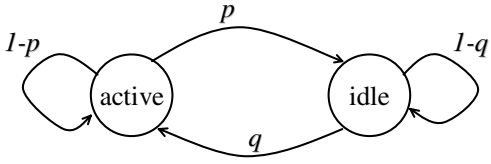

Fig. 2. The 2-state Markov chain process.

And the mean idle period is given by

$$E[I] = \sum_{i=1}^{\infty} j \cdot \Pr(I = j) = (1-q)/q \tag{5}$$

Given p and q, the offered traffic load ρ can be found by

$$\rho = \frac{E[B]}{E[B]+E[I]} = \frac{q}{p+q-pq} \tag{6}$$

We assume there is no correlation between different bursts and the destination of each burst is uniformly distributed among the output ports.

5 Numerical Results

Our simulation experiments are divided into three parts. In the first two parts, we measure the latency of cells which are simultaneously fed to both the shadow OQ switch and the CIOQ switch with infinite input buffer and finite output buffer under uniform and correlated traffic models, respectively. We define $d(x) = | d_{OQ}(x) - d_{CIOQ}(x) |$ as the deviation index of cell x. Here $d_{OQ}(x)$ and $d_{CIOQ}(x)$ denote, respectively, the departure times of cell x for the shadow OQ switch and the CIOQ switch. Given a fixed d, we measure the percentage of cells that has a deviation index smaller than or equal to d. This percentage is denoted by P_d. For example, $P_{d=0}$ equals 100% means that the CIOQ switch exactly emulates the shadow OQ switch. In the third part, we study the effect of finite input buffer. We measure the cell-loss probability (denoted by P_{loss}) of a CIOQ switch with finite input buffer and finite output buffer.

Simulations are performed for N =4, 8, 16, and 32, with 4 priority levels under both uniform and correlated traffic models. For correlated traffic, the mean burst length ℓ = 4, 8, 16, 32, and 64 are considered at the same time. The statistics of about one million cells are collected (over all queues in the switch) and thus loss probabilities smaller than 10^{-6} are not measurable.

5.1 Results of Uniform Input Traffic

For uniform input traffic load, simulation results are shown in Fig. 3 to Fig. 5. Fig. 3 shows that the CIOQ switch almost behaves like an OQ switch under moderate loads if sufficiently many output buffers are installed. For N =16 with infinite input buffer ($B^i = \infty$) and finite output buffer B^o = 9 cells, the value of $P_{d=0}$ is larger than 90% for the 1st priority traffic cells under an offered load up to 0.8. Fig. 4 shows the values of $P_{d=0}$ and $P_{d<=2}$ for all the four priority cells. We observe that the performance behaves similarly for all the four priority cells and the values of $P_{d<=2}$ are close

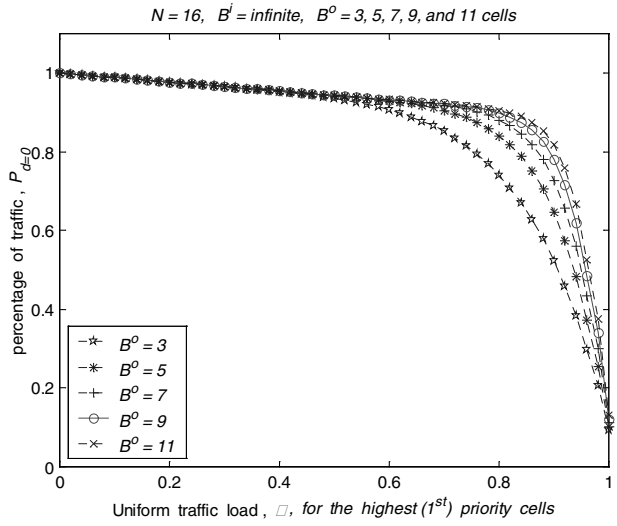

Fig. 3. Performance of the 16×16 CIOQ switch with $B^i = \infty$ and B^o = 3, 5, 7, 9, and 11 cells. The curve is shown for the highest priority cells only.

