ELSEVIER

Computer Communications 22 (1999) 1540–1548

computer communications

www.elsevier.com/locate/comcom

A fault-tolerant architecture for ATM networks

C.-C. Lo*, C.-Y. Chiou

Institute of Information Management, National Chiao-Tung University, 1001 Ta Hsueh Road, Hsinchu 300, Taiwan

Received 2 February 1999; received in revised form 29 April 1999; accepted 29 April 1999

Abstract

The asynchronous transfer mode (ATM) is the transfer mode recommended for the broad integrated service digital network (B-ISDN) by ITU-T. In this paper, we propose a self-routing fault-tolerant switching architecture for ATM networks. The proposed architecture uses subswitches and extra links to provide alternative paths; hence, can tolerate multiple faults. Analytical results show that the total number of redundant paths increases exponentially as the size of the network increases. A simulation model is developed. Simulation results indicate that the proposed architecture is much more fault-tolerant and cost-effective than those architectures found in the literature. Simulation results also illustrate that the proposed architecture still maintains a high throughput with an acceptable cell delay time, even when the number of faulty elements increases. © 1999 Elsevier Science B.V. All rights reserved.

Keywords: Fault-tolerant; Redundant path; Survival probability; Cost-effectiveness ratio; Throughput; Cell delay

1. Introduction

The broadband integrated services digital network (B-ISDN) provides end-to-end transport for a wide range of broadband services in a flexible and efficient manner. To fulfill the requirements of B-ISDN, the asynchronous transfer mode (ATM) is recommended for B-ISDN by ITU-T. An ATM switch should provide considerable capacity that may only be achieved using a distributed and highly parallel architecture [1–3]. Therefore, the multistage interconnection network, such as the Banyan network, will play an important role in building ATM switches. Banyan networks [4] are self-routing, cost-effective, and very efficient at handling uniform traffic. However, because of the existence of a unique path for each input–output pair, any single fault in a link or a switching element (SE) may render some output ports unreachable from certain input ports.

Interconnection networks, which continue to provide service even when they contain faulty components, are known as *fault-tolerant* networks [5,6]. Although fault tolerance is essential to ATM systems, few fault-tolerant switching networks have been proposed for ATM. Adam and Siegel proposed the extra stage cube (ESC) network [7] which adds an extra stage to the input side of the cube network. Although the ESC network is robust in the presence of multiple faults, it requires extra logic and extra computation effort. The multipath omega network proposed by Padmanabham and Lawrie [8] provides multiple paths by adding redundant switching stages to the omega network, but the switching elements become more complicated. The augmented C-network proposed by Reddy and Kumar [9] is derived from the C-network. Its routing scheme is far more complicated than the C-network, and it loses the self-routing property of the C-network. The augmented shuffle-exchange network (ASEN) proposed by Kumar and Reibman [10] is also fault-tolerant, but the throughput of ASEN degrades rapidly when the number of faulty components increases. Tzeng et al. [11] proposed a simple scheme to enhance the fault-tolerance of multistage interconnection networks by creating multiple paths between each input–output pair through extra links between SEs in the same stage. This scheme requires a simple routing algorithm and allows a low level of fault tolerance with reasonable cost. The switching throughput of the network degrades quickly when the number of faulty components increases. Itoh [12] presented a self-routing fault-tolerant ATM switching network which adds many subswitches among switching stages. This network maintains a high throughput when the number of faulty elements increases. However, a large number of redundant SEs is required.

In this paper, we propose a self-routing fault-tolerant ATM switching network whose design idea originates from Tzeng et al. [11] and Itoh [12]. Simulation results indicate that the proposed network not only keeps a high

* Corresponding author. Tel.: + 886-3-5731909; fax: + 886-3-5723792.

E-mail address: cclo@cc.nctu.edu.tw (C.-C. Lo)

