IP 스위칭 시스템을 위한 iSLIP 스케줄링에 기반의 Dual 라운드로빈 설계 및 분석

(Design and Analysis of a Dual Round-Robin based iSLIP (DiSLIP) Scheduling Scheme for IP Switching System)

최 진 식*, 양 미 정**, 김 태 일**

(Jin Seek Choi, MiJung Yang, and Tae Il Kim)

요 약

본 논문에서는 IP 스위칭 시스템을 위한 iSLIP 기반의 Dual Round Robin(DRR) 방식 스케줄링 알고리즘을 제시한다. 기존의 iSLIP 방식과 DRR 방식을 결합하여 단점은 최소화하면서 DRR이 가진 분배기능의 장점과 iSLIP 기능이 가진 고성능의 장점을 이용할 수 있게 한다. 불균일한 트래픽 환경에서 스위치에서의 처리율과 지연성능을 컴퓨터 시뮬레이션을 통해 확인한다. 또한 제안된 방식은 기존의 iSLIP과 비교하여 구현 복잡성을 매우 낮아짐을 보인다. 이러한 실험을 통해 본 논문에서 제안한 스케줄링 알고리즘이 기존의 DRR에 비해 우수한 성능을 iSLIP에 비해 구현의 복잡성에서 우수함을 알 수 있었다.

Abstract

In this paper, a new Dual Round-Robin (DRR) based iterative SLIP (iSLIP) scheduling scheme, called DiSLIP is proposed for IP switching systems. By using DRR followed by iSLIP, DiSLIP can exploit desynchronization effect of DRR and high performance of iSLIP, while the drawbacks of two schemes are minimized. Through computer simulation, we verify the switch throughput and total waiting time of the proposed scheme under nonuniform and correlated self-similar traffic. Moreover, the proposed scheme can considerably reduce the complexity of parallel matching logics compared to iSLIP. From the result, we observe that the proposed scheme outperforms DRR on throughput as well as iSLIP schemes on complextiy.

I. INTRODUCTION

Next Generation Networks (NGN) can be characterized by the large number of connections and very-high-speed transmission medias. In constructing

* 정회원, 한양대학교
  (Hanyang University),
** 정회원, 한국전자통신연구원
  (ETRI)

NGNs, one of the essential elements is a switching system. Switching systems are characterized by a software- and hardware-based system that provides both switching and exchanging of IP packets. Switching system needs to handle the large number of connections and very high-speed transmission links. With the emergence of asynchronous transfer mode (ATM), a cell-based switching system provides very high-speed transmission links up to terabit per second and simple hardware switching with a scalable architecture, while minimizes the overhead due to the processing of IP packet header, since the

length of ATM cell is fixed [1]. Therefore, researchers have envisioned that the cell-based switching system is applied to achieve high speed IP switching.

The cell-based IP switching system has substantial advantages of price/performance over packet-based IP switching system by merging IP packet forwarding with cell-based switching. However, there are several challenges in merging IP packet forwarding with cell-based switching. One important issue is the bursty traffic with highly correlated destinations due to the fragmentation of IP packet[2,3,8]. Another important issue is traffic imbalance[4]. In Internet and telecommunication networks, traffic imbalance is inherent since some particular destinations such as popular databases, communication servers or outgoing trunks can cause traffic concentration. The concentrated traffic may cause the traffic imbalance for the specific port(s). The traffic of the concentrated input is likely to present short-term traffic imbalance. This imbalance refers to a traffic model with unevenly distributed routing and different intensity at certain output port, which is called nonuniform traffic.

Researchers have made various attempts to solve the problems of fragmentation of IP packet and traffic imbalance in switching systems. Considerable work has been done on virtual output queue (VOQ)based scheduling algorithms for IP switching system[5~7,9,10]. Marsan et al. developed novel scheduling algorithms to deal with variable length packets for IP switching system, and proved that no throughput limitations exist by operating input queueing switches in packet mode compared to output queueing switches[9]. Nong et al., evaluated the maximum throughput of the cell-based IP switching systems for the parallel iterative matching (PIM) algorithm under bursty traffic[6]. Note that all of these works are based on VOQ based maximal matching algorithms, which can achieve 100% throughput even under nonuniform traffic. Maximal matching algorithms can be accomplished through multiple iterations of either PIM or SLIP. However, in these schemes, inputs send a total of up to $N2$ requests to outputs and each outputs arbitrates up to N requests. However, these matching algorithms are not practical due to the implementation complexity of parallel matching logics.

