RCES : A Replication/Contention/Elimination Strategy for Replication Baseline ATM Switch Architectures

Shiann-Tsong Sheu and Yueh-Ru Chuang
Department of Electrical Engineering, TamKang University
Tamsui, Taipei Hsien, Taiwan 251, Republic of China
E-mail : stsheu@ee.tku.edu.tw

Abstract

A Baseline switch architecture with several independent network planes has been proposed to enhance the switching efficiency. This paper proposes a replication/competition/elimination strategy (RCES) for the Dual-baseline switch architecture with two parallel network planes to increase the throughput and decrease the effect of the head of line (HOL). For each incoming cell, the RCES first duplicates it and forwards the copy cell into a different inlet in another plane to find another route to pass through network plane. Based on this strategy, cells will have a better chance to be switched out successfully even when the head of line occurs or burst traffic crowded toward a hot spot. When congestion occurs in the buffers, the cell behind another identical cell will be dropped immediately to reduce the cell loss probability. The performance of the proposed strategy is evaluated by simulations. Simulation results show that the proposed method obtains a higher throughput and lower cell loss probability than the traditional baseline network does.

Keywords: Head of Line (HOL), Hot Spot (HS), Replication/ Competition/ Elimination Strategy (RCES), Replication Baseline Network (RBN),.

1. Introduction

Telecommunication networks of today are passing through a rapid evolution. In Broadband Integrated Service Digital Networks (B-ISDN), the video, voice, and data services are transported over the same network. In order to provide different quality of services (QoS), asynchronous transfer mode (ATM) is introduced as a packing switching, transport, and multiplexing technique [1],[2]. However, even though its optical fiber network is capable of supporting tremendous bandwidth, the ATM network still suffers from the inefficient switching performance. It is desirable to design a high performance ATM switch to handle the flood of various data arriving at input ports. Recently, there are many researches about the switch architectures have been proposed [3],[4],[5],[6].

In ATM networks, the Baseline architecture has been used widely for supporting a multistage fast packet switches [7]. However, in a Baseline network, there is only one path for each pair of input/output ports, and any two paths may partially overlap each other in some stages. If there are two data flows requesting to pass through an overlapped link at the same time. Based on some competition or priority scheduling strategy, one of them will be queued in the buffer and the cells behind it are also queued in the buffer, this introduces the well-known head of line (HOL) problem. As a result, it decreases the network throughput significantly. In order to overcome the performance limitations of the category networks, various performance enhancing techniques have been introduced [4],[5],[8],[9],[10],[11],[12].

To realize a high speed asynchronous transfer mode (ATM) switch, several architectures for terabits per second (Tb/s) throughput have been proposed [13],[14],[15]. The input buffer and output buffer switches are the kind of crossbar switch, and they have respective buffer to connect to the input port or the output port. The HOL problem described above can be avoided by using the internal speed up technique, the parallel switch technique [16],[17], or employing another new switch architecture [18]. However, some of them are very difficult to implement or need to design a whole new architecture, the costs are very high. At some times, some algorithms of new switch architectures are not fair enough for cells destine to a same output port.

The concept of replicated network planes has been proposed in recent years. It is the kind of the parallel switch technique. The Baseline network is also the kind of the crossbar switch. The Replication Baseline Network (RBN) is constructed from two independent Baseline networks to distribute incoming cells into two planes to release the contention during the switching process. In this paper, we propose a replication/competition/elimination strategy (RCES) for the RBN architecture. Based on the proposed strategy, when a cell arrives at an input port, an identical cell will be replicated to the input port with the shortest queue length in another plane. Both cells are forwarded separately in two independent network planes with different priorities. The major concept is to use another identical cell to create or seek the other faster route (or uncontested route) to arrive the same output port. Moreover, if there is a data flow which is suffering from

This work was supported by National Science Council, R.O.C., under Contract NSC 88-2213-E-032-007.

the head of line (HOL) or this route is more crowded (such as a hot spot), the other identical data flow may be lucky enough to arrive the output port. For simplicity, this paper only discusses two independent planes in RBN. The proposed strategy can be easily extended for the multiple planes. The rest of the paper is organized as following. Section 2 describes the architecture of the Replication Baseline Network (RBN) with two parallel planes. Section 3 presents the competition and elimination strategy for the Baseline network. Simulation models and results of the RCES are shown in Section 4. Finally, section 5 summarizes the conclusions.

2. Replication Baseline Network (RBN) Architecture

The RBN architecture is composed of three major parts: 1) distribution and copy network (DCN), 2) router, 3) output-port plane selector (OPS) as shown in Fig. 1.