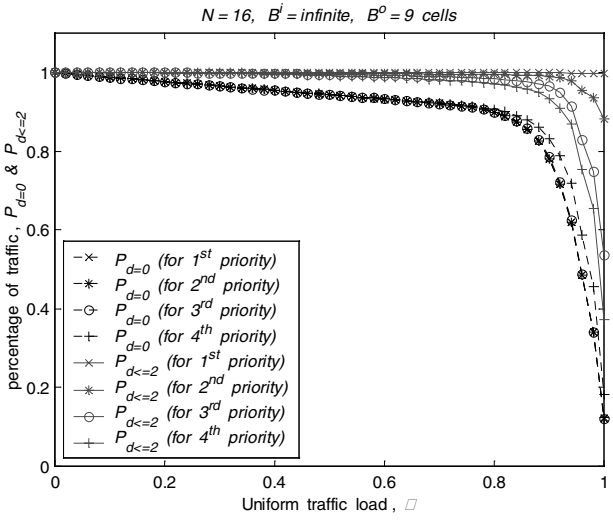

Fig. 4. Performance of the 16×16 CIOQ switch with $B^i = \infty$ and B^o = 9 cells. The curves show the performance of values of the $P_{d=0}$ and $P_{d<=2}$ for all the four priority cells.

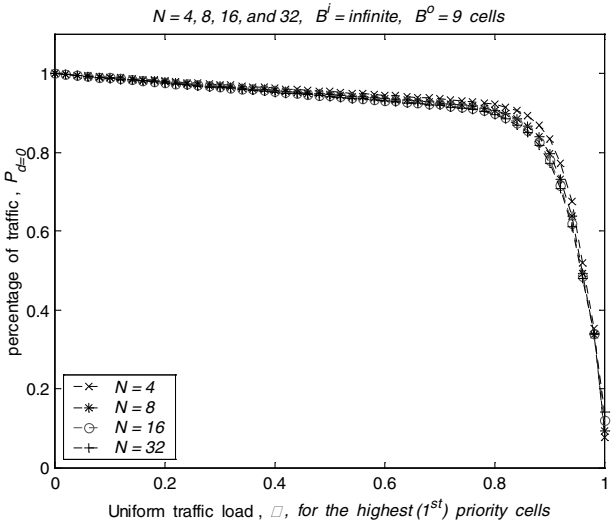

Fig. 5. Performance of the CIOQ switch with $B^i = \infty$ and $B^o = 9$ cells. The curve shows the performance of the highest priority cells as a function of the switch size N=4, 8, 16, and 32.

to 100% even under an offered load up to 0.8. As is shown in Fig. 5, the performance degrades as the number of ports increases. But it degrades slowly when N is larger than 8. Based on these results, we conclude that a CIOQ switch performs like an OQ switch if output buffer size is at least 9 cells.

Note that, the maximum queue length at all input ports is also estimated in these simulations. The input buffer size of 60 cells is sufficient under the uniform traffic model.

5.2 Results of Correlated Input Traffic

For the correlated traffic model, results depend on the mean burst size ℓ. In Fig. 6, we show the results for ℓ = 16 cells for N =16 with $B^o = 5\ell$, 7ℓ, 9ℓ, 11ℓ and 13ℓ cells. The value of $P_{d=0}$ is larger than 90% for the 1st priority traffic under an offered load up to 0.8 when $B^o = 11\ell$ cells are installed. From Fig. 7, one can see that the performances are roughly the same for various switch sizes. In Fig. 8 we performed similar simulations for other values of ℓ. Our observation is that the performance degrades as a function of the value of ℓ. But the degradation is became slow once ℓ is larger than 16. Based on these results, we suggest to allocate $B^o = 11\ell$ cells at each output port for correlated traffic.

Note that the maximum queue length at all input ports is smaller than 237 cells (about 15ℓ cells) which is also estimated in simulations.