PII: S0140-3664(99)00118-8

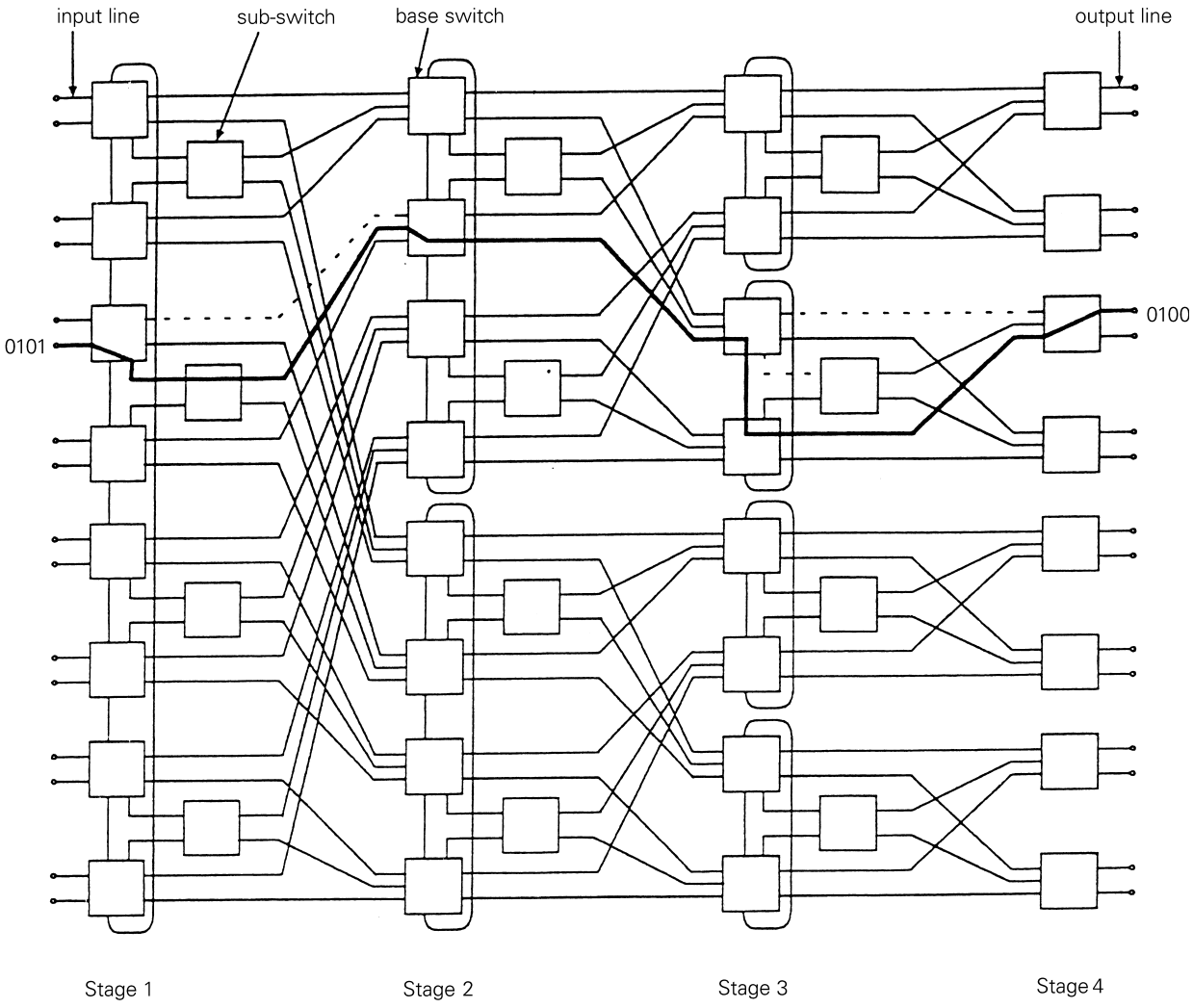

Fig. 1. Proposed architecture.

level of fault tolerance but also maintains a high throughput with an acceptable cell delay time, even when the number of faulty components increases.

In Section 2, we describe the enhanced fault-tolerant switching architecture and its routing algorithm, followed by the analysis of the total number of elements and redundant paths of this architecture. Section 3 defines the selected metrics for measurements. In Section 4, simulation results are presented and analyzed. Section 5 concludes this paper with possible future directions.

2. Proposed architecture

Tzeng's network provides low fault tolerance and throughput, however the cost is reasonable. In contrast, Itoh's network supports high fault tolerance and throughput, but it requires substantial amount of redundant SEs. The motivation behind this research is to examine the pros and cons of these two schemes, and then derive a new fault-tolerant architecture.

The proposed network provides multiple paths by adding subswitches between switching stages as Itoh's network proposed and adding extra links between SEs in the same stage which is similar to Tzeng's network.

2.1. The fault-tolerant ATM switching architecture

The fault-tolerant ATM switching network we propose is an $N \times N$ network, with N input ports and N output ports. The network is constructed by adding Rank-1 subswitches between switching stages of the network as described in Ref. [12]. Subswitches are added between base switches, so each switching stage except the last one contains both base switches and subswitches. Except for those in the right-most stage, each base switch has a chain-in link, a chain-out link, and an extra link to the subswitch, in addition to the original input and output links. To simplify the discussion, a 16×16 ATM network, shown in Fig. 1, is used to illustrate