A number of practical algorithms such as windowing and DRR schemes have also been proposed[11-13]. In particular, DRR scheme is a simple practical solution based on VOQ queue[11]. It can achieve the desynchronization effect without using iterative mechanism. Moreover, DRR scheme can achieve lower implementation complexity of a simple contention logic that can handle one cell at each contention. It has also been demonstrated that the DRR scheme can get 100% throughput under uniform traffic[11,15]. However, severe performance degradation of delay takes place in the presence of nonuniform IP traffic.

In this paper, we propose a new DRR-based iSLIP scheme and analyze the performance under nonuniform IP traffic. By implementing DRR followed by iSLIP, DiSLIP can reduce the complexity of parallel matching substantially while achieving the high throughput of iSLIP. Finally, we compare the performance of the proposed scheme to those of iSLIP and DRR under nonuniform and correlated bursty traffic.

The remainder of this paper is organized as follows. In Section Ⅱ, we describe the switch model and the traffic model at the input queue. In Section Ⅲ, we present the numerical results and compare with simulation. Finally, we conclude in Section Ⅳ.

Ⅱ. The Proposed Scheduling Scheme

A. Switch Architecture

The switch architecture considered in this paper is a NN input queueing cell-based switch with VOQs [14]. The switch fabric is nonblocking and has no internal speed-up. The variable length packets are internally segmented into ATM-like cells that are switched. Cells are of fixed-length and the buffer size of each VOQ is infinite. The switch operates synchronously so that the cells are received and
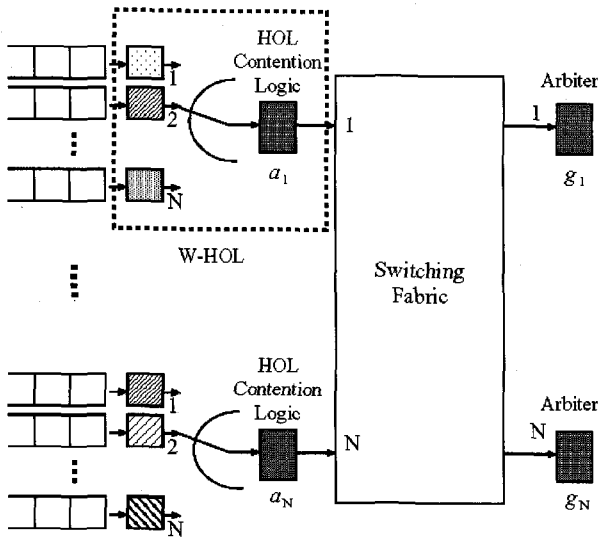
그림 1. DiSLIP 스케줄링 방식의 스위치 구조
Fig. 1. Switch architecture with DiSLIP scheduling scheme.

transmitted within a fixed time interval called a slot. Each input can transmit at most one cell to any output and each output can receive at most one cell in each slot time, t. However, multiple cells arrive in the input queue as a train of cells.

