ANALYSIS OF A FAULT-TOLERANT ATM SWITCH BASED ON A PARALLEL ARCHITECTURE

S. Segkhoonthod and M.C. Sinclair[1]

1 Introduction

B-ISDN is regarded as the future telecommunications system [1]. It will be capable of supporting all kinds of information such as voice, video and data. To realise this, Asynchronous Transfer Mode (ATM) has been selected as a means of switching and multiplexing [1]. One of the most important issues in ATM-based systems is how to construct a switch capable of offering high performance and reliability [2]. Many ATM switches have been proposed in the literature, most of them based on a class of networks called Multistage Interconnection Networks (MINs) [3, 6]. These are favoured because of several desirable features including distributed control, low complexity, modularity and their suitability for VLSI implementation. However, two main drawbacks of MINs have been noted: their low throughput and lack of fault tolerance. The former is due to the fact that a MIN is internally blocking, whereas the latter is due to the single-path property possessed by most MINs.

To improve both the performance and the reliability of MINs, one approach is to use a number of MINs in parallel to provide multiple paths for every source-destination pair [3, 4, 6]. This approach, also called K-replicated MIN (K is the number of MINs), offers a number of advantages. In terms of performance, the offered traffic is shared among MINs, therefore reducing internal blocking in each MIN. Moreover, as a result of multiple-paths, queuing at the output can be achieved—alleviating the output contention problem. From the reliability point of view, both redundancy and fault-tolerance are inherently provided. However, with this configuration, a great deal of complexity is placed in the input port controllers (IPC), used to dispatch the incoming packets among MINs, as they have to decide to which MIN packets should be sent. The easiest approach is to make the selection randomly, but as is shown in [4], this results in poor performance. A more sophisticated mechanism, proposed in [6], achieves a significant improvement in performance, but at the expense of increased complexity in the IPC dispatching algorithm. Another point is that when a fault occurs in one of the routing networks, the IPC also has to keep track of the faulty network in order to send packets only to the good ones—a further increase in the complexity of the IPC.

This paper presents a fault-tolerant ATM switch which adapts the idea of using MINs in parallel. It has a distribution network to distribute incoming packets to several routing networks arranged in parallel. Consequently, the IPC is simply required to submit packets to the distribution network. As the latter is designed to perform some of the routing itself, the routing networks are not required to carry out the whole routing process; consequently incomplete (*i.e.* truncated) routing networks are used. As a result, the overall complexity of the switch is not further increased by the introduction of the distribution network. The distribution network is also in charge of re-routing packets in the presence of a fault. Unlike other fault-tolerant networks where the routing algorithm is usually modified in order to avoid faults, in the proposed switch there is no modification to the routing algorithm in fault conditions: hence the re-routing process is simple and the complexity of switching elements is kept low. Further, it will be shown that the proposed switch gives both improved performance and lower complexity than replicated MINs.

[1]S. Segkhoonthod and M.C. Sinclair are with Dept. of Electronic Systems Engineering, University of Essex, Wivenhoe Park, Colchester, Essex C04 3SQ

The paper is organised as follows. In section 2, the switch architecture and operation are presented. Section 3 gives the evaluation of the switch performance. A complexity comparison between the proposed switch and replicated networks is made in section 4. The results obtained from analytical models and simulation are presented and discussed in section 5. Finally, conclusions and further work are given in section 6.

2 Switch Architecture and Operation

Fig. 1 depicts the general configuration of an l-level $N \times N$ version of the proposed switch which consists of an $N \times KN$ distribution network (DN), K $N \times N$ routing networks (RNs) and N $K \times 1$ output modules (OMs) where $K = 2^l$. The output of each IPC is connected to the input of the DN. The IPC is in charge of receiving packets from the incoming lines and submitting them to the DN. As shown in Fig. 2, the DN is constructed based on an l-level binary tree where each node is a plane of $N/2$ 2×4 switching elements (SEs). In stage i, there will be 2^i such planes. The 2×4 SE used in the DN is not a typical 2×4 crossbar. It behaves like a 2×2 SE but with two identical sets of two outputs, instead of only one set as in the case of a 2×2 SE. At a given node (a plane that belongs to stage i), one set of outputs of each SE are connected to the inputs of the SEs in its left child node (a plane in stage $i+1$), whereas the other set are for the right one (another plane in stage $i+1$). In addition, link patterns from a parent node to both its left and right child nodes are the same. The link pattern between a node in stage i and another in stage $i+1$, is identical to the link pattern between stages i and $i+1$ in an $N \times N$ Banyan network. At stage $l-1$, the outputs of the 2×4 SEs of the DN are connected to the inputs of the 2×2 SEs of the RN. An $N \times N$ Banyan network is used as a RN. However, it is not a complete one, instead each routing network is a truncated Banyan where stages 0 to $l-1$ are absent, because these stages are already embedded in the DN. Each output of the RNs is connected to one of the inputs of the OMs. Here the OM can accept K cells at the same time. The 32×32 proposed switch with $l = 2$ and $K = 4$ is shown in Fig. 3. It can be seen that there are 4 Banyan networks embedded with their stages 0 and 1 overlapped in the DN.