2.1 Distribution and Copy Network (DCN)

The DCN contains N distribution and copy blocks (DC Blocks) and one for each input port as shown in Fig. 2 and Fig. 3. The DC Block is responsible for performing distribution and replication. When a cell arrives at a DC, it will be forwarded into a plane selector to randomly select a plane which this cell will be passed through. To maintain the cell sequence, the incoming cell in inlet *i* will be forwarded to input queue *i* in the selected plane. For simplicity, this cell is referred as original cell *O*. If cell *O* selects the primary plane, the secondary plane port selector (SS) will replicate an identical cell (which is referred as replicated cell *R*) and forward it into secondary plane. To avoid cell *R* passing along a same route as cell *O*, cell *R* is forwarded to the queue with the shortest queue length. To do this, a set of signals feedback from the router part is necessary to indicate the queue length of each input buffer in the router. On the other hand, if cell *O* selects the secondary plane, the primary plane port selector (PS) will be activated to replicate and forward as mentioned above. Before a cell forwards into a plane, it will be first added three fields of stage, port and position of the other identical cell. When a cell forwards to the next stage, it will inform its identical cell to modify the content of the three fields. By this way, the two identical cells are able to compare their locations right away. Because there are two identical cells trying to find a fast route to the destination port, this approach is most likely to confuse the cells sequence. Therefore, when the cells arrive at a destination port, they should be rescheduled to maintain the original cell sequence. Moreover, DCN needs to add the sequence number and the inlet identifier (inlet ID) for each cell to perform rescheduling.

2.2 Router

The router consists of two parallel Baseline network planes. The property of a Baseline network is briefly described as follows. Shown in Fig. 4, in a Baseline network, let *N* and *P* denote the number of the input/output ports and the number of stages in the network, respectively. Therefore, the relation between *N* and *P* is $P=log_2N$. Each stage contains *N/2* 2×2 switching elements (SE). For each SE, there are two basis states: straight forwarding state and crossed forwarding state. According to the property of self routing, an arriving cell with the binary destination address $D=d_1d_2d_3\cdots d_P$ (where $d_i\in\{0,1\}$) will be switched according to the corresponding binary values in the destination address field. That is, if d_i=0 (d_i=1), the corresponding switching element in stage *i* will state in the straight forwarding state (crossed forwarding state).

A 2×2 switch element is the essential forwarding mechanism in Baseline network. For each input inlet, one internal buffer is allocated for temporarily storing incoming cells as shown in Fig. 5. The timing of forwarding a cell to the next stage is controlled by a switch controller. The switch controller can compare the positions between the HOL cell in this queue and the other identical cell in the other plane by checking the three fields of the HOL cell. Moreover, according to the three fields, the switch controller can find the precise position of the other identical cell, then the priority swapper can execute the contention resolution algorithm. The detailed contention resolution algorithm will be described in section 3. Because there are two identical cells in different planes to compete each other to the same destination port. One of them should be eliminated when the other cell already reaches the destination port or its site is far behind of the other cell (i.e., it has a little chance to catch up with the other cell). Let leading degree *LD* is the specified distance threshold value. By the threshold value (*LD*), a priority swapper (PS), as shown in Fig. 5, can decide a HOL cell should be dropped, just be exchanged the priority with its identical cell or keep the original condition. If the distance between cells *O* and *R* is larger than *LD*, the lagged cell will be dropped immediately. If another identical cell is eliminated, the left cell is denoted as single cell *S* (the signal cell *S* has the highest priority).

To accomplish the switching process, each internal buffer needs two extra mechanisms : a contention counter (CC) and a priority swapper (PS) as shown in Fig. 5. The CC is used to resolve the switching priority when two competing cells have a same priority. Each time the HOL cell fails in contention, the corresponding CC is incremented by one. On the contrary, the other CC will be reset to zero. If the switch controller can not determine the winner via their priorities, the cell with a higher counter value will be selected. Recall that two identical cells are switched on different planes in parallel. If cell *R* moves further than its original cell *O*, the priorities of them will be swapped right away to speed up the switching process. The all possible operation conditions between cell *O* and its replication cell *R* according to the leading degree *LD* are shown in Table1.

We note that the considered ATM architecture is cell loss free during forwarding in the router part. That is, before forwarding the front cell into the next stage, the controller will check the buffer space in the next stage. If there is available space, this cell will be forwarded.

Otherwise, this cell must wait. Therefore, the cell loss in the RBN architecture only happens at the input ports in the first stage.

Table1. The possible operations between cell *O* and its replication cell *R*.