Fig. 1 shows the switch structure. As shown in this figure, each input has a separate First-in-First-out (FIFO) queue for each output called a VOQ. Input port i has N VOQs, says from 1 to N, where $VOQi,j$ stores cells arriving at input port i with the destination of output port j. W-HOL queue consists of N HOL cells at all VOQs. Each input has its own contention logic, and operates independently from the others. Each output also has its own arbiter, and operates independently. The contention logic maintains a contention pointer ai pointing at the output that has the highest priority VOQ in which one cell is transmitted in the slot, and each arbiter also maintains a grant pointer gi pointing at the input that has the highest priority input that will be granted to transmit.
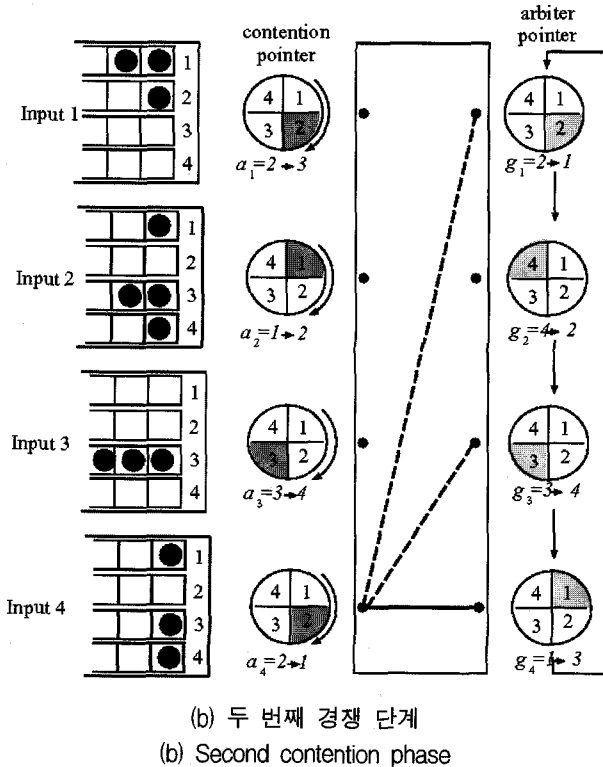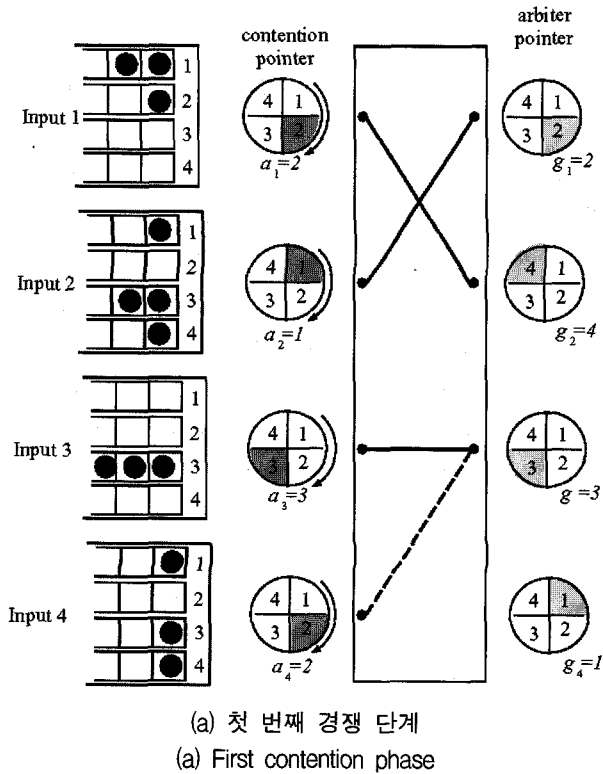
B. Scheduling Scheme

Similar to the iSLIP, the proposed DiSLIP scheme divides the contention procedure into w contention

phases $t = \sum_{i=1}^{w} c_i$ where ci is the time duration of ith contention phase. The DiSLIP scheduling scheme starts the desynchronizing procedure followed by parallel matching procedures. The first contention is DRR scheduling procedure and the others are the same as the iSLIP scheduling procedure.

In the DRR scheduling procedure, each contention logic at input queues selects which VOQ will contend to the outputs as a round-robin service discipline, starting from the contention pointer ai. After selection, each input sends a request to the designated output. When an output receives requests, its arbiter chooses one of them as a round-robin service discipline starting from the arbiter pointer gi, and sends a grant to the winner input. In the following iSLIP scheduling procedure, only unmatched inputs send a request to every output for which VOQs have a queued cell. If an unmatched output receives multiple requests, the arbiter picks a request from the contending requests according to the round-robin service discipline starting from the pointer gi, and sends a grant to the winner input port. If an input receives a multiple grant, the contention logic accepts the one that appears next in its round-robin scheduler starting from the contention pointer ai, and notifies the selected output whether or not its grant was accepted. In the following iterations, only the unmatched inputs and outputs repeat the same procedure.

If the input is allowed to transmit a cell, the contention pointer ai (VOQ identifier) is incremented (modular N) to one location beyond the accepted output. On the other hand, the grant pointer gi is forwarded to the next pointer $gi+1$ (modular N). When no cell has been transmitted in the slot, the forwarded grant pointer remains at the same pointer. When it gets a grant to transmit, the forwarded pointer becomes the corresponding input port. By forwarding the pointer to the next port, the arbiter pointer can be tightly matched with the contention pointer.

If the input port looses output contention at all

(a) 첫 번째 경쟁 단계
(a) First contention phase

(b) 두 번째 경쟁 단계
(b) Second contention phase

그림 2. w=2일 때 DiSLIP 스케줄링 절차 예
Fig. 2. An example of the DiSLIP scheduling procedure when $w=2$.

previous contentions, then the next contentions is performed similar to windowing scheme[15]. The

matched input port (i.e., the port won the output contention) is not allowed to contend for the remaining contention phases, and the output ports will not arbitrate the contending requests after an input has been matched. However, at most one cell is allowed to be matched from/to input/output. Consequently, the number of contention requests to be arbitrated can be considerably reduced since only unmatched inputs and outputs at the DRR scheduling scheme are eligible to participate the iSLIP scheduling scheme.