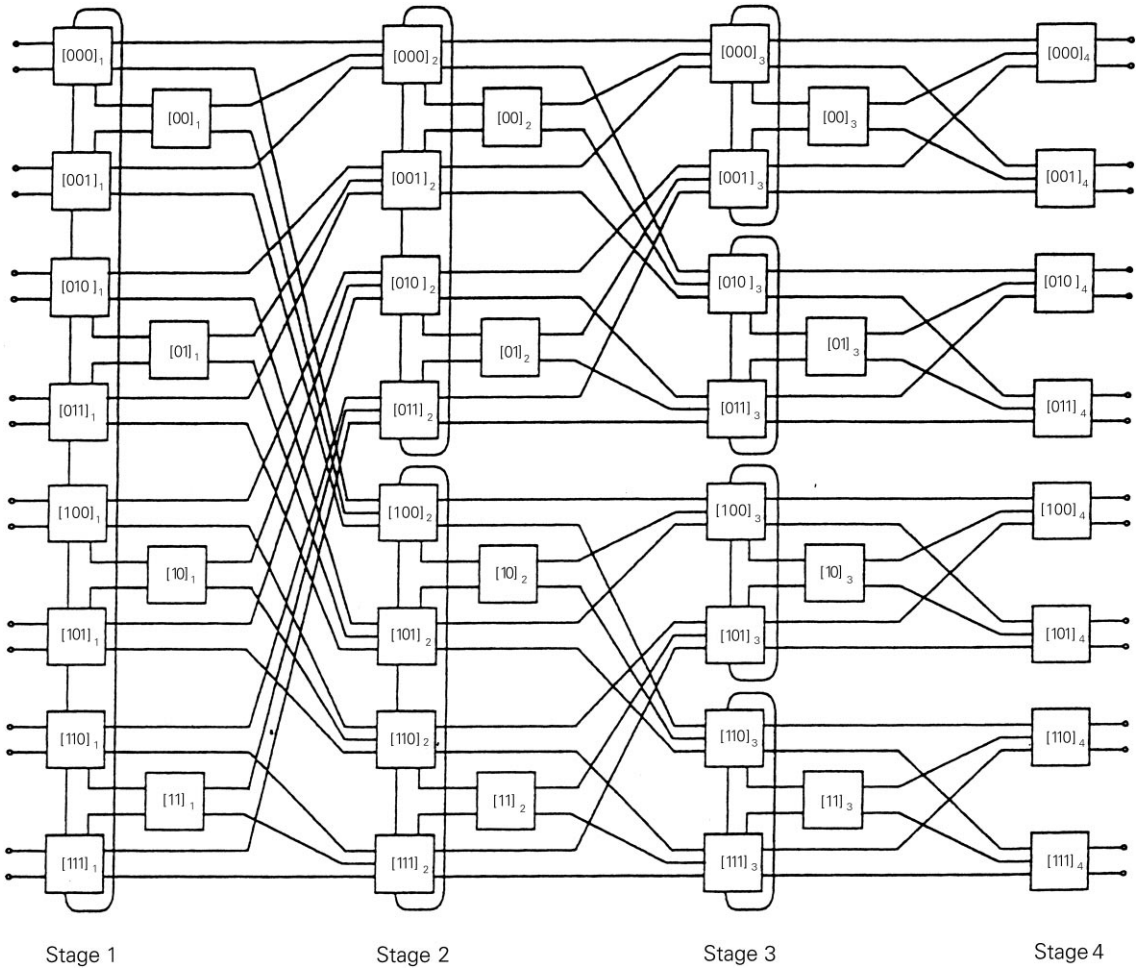

Fig. 2. Labelling scheme of the proposed architecture.

the proposed architecture. There are three types of base switches: 3×4, 4×4, and 3×2. The subswitches are implemented with 2×2 crossbars.

Fig. 2 shows the labeling scheme used for the SEs of the exemplary 16×16 network. The stages are labeled from 1 to 4, with 1 for the leftmost stage. SEs are named by binary bits, $[p_{n-1}\,p_{n-2}\ldots p_k]_s$. The symbol s represents the stage number. The symbol $p_{n-1}\,p_{n-2}\ldots p_k$ denotes the location of the SE in stage s. k is equal to 1 for a base switch, and 2 for a subswitch. For instance, $[p_{n-1}\,p_{n-2}\ldots p_2]_1$ indicates a subswitch in stage 1, and $[p_{n-1}\,p_{n-2}\ldots p_1]_3$ indicates a base switch in stage 3.

Using this labeling scheme, we can describe a link wiring algorithm for the proposed network. The two output links of a base switch $[p_{n-1}\,p_{n-2}\ldots p_1]$ in stage s $(s<n)$ are connected to two base switches in the next stage: $[p_{n-1}\,p_{n-2}\ldots p_{n-s+1}\,0p_{n-s}\ldots p_2]_{s+1}$ for the upper output link, and $[p_{n-1}\,p_{n-2}\ldots p_{n-s+1}1p_{n-s}\ldots p_2]_{s+1}$ for the lower one. The redundant path from the base switch $[p_{n-1}\,p_{n-2}\ldots p_1]_s$ in stage s is connected to a subswitch $[p_{n-1}\,p_{n-2}\ldots p_2]_s$ in the same stage. The chain-out link of a base switch $[p_{n-1}\,p_{n-2}\ldots p_1]_s$ in stage s is connected to a base switch $[m+((p_{n-1}\,p_{n-2}\ldots p_1/k) \bmod k)]_s$ in the same stage, where $k=N/2^s$, $m=[p_{n-1}\,p_{n-2}\ldots p_1/k]\times k$. Thus the chain-out links in the same stage form chains as shown in Fig. 2.

In this paper, we regard an element as the basic network unit. An *element* in the ith stage $(1\le i\le n)$ is formed by the output module of an SE in the ith stage, its connecting link, and the input module of the next SE in the (i or $i+1$)th stage to which the link is connected. The chain-out module of an SE, its connecting link, and the chain-in module of the next SE also constitute an element. We consider an element as the basic network unit, because when an element fails, the path between an input–output pair using a faulty element is broken. We can determine the number of elements in Tzeng's, Itoh's, and the proposed networks, as

$$L_{\text{Tzeng}}=\frac{3}{2}N(n-1), \tag{1}$$

$$L_{\text{Itoh}}=3nN-5N+4, \tag{2}$$

$$L_{\text{proposed}}=\frac{5}{2}N(n-1), \tag{3}$$

respectively, where L is the total number of elements, n the

Table 1
Total number of elements of each of the three networks

n	4	5	6	7	8	9	10
Tzeng's network	72	192	480	1152	2688	6144	13 824
Itoh's network	116	324	836	2052	4868	11 268	25 604
Proposed network	120	320	800	1920	4480	10 240	23 040

number of stages of the network, and $N(=2^n)$ the network size.

Table 1 compares the total number of elements in each of the three networks. It is clear that Tzeng's network requires the least number of elements, since it is designed to provide limited fault tolerance with low cost. The proposed network requires less elements than Itoh's network. Notice that when the number of stages (n) is 10, the proposed network uses 10% less elements than Itoh's network.