Cell type/ position		Distance	Operation
cell *O*	cell *R*		
Rear	Front	< LD	Exchange priority ($O \leftrightarrow R$)
		≥ LD	Drop original cell *O* and change cell *R* into single cell *S*
Front	Rear	< LD	Do not change
		≥ LD	Drop replication cell *R* and change cell *O* into single cell *S*

2.3 Output-port Plane Selector (OPS):

When a cell leaves router, it will be forwarded into the output-port plane selector (OPS), as shown in Fig. 6. In OPS, a selector, which connects two links from different planes, is used to select which cell will be first passed into the rearrangement mechanism. As shown in Fig. 7, the rearrangement mechanism contains an inserter, a counter, an output buffer and a switch unit. These queued cells will be rescheduled according to the additional sequence number and the inlet ID by DCN. A counter is to record the number of cells queued in the output buffer at present. An output buffer will be dynamic divided into several groups to store these cells from different inlets and with different sequence number. For example, when a cell with a sequence number 3 and inlet ID 5 arrives the inserter, the inseter first checks the sequence number and inlet ID of the cell. Then the inserter will search the group with the same sequence number 3 and insert the cell in it (the group). If the inserter can not find the group with the same sequence number 3, then it will add a group 3 and insert in the output buffer in order. (because these groups are located in the output buffer in order) The search adopts the hash method.

A switch unit contains a switching lookup table which is shown in Fig. 7. Each entry in this table records two information : the inlet ID and the sequence number of the last switching cell. The switch unit will search all output buffers one by one to find out the next outgoing cell. The determining process is described as follows. If the cell with the next sequence number is found, it will be forwarded in the next slot time. Otherwise, if the output buffer is not full, the switch unit will wait for the next cell. In the case of buffer full, the switch unit will select the cell whose sequence number is the closet one to the recorded sequence number in table. Each time a cell is being switched out, the corresponding entry will be modified to maintain the switching sequence. We note that this scheme may solve the HOL problem in the output buffer.

3. The Contention Resolution Algorithm

When two cells content for a same outlet of SE in the next stage, only one of them will be switched successfully. The way to select a proper cell to forward is described in this section. For simplicity, let two internal buffers are denoted as Ba and Bb, respectively (see Fig. 8). Before describing this strategy, each cell is given the priority according to its relative position and the role it plays. Owing to cell type *S* must complete the switching process by itself, it is assigned the highest priority. On the other hand, cells *O* and *R* are assigned the second and the lowest priorities, respectively.

When the contention happens on two cells with different priorities, the highest priority cell will be selected. If two contention cells are *S* cells, the RCES will select the one with a higher CC. However, if these two cells are either type *O* or type *R*, the decision becomes more complicate. To make a proper decision, the distance between each cell *O* and its cell *R* is taken into considerations. If both competing cells are type *O* (*R*), the cell leading (lagging) farther its copy should be selected. If the estimated distances of them are still equal, the CC is referred as mentioned before. Based on this strategy, the leading cell is trying to go further to drop its *R* cell. Contrarily, the lagging cell will do its best to catch up with its *O* cell. The detailed flow chart of proposed contention resolution algorithm is shown in Fig. 9.

4. Simulation Model and Simulation Results

The performance of proposed RCES was investigated by simulation. The simulation model considers a $N \times N$ ($P=log_2N$) RBN with two network planes. Let *IB* denote the internal buffer size of each SE. Moreover, let *OB* denote the buffer size of the rearrangement mechanism in OPS. The traffic arrival rate in each input port is a Poisson distribution with a mean λ. The frame length is an exponential distribution with a mean of *L* cells. Therefore, the total traffic load Λ for the switch is defined as $\Lambda = N \times \lambda \times L$. We note that the heavy loaded condition (saturated traffic load) occurs when $\Lambda = N$ ($\lambda \times L = 1$).

To investigate the effect of proposed strategy, the hot spot scenario was being considered in our simulation models. Let *HS* ($0 \leq HS \leq 1$) denote the proportion of incoming traffics destine to a hot spot output port. The hot spot output port is randomly selected from *N* output ports. The destinations of remainders are uniform distributed among the other *N-1* output ports. It is intuitive that a higher *HS* occurs, a much serious congestion will occur on SEs and the HOL problem will become much obvious. To investigate the effect of proposed strategy, the switch throughput and the cell loss probability are measured. Figures 10(a) and 10(b) show the obtained throughput and cell loss probability of RCES under different *HS*s and mean length *L* when N=8, $P=3$, *IB*=30, *OB*=30, λ=0.2, *LD*=45. In this simulation, *HS* is considered from 0 (uniform distribution) to 0.6 (highly hot spot) in a step of 0.2. Moreover, the mean length is considered from 2.0 to

6.0 in a step of 0.5. (That is, the total network load is investigated from light load 3.2 to heavy load 9.6 and the saturated traffic load occurs when L=5) We can see that the RCES obtains a higher saturated throughput than the traditional dual-Baseline networks when the traffic load is more higher and the hot spot degree is lower ($HS<0.4$). We note that the RCES does not improve system throughput and cell loss probability obviously when $HS \geq 0.4$. This is because that lots of cells are congested on the hot spot output port. Though the RECS has the ability to find another route from different plane to reach destination, the total throughput still can not be improved. Intuitively, this drawback can be solved by employing a large output buffer size in the OPS.