Fig. 2 shows an example of the DiSLIP algorithm when $w=2$. In this example, $a1$, $a2$, $a3$, and a4 are designated contention pointers for input 1, 2, 3, and 4, respectively. $g1$, $g2$, $g3$, and $g4$ are designated arbiter pointers for output 1, 2, 3, and 4, respectively. At the beginning of the contention phase 1, each contention logic finds one active VOQ starting from ai: VOQ 2 for input 1, 1 for 2, 3 for 3, and 2 for 4, respectively. Since VOQ 2 for input 4 is empty, the next VOQ 3 instead of 2 has to contend at this phase. Each arbiter selects one contending cell starting from gi: input 2 for arbiter 1, 1 for 2, 3 for 3, and none for 4, respectively.

In the first contention phase, input 4 failed to transmit. As a result, all VOQs of input 4 (VOQ 1, 3, and 4) contend for their outputs at the second phase while inputs 1, 2, and 3 notify transmitting of 2, 1, and 3. The output arbiters 1 and 3 were matched at the first contention phase. Only output arbiter 4 can send a grant to the input 4 at the second contention phase. Therefore, input 4 can receive a grant from output 4.

At the end of the second contention phase, the contention pointer updates by an increment of one from the transmitting VOQ, and each arbiter pointer also updates with the arbiter pointer at the prior port as a round-robin discipline: $aiai+1$, and $gigi-1$. After then, the contention pointers $a1$, $a2$, $a3$, and $a4$ become 3, 2, 4, and 1, respectively. The arbiter pointers $g1$, $g2$, $g3$, and $g4$ become 1, 2, 4, and 3, respectively.

III. Performance Study And Simulation Results

The switch throughput and total waiting time cannot be explicitly analyzed since the service probability of a contending cell depends on the rank of the position and the results of previous contentions. Moreover, the state space of the queueing system is greatly increased as the number of input and window size increase. Therefore, we evaluate the switch throughput and total waiting time of the switch through computer simulation.

A. Traffic Model

The traffic intensity in the switch can be represented by means of a rate matrix describing traffic passing from input i to output j. The particular form of the rate matrix used in previous studies is given by

$$\lambda_{i,j} = \lambda_i Q_j \qquad (1)$$

where λ_i is the average arrival rate of cells at input i, and Q_j is the probability of a cell at any input passing to output j[15].

The arrival statistics considered in this paper are nonuniform and correlated bursty traffic. The correlated bursty traffic model represents realistic IP traffic since real IP packets tend to be fragments of variable length packet, corresponding to arrival in bursts. The input traffic alternates between burst and idle with geometrically distributed mean lengths, while the output address of each burst tightly correlates with the same output. We can assume this input traffic as a simple on/off arrival process modeled by the interrupted Markov process. For the input arrival statistic, we also consider the self-similar arrival process modeled by Pareto-distributed ON/OFF traffic with Hurst parameter H=(3-)/2. It can be used to characterize probability densities that describe packet interarrival time with heavy-tailed distribution.

Here, the probability of a burst consisting of k slots is given by

$$B(k) = b(1-b)^{k-1}, \qquad k \geq 1 \qquad (2)$$

where $1/b$ is the average burst length. If the input is in the idle state, it will stay in the same state. The idle period with k slots is also given by

$$I(k) = v(1-v)^k, \qquad k \geq 0 \qquad (3)$$

where v is the average idle length. Here, a burst can be followed immediately by another burst in which case there is no idle period ($k=0$). The offered load can be calculated as

$$\rho = \frac{E[busy]}{E[busy] + E[idle]} = \frac{v}{b+v-bv} \qquad (4)$$

Next, we consider outgoing traffic intensity. In this paper, we do not consider input imbalance traffic. We only consider output imbalance traffic where the output addresses are not uniformly distributed. The output imbalance factor is defined by Q_j such that the probability of an output address being j as follows

$$Q_j \neq 1/N \qquad \text{for all} \quad j \qquad (5)$$

and

$$\sum_j Q_j = 1 \qquad \text{for all} \quad j \qquad (6)$$

The most general nonuniform traffic pattern is the output imbalance traffic consisting of two output groups. In this case, the outputs are divided into two groups N_1^o and N_2^o. The output imbalance factor for each output group is given by

$$Q_j = \begin{cases} P_1 1/N_1^o & \text{if } j \in N_1^o \\ (1-P_1)1/N_2^o & \text{if } j \in N_2^o \end{cases} \qquad (7)$$

where P_1(or $1-P_1$) means the portion of input traffic going to group $N_1^o(N_2^o)$ and $1/N_1^o(1/N_2^o)$ means the portion of a specific output in the same output group, respectively. From now on, we regard P_1 as the bi-group coefficient.