The DN and RNs are both unbuffered. The switch operation is based on a request/acknowledgement protocol, consisting of two phases: path-establish and packet-transfer [5]. At the beginning of the first phase, an IPC that has a packet generates a request, a form of short packet, and then submits it to the DN and waits for an acknowledgement (ACK). If the returned ACK is a positive one, the IPC can transfer the packet to the appropriate output module during the second phase. If not, it waits for the next path-establish phase and starts the same procedure again. After being admitted to the DN, the request goes along stage by stage attempting to set up a path across the DN and RN until reaching its required output module where a positive ACK is now generated and sent backward following the path that the request has set up. However, in some cases when setting up the path, the request might be unable to get through because of losing contention to another request that also wants the same output of an SE at that stage. In this case, a negative ACK is generated from there.

The DN is in charge of distributing, routing and, if required, re-routing. Basically, it distributes an incoming request to one of the RNs according to some rule, and at the same time routes the request. After receiving a request from the DN, the RN then directs it to an appropriate output module. The routing algorithm used is the same as that of a Banyan network. The 2×4 SEs employed in the DN have two sets of outputs, referred to here as first and second choices. In each set, there are two outputs—upper and lower outputs like those of a 2×2 SE. As an example, when a request arrives at the upper input of a 2×4 SE in stage i, if bit i of its routing tag is 0, then the request will be routed to the upper output. But which upper output? There are at least two ways of selecting the choice: *random* and *partially-deterministic* selections. By using the random selection, a request is randomly granted one of the choices; while with the partially-deterministic policy, a request is forced to take the first choice, but if two requests want the same output, say the upper one, one gets the first choice and the other the second on a random basis. Another option that also concerns the function of a 2×4 SE is how it behaves after receiving a negative ACK. In *no re-routing mode*, the 2×4 SE simply returns a negative ACK to a 2×4 SE in the previous stage. In *re-routing mode*, the SE attempts to route the request through the other choice. If still unsuccessful (because the request has been blocked or there is a faulty SE in a further stage), a negative ACK is generated back to the previous stage. In order to distribute requests evenly among the RNs when partially-deterministic selection is used, the arrangement of the first and second choices of 2×4 SEs in the DN is done in such a way that the first choice of half of the SEs in stage i are connected to their left child node and the other half to their right.

3 Performance Evaluation

The proposed switch is analysed under the following assumptions.

- Packets arriving at different IPCs in the same cycle are mutually independent.
- Packets arriving at an IPC or at different IPCs in different cycles are mutually independent.
- The destination of a packet is uniformly addressed to one of the OMs with equal probability.
- Packets arrive at an IPC in a given time slot with probability p. This is the input rate of the switch.
- A blocked packet, after returning to its corresponding IPC, is discarded without being re-submitted.
- The buffer size of an OM is infinite [3].

It was observed in [7, 8, 5] that the above assumptions taken together imply that the states of the IPCs are statistically independent and SEs in the same stage or different stages also behave independently. It follows that it is sufficient to study only one SE per stage, as it represents the behaviour of all SEs in the same stage. The first two assumptions are unrealistic in practice but, as it is shown in [8] by means of simulation, the results obtained are only slightly different with those two relaxed. The third assumption is a uniform traffic pattern, which is widely used when comparing different networks. The fifth assumption is in fact an implication of the second one. The last assumption is there to simplify the analysis. The performance measure used is the probability of acceptance (P_{A}) which is defined as the probability that an arbitrary request will be accepted [8].