2.2. Routing algorithm

The routing scheme we propose is essentially the same as the one used in a regular Banyan network. Assume that an input port, labeled $s_1\ s_2 \ldots s_n$, is to be connected to a network output port, labeled $d_1\ d_2 \ldots d_n$. At stage i, bit d_i of the destination address is used to route a request in an "augmented" SE. If $d_i = 0$, it will be routed to the upper outlet of the SE, and if $d_i = 1$, to the lower one. If the desired outlet is blocked due to a conflict, a link failure, or a failure of the SE in the next stage to which the output link is connected, the request will be routed through the redundant path to the subswitch. The d_i bit is used in exactly the same way in the subswitch. If the extra link to the subswitch also fails or is blocked, the request will be routed through the chain-out link to another SE within the chain. The same d_i bit will be utilized with the identical routing scheme in the new SE. If the destined outlet in the new SE is again blocked, this request will be routed to yet another new SE in the same chain. A request can be routed through as many SEs within the chain as required. Eventually, an appropriate output link can be found and the request will be able to proceed to the next stage. For example, the alternative routing path from the input port 0101 to the output port 0100, as shown by the bold line in Fig. 1, is selected under the condition that link faults, as indicated by the broken line in Fig. 1, occur in some parts of the network.

2.3. Redundant path

With the routing algorithm aforementioned, we can calculate the total number of redundant paths in the proposed network by induction. We define *the number of redundant paths as the total number of paths minus 1*. For each SE, there are $R(i, j)$ paths to any reachable output port in the last stage, where i indicates the number of stages from the SE to the output port, and j indicates the type of SE ($j = 0$ for base switches; 1 for subswitches). Hence $R(n, 0)$ represents the number of paths from a base switch in the leftmost stage to a reachable output port, and $R(n, 1)$ indicates the number of paths from a subswitch in the leftmost stage to a reachable output port. Now, we can calculate the number of paths as follows:

For a 2×2 switch network:

$$R(1,0) = 1.$$

For a 4×4 switch network:

$$R(2,1) = R(1,0) = 1,$$

$$R(2,0) = [R(1,0) + R(2,1)] \times 2 = 2^2 \times R(1,0).$$

For a 8×8 switch network:

$$R(3,1) = R(2,0),$$

$$R(3,0) = [R(2,0) + R(3,1)] \times 4 = 2^3 \times R(2,0),$$

and so on.

By induction, the total number of paths can be obtained for any network size. When the network sizes (N) is 2^n, the total number of paths can be expressed as follows:

$$R(n,1) = R(n-1,0),$$

$$\begin{aligned} R(n,0) &= [R(n-1,0) + R(n,1)] \times 2^{n-1} \\ &= 2^n \times 2^{n-1} \times 2^{n-2} \times \cdots \times 2^2 \times R(1,0) \\ &= 2^{(1/2)((n-1)(n+2))} \end{aligned}$$

$R(n, 0)$ indicates that the total number of redundant paths increases exponentially as the size of the network increases.

Table 2 shows the total number of paths for n, where n is the number of stages in the network. For comparison, we

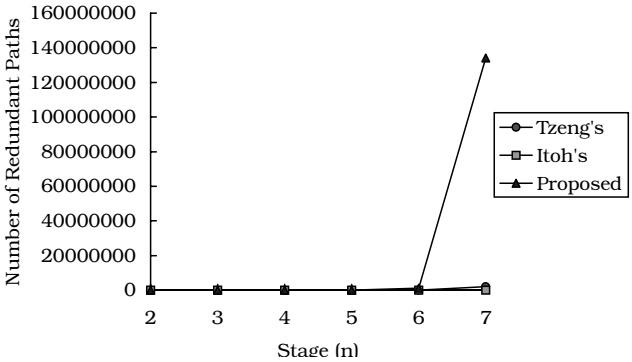

Fig. 3. Number of redundant paths.

Table 2
Total number of paths of each of the three networks

n	2	3	4	5	6	7
Tzeng's network	2	8	64	1024	32 768	2 097 152
Itoh's network	2	5	14	42	132	429
Proposed network	4	32	512	16 384	1 048 576	134 217 728

