Achieving 90% Throughput in A Flow-Oriented Input-Queued Switching Router System

Geng-Sheng (G.S.) Kuo and Po-Chang Ko

National Central University

Department of Information Management

Information Technology Group

Chung-Li, Taiwan 32054 R.O.C.

Tel: +886 3 4263086; Fax: +886 3 4262309

e-mail: gskuo@imrnet.mgt.ncu.edu.tw

Abstract

Due to the head-of-line (HOL) blocking, the throughput limitation of an input-queued packet switching system with FIFO queues is $(2 - \sqrt{2}) = 0.586$ (Karol, 1986) (Karol, 1987). Previous results show that for all independent and identical arrival processes, or packets, the throughput can possibly achieve 100%, if some suitable queueing policy and scheduling strategy are used (Anderson, 1993) (Karol, 1992) (McKeown, 1998). However, in the *real situation* all arrival processes are absolutely not independently and identically distributed, they are *flow-oriented*. The main purpose of this paper is to study the throughput in a flow-oriented input-queued switching router system utilizing the characteristics of a flow, and to design three scheduling strategies - SFROCF, SFRF, and LFRF. They can achieve 100% throughput for all independent and identical arrival processes and obtain a better throughput than that of McKeown et al.'s OCF (McKeown, 1993) (McKeown, 1998) in real flow-oriented input-queued switching router system on future *broadband* Internet. Among them, SFRF can achieve 90% throughput approximately, the best one, in the situation. In addition, the three strategies do not lead to the permanent starvation of a non-empty queue, because the weight value of a cell will increase if its waiting time increases. In closing, McKeown et al.'s OCF is a special case of our proposed scheduling strategies when the flow has one cell.

Keywords

Flow-oriented, switching router, throughput, OCF, SFROCF, SFRF, LFRF

The original version of this chapter was revised: The copyright line was incorrect. This has been corrected. The Erratum to this chapter is available at DOI: [10.1007/978-0-387-35388-3_42](https://doi.org/10.1007/978-0-387-35388-3_42)

1 INTRODUCTION

Recently, the Internet is widely used and growing rapidly. The number of its hosts, www servers, and subnetworks has increased considerably (Gray, 1998) (Zakon, 1997). Much more multimedia-based *universal* information services will be developed on the Internet in the coming years. It implies that the traffic on future *broadband* Internet will be much busier than that of the current situation. Obviously, the router is going to be the most crucial bottleneck for future *global* broadband Internet. Therefore, it is necessary to improve the processing and performance of the router from an architectural viewpoint. The switching router is a new hot topic attracting attention in communications industry worldwide, which *integrates* switching and routing functions together to achieve better performance, e.g., IP switch, gigabit router, label switch, etc. (Newman, 1997).

However, it is well known that the maximum throughput of an input-queued switching router system is limited to $(2 - \sqrt{2}) = 0.586$ derived by Karol et al. (Karol, 1986) (Karol, 1987), if the following five conditions are satisfied. 1. FIFO queues are used in an $N \times N$ switching router, where $N \rightarrow \infty$. 2. Arriving packets are independently and identically distributed for each of N input lines during any one time slot. 3. The destined output line is uniformly distributed for each input line. 4. All arriving packets are of fixed and equal length, called cells. 5. Each input line has the same input probability distribution function (*p.d.f.*). Under these assumptions, the throughput limitation of an input-queued switching router system is due to the head-of-line (HOL) blocking.

Up to now, to the best of our understanding all studies and efforts on the throughput of the packet switching system are only concentrated on an input-queued switch, whose arriving cells are all independently and identically distributed. In the literature, many techniques, algorithms, and strategies have already been designed and proposed for the situation to reduce the HOL blocking and to increase its throughput (Anderson, 1993) (Chen, 1994) (Hluchyj, 1988) (Karol, 1992) (Obara, 1990). Both "*longest queue first*" (LQF) and "*oldest cell first*" (OCF) scheduling algorithms can achieve 100% throughput (McKeown, 1998). And, the *iSLIP* algorithm proposed in (McKeown, 1993) achieves 100% throughput by using some simulation method.