3.1 Probability of Acceptance of the Proposed Switch: No Re-routing

The analysis of the no re-routing version of the proposed switch is based on an approach used in [8]. As the proposed switch consists of a number of stages of SEs, the output of an SE in stage i being the input of a SE in stage i+1, here P_{A} can be thought of as the ratio of the output rate and the input rate of the switch. Consequently, derivation of P_{A} for the switch is done by recursively evaluating the output rate of a SE in each stage. In this approach, the expression of interest is the output rate of a SE in each stage given its input rate. As the proposed switch employs three forms of switching devices (*i.e.* 2×4 SEs, 2×2 SEs and K×1 OMs), the task here is to derive an expression for the output rate of each device.

3.1.1 Derivation of Output Rate Expressions

I. Output Rate Expression for a 2×4 SE

In Fig. 4(a), denoting $P_{\mathrm{out,u,1}}(P_{\mathrm{out,u,2}})$ as the output rate at the first (second) choice of the upper

outputs of a 2×4 SE and $P_{\mathrm{in,u}}$ ($P_{\mathrm{in,l}}$) as the input rate at the upper (lower) input of a 2×4 SE, then:
1) Partially-deterministic Selection

$$P_{\mathrm{out,u,1}} = P_{\mathrm{out,l,1}} = (P_{\mathrm{in,u}}/2)(P_{\mathrm{in,l}}/2) + (1 - P_{\mathrm{in,u}}/2)(P_{\mathrm{in,l}}/2) + (P_{\mathrm{in,u}}/2)(1 - P_{\mathrm{in,l}}/2) \tag{1}$$

$$P_{\mathrm{out,u,2}} = P_{\mathrm{out,l,2}} = (P_{\mathrm{in,u}}/2)(P_{\mathrm{in,l}}/2) \tag{2}$$

2) Random Selection

$$P_{\mathrm{out,u}} = P_{\mathrm{out,l}} = (0.5)\big((2)(P_{\mathrm{in,u}}/2)(P_{\mathrm{in,l}}/2) + (1 - P_{\mathrm{in,u}}/2)(P_{\mathrm{in,l}}/2) + (P_{\mathrm{in,u}}/2)(1 - P_{\mathrm{in,l}}/2)\big) \tag{3}$$

Eq. 1 is the probability that a request is present at the first choice output of a 2×4 SE; this is when there is at least one of the two requests which is addressed to the upper output (second and third terms) or both of them (first term). While in Eq. 2, there will be a request at the second choice only if the two packets are destined to the same output.

II. Output Rate Expression for a 2×2 SE
In Fig. 4(b), denoting $P_{\mathrm{out,u}}(P_{\mathrm{out,l}})$ as the output rate at the upper (lower) output of a 2×2 SE and $P_{\mathrm{in,u}}$ ($P_{\mathrm{in,l}}$) as the input rate at the upper (lower) input of a 2×2 SE, then:

$$P_{\mathrm{out,u}} = P_{\mathrm{out,l}} = 1 - (1 - P_{\mathrm{in,u}}/2)(1 - P_{\mathrm{in,l}}/2) \tag{4}$$

III. Output Rate Expression for a K×1 OM
In Fig. 4(c), denoting P_{out} as the output rate of a K×1 OM and $P_{\mathrm{in},i}$ as the input rate at the i^{th} input of a K×1 OM, then:

$$P_{\mathrm{out}} = \sum_{i=1}^{K} P_{\mathrm{in},i} \tag{5}$$

3.1.2 Derivation of P_{A} of the Proposed Switch: No Re-routing

Consider an l-level N×N switch with $n = \log_2 N$ stages and $K = 2^l$ RNs. First of all, as a result of the arrangement of first and second choices in the DN, one input of a 2×4 SE in stage i is loaded by traffic from the first choice output of a 2×4 SE in stage $i-1$, while the other input is loaded by traffic from the second choice output of another 2×4 SE in stage stage $i-1$ where $(1 \leq i \leq l-1)$ (see Fig.5). Define P_{in} as the input load of the switch (or p), $P_{\mathrm{out1},j}$ ($P_{\mathrm{out2},j}$) as the output load of the first (second) choice output of a 2×4 SE in stage j, $P_{\mathrm{out},j}$ as the output load of a 2×2 SE in stage j and P_{out} as the output load of an OM. Then, the Probability of Acceptance (P_{A}) of the switch is expressed as below:

I. Partially-Deterministic Selection

$$\left.\begin{aligned} P_{\mathrm{out1},j} &= 2\left(1 - \tfrac{P_{\mathrm{in}}}{2}\right)\tfrac{P_{\mathrm{in}}}{2} + \tfrac{{P_{\mathrm{in}}}^2}{4} \\ P_{\mathrm{out2},j} &= \tfrac{{P_{\mathrm{in}}}^2}{4} \end{aligned}\right\} \quad j = 0 \tag{6}$$

$$\left.\begin{aligned} P_{\mathrm{out1},j} &= \tfrac{P_{\mathrm{out1},j-1}}{2}\left(1 - \tfrac{P_{\mathrm{out2},j-1}}{2}\right) + \\ &\quad \tfrac{P_{\mathrm{out2},j-1}}{2}\left(1 - \tfrac{P_{\mathrm{out1},j-1}}{2}\right) + \tfrac{P_{\mathrm{out1},j-1}P_{\mathrm{out2},j-1}}{4} \\ P_{\mathrm{out2},j} &= \tfrac{P_{\mathrm{out1},j-1}P_{\mathrm{out2},j-1}}{4} \end{aligned}\right\} \quad 1 \leq j \leq l-1 \tag{7}$$

$$P_{\mathrm{out},j} = 1 - \left(1 - \frac{P_{\mathrm{out1},j-1}}{2}\right)\left(1 - \frac{P_{\mathrm{out2},j-1}}{2}\right) \quad j = l \tag{8}$$

$$P_{\text{out},j} = 1 - \left(1 - \frac{P_{\text{out},j-1}}{2}\right)^2 \quad l+1 \leq j \leq n-1 \tag{9}$$

$$P_{\text{out}} = \sum_{i=1}^{K} P_{\text{out},n-1} \tag{10}$$

$$P_{\text{A}} = \frac{P_{\text{out}}}{P_{\text{in}}} \tag{11}$$

II. Random Selection

$$P_{\text{out}1,j} = P_{\text{out}2,j} = (0.5)\big((2)(1 - \frac{P_{\text{in}}}{2})\frac{P_{\text{in}}}{2} + (2)\frac{P_{\text{in}}{}^2}{4}\big) \quad j = 0 \tag{12}$$

$$\left.\begin{aligned} P_{\text{out}1,j} = P_{\text{out}2,j} \quad = \quad & (0.5)\Big(\tfrac{P_{\text{out}1,j-1}}{2}\left(1 - \tfrac{P_{\text{out}2,j-1}}{2}\right) + \\ & \tfrac{P_{\text{out}2,j-1}}{2}\left(1 - \tfrac{P_{\text{out}1,j-1}}{2}\right) + \\ & (2)\tfrac{P_{\text{out}1,j-1}P_{\text{out}2,j-1}}{4}\Big) \end{aligned}\right\} \quad 1 \leq j \leq l-1 \tag{13}$$

where the rest of the equations are the same as those in the case of partially-deterministic selection, *i.e.* Equations 8, 9, 10 and 11.

3.2 Probability of Acceptance of the Proposed Switch: Re-Routing

Simulation was used to obtain the P_{A} of the switch in case of re-routing. The switch was simulated under the same assumptions as those adopted in the analytical model. Two independent runs were carried out, each consisting of five batches—altogether giving ten values to be averaged for each simulation point. Each batch consisted of 10,000 cycles. In each run, an additional 1,000 cycles were completed before data was collected in order to avoid transient affects. The 90% confidence intervals are very tight and therefore will not be shown. However, due to time limitations, only 16×16 and 32×32 switches were modelled.

4 Complexity Comparison

Denoting the number of SE crosspoints (excluding demultiplexers and multiplexers) by N_c, then for the proposed switch this is $8\frac{N}{2}(\sum_{i=0}^{l-1} 2^i) + 4(2^l)\frac{N}{2}(\log_2 N - l)$. For a K-replicated MIN, N_c is $4K\frac{N}{2}\log_2 N$. Thus with $l > 1$, the proposed switch has lower complexity for all values of N. If we take $N = 1024$ and $K = 4$, then for the proposed switch, $N_c = 77824$, whereas for a 4-replicated MIN, $N_c = 81920$.