The other nonuniform traffic pattern is the

hot-spot imbalance where a single hot-spot is super-imposed on the background of uniform traffic. This is a special case of bi-group imbalance model as $N_1^o \to 1$ and the output imbalanced factor becomes

$$Q_j = \begin{cases} h + \dfrac{1-h}{N} & \text{if } j \in N_1^o \\ \dfrac{1-h}{N} & \text{otherwise} \end{cases} \qquad (8)$$

where h is called the hot-spot coefficient.

B. Performance Results

Based on the above traffic model, we present the switch throughput and complexity as simulation results. In the simulation, we assume that the size of the switch is 128128. All VOQs for each destination are composed of 50 buffers. In the switch, there is no loss and the switch operation is slotted of running for 106 times. Input load is balanced and overhead is not taken. In this study, we are interested in the switch throughput and the switch complexity to observe the effects of scheduling scheme on the switch performance under various window sizes under correlated bursty and nonuniform self-similar traffic. The switch throughput is derived under various load conditions, for instance, $\rho \le 1$. The maximum throughput indicates the upper bound of the average utilization of the N outputs without driving the switch stability. The switch complexity can be derived from the average number of contending cells at each output. The average number of contending cells indicates the arbitration complexity of the parallel matching which is dominated in switch complexity.

In this study, we consider the switch performance while considering the following two factors. One is the average burst length of an incoming packet and the other is the imbalance of output address distribution. For the bursty traffic, the average burst length of an incoming packet $1/b$ is fixed to 20. The cells in the same burst have the same destination address. The arrival process of all inputs is assumed to be an independent process. For the imbalance t
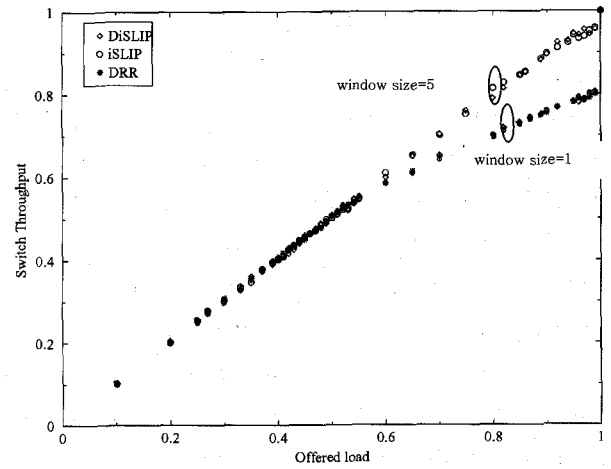

그림 3. 핫 스팟 상관관계를 갖는 버스트한 트래픽 (h=0.005, H=0.8) 상에서 처리율 비교

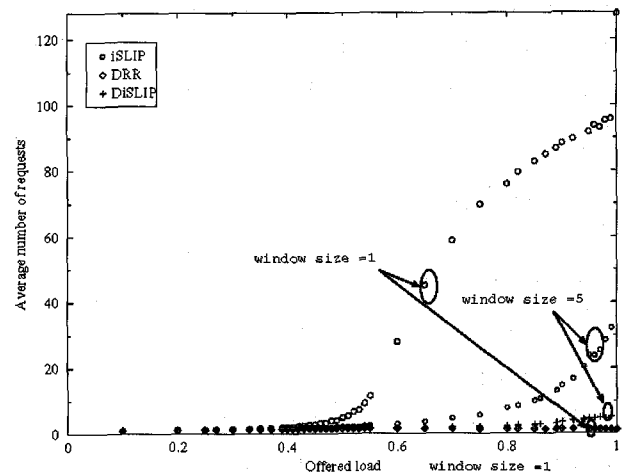Fig. 3. Throughput comparison under the hot-spot correlated bursty traffic (h=0.005, H=0.8).