It is much more reasonable and realistic that the cells transmitted on real broadband Internet should be treated as many different sequences of cells. All cells in a *specific* sequence are coming from the same source node and destined for the same destination node, and are coined a *flow*. In other words, every cell arriving to a switching router can be treated as one cell belonging to some specific flow. All input flows are assumed to be independently and identically distributed. Every flow is labelled by one label number. Different cells belonging to the same flow are arriving to the same input line and destined for the same output line. The main purpose of this paper is to study the throughput in a flow-oriented input-queued switching router system, and to propose three scheduling strategies achieving 100% throughput in a cell-oriented input-queued switching router and obtaining

better throughput in a flow-oriented input-queued switching router by utilizing the above characteristics of a flow.

In Section 2, we briefly describe our switching router architecture, which is illustrated in Figure 1, and discuss the characteristics of a flow used in this paper. In Section 3, the scheduling strategies, LQF and OCF, are briefly mentioned. Although both two throughputs achieve 100% goal, they are *ideally* a maximum value. Some weaknesses need to be considered seriously on real Internet. At the same time, we propose our own scheduling strategies - "smallest flow rate and oldest cell first" (SFROCF), "smallest flow rate first" (SFRF), and "largesst flow rate first" (LFRF). In Section 4, we specify our simulation model and method, and compare the simulation results derived from our scheduling strategies with the results obtained in (McKeown, 1993) and (McKeown, 1998). Finally, some conclusions are made that our proposed scheduling strategies are better than McKeown et al.'s OCF in a flow-oriented input-queued switching router system on future broadband Internet.

2 FLOW-ORIENTED INPUT-QUEUED SWITCHING ROUTER ARCHITECTURE

Figure 1. Every arriving cell is treated as one cell belonging to some specific flow in a flow-oriented input-queued switching router architecture.

There are M input lines and N output lines in our proposed flow-oriented input-queued switching router architecture illustrated in Figure 1. If one cell arrives from input line m , where $1 \leq m \leq M$, and is destined for output line n , where $1 \leq n \leq N$, this cell will be placed in queue $Q_{m,n}$. Because every arriving cell is treated as one cell

belonging to some specific flow, each cell entering $Q_{m,n}$ is immediately transmitted to the corresponding flow queue according to its flow label number, FL. Let $FQ_{i,n}(t)$ denote the i th flow queue destined for the n th output line during the t th time slot. From the viewpoint of output line, there are k flow queues, $FQ_{1,n}(t)$, $FQ_{2,n}(t)$, ..., $FQ_{k,n}(t)$, serving the n th output line, where $1 \leq n \leq N$, during the t th time slot.

During the t th time slot, for every output line at most one cell is selected from one of these k flow queues, $FQ_{i,n}(t)$ for some i , where $1 \leq i \leq k$, and transmitted through the architecture, if there is at least one cell existing in $FQ_{1,n}(t)$ to $FQ_{k,n}(t)$. Let $FT_{i,n}(t)$ denote the number of time slots spent on the cell arrival to the HOL of $FQ_{i,n}(t)$ and at it during the t th time slot. In the proposed flow-oriented input-queued architecture, every flow has its own flow rate during the t th time slot, $FR(t)$. And, $FT_{i,n}(t) = (1 / FR_{i,n}(t))$. Let $w_{i,n}(t)$ denote the *weight value* of the cell at the HOL of $FQ_{i,n}(t)$ during the t th time slot, and $W_{i,n}(t)$ be the *waiting time* of the cell at the HOL of $FQ_{i,n}(t)$ during the t th time slot. In this paper, the weight value is defined by its scheduling strategy. And, the cell selected from these k flow queues and transmitted through this switching router for each output line is decided by the scheduling strategy accordingly.

3 SCHEDULING STRATEGIES – SFROCF, SFRF, AND LFRF

3.1 LQF scheduling strategy

The LQF scheduling algorithm was proposed in (McKeown, 1993) and (McKeown, 1998). The algorithm can ideally achieve 100% throughput only under the input cell distribution is uniform. Otherwise, its throughput can not achieve the 100% goal. Furthermore, this makes the permanent starvation situation occur. This situation is demonstrated in (McKeown, 1998), which needs more attention to consider carefully.