5 Results and Discussions

Fig. 6 shows the P_{A} of the switch in case of no re-routing with both selection modes. It is clear that the partially deterministic mode gives better performance. Nevertheless, there is no significant difference between P_{A} in different selection modes in the case of re-routing mode as depicted in Fig. 7. Although, as Fig. 8 shows, P_{A} in the case of re-routing is significantly higher than in case of no re-routing for partially deterministic selection. This is due to the fact that in the re-routing version, requests have more choice to get to their requested outputs as a result of re-routing. However, this comes at the expense of a longer set up time. Curves of P_{A} for a no re-routing version of the proposed switch with partially deterministic selection and a replicated MIN, as a function of the number of stages and the number of routing networks, are plotted in Fig. 9, with the results for the K-replicated MIN (multiple acceptance) taken from [3]. Clearly, Fig. 9 indicates that the proposed switch offers better performance than this replicated MIN.

6 Conclusions and Further Work

In this paper, a fault tolerant ATM switch design based on the use of a number of routing networks in parallel has been proposed. The design has a distribution network which is used to distribute the incoming packets among the routing networks. As a result, the input port controller simply submits the packets to the distribution network—in contrast to typical replicated networks where the input port controller has to decide which routing network the packets are to be sent to. The distribution network is also in charge of re-routing when there is failure in the switch. Since the re-routing process requires no additional modification, the routing algorithm is simple and the complexity of switching elements is low. Analytical models of the switch have been developed. Finally, it was shown that the proposed design gives better performance but requires less complexity when compared to a replicated network if the number of routing networks is greater than 2.

There are several further aspects of the proposed design to be investigated. These include an analytical model for re-routing mode, performance evaluation in the presence of faults and reliability analysis of the switch.

Acknowledgements

The research described in this paper was undertaken by the first author (supervised by the second) as a Ph.D. project in the Dept. of Electronics Systems Engineering, University of Essex. The first author was initially supervised by Dr. E.A. Medova, and is sponsored by the Thai Government.

References

[1] MINZER S. E. : 'Broadband ISDN and Asynchronous Transfer Mode (ATM)', *IEEE Communication Magazine*, Sep. 1989, pp. 17–57

[2] TOBOGI F.A. : 'Fast packet switch architectures for broadband integrated services networks', *Proceedings of the IEEE*, Vol. 78, 1990, pp. 133–167

[3] VENKATESAN, R. and MOUFTAH, H.T.: 'Performance analysis of multipath Banyan networks', *Proc. IEEE International Conference on Communications*, 1992, pp. 912–915

[4] KUMAR, M. and JUMP, J.R.: 'Performance of unbuffered shuffle-exchange networks', *IEEE Trans. Computers*, 1986, **C-35**, (6), pp. 573–577

[5] PADMANABHAN, K and LAWRIE, D.H.: 'Performance analysis of redundant-path networks for multiprocessor systems', *ACM Trans. Computer Systems*, 1985, **3**, (2), pp. 117–144

[6] WENG, C.-M. and LI, J.-J.: 'Solution for packet switching of broadband ISDN', *IEE Proc. I*, 1991, **138**, (5), pp. 394–400

[7] WU, C.-L and LEE, M.-L.: 'Performance analysis of multistage interconnection network configurations and operations', *IEEE Trans. on Computers*,1992, **C-41**, (6), pp. 18–2

[8] PATEL, J.H.: 'Performance of processor-memory interconnections for multiprocessors', *IEEE Trans. on Computers*,1981, **C-30**, (10), pp. 771–780

