Reliability Study of Multilayer Multistage Interconnection Networks Equipped with Internal Path Redundancy

Eleftherios Stergiou (1), Dimitrios Liarokapis (1), Euripidis Glavas (1)

(1) Dept. of Computer Engineering, TEI of Epirus
Arta, Greece
e-mail: ster@teiep.gr, dili@teiep.gr, eglavas@teiep.gr

Georgios E. Rizos (2), D. C. Vasiliadis (2)

(2) Network Operations Center, TEI of Epirus
Arta, Greece
e-mail: georizos@teiep.gr, dvas@teiep.gr

Abstract—Multilayer multistage interconnection networks have introduced multiple parallel layers for enhancing performance metrics over traditional multistage interconnection networks. In this work, we propose an innovative multilayer multistage interconnection network fabric, which improves the switch efficiency by allowing multiple internal paths. The proposed fabric has demonstrated improved performance metrics compared to other existing fabrics and acceptable reliability in terms of fault tolerance. This type of fabric can considerably enhance the connection points of a modern network and improve its data flow. The configuration of the network allows for a conflict drop resolution mechanism or a classic backpressure blocking mechanism. This novel fabric can also handle multicast traffic, hotspot traffic or a combination of them.

Keywords— Reliability; Multilayer Multistage Interconnection Networks; Quantitative Analysis; Multistage Architecture; Performance Evaluation.

I. INTRODUCTION

Multilayer Multistage Interconnection Networks (MLMINs) are devices that improve performance metrics when transferring data. The use of MLMINs avoids the necessity for the crossbar type of interconnection device, which is expensive to construct. However, although this shared-bus type of switching device is of low cost, it has low performance and therefore there is a need for interconnection devices, which strike a balance between efficient performance, reliability and reasonable cost. MLMINs were proposed by Tutsch and Hommel [1], after it was found in various studies that more switching power was needed in the last stages of a multistage interconnection network (MIN) than in the first stages [1]-[5].

The performance of MLMINs has been studied thoroughly by Garofalakis et al. [3] [6].

A typical MLMIN consists of an $N \times N$ MIN and $L = \log_k N$ stages and $k \times k$ or $k \times n$ switching elements (SEs), where $k$ and $n$ are the number of inlets/outlets of the SEs.

A typical MLMIN has two seriati segmentation: the single-layer segment followed by the multilayer segment.

In the first segment, each stage consists of $(N / k)$ SEs of size $k \times k$, while the last stage consists of SEs with size $k \times n$, where $n \geq k$.

In the second stage, however, the multilayer segment consists of $k \times k$ SEs in parallel rows. If the single layer segment has S stages, the second part has $(L-S)$ stages. The size of S is a matter of engineering choice, and depends on the degree of reduction to be implemented in the last stages of construction.

Figure 1 illustrates several MLMINs and corresponding definitions.

In addition to the above, in order to accurately determine the structure of a MLMIN, three further parameters are required [3].

The first parameter is called the “Start replication factor” ($G_S$), and denotes the stage number at which the replication of layers starts. The second parameter is known as the “Growth factor” ($G_F$), and denotes the number of layers that can be developed at one stage by each SE. Finally, the third parameter is the “Layer limit factor” ($G_L$), and this denotes the maximum number of layer replications. Structures which can be described by the above three parameters are known as semi-layer MINs [6].

Figure 1 illustrates several MLMINs and gives the three factors that describe them. Of all the possible types of multilayer MINs, semi-layer MINs are the most suitable for use in the information technology industry due to the simplicity of their construction compared to other multilevel devices.
In addition to their performance characteristics, their fault tolerance or reliability is of interest to the scientific community, since the requirements of modern applications involve a constant demand for improved solutions [7, 9-10, 12-13].

In this work, in order to improve the behaviour of MLMINs even further, the number of possible paths between each pair (input-output) is multiplied. In this architecture, in cases of internal blocking, the system has the ability to send the packet via another, ‘parallel’, alternative route. This operation reduces the stacking of packets in internal buffers, increasing forwarding speed.