3.2 OCF scheduling strategy

The OCF scheduling algorithm was proposed in (McKeown, 1998). It considers the waiting time and assigns the weight value, $w_{i,n}(t) = W_{i,n}(t)$, to each cell at the HOL. If it has waited the longest time, the cell has the maximum weight value. During each time slot, OCF selects the cell with the maximum weight value to be transmitted. Therefore, no cell is starved for transmission indefinitely within a switching router system. Although it can overcome the permanent starvation of a non-empty queue, OCF still has one weakness behind. In a real physical situation, the traffic in a switching router system is flow-oriented. A flow is transmitted when one (VPI, VCI) pair exists, which is established between source node and destination node, in the switching router system. Two different flows may be established by two different nodes and have different flow rates, respectively.

Now, consider the following example. One 2x2 switching router system is illustrated in Figure 2. Suppose two arriving flows are transmitted to output line 1. One flow arrives from input line 1 with $FR=1$ and $FL=1$, and the other arrives from input line 2 with $FR=0.5$ and $FL=3$. On the other hand, there is only one flow destined for output line 2, which also arrives from input line 2 with $FR=0.5$ and $FL=2$. These three input flow orders and output cell orders by using OCF are shown in Figure 3. The number labelled in each time slot is the flow label number FL .

Figure 2. One 2x2 switching router system.

In Figure 3, it is obvious that there are many idle time slots wasted in output line 2. If the cells of Flow 2 are selected prior to the cells belonging to Flow 1 or 3, a better throughput can be achieved than that of OCF. If FR of Flow 2 is much smaller than that of Flow 1 or 3, much more cells of Flow 1 or 3 can exist between two cells of Flow 2. Because cells belonging to the same flow arrive from the same input line and destined for the same output line, some cells, destined for different output lines, must belong to different flows. Therefore, if the cells of some flow with smaller FR can be transmitted with higher priority, a better throughput can be achieved. Based on this important idea pointed out here, the following three new scheduling strategies are constructed.

Figure 3. Three flows are transmitted through the 2x2 switching router system using OCF scheduling strategy.

3.3 SFROCF scheduling strategy

The SFROCF is a new scheduling strategy proposed in this paper by us. In SFROCF, the weight value of the cell in HOL of a queue during the t th time slot is defined in terms of the following two components.

- $(1/FR(t))$: If the cell belongs to a flow with a specific flow rate at the t th time slot, $FR(t)$, the *initial* weight value of this cell at the t th time slot is defined as $(1/FR(t))$.
- $W(t)$: If the cell is destined for output line n , but is not selected to be transmitted during the t th time slot, its weight value will be increased 1 for the time slot.

The weight value $w(t)$ of the cell is defined as $(1/FR(t)) + W(t)$ by SFROCF. Its selection strategy is designed to select the cell having the maximum weight value. Although $w(t)$ is assigned as $(1/FR(t))$ initially, the cell, that has waited the longest time, will have the maximum weight value. It is very clear that no cell in a queue will be starved for transmission indefinitely.

Considering the example illustrated in Figures 2 and 3 again by using SFROCF, the input and output cell sequences are shown in Figure 4.

Figure 4. Three flows are transmitted through the 2x2 switching router system using SFROCF scheduling strategy.

Comparing Figures 3 and 4, it is obvious that the OCF spends 12 time slots to transmit these three flows, but SFROCF only spends 11 time slots. This situation is much clearer, if the difference between the maximum and minimum of all $(1/FR(t))$ increases and the number of input lines increases.

3.4 SFRF scheduling strategy

Another new scheduling strategy SFRF is proposed in this paper by us too. Its weight value of the cell in HOL of $FQ_{i,n}(t)$ during the t th time slot is defined in terms of the following two components.

- $(1/FR_{i,n}(t))$: If the cell belongs to a flow with a specific flow rate at the t th time slot, $FR_{i,n}(t)$, the *initial* weight value of this cell at the t th time slot is defined as $(1/FR_{i,n}(t))$.
- $W_{i,n}(t)$: If the cell is destined for output line n , but is not selected to be transmitted during the t th time slot, its weight value will be increased 1 for the time slot.