그림 4. 핫 스팟 상관관계를 갖는 버스트한 트래픽 (h=0.005, H=0.8) 상에서 트래픽 부하와 스위치 복잡도와의 관계

Fig. 4. Switch complexity versus offered load under the hot-spot correlated bursty traffic (h=0.005, H=0.8).

raffic, the outputs are divided into 1 and 127 for group 1 and 2, respectively. The traffic intensity of each output group is not the same. The additional traffic intensity of output 1, called the hot-spot coefficient h, will be added to the uniform traffic intensity. When $P_1 = 0.25$ or $h=0$ both cases become the uniform traffic case (i.e., addresses of incoming cells are uniformly distributed to all outputs).

Fig. 3 shows the throughput comparison with iSLIP and DRR. As shown in this figure, the switch throughput is bounded by the window size. When the

window size is 1, the throughput is identical to that of DRR. On the other hand, the throughput increases as much as that of iSLIP when window size becomes greater than 5.

Fig. 4 shows the complexity of the parallel matching under hot spot correlated bursty and self-similar traffics (Hurst parameter= 0.8, h=0.005, $1/b$=20). As shown in this figure, the average number of contending cells for the iSLIP is rapidly increased after the traffic load is greater than 0.6 for the window size 1. Compared to iSLIP, the DiSLIP further reduces the average number of contending cells as much as that of DRR. Moreover, the number of contending cells remains a little higher than DRR when the window size is 5.

Fig. 5 shows the number of contending cells at each contention phase. The number of contending cells for iSLIP is very high at the first contention phase. This number indicates that every input is contending for every output. In the following phases, the number of contending cells rapidly decreases. On the other hand, the number of contending cells for DRR remains at low the first contention phase. The number of contending cells for the proposed scheme adapts the number of contending cells for DRR at the first contention phase, and follows that of iSLIP after
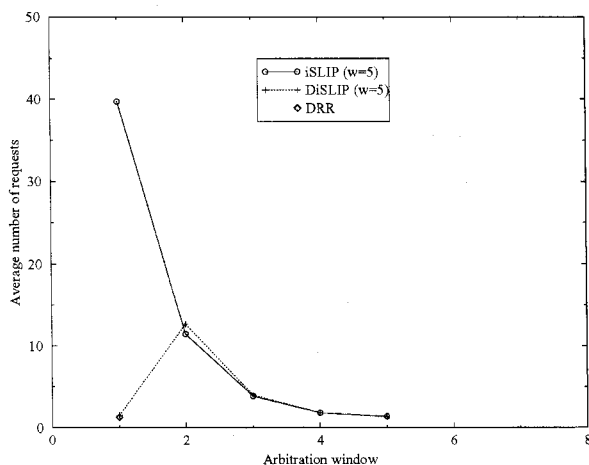
second phase.

This implies that the complexity of the DiSLIP scheme can be considerably reduced as much as the DRR while the switch throughput can be as great as that of the iSLIP. In addition, the nonuniform or correlated bursty traffic has no impact on the performance of the switch. From the results, we observe that the DiSLIP scheme can be useful for IP switching system modeled on nonuniform and correlated bursty self-similar traffic.
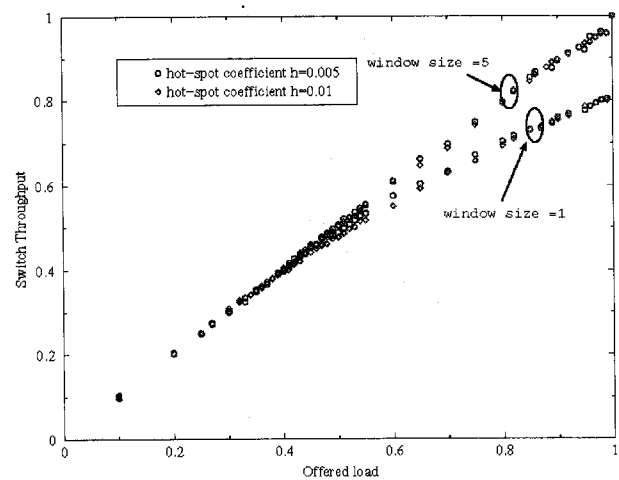