In RBN, because a cell has two routes to the OPS, the arrival cells need to be rescheduled in the rearrangement mechanisms. The amount of allocated output buffers will affect the system performance. Figures 11(a) and 11(b) show the obtained results under different output buffer sizes OB when N=8, $P=3$, IB=30, λ=0.2, LD=45, HS=0.0. It is clear that given a larger buffer size, a higher throughput and a lower cell loss probability will obtain.

To observe how the frame length L and arrival rate λ affect the performance of proposed strategy, we consider the case that the total traffic loaded is constant (Λ=0.8). In this simulation, the other parameters are N=8, $P=3$, IB=40, OB=60, LD=45, and HS=0.0. The frame length is considered from 2 to 20 in a step of 2. That is, the arrival rate varies from 0.05 to 0.005 to meet the constant traffic load. In Figures 12(a) and 12(b), we can see that the system throughput is degraded when incoming frame length grows. From these derived curves, the throughput and cell loss probability of the dual-Baseline network with RCES are being improved about 7% and 10%, respectively, no matter what frame size is.

In this simulation model, we always select the value of leading degree (LD) is 45. Because the major purpose of LD just decides the lagging cell should be dropped or not, it does not have much significance to improve the throughput and cell loss probability, as shown in Fig. 13(a) and 13(b). But we still can increase the throughput and reduce the cell loss probability by selecting a proper LD.

Finally, Table 2 shows the performance obtained by employing different switch sizes (N) of RBN when IB=20, OB=40, LD=45, Λ=0.8, λ=0.2 and HS=0.0. We can see that the RCES is able to enhance the throughput and cell loss probability dramatically.

5. Conclusion

In this paper, we proposed the replication /competition/elimination strategy (RCES) for the RBN ATM architecture to improve the performance. Based on RCES, incoming cell will be duplicated into another plane to find another route to its destination. Such strategy will shorten the switching delay and both switch throughput and the cell loss probability are being further improved. To release the generated double traffic load, a threshold is referred to drop the lagged cells. In addition, the RCES provides an efficient contention scheme instead of the traditional random selection. Simulation results shown that the RCES improve the throughput as well as the cell loss probability.

References

[1] M. De Prycker, "Asynchronous Transfer Mode Solution for Broadband ISDN", U.K. : Ellis Horwood Ltd., 1991.

[2] J. Le Boudec, "The Asynchronous Transfer Mode: A Tutorial", Computer Networks ISDN Systems, vol. 24, pp. 279-309, 1992.

[3] R. Y. Awdeh and H. T. Mouftah, "Survey of ATM Switch Architectures", Computer Networks ISDN Systems, vol. 27, 1995.

[4] J. J. Degan, G. W. R. Luderer, and A. K. Vaidya. "Fast packet technology for future switches", AT&T Tech. J., Mar./Apr. 1989.

[5] A. Pattavina, "Nonblocking Architectures for ATM Switching", IEEE Communication Magazines, vol. 31, Feb, 1993.

[6] E. W. Zegura, "Architecture for ATM Systems", IEEE Communication Magazine, vol. 31, Feb. 1993.

[7] C. L. Wu and T. Y. Feng, "On a Class of Multistage Interconnection Networks," IEEE Transactions on Computer, vol. 29, pp. 694-702, 1980.

[8] C. Catania and A. Pattavina, "Analysis of Replicated Banyan Networks with Input and Output Queueing for ATM Switching", in Proceedings of IEEE ICC'96, pp. 1685-1689, Dallas, Texas, June 1996.

[9] C. A. Fun and J. Silvester, "A New Parallel Banyan ATM Switch Architecture", in Proceedings of IEEE ICC'95, pp. 523-527, Seattle, Washington, June 1995.

[10] J. Y. Hui, "Switching and Traffic Theory for Integrated Broadband Networks", Kluwer Academic Publishers, Norwell, MA (1991).

[11] Sema F. Oktug, and Mehmet U. Caglayan, "Design and Performance Evaluation of a Banyan Network Based Interconnection Structure for ATM Switches", IEEE Journal on Selected Areas in Communications, vol. 15, no. 5, June 1997.

[12] Christos Kolias and Leonard Kleinrock, "The Dual-Banyan (DB) Switch: A High-Performance Buffered-Banyan ATM Switch", in Proceedings of IEEE ICC'97, June, 1997.

[13] T. Chaney, J. A. Fingerhut, M. Flucke, and J. S. Turner, "Design of a Gigabit ATM Switch", in Proceedings of IEEE INFOCOM'97, pp. 2-11.