The weight value $w_{i,n}(t)$ of the cell is defined as $(1/FR_{i,n}(t)) + W_{i,n}(t)$ by SFROCF. Its selection strategy is designed to select the cell having the maximum weight value. The assigned weight value of the cell in HOL of $FQ_{i,n}(t)$ during the t th time slot is the same as that of SFROCF. Its selection strategy is the same as that of SFROCF. The only difference between SFROCF and SFRF is the switching router architecture, where SFROCF uses the FIFO input-queued architecture, and the other uses the FIFO flow-oriented input-queued architecture shown in Figure 1. In SFRF, every arriving cell enters its corresponding flow queue. According to the assigned weight value of the cell in HOL of these flow queues, SFRF can select one cell with the maximum weight value and transmit it during each time slot.

Using SFRF to review the example illustrated in Figures 2 and 3 again, the input and output cell sequences are shown in Figure 5.

Figure 5. Three flows are transmitted through the 2x2 switching router system using SFRF scheduling strategy.

3.5 LFRF Scheduling Strategy

Figure 6. Three flows are transmitted through the 2x2 switching router system using LFRF scheduling strategy.

In LFRF, the assigned weight value of the cell in HOL of $FQ_{i,n}(t)$ during the i th time slot is the same as that of SFRF. The switching router architecture used in this scheduling strategy is the same as that of SFRF too. The difference between SFRF and LFRF is that the FR of the flow, to which the cell selected by SFRF belongs, has the smallest value. On the other hand, if one cell is selected by LFRF, and the cell belongs to Flow x , the FR of Flow x has the largest value.

Similarly, using LFRF to review the example illustrated in Figures 2 and 3 again, the input and output cell sequences are shown in Figure 6.

4 SIMULATION RESULTS

In this section, we specify our simulation model and illustrate the simulation results in Figure 7 by using OCF, SFROCF, SFRF, and LFRF scheduling strategies, respectively. This simulation program was written in VB ver. 5.0 programming language and run in Microsoft Window 95 environment.

Figure 7. The simulation results using OCF, SFROCF, LFRF, and SFRF, respectively.

The arrival process used in this simulation model satisfies the following assumptions.

1. There is an arbitrary number of flow arrivals in each input line. All input flows are independently and identically distributed.
2. The (1/FR) of each flow is a random integer.
3. The cells belonging to the same flow have the same destination, that is, their output line numbers are the same.
4. The output line number, for which every flow is destined, is random.
5. Each cell in the HOL of a queue has an initial weight value according to the used scheduling strategy.
6. Considering the permanent saturation case, there is at least one cell arriving at each input line.
7. The calculation formula of effective throughput is the ratio of the total number of transmitted cells over the total number of arriving cells.
8. The switching router system used in this simulation model is 16x16.
9. The maximum number of traffic flows is 65535.
10. There may be one flow is ended and one new flow is generated during each time slot.
11. In this simulation model, the saturation case is considered.

If we neglect the transition state at the beginning of time axis (< 200 time slots approximately), it is clear that both effective throughputs of LFRF and SFRF are 90%, and both effective throughputs of SFROCF and OCF are 60% approximately. Precisely speaking, the effective throughput of SFROCF is better than that of OCF. It is a fact that OCF can achieve 100% throughput under one important assumption that all arrival processes are independently and identically distributed. For the real situation, the flow-oriented approach is much more suitable. The cell currently transmitted on the Internet may have some relationship with the cells transmitted previously. This relationship is much more obvious in an input-queued switching router system. And, in this flow-oriented situation, our proposed three scheduling strategies can achieve better throughput than that of OCF. Among them, SFRF can achieve 90% throughput, the best one at this moment.

5 CONCLUSIONS

From the results in (McKeown, 1998), we have known that 100% throughput is achievable in an input-queued switching router system, if the specific assumptions are satisfied and a suitable queueing policy and scheduling strategy are used. The OCF proposed in (McKeown, 1998) can achieve the goal of 100% throughput under the assumptions. Among them, one very important assumption is that all arrival cells are independently and identically distributed. However, in the real situation all arrival cells are absolutely not independently and identically distributed, they are flow-oriented, especially in future broadband Internet. Therefore, it is more reasonable and realistic to consider the throughput in a real flow-oriented input-queued switching router system.