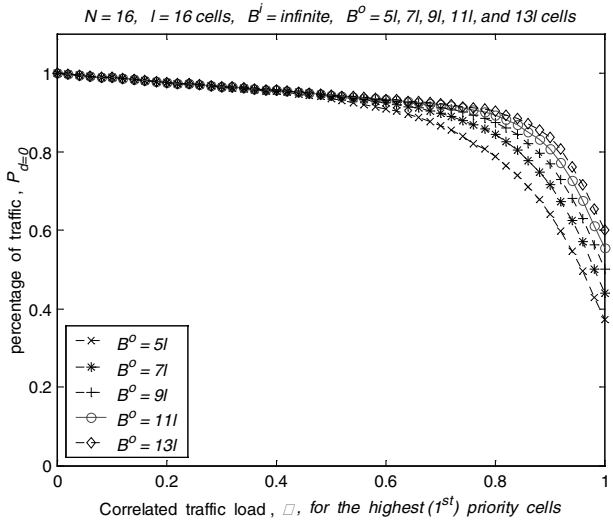

Fig. 6. Performance of the 16×16 CIOQ switch with $B^i = \infty$ and $B^o = 5\ell$, 7ℓ, 9ℓ, 11ℓ and 13ℓ cells (ℓ =16). The curve is shown for the highest priority cells only.

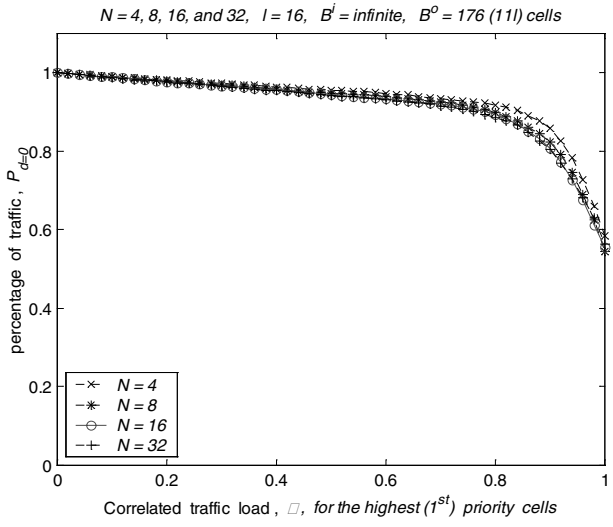

Fig. 7. Performance of the CIOQ switch with $B^i = \infty$ and $B^o = 11\ell$ cells (ℓ =16). The curve shows the performance of the highest priority cells as a function of the switch size N=4, 8, 16, and 32.

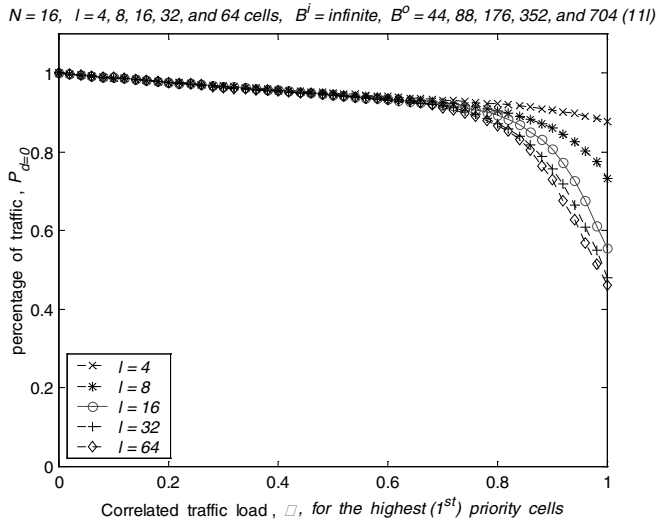

Fig. 8. Performance of the 16×16 CIOQ switch with $B^i = \infty$ and $B^o = 11\ell$ cells. The curve shows the performance of the highest priority cells as a function of the mean burst length ℓ = 4, 8, 16, 32 and 64 cells.