[14] K. Y. Eng, M. J. Karol, G. J. Cyr, and M. A. Pashan, "Design and Prototype of a Terabit ATM Switch Using a Concerntrator-based Growable Switch Architecture", in Proceedings of ISS'95, vol. C3.7, 1995.

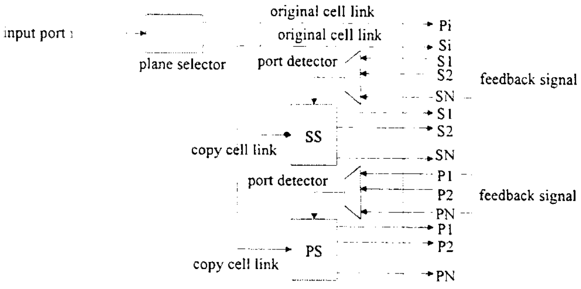
[15] N. Yamanaka, K. Endo, K. Genda, H. Fukuda, T. Kishimoto, and S. Sasaki, "320 Gb/s High-speed ATM Switch System Hardware Technologies Based on Copper-polyimide MCM", IEEE Transactions on

Computer Package Manufacture Technology, vol. 18, pp. 83-91, 1995.

[16] Y. Oie, T. Suda, M. Murata, D. Kolson, and H. Miyahara, "Survey of Switching Techniques in High-speed Networks and Their Performance", in Proceedings of IEEE INFOCOM'90, pp. 1242-1251, 1990.

[17] J. S. Turner, "Design of a Broadcast Packet Switching Networks", IEEE Transactions on Communications,
vol. 36, pp. 734-743, 1988.

[18] Eiji Oki, Naoaki Yamanaka, "Tandem-crosspoint ATM Switch with Input and Output Buffers", IEEE Communications Letters, vol. 2, no. 7, July 1998.



P : primary plane
S : secondary plane
PS : primary plane port selector
SS : secondary plane port selector

Figure 3. The detailed block diagram of the distribution and copy block (DC Block).

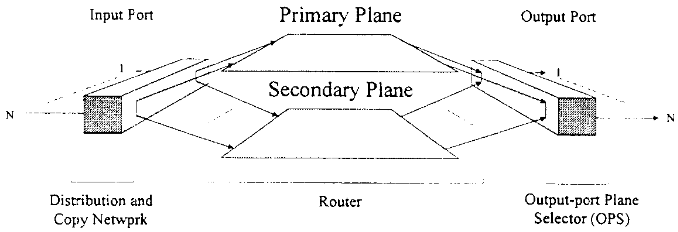

Input Port
Primary Plane
Output Port
1
Secondary Plane
1
N
N
Distribution and Copy Netwprk (DCN)
Router
Output-port Plane Selector (OPS)

Figure1. The architecture of RBN

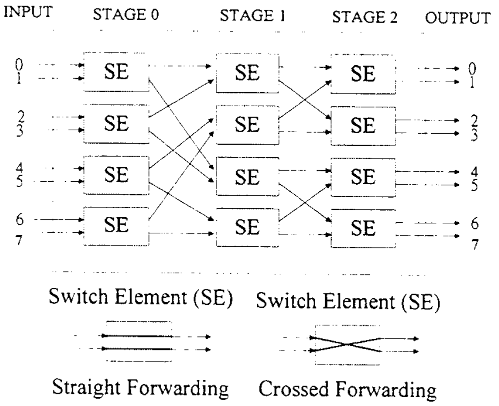


Figure 4. A 8*8 Baseline network architecture (N=8, P=3).

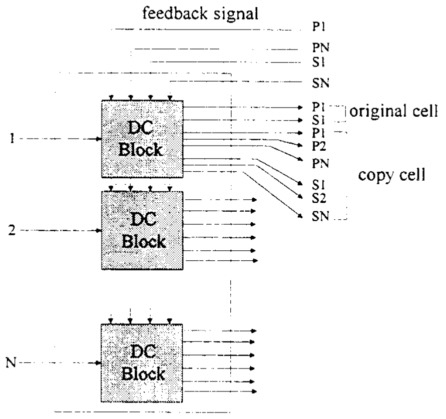


Figure 2. The block diagram of the distribution and copy network (DCN).