A series of multiplexers (MUXs) at the beginning of the fabric is introduced to implement this concept. This mechanism can dissipate the packets using multiple paths. The diffusion of packets increases (since more parallel paths are used) as the blocking phenomenon develops.

This forms the basic concept underlying the current work. To implement this idea, a new architecture is introduced which significantly raises both performance metrics and the reliability factor to acceptable values. Moreover, this system provides the ability to handle special load cases such as hotspot or multicast traffic.

Hence, the novel contribution of this work can be briefly summarized as:

- A new, productive and therefore promising MLMIN architecture is introduced with finite buffers, multiple internal routes and MUXs.
- A simple but detailed study of the proposed fabrics is carried out, showing that these fabrics have many benefits; the results from the proposed fabric in terms of metrics, the reliability factor and packet latency are encouraging.
- A comparative study of other similar modern architectures indicates that the new fabric outperforms these in permutation capability and has an acceptable level of reliability.

The remainder of this paper is organized as follows: In Section II, the proposed MLMIN fabric is introduced and several details of its operation (e.g., internal paths and routing) are presented. In Section III, the basic definitions are given and the necessary analysis carried out. In Section IV, the reliability and performance capability are analysed, and several numerical outcomes are described for the implemented MLMINs in terms of network size. Finally, in Section V, the conclusions and anticipated future work are presented.

II. THE PROPOSED FABRIC

A. Proposed fabric: Semi-layer MIN with internal path redundancy

To establish internal paths and to disperse the load uniformly, a preamble block composed of MUXs is located in front of the MIN. Thus, the proposed MIN consists of three segments in sequence, as presented below (Fig. 2).

![Fabric with internal path redundancy consisting of three blocks](image)

1) Preamble of the MIN redundancy consisting of three blocks: In the preamble block, a column of \(m \times 1\) MUXs are used to connect the inputs of the first stage of the MIN. Thus, a \(N \times N\) network needs \((3N/2)\) multiplexers at the input stage (See Figure 3). This part is introduced because we want to have the possibility to make diffusion packages to alternative paths, when the first of them are busy. The main body of the MLMIN follows the preamble segment.

2) First segment of the MIN (single-layered segment): The first segment of the MIN contains only a single layer, which employs multiple internal paths (see Figure 3).

![Detail of the connection between the preamble and single-layer segments, shown here for a network of diameter 3](image)

A network with size \(N \times N\) has \((\log_2 N - S)\) stages with \((3 \cdot N/4)\) switches per stage, where \(S\) is the stage number of the single-layer segment. This part consists of \(2 \times 2\) switches, except for the final stage, which employs \(2 \times 2^m\) switches, where \(m = 1, 2, \ldots\) depends on the fan-out configuration to which they are connected (Figure 4).

3) Second segment of the MIN (multilayered segment or fan-out): The multilayered segment or fan-out is accommodated in the second segment (Fig. 4). In the case of a double layer (one stage multiplied) the fan-out has a quantity \((2N/2) = N\) of \(2 \times 2\) switches, while the triple-layered (one stage multiplied) network is composed of \((3N/2) \times 2 \times 2\) switches.
The number of SEs in a $N \times N$ network with one final stage multiplied can be calculated as

$$\left( \frac{3 \cdot N}{4} \right) \cdot (\log_2 N - 1) + N$$

for a double-layered network, and

$$\left( \frac{3 \cdot N}{4} \right) \cdot (\log_2 N - 1) + \frac{3}{2} \cdot N$$

for the corresponding fabric with a triple-layered network and so on.

For simplicity, the proposed device is referred to here as SemiL (Semi-Layer MIN with multiple internal paths).

B. Multiple paths between two points

Fig. 5 shows all the possible routes between two points of the MIN (inlet: 000; outlet: 000), e.g., in a fabric with eight inputs and equal outputs. All alternative paths have the same length, which means that while the network is operating in real time, many equivalent actions for bypassing blocks are generated when faults arise in the system.