5.3 Effect of Finite Input Buffer

In this sub-section, we study the effect of finite input buffer. The output buffer size is chosen to be 9 cells for uniform traffic and 11ℓ cells for correlated traffic. The switch size and the mean burst length investigated in our simulations are N=16 and ℓ=16. We measure cell loss probability (P_{loss}) due to finite input buffer. Fig. 9 shows the curves of P_{loss} versus ρ for input buffer size B^i=2, 3, 5, 7 and 9 cells under uniform traffic. Fig. 10 shows the curve of P_{loss} versus ρ for input buffer size B^i=3ℓ, 5ℓ, 7ℓ, 9ℓ and 11ℓ cells under correlated traffic. It can be seen that P_{loss} of the CIOQ switch with B^i=3 and B^o=9 is smaller than 10^{-5} under the uniform traffic model for an offered load up to 0.8. For correlated input traffic, one can achieve similar performance with B^i=7ℓ and B^o=11ℓ.

Note that, based on our simulation results, the CIOQ switch needs to maintain a buffer of 60 cells at each input port to achieve zero cell loss probability under uniform traffic (or 15ℓ cells under correlated traffic) for an offered load up to 0.9.

6 Conclusions

We have investigated the performance of a CIOQ switch with finite input and output buffers. The LCF/MUF algorithm is selected as the matching algorithm to deliver cells from input ports to output ports. To implement the LCF/MUF algorithm, a switch has to know the cushion of all cells and the relative departure order of cells destined to the

same output port. The complexity of cushion calculation strongly depends on service discipline. In this study, we choose strict priority service discipline because it allows cushions to be easily calculated. Based on simulation results, we found that a CIOQ switch can mimic an OQ switch quite well for both uniform and correlated traffic up to ρ=0.8.

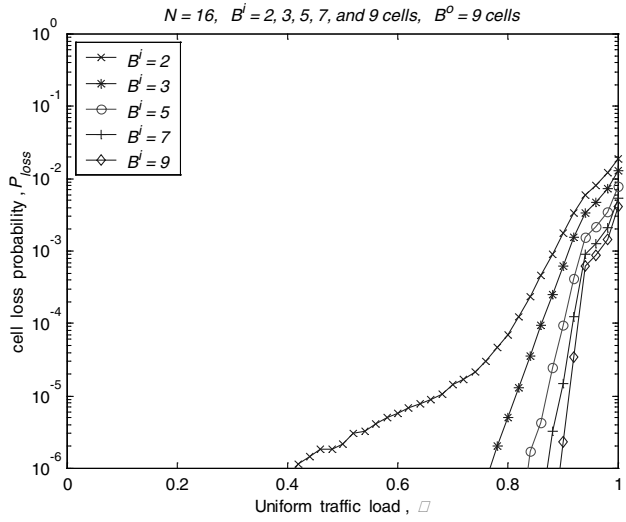

Fig. 9. P_{loss} of a 16×16 CIOQ switch with B^i=2, 3, 5, 7, and 9 cells, and B^o=9 cells under uniform input traffic.

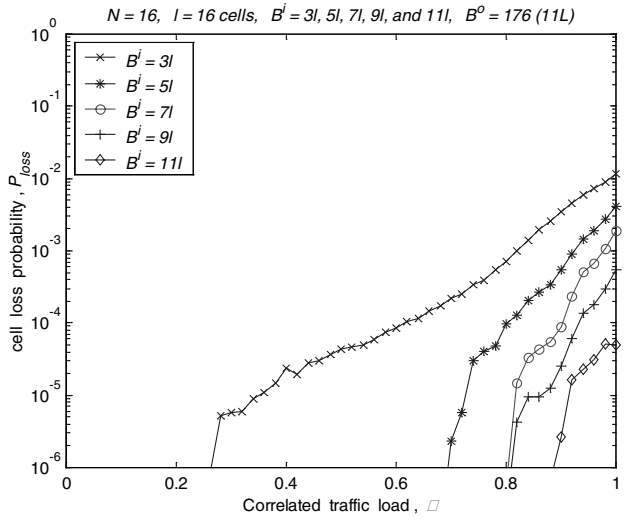

Fig. 10. P_{loss} of a 16×16 CIOQ switch with B^i=3 ℓ , 5 ℓ , 7 ℓ , 9 ℓ , and 11 ℓ , and B^o=11 ℓ cells (ℓ =16 cells) under correlated input traffic.