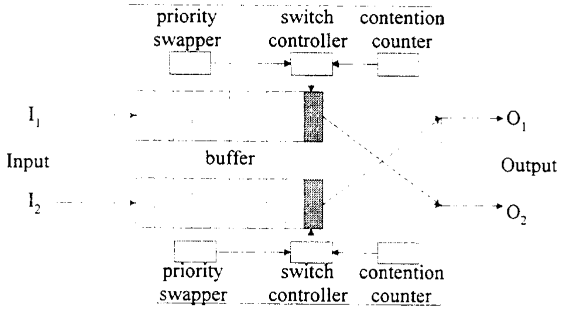
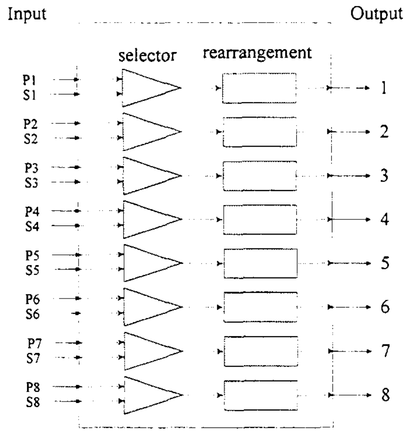


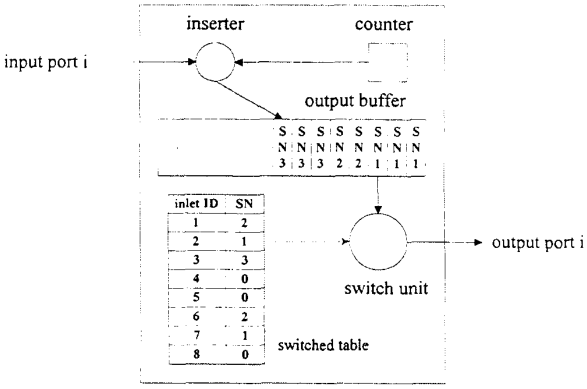
Figure 5. The block diagram of a switch element (SE).



P : Primary Plane

S : Secondary Plane

Figure 6. The block diagram of the output-port plane selector (OPS) (N=8).



inlet ID	SN
1	2
2	1
3	3
4	0
5	0
6	2
7	1
8	0

Figure 7. The structure of the rearrangement.

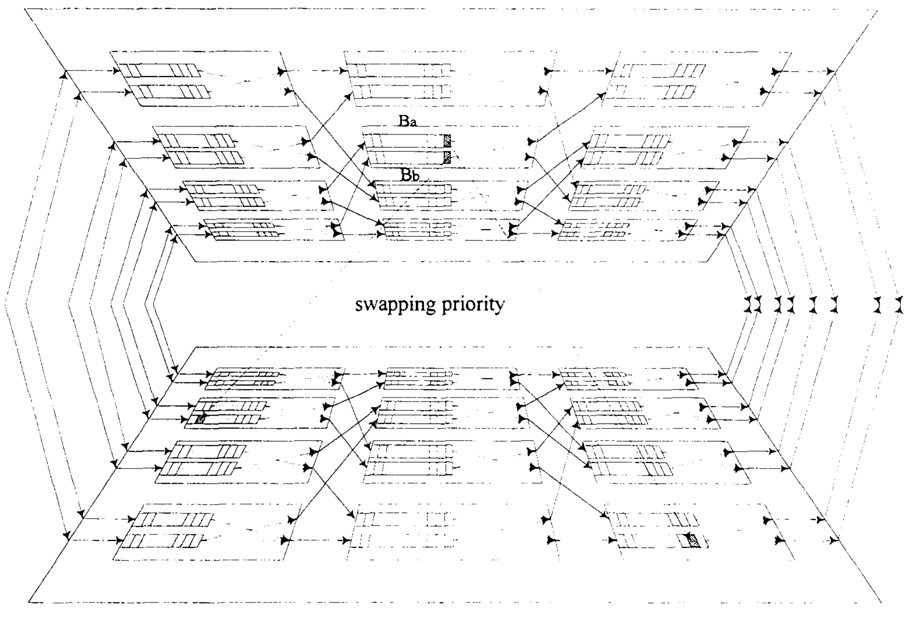


Figure 8. An example of priority swapping .