In standard $N \times N$ MLMINS, there are $2^{\log_2 N}$ distinct routes from one input point to all outputs, and thus a total of possible routes of $N \cdot 2^{\log_2 N}$ distinct paths. However, the equivalent novel MLMIN fabric, which contains internal redundancy paths, provides $\frac{3 \cdot N}{4} \cdot N \cdot 2^{\log_2 N}$ distinct paths.

Hence, for example, a classic MLMIN with $N = 8$ provides 64 distinct paths, while the equivalent novel MLMIN fabric provides six times more (6x64) distinct paths.

C. Marking paths

Marking paths and establishing a hierarchy among the different paths aids in the implementation of an automatic routing mechanism. Taking an arbitrary input-output pair, then according to Fig. 5, packets can be routed from an input (e.g., 000) via three different SEs (SE01, SE21 and SE41). These links can be considered as the first link (FL) of the path.

\begin{align*}
1\text{-path:} & \quad 000 \rightarrow \text{SE01} \rightarrow \text{SE02} \rightarrow \text{MSM0} \rightarrow 000 \\
2\text{-path:} & \quad 000 \rightarrow \text{SE01} \rightarrow \text{SE22} \rightarrow \text{MSM0} \rightarrow 000 \\
3\text{-path:} & \quad 000 \rightarrow \text{SE21} \rightarrow \text{SE02} \rightarrow \text{MSM0} \rightarrow 000 \\
4\text{-path:} & \quad 000 \rightarrow \text{SE21} \rightarrow \text{SE42} \rightarrow \text{MSM0} \rightarrow 000 \\
5\text{-path:} & \quad 000 \rightarrow \text{SE41} \rightarrow \text{SE22} \rightarrow \text{MSM0} \rightarrow 000 \\
6\text{-path:} & \quad 000 \rightarrow \text{SE41} \rightarrow \text{SE42} \rightarrow \text{MSM0} \rightarrow 000
\end{align*}

Six different internal routes between a pair input-output.

Thus, FL1, FL2 and FL3 links are marked. Each request at Stage 1 has to choose one of these three links. Subsequently, from Stage 1 to the next stage, each request has two available links to choose from. The first is marked as the main link (ML), while the other can be marked as the auxiliary link (AL).

The same pattern of marking of links is continued for all the following stages of the single-layer segment. As the number of stages is increased, the number of distinct paths is multiplied. Thus, we have links marked MLn and ALn where $n \in [1, +\infty]$.

As a whole, this approach leads to distinct routes which constitute a hierarchy at every stage. Fig. 5 shows all the possible paths between an arbitrary input-output pair. In addition, it depicts the intermediate nodes (e.g., switches) that are involved in a specific input-output pair.

D. Routing technique and blocking mechanism

The clocking of the internal fabric guides the whole routing operation. The packets are forwarded synchronously at every stage in a parallel manner.

In MINs, routing and forwarding decisions are made independently for each packet by computation for each outgoing link. When a packet enters the fabric, it receives the ‘routing address’ (RA). The RA has as many bits as the number of stages in the MIN. The bits of the RA guide the packet-forwarding through the MIN, from SE to SE.

The RA is used by a packet in order to reach its unicast destination (including in the case of hotspot traffic). Moreover, in the case of multicast distribution, beyond the RA, a ‘multicast tree number address’ (MTNA) is assigned to each packet. Both types of address have as many bits as the number of stages in the MIN. The multicast address...
indicates at which stage a packet has to act as a multicast operation (generating a copy packet), thus creating the so-called ‘multicast tree’ if this mechanism is implemented in the fabric.

The multistage construction described above can work either with backpressure or with a drop resolution mechanism. For simplicity, the accommodation of the backpressure technique in the system is considered here. This backpressure mechanism is appropriate in a network where systems employ non-deterministic protocols such as the UDP protocol; however, the backpressing operation is completely unsuitable for deterministic protocols, such as the TCP protocol. Nowadays, although the backpressure mechanism is considered responsible for phenomena such as packet looping, large packet delays, and scalability, it continues to be implemented due to a lack of credible alternatives.