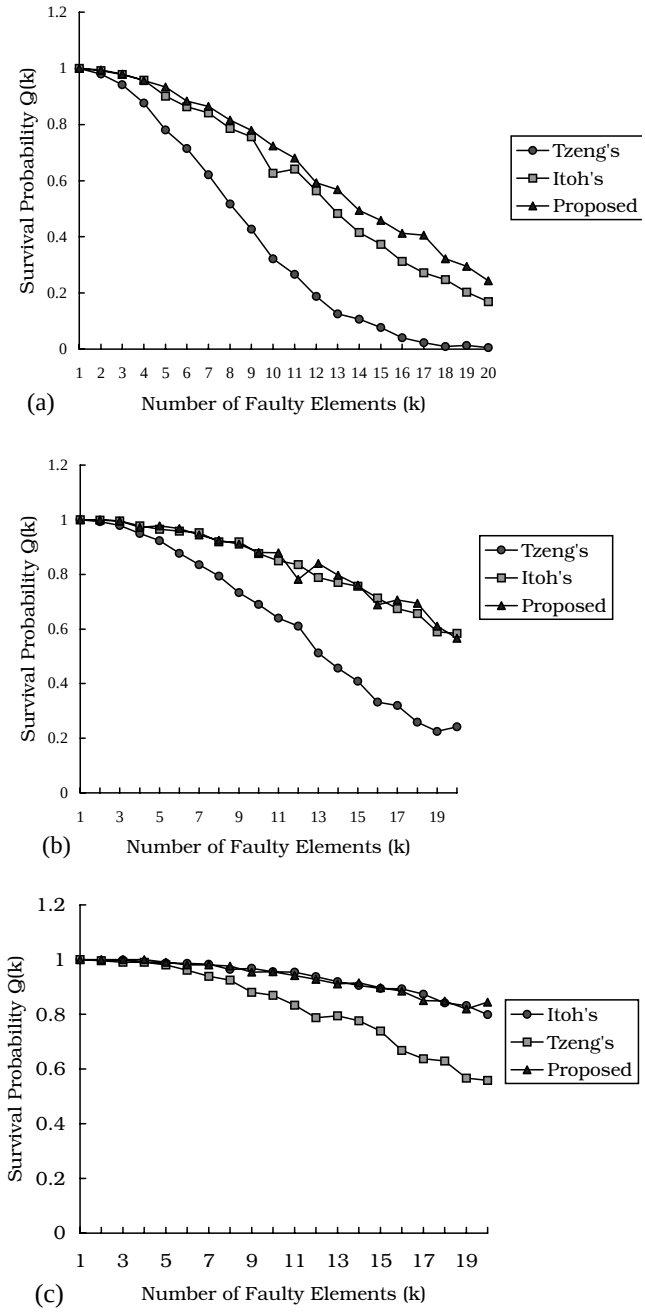

Fig. 4. $Q(k)$ versus number of faulty elemets for: (a) $n = 4$ (16×16); (b) $n = 5$ (32×32); and (c) $n = 6$ (64×64).

have included the total number of paths calculated from Tzeng's network and Itoh's network in Table 2.

Fig. 3 presents an analysis of data given in Table 2. It illustrates that the proposed architecture has a dramatic improvement over Tzeng's and Itoh's networks in terms of the redundant path.

3. Selected metrics

In order to evaluate the level of fault tolerance and the performance of the proposed architecture, we have chosen five metrics for measurements. The definitions of these five metrics are given below:

- Survival probability ($Q(k)$): Survival probability is defined to be the probability of an entire network surviving with k faulty elements. In other words, $Q(k)$ is the probability, given that k elements in the network are faulty but they do not cause the network to fail.
- Level of fault tolerance (α): The probability that k or less faults cause failure of the network, denoted as $P(k)$, is computed from:

$$P(k) = 1 - Q(k). \tag{4}$$

Let $p(i)$ be the probability that the ith fault will cause the entire network to fail. We get

$$P(k) = \sum_{i=2}^{k} p(i). \tag{5}$$

From (4) and (5), we can obtain all $p(i)$'s recursively. Let α be the expected number of faulty elements that will cause the entire network to breakdown, then

$$\alpha = \sum_{i=2}^{L} ip(i), \tag{6}$$

where L is the total number of elements. In essence, α represents the maximum faults the network is expected to tolerate. It is an indicator of the level of fault tolerance. The greater the value of α, the larger the level of fault tolerance.

- Cost-effectiveness ratio (CER): At a first glance, it seems

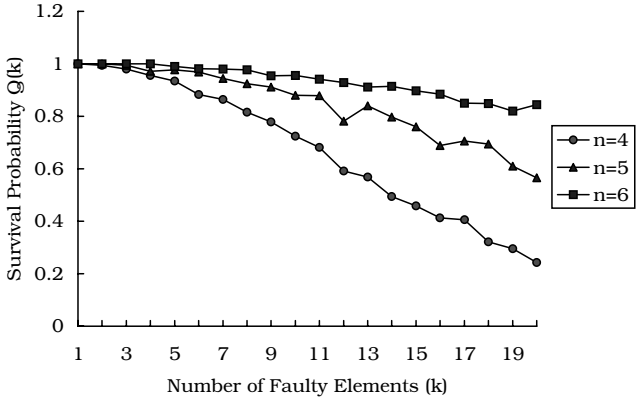

Fig. 5. $Q(k)$ versus number of faulty elements for the proposed network.

Table 3
Level of fault tolerance (α) of each of the three networks

n	4	5	6
Tzeng's network	9.04	14.63	23.11
Itoh's network	14.06	23.76	38.57
Proposed network	15.59	25.25	40.18

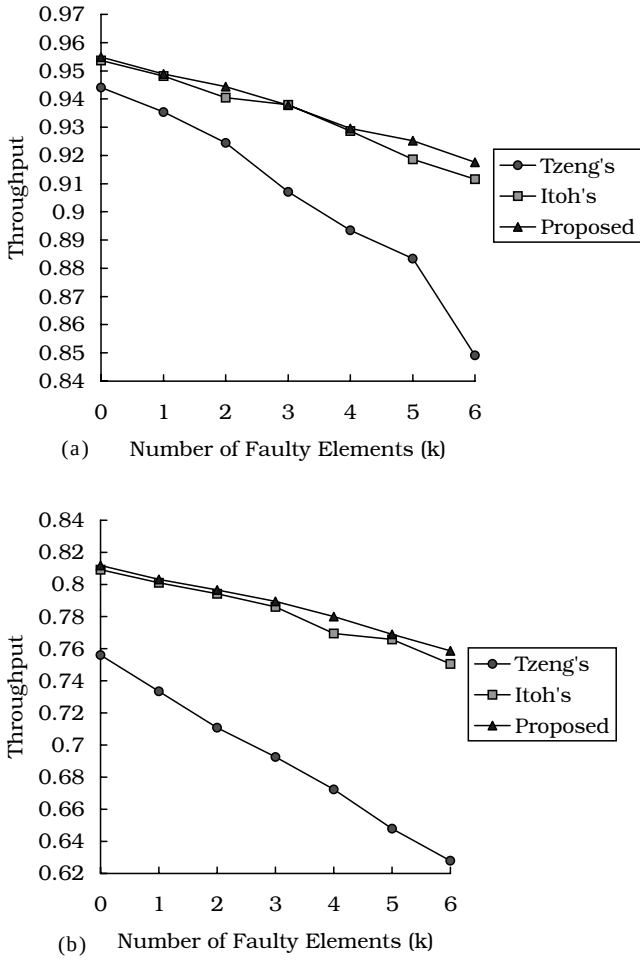

Fig. 6. Throughput versus number of faulty elements for: (a) $n = 4$ (16×16) and load = 0.4; (b) $n = 4$ (16×16) and load = 0.8.