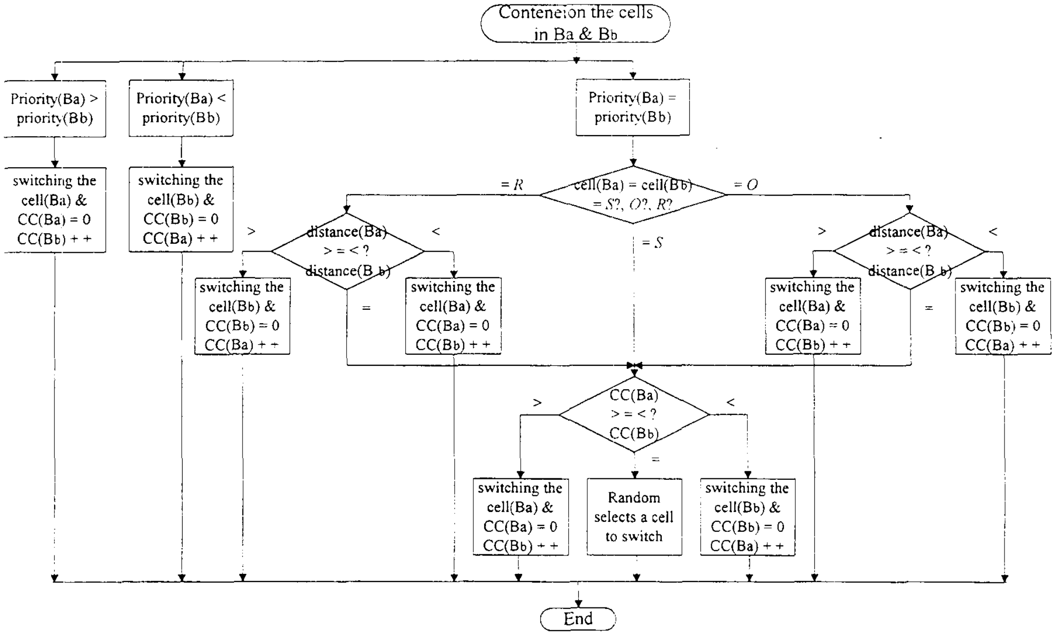
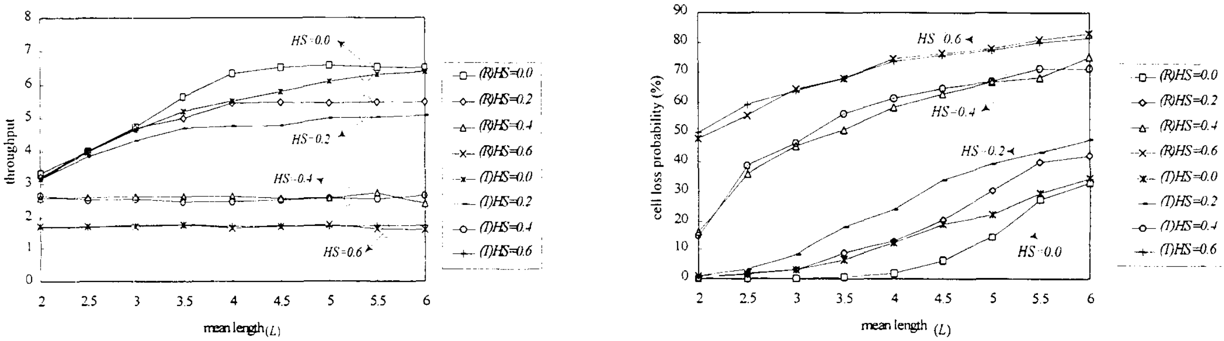
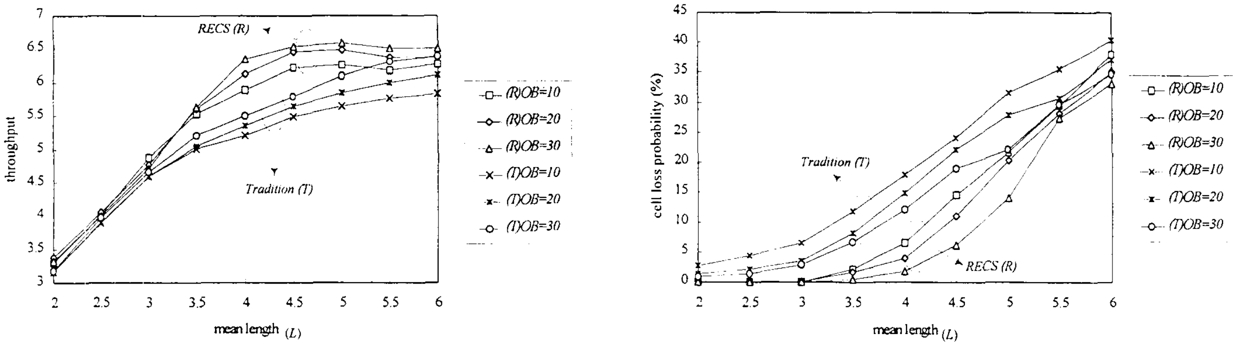

Conteneion the cells in Ba & Bb
Priority(Ba) > priority(Bb)
Priority(Ba) < priority(Bb)
Priority(Ba) = priority(Bb)
switching the cell(Ba) & CC(Ba) = 0 CC(Bb) + +
switching the cell(Bb) & CC(Bb) = 0 CC(Ba) + +
= R
cell(Ba) = cell(Bb) = S?, O?, R?
= O
= S
distance(Ba) > = < ? distance(Bb)
>
<
=
switching the cell(Bb) & CC(Bb) = 0 CC(Ba) + +
switching the cell(Ba) & CC(Ba) = 0 CC(Bb) + +
CC(Ba) > = < ? CC(Bb)
Random selects a cell to switch
End

Figure 9. The flow chart of the contention resolution algorithm when the front cells in Ba and Bb are destined to a same outlet of SE.