그림 6. 핫 스팟 상관관계를 갖는 버스트한 트래픽 상에서 트래픽 부하와 스위치 처리율과의 관계

Fig. 6. Switch throughput versus offered load for various hot-spot coefficients under correlated bursty traffic.
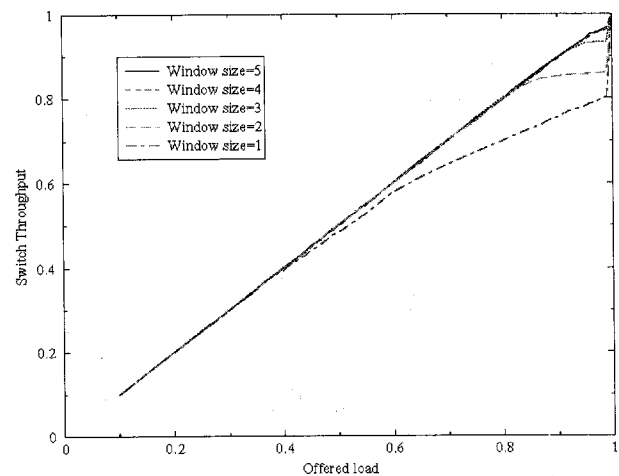

그림 5. 핫 스팟 상관관계를 갖는 버스트한 트래픽 (h=0.005, H=0.8) 상에서 각 경쟁 단계별 경쟁 셀의 수

Fig. 5. The number of contending cells at each contention phase under the hot-spot correlated bursty traffic (h=0.005, H=0.8).

그림 7. 핫 스팟 버스트한 트래픽에서 다양한 위도우 크기를 갖는 스위치 처리율과 트래픽 부하와의 관계

Fig. 7. Switch throughputversus offered load for various window sizes (h=0.005, 1/b=20).

Fig. 6 and 7 show the switch throughput of the proposed scheme versus offered load for various hot-spot coefficients and window sizes under correlated bursty traffic ($1/b$=20 and h=0.005~0.01), respectively. As shown in these figures, correlated bursty and nonuniform traffic has no impact on the switch throughput. On the other hand, the switch throughput is bounded by the window size. Moreover, the DiSLIP scheme can linearly increase the switch throughput at the certain point. At around $\rho = 1$, the
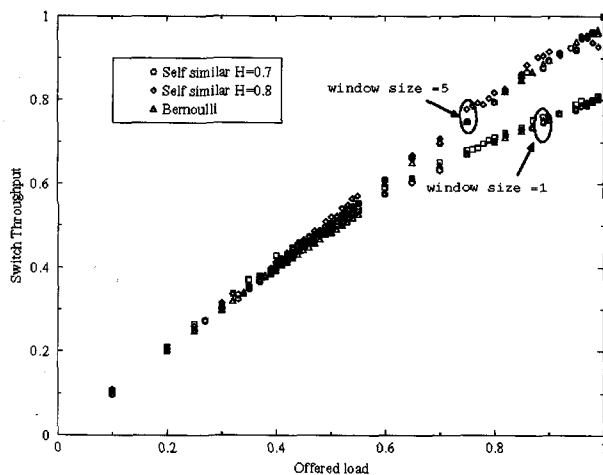

그림 8. 핫 스팟 상관관계를 갖는 버스트한 특성과 비균일 셀프시밀러 트래픽 특성을 갖는 트래픽 부하와 스위치 처리율과의 관계

Fig. 8. Switch throughput versus offered load under correlated bursty and nonuniform self-similar traffic.
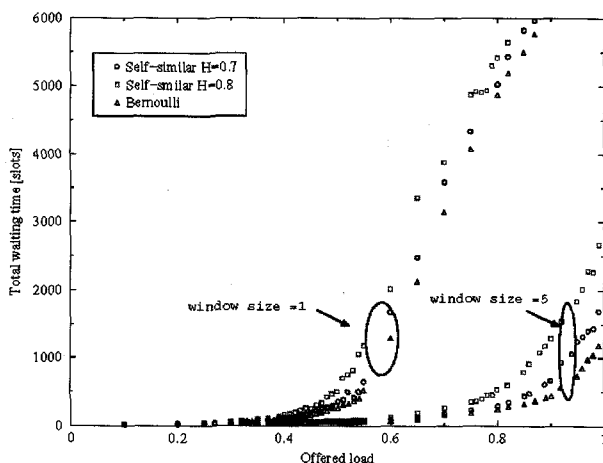

그림 9. 핫 스팟 상관관계를 갖는 버스트한 특성과 비균일 셀프시밀러 트래픽 특성을 갖는 트래픽 부하와 총 대기 시간과의 관계

Fig. 9. Total waiting time versus offered load under correlated bursty and nonuniform self-similar traffic.

switch throughput steeply increased into 1. This is because the switch receives a desynchronization effect around the saturation point.