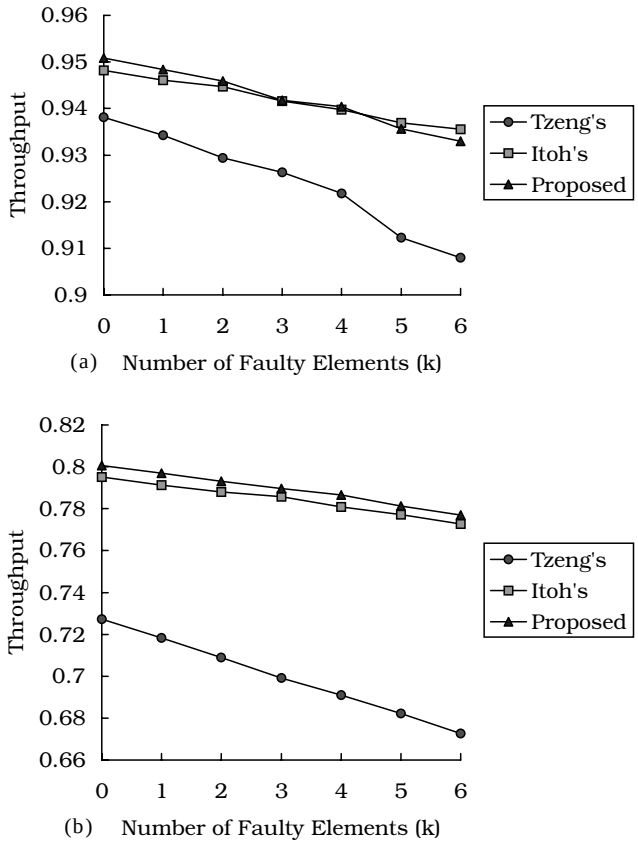

Fig. 7. Throughput versus number of faulty elements for: (a) $n = 5$ (32×32) and load = 0.4; (b) $n = 5$ (32×32) and load = 0.8.

reasonable to compare the networks with the same size. However, the total number of elements differs greatly. To compare the networks fairly, we define the cost-effectiveness ratio (α/L') as the ratio of the expected number of faults, α, to the number of redundant elements, L'. Note that L' is derived by the following equality:

$$L' = L - L_B,$$

where L is the total number of elements in the fault-tolerant network and L_B the total number of elements in the original Banyan network.

- Throughput (T): Throughput can be defined as the average number of output cells per unit time. However, for the sake of comparing cases with different input loads, we redefine the throughput as the ratio of the total number of output cells to the total number of input cells. Conceptually, the latter definition is a normalized value.
- Cell delay (D): Cell delay is defined as the total time for a cell to pass the entire network. Further, we define the additional delay time as the cell delay time minus the total time for a cell to pass the entire network when there are no fault and no contention.

Table 4
Comparison of cost-effectiveness ratio (n: number of stages; α: expected number of faulty elements causing network failure; L': number of redundant elements; and CER = α/L')

n	Itoh's network			Proposed network			(CERP − CERI)/CERI (%)
	α	L'	CERI (%)	α	L'	CERP (%)	
4	14.06	68	20.67	15.59	72	21.65	4.72
5	23.76	196	12.12	25.25	192	13.15	8.50
6	38.57	516	7.47	40.18	480	8.37	11.98

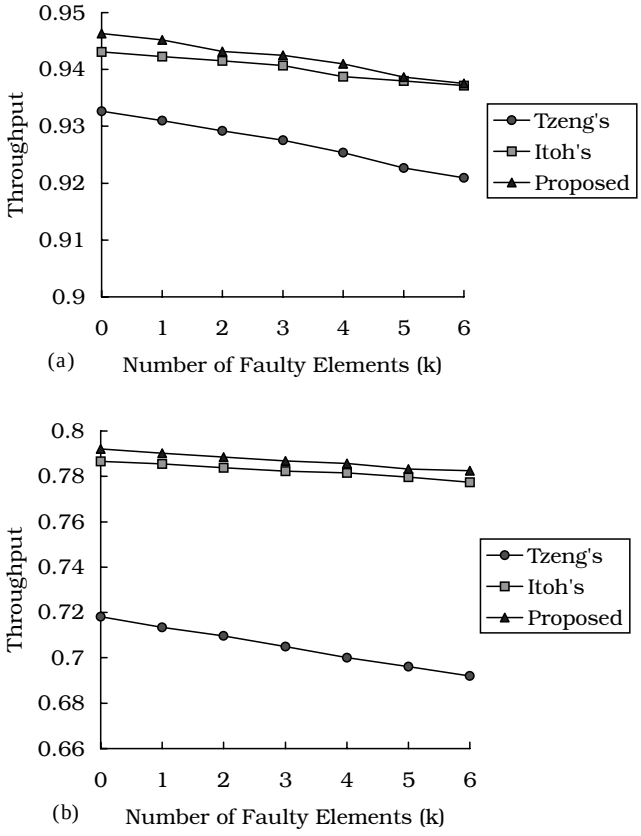

Fig. 8. Throughput versus number of faulty elements for: (a) $n = 6$ (64×64) and load = 0.4; (b) $n = 6$ (64×64) and load = 0.8.