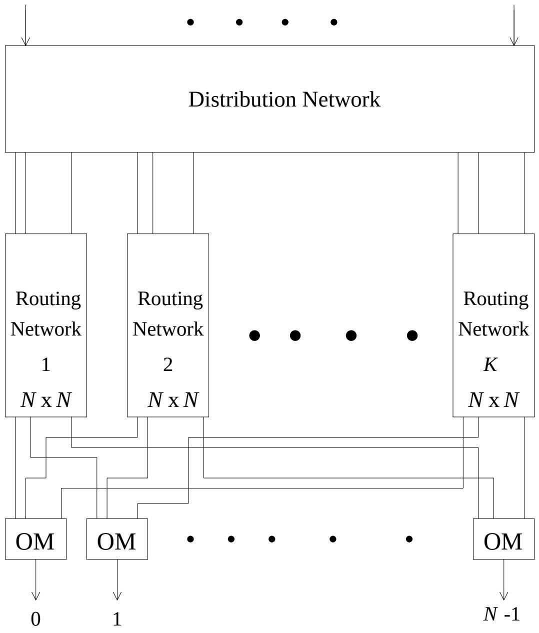

Figure 1: Configuration of the proposed switch

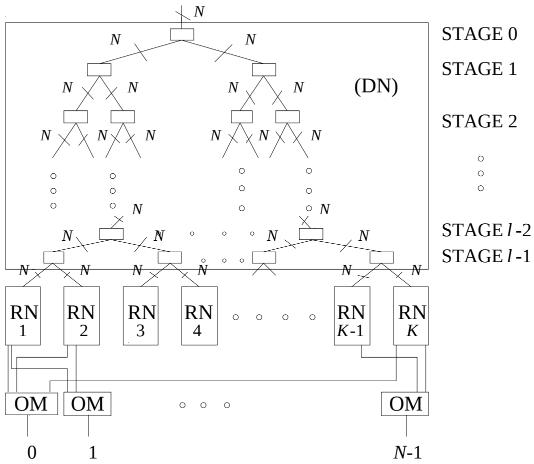

Figure 2: The distribution network

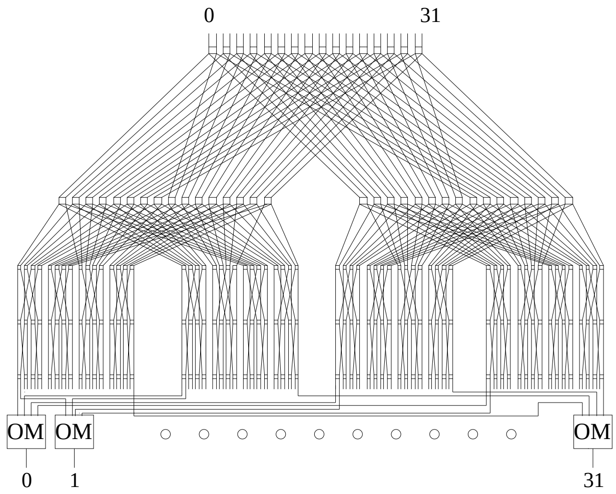

Figure 3: A 32×32 switch with $l = 2$

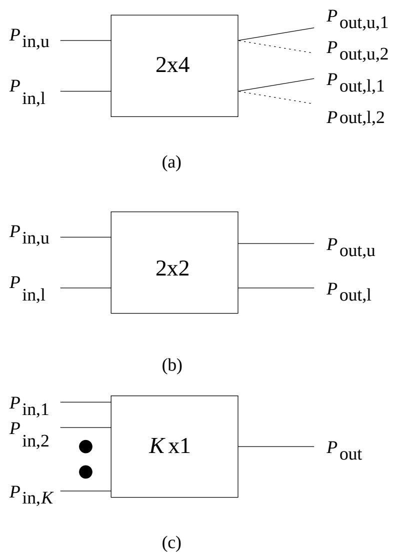

Figure 4: Output expression of (a) 2×4 SE, (b) 2×2 SE and (c) K×1 OM

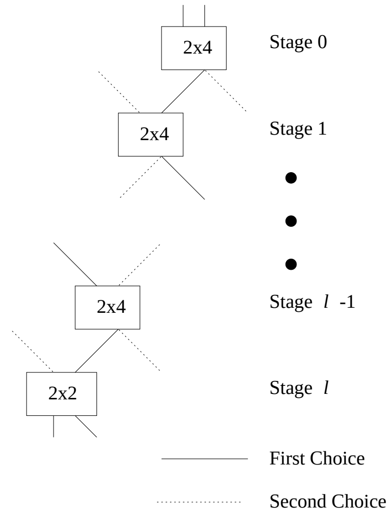

Figure 5: Arrangement of first and second choices in the distribution network

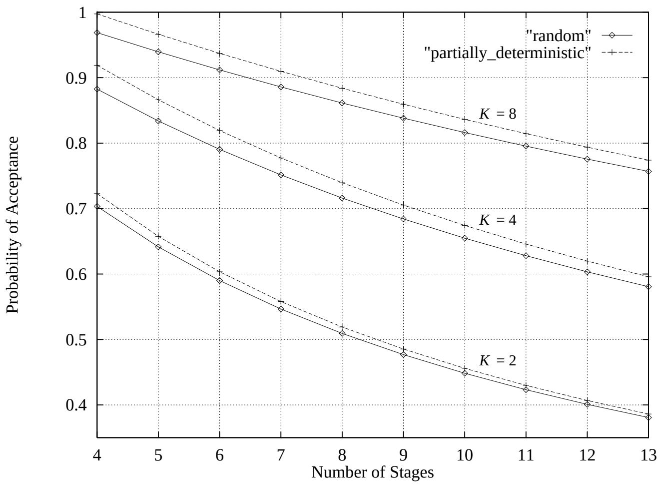

Figure 6: P_{A} of the proposed switch in no re-routing mode with random and partially deterministic selections (analytical results with Input Load = 1.00)