Figs. 8 and 9 show the switch throughput and total waiting time under correlated bursty and nonuniform self-similar traffics (Hurst parameters= 0.5, 0.7 and 0.8, h=0.005, $1/b$=20). These results show that the self-similar traffic has little impact on the switch throughput and waiting time of the switch with the DiSLIP scheme. Moreover, the DiSLIP scheme can achieve 100% maximum throughput under nonuniform and correlated burst self-similar traffic. From the results, we observe that the DiSLIP scheme can be useful for an IP switching system modeled by nonuniform and correlated bursty self-similar traffic.

IV. Conclusions

The objective of this paper is to show the high performance of the DiSLIP scheme compared to the DRR and the iSLIP under the correlated bursty and nonuniform traffic. We simulated the switch throughput that takes into account the influence of correlated bursty and nonuniform traffic. The results show that the nonuniform or correlated bursty traffic has no impact on the performance of the switch with DiSLIP. In addition, the DiSLIP scheme can considerably increase the switch throughput and reduce total waiting time even under nonuniform traffic. In addition, the DiSLIP considerably reduces the arbitration complexity compared to the iSLIP. Consequently, we concluded that the DiSLIP scheme is useful for the design of scheduling schemes for IP switches.

References

[1] P. Newman, G. Minshall, and T. L. Lyon "IP switching-ATM under IP," IEEE/ACM Trans. on Networking, vol. 6, No. 2, pp. 117~129, April 1998.

[2] A. Adas, "Traffic models in broadband

networks," IEEE Commun. Mag., vol. 35, pp. 82~89, July 1997.

[3] X. -R. Cao and D. Towsley, "A performance model for ATM switches with general packet length distributions," IEEE/ACM Trans. on Networking, vol. 3, No. 3, pp. 299~309, June 1995.

[4] N. Mir, "Analysis of nonuniform traffic in a switching network," in ICCCN '98, 1998.

[5] P. Gupta, "Scheduling in input queued switches: a survey,"in citeseer.nj.nec.com/246798.html

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

[7] A. Mekkittikul and N. Mckeown, "A practical scheduling algorithm to achieve 100% throughput in input-queued switches," IEEE INFOCOM '98, pp. 792~799, 1998.

[8] D. Manjunath and B. Sikdar, "Variable length packet switches: delay analysis of crossbar switches under Poisson and self similar traffic," in IEEE INFOCOM '00, 2000.

[9] M. A. Marsan, A. Bianco, P. Giaccone E. Leonardi and F. Neri, "Packet scheduling in inputqueued cell-based switches," in IEEE INFOCOM '01,

[10] N. Mckeown, A. Mekkittikul, V. Anantharam, and J.Walrand, "Achieving 100% throughput in an input-queued switch," IEEE Trans. Commun., vol. 47, pp. 1260~1267, Aug. 1999.

[11] Y. Li, S. Panwar, H. J. Chao, "On the performance of a Dual Round-Robin switch," IEEE INFOCOM '01., vol. 3, pp. 1688~1697, April 2001.

[12] M. G. Hluchyj and M. J. Karol, "Queueing in high performance packet switching," IEEE J. Sel. Areas Commun., vol. 6, pp. 15871597, Dec. 1988.

[13] K. W. Sarkies, "The bypass queue in fast packet switching," IEEE Trans. on Commun., vol. 39, pp. 766~774, May 1991.

[14] A. Santhanam and A. Karandikar, "Window-based cell scheduling algorithm for VLSI implementation of an input-queued ATM switch, in IEE Proc.-Commun. Vol. 147, No. 2, April 2000.

[15] J. S. Choi and H. H. Lee, "Performance Study of an Input queueing ATM switch with windowing scheme for IP switching system, in IEE Proceeding of HPSR 2002, Kobe, Japan, May 2002.

─── 저 자 소 개 ───

최 진 식(정회원)-교신저자
1985년 서강대학교 전자공학과
      학사 졸업
1987년 한국과학기술원
      전자공학과 석사 졸업
1995년 한국과학기술원
      전자공학과 박사 졸업
1995년~2001년 공주대학교 정보통신과 교수
2001년~2004년 한국정보통신대학교 교수
2004년~현재   한양대학교 교수
<주관심분야: Optical Ethernet, Switching/ Routing, Personalized Multimedia network, Mobility management>

김 태 일(정회원)

양 미 정(정회원)