4. Simulation and analysis

Simulation programs are written in C and are running on the SUN SPARC workstation.

4.1. Simulation assumptions

The following assumptions are used in the simulations:

1. An element is considered to be faulty if any of its components is faulty; a faulty element is permanently unusable.
2. The elements considered are only those between the first stage and the last stage. This implies that the input ports of the fist stage and the output ports of the last stage are not considered.
3. Events in which an element becomes faulty are independent and occur randomly.
4. A network is considered to have failed when the faulty elements prevent the connection of any path between an arbitrary input–output pair. This implies that the criterion of fault tolerance is the retention of full access capability.
5. Cell arrivals are random and independent, and input loads are uniformly distributed over all output ports of the network.

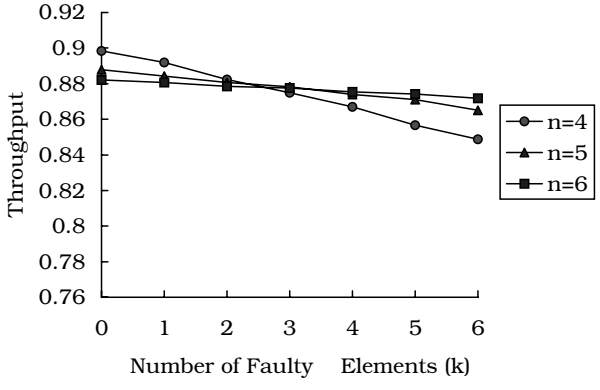

Fig. 9. Throughput of the proposed network with different number of network stages (n) for load = 0.6.

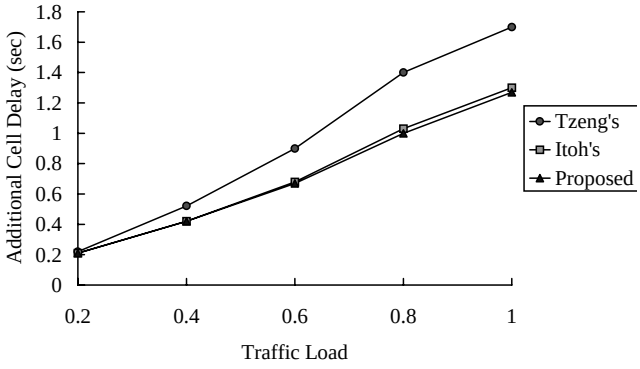

Fig. 10. Additional cell delay time versus traffic load for $n = 4$ (16×16).

4.2. Results and analyses

4.2.1. Survival probability

The three fault-tolerant networks, Tzeng's network, Itoh's network, and the proposed network, are considered. $Q(k)$s are obtained by computer simulation. $Q(k)$s for networks of different stages $4(N = 16)$, $5(N = 32)$, and $6(N = 64)$ are shown in Fig. 4(a)–(c), respectively.

As shown in Fig. 4(a)–(c), the proposed network performs far more reliably than Tzeng's network, and its survival probability is close to Itoh's network. Moreover,

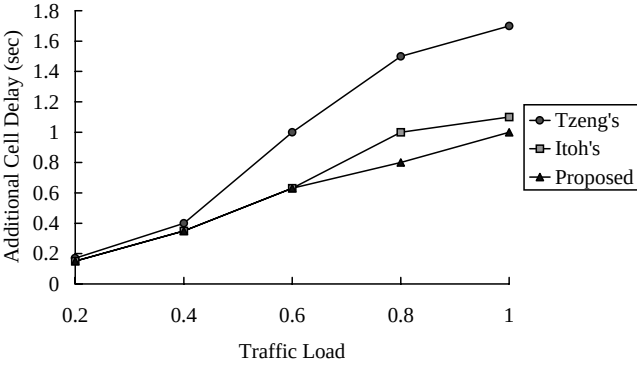

Fig. 11. Additional cell delay time versus traffic load for $n = 5$ (32×32).

Fig. 5 shows that as the network size increases, so is the probability of survival of the proposed network. This is because a larger network will have a larger number of elements; hence, can provide more redundant paths. This is an attractive property since the probability of survival of most fault-tolerant networks deteriorates as their sizes grow. For an ATM network which requires a large number of switching elements, this property is essential.

4.2.2. Level of fault tolerance

The values of α for the three networks are compared in Table 3. As expected, α increases when the number of stages, n, increases. This is because a larger network will provide more redundant paths, and therefore will be more fault-tolerant. This is in sharp contrast to a network with a unique path between each input–output pair, such as the regular Banyan network.

It is shown in Table 3 that the proposed network has the highest fault tolerance among the three networks. The level of fault tolerance of the proposed network is much higher than Tzeng's network (about 74% higher for each network size).

4.2.3. Cost-effectiveness ratio

Table 4 compares the cost-effectiveness ratio between the proposed network and Itoh's network. It shows that the proposed network is more cost-effective than Itoh's network. Further, as the size of the network increases, the cost-effectiveness of the proposed network increases.

4.2.4. Throughput

Throughputs are illustrated in Figs. 6–8, for networks of different stages and various traffic loads.

From simulation results, we notice that throughput decreases as the number of faulty elements increases for all three networks. This is due to the decrease of the number of redundant paths. Also, the throughput of the three networks decreases as their traffic loads increase, since the internal contention is more likely to occur when the traffic load is heavy. Nevertheless, the proposed network is able to keep the highest throughput among the three networks.