Further applications may require a more complicated service discipline (such as WFQ) than strict priority. Unfortunately, a complicated service discipline may induce dramatic complexity in cushion calculation. Therefore, how to design a scheme which simplifies cushion calculation and mimics the complicated service discipline is worth to be further studied.

References

1. B. Prabhakar and N. McKeown, "On the speedup required for combined input and output queued switch," Stanford University, Stanford, CA, *Tech. Rep., STAN-CSL-TR-97-738*, 1997.
2. A. Charny, P. Krishan, N. Patel, and R. Simcoe, "Algorithms for providing bandwidth and delay guarantees in input-buffered crossbar with speedup," *Proc. of IWQoS'98*, pp.235-244.
3. I. Stoica and H. Xhang, "Exact emulation of an output queueing switch by a combined input output queueing switch," *Proc. of IWQoS'98*, pp.218-224.
4. P. Krishna, N.S. Patel, A. Charny, and R. Simcoe, "On the speedup required for work-conserving crossbar switches," *Proc. of IWQoS'98*, pp.225-234.
5. S.T. Chuang, A. Goel, N. McKeown, and B. Prabhakar, "Matching output qeueueing with a combined input/output-queued switch," *IEEE J. Selected Area in Commun.*, Vol.17, No.6, pp.1030-1039, June 1999.
6. T. H. Lee, Y. W. Kuo, and J. C. Huang, "Quality of Service Guarantee in a Combined Input Output Queued Switch," *IEICE Trans. on Commun.*, pp.190-195, Feb. 2000.
7. H. Zhang, "Service disciplines for guaranteed performance service in packet-switching networks," *Proc. of the IEEE*, 83(10), pp.1374-1399, October 1995.
8. D. Ferrari and D.C. Verma, "A scheme for real-time channel establishment in wide-area networks," *IEEE J. Selected Area in Commun.*, Vol. 8, No. 3, pp.368-379, April 1990.
9. A.K. Parekh and R. G. Gallager, "A generalized processing sharing approach to flow control in integrated services networks: the single node case," *IEEE Trans. on Networking*, Vol.1, No.3, pp.344-357, June 1993.
10. J.C.R. Bennett and H. Zhang, "WF^2Q: worst-case fair weighted fair queueing," *Proc. of IEEE INFOCOM'96*, pp.120-128.
11. M. Karol, M. Hluchyj, and S. Morgan, "Input verse output queueing on a space division switch," *IEEE Trans. on Commun.*, Vol.35, pp.1347-1763, Dec. 1987.
12. N. McKeown, V. Anantharam, and J. Walrand, "Achieving 100% throughput in an input-queued switch," *Proc. of IEEE INFOCOM'96*, pp.296-302.
13. T.E. Anderson, S.S. Owicki, J.B. Saxe and C.P. Thacker, "High speed switch scheduling for local area networks," *IEEE/ACM Trans. on Computer Systems*, Vol.11, No.4, pp.319-352, Nov. 1993.
14. A. Mekkittikul and N. McKeown, "A practical scheduling algorithm to achieve 100% throughput in input-queued switches," *Proc. of IEEE INFOCOM'98*, pp.792-799.
15. N. McKeoen, "The *i*SLIP scheduling algorithms for input-queued switches," *IEEE/ACM Trans. on Networking*, Vol.7, No.2, pp.188-201, April 1999.
16. S.Q. Li, "Performance of a non-blocking space-division packet switch with correlated input traffic," *Proc. of IEEE Globecom'89*, Vol.3, pp.1754-63.
17. S.C. Liew, "Performance of various input-buffered and output-buffered ATM switch design principles under bursty traffic : simulation study," *IEEE Trans. on Commun.*, Vol.42, pp.1371-79, Feb/Mar/Apr. 1994.