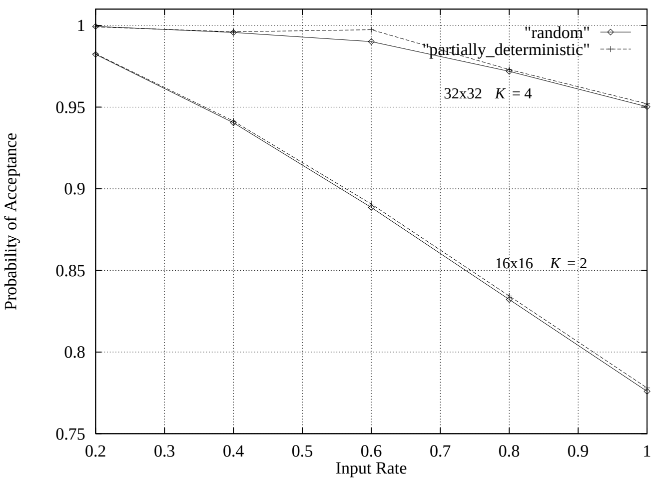

Figure 7: P_{A} of the proposed switch in re-routing mode with random and partially deterministic selections (simulation results)

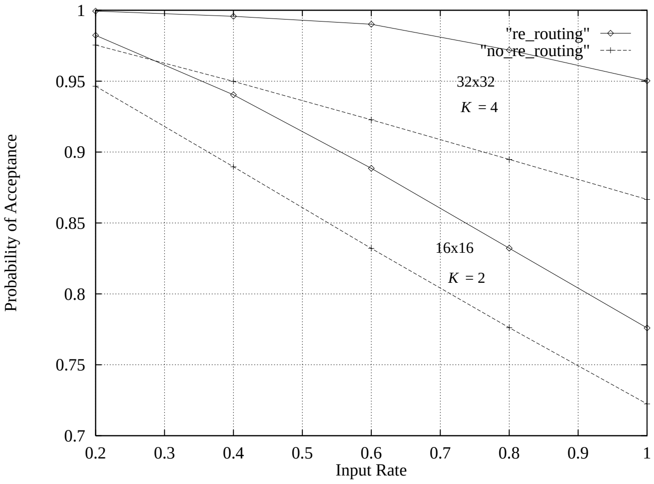

Figure 8: P_{A} of the proposed switch in no re-routing and re-routing modes with partially deterministic selection (simulation results)

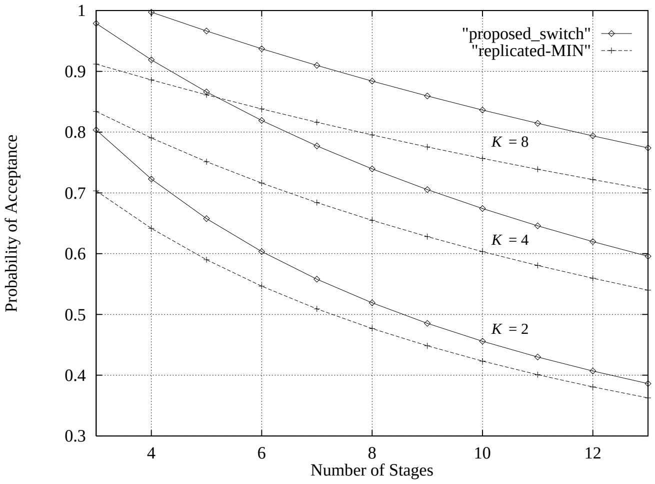

Figure 9: P_{A} of the proposed switch in no re-routing mode and with partially deterministic selection and a replicated MIN (analytical results with Input Load = 1.00)