In this paper, we consider the flow as a unit for transmission in the real switching router system, and propose three new scheduling strategies - SFROCF, SFRF, and LFRF, in order to achieve a better throughput. Every arriving flow has a specific flow rate. If the flow rate is much smaller than that of other flows, much more cells of other flows can exist between two cells of the flow. And, different flows selected to be transmitted might obtain the maximum output. After careful and real simulation study, some meaningful results are obtained, which are very impressive to us. Our proposed three new scheduling strategies can achieve a better throughput than that of OCF. SFRF utilizes the characteristics of a flow, and achieves 90% throughput approximately, which is the best one at this moment. Furthermore, the permanent starvation situation does not occur in using SFROCF, LFRF or SFRF, because the weight value of the cell in HOL of a queue is composed of two components, $(1/(\text{flow rate}))$ and *waiting time*. If one cell is not selected for transmission, its weight value will be increased. Eventually, the cell will have the maximum weight value and then be selected for transmission through the switching router system. In the future, it is our strong intention to improve the scheduling strategy for achieving 100% throughput in the real switching router system.

6 REFERENCES

- Anderson, T.E., Owicki, S.S., Saxe, J.B. and Thacker, C.P. (1993) High-Speed Switch Scheduling for Local-Area Networks. *ACM Trans. on Computer Systems*, **11**, no. 4, 319-52.
- Chen, M., Georganas, N.D. and Yang, O.W.W. (1994) A Fast Algorithm for Multi-Channel/Port Traffic Assignment. *Proc. of ICC '94*, 96-100.
- Gray, M.K. (1998) <http://www.mit.edu/people/mkgray>.
- Hluchyj, M.G. and Karol, M.J. (1988) Queueing in High-Performance Packet Switching. *IEEE J. Select. Areas Commun.*, **6**, no. 9, 1587-97.
- Karol, M.J., Eng, K.Y. and Obara, H. (1992) Improving the Performance of Input-Queued ATM Packet Switches. *Proc. of INFOCOM '92*, 110-5.
- Karol, M.J., Hluchyj, M.G. and Morgan, S.P. (1986) Input Versus Output Queueing on a Space-Division Packet Switch. *Proc. of GLOBECOM '86*, 659-65.
- Karol, M.J., Hluchyj, M.G. and Morgan, S.P. (1987) Input Versus Output Queueing on a Space-Division Packet Switch. *IEEE Trans. Commun.*, **COM-35**, no. 12, 1347-56.
- McKeown, N., Mekkittikul, A., Anantharam, V. and Walrand, J. (1998) Achieving 100% Throughput in an Input-Queued Switch (extended version). submitted to *IEEE Trans. Commun.*
- McKeown, N., Varaiya, P. and Walrand, J. (1993) Scheduling Cells in an Input-Queued Switch. *Electron. Lett.*, **29**, no. 25, 2174-5.
- Newman, P., Minshall, G., Lyon, T. and Huston, L. (1997) IP Switching and Gigabit Routers. *IEEE Commun. Mag.*, 64-9.

- Obara, H., Sasagawa, M. and Tokizawa, I. (1990) An ATM Cross-Connect System for Broadband Transport Networks Based on Virtual Path Concept. *Proc. of ICC '90*, 839-43.
- Zakon, R. (1997) Hobbes' Internet Timeline. *IETF RFC 2235*.

7 BIOGRAPHY

Geng-Sheng (G.S.) Kuo received his Ph.D. degree in systems engineering from Case Western Reserve University, Cleveland, Ohio, USA, in 1982. He then worked with R&D laboratories of the telecommunications industry in the United States, such as AT&T Bell Laboratories. In 1990, he came back to Taiwan and joined the Department of Information Management, Information Technology Group, at National Central University, where he is a professor. His current research interests include ATM-based broadband switching and networking technologies, and Internet service technologies. Currently, he is secretary of the IEEE Communications Society *Communications Switching Technical Committee*, editor for *Communications Architecture of IEEE Transactions on Communications*, a technical editor of *IEEE Communications Magazine*, etc.

Po-Chang Ko received his B.S. and M.S. degrees in electrical engineering from National Taiwan University, Taiwan, in 1989 and 1991 respectively. Since 1993, he has been an instructor at Van Nung Institute of Technology, where he teaches Computer Networks and Programming Languages. Currently, Mr. Ko is pursuing his Ph.D. degree in Information Management at National Central University, Taiwan. And, his research interests include design and implementation of switching router system, design of MAC for wireless LAN, and broadband Internet.

ACKNOWLEDGMENT

This work was supported by the National Science Council of the Republic of China under Grant NSC 87-2416-H-008-006.