Fig. 9 shows that for the same traffic load and different numbers of faulty elements, the larger the network size, the less the decreasing rate of the throughput. This is because a larger network has more redundant paths; hence, can alleviate internal contention caused by faulty elements.

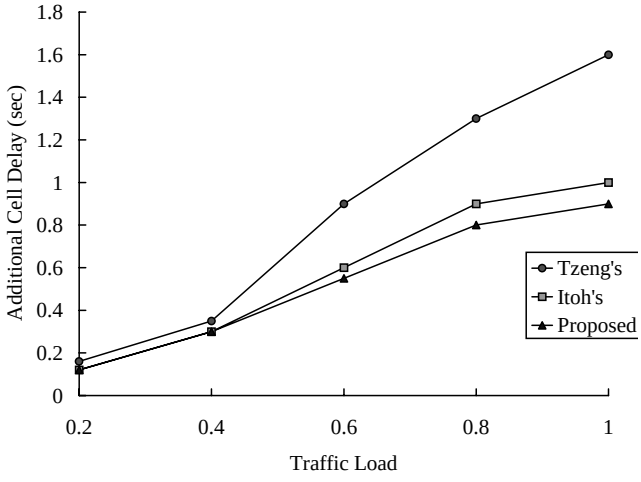

Fig. 12. Additional cell delay time versus traffic load for $n = 6$ (64×64).

4.2.5. Cell delay

We evaluate the cell delay time in terms of a uniform traffic pattern. Every SE has an internal buffer slot. Simulation results are shown in Figs. 10–12, for networks of different stages $4(N = 16)$, $5(N = 32)$, and $6(N = 64)$, respectively. The number of faulty elements, from one to six, is generated randomly. We notice that Tzeng's network has the highest cell delay time. Cell delay of the proposed network and that of Itoh's network are close. However, for the proposed network, the rate of decrease of cell delay is higher than that of Itoh's network, as network size increases.

4.2.6. Summary

Table 5 gives a comparison among Tzeng's, Itoh's, and the proposed networks, with respect to the selected metrics. Overall, the proposed network has the best performance and the highest level of fault tolerance, when network has faulty elements.

5. Conclusions

In this paper, we proposed a fault-tolerant ATM switching architecture. In essence, it is a self-routing Banyan

Table 5
Comparison between each of the three networks with respect to the selected metrics

Metrics	Tzeng's network	Itoh's network	Proposed network
No. of elements	Small	Large	Medium
No. of redundant paths	Medium	Small	Large
Survival probability $Q(k)$	Low	Middle	High
Level of fault tolerance (α)	Low	Middle	High
Cost-effectiveness ratio (α/L')	Low	Middle	High
Throughput (T)	Small	Medium	Large
Cell delay (D)	Long	Medium	Short

network. This architecture is an enhancement of Itoh's and Tzeng's networks. The routing algorithm is quite simple and no extra computation is necessary. It provides multiple paths by adding subswitches between switching stages and adding interstage links between switches within each stage. The total number of redundant paths between each input–output pair in the proposed network is far greater than those of the fault-tolerant networks found in the literature. Thus, the network can tolerate more faulty components. Subswitches and interstage links can also enhance the network performance by using redundant paths, which alleviate internal contention among cells. Simulations results indicate that the proposed network has the following merits: its level of fault-tolerance is high; it maintains a high throughput and a low cell delay time, even when the number of faulty elements increases; and it is very cost-effective.

A possible future research is to investigate the possibility of chaining up the subswitches. These additional chains will increase the number of redundant paths substantially; hence, may improve the fault tolerance of the proposed network.

References

[1] J.S. Turner, New direction in communications, IZS'86, A3, 1986, pp. 1–8.
[2] E.P. Rathgeb, T.H. Theimer, M.N. Huber, ATM switches—basic architecture and their performance, International Journal of Digital and Analog Cabled Systems 2 (1989) 227–236.
[3] F.A. Tobagi, Fast packet switch architecture for broadband integrated services digital networks, Proceedings of the IEEE 78 (1) (1990) 133–166.
[4] C. Wu, T. Feng, On a class of multistage interconnection networks, IEEE Transactions on Computers C-29 (8) (1980) 694–702.
[5] G.B. Adams III, D.P. Agrawal, H.J. Siegel, A survey and comparison of fault-tolerant multistage interconnection networks, IEEE Computer (1987) 14–27.
[6] R.J. McMillen, A survey of interconnection networks, IEEE Globecom 84, 1984, pp. 105–113.
[7] G.B. Adams, H.J. Siegel, Modifications to improve the fault-tolerance of the extra stage cube interconnection network, Proceedings of the International Conference on Parallel Processing, 1984, pp. 169–173.
[8] K. Padmanabham, D.H. Lawrie, A class of redundant path multistage interconnection networks, IEEE Transactions on Computers C-32 (12) (1983) 1099–1108.
[9] S.M. Reddy, V.P. Kumar, On fault-tolerant multistage interconnection networks, Proceedings of the International Conference on Parallel Processing, 1984, pp. 155–164.
[10] V.P. Kumar, A.L. Reibman, Failure dependent performance analysis of a fault-tolerant on fault-tolerant multistage interconnection network, IEEE Transactions on Computers C-38 (12) (1989) 1703–1713.
[11] N. Tzeng, P. Yew, C. Zhu, A fault-tolerant scheme for on fault-tolerant multistage interconnection networks, 12th International Symposium on Computer Architecture, 1985, pp. 368–375.
[12] A. Itoh, A fault-tolerant switching architecture for ATM networks, IEEE International Conference on Communications, vol. 3, 1992, pp. 1639–1645.