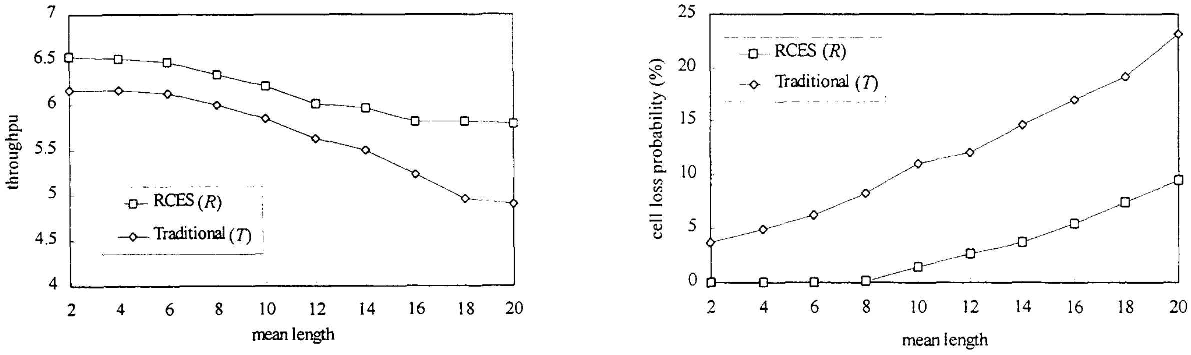
(a) (b)

Figure 10. Comparisons of throughput and cell loss probability obtained by RECS (*R*) and traditional approach (*T*) under different hot spot conditions (*HS*) when *N*=8, *P=3*, *IB*=30, *OB*=30, λ=0.2 and *LD*=45.



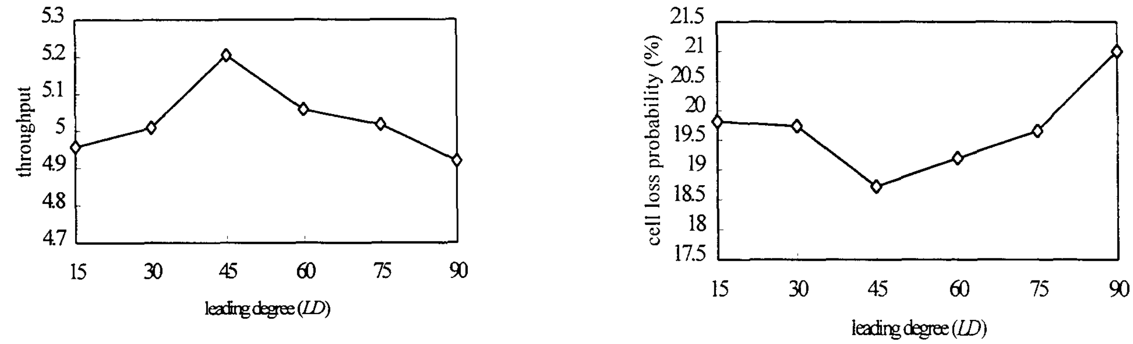
(a) (b)

Figure 11. Comparisons of throughput and cell loss probability obtained by RECS (*R*) and traditional approach (*T*) under different output buffers (*OB*), when *N*=8, *P=3*, *IB*=30, λ=0.2, *LD*=45 and *HS*=0.0.



(a) (b)

Figure 12. Comparisons of throughput and cell loss probability obtained by RECS (*R*) and traditional approach (*T*) under different arrival rates (λ) and mean length (*L*), when *N*=8, *P=3*, *IB*=40, *OB*=60, Λ=0.8, *LD*=45 and *HS*=0.0.



(a) (b)

Figure 13. Comparisons of throughput and cell loss probability obtained by RECS (*R*) under different leading degree (*LD*) and mean, when *N*=8, *P=3*, *IB*=30, *OB*=30, Λ=0.8, λ*=0.16* and *HS*=0.2.

Table2. Comparisons of obtained performances between the traditional method and the RCES under different switch system, when *IB*=20, *OB*=40, Λ=0.8, λ=0.2, *LD*=45 and *HS*=0.0.

Network Size	Strategy	Throughput	Cell loss Probability (%)
4×4	RCES	3.2	3.1
	Tradition	2.6	17.3
8×8	RCES	6.5	1.1
	Tradition	5.2	18.2
16×16	RCES	12.7	0.7
	Tradition	10.5	17.9