For the selection of an internal path which is not overburdened, the concept of backpressure weight per link may be introduced at the start of the device and constantly calculated for each input-output pair. A backpressure weight is an operation (function) based on local queue conditions, and gives information about the link state.

In contrast, the drop resolution mechanism, also known as the relaxing blocking mechanism, may be used when the switching element suffers from blocking; it sends back a request releasing a signal to the sender along a specific path. The blocked request, now lost, repeats the same process (starting again from the beginning) until a new path is established.

III. BASIC DEFINITIONS AND ESSENTIAL ANALYSIS OF THE FABRIC

- **Probability of packet arrivals** (\(\lambda\)) represents the offered load at an input; in our experiment, the probability of packet arrivals is ranked from 0.1 to 1.0, with steps equal to 0.1. This probability can be expressed as \(\lambda_{Norm}\).

  The arrival process for packets at the output queues of the first stage of the network is given by a binomial distribution \(bin(c,\lambda/c)\), where \(\lambda\) is the fixed probability of a packet being generated by a processor at each cycle.

- The **arrival process of packets** at the output queues of Stage \((i)\) (for \(i = 2\) to \(i = S-1\) of the network, where \(S\) is the number of stages in the single-layer segment) is approximated by a binomial distribution \(bin(c,u_{(i-1)})\), where \(u_{(i-1)}\) is the utilization of a queue (buffer) of Stage \((i-1)\), which we assume plays the role of the fixed probability of packets generated by processors at each cycle, feeding Stage \((i)\). A similar forward technique is applied at the fan-out.

- **Buffer size** \((b)\) represents the maximum number of packets that can be held by an input buffer of an SE.

- **Reliability** \((r)\) of a component represents the probability of a switch component being operational. Each switch component has its own reliability. The reliability symbol can be denoted by \(r_{type-n}\), where type signifies the switch type and \(n\) depicts the number of gates of the switch. Thus, for example, \(r_{MUXn}\) and \(r_{DEMUXn}\) signify the reliability of the nx1 multiplexer and the 1xn de-multiplexer respectively, and \(r_{SEx2x2}\), \(r_{SEx2x4}\) and \(r_{SEx2x6}\) represent the reliabilities of \(SE_{2x2}\), \(SE_{2x4}\) and \(SE_{2x6}\) correspondingly.

- **Average throughput** \((T_{avg})\) signifies the average number of packets accepted by all destinations per network cycle. Because the last stage of a single layer is never blocked, the average throughput can be expressed as \(T_{avg} = u^{(s)}\), where \(u\) is the average utilization of the S stages.

  Moreover, \(T_{Norm}\) is the normalized throughput appearing at the end of the fabric.

- **Normalized packet latency** \((D_{Norm})\), of a MIN with \(L\) stages is defined as the ratio of average latency \((D_{avg})\) to the minimum packet delay. As minimum packet delay we consider the minimum number of time slots needed by a packet to be transmitted to its destination, i.e. when packets don’t face any blocking during their routes. For \(L\)-stage MINs the minimum delay is equal to \((L)\) time slots. Thus \(D_{Norm} = D_{avg}/L\).

- The normalized packet loss probability \(\lambda_{loss(Norm)}\) of typical MLMINs (with \(S\) stages in a single layer segment), with or without multiple internal routes that apply the backpressure blocking mechanism at the inputs of the MINs before they enter the fabrics, can be expressed as

  \[
  \lambda_{loss(Norm)} = \lambda_{loss(Norm)}^{(0)} = \lambda_{Norm} - T_{Norm} \tag{1}
  \]

  where \(\lambda_{Norm}\) depicts the normalized packet arrivals at the beginning of the fabric. This can also be written

  \[
  \sum_{i=1}^{S-1} \lambda_{loss(Norm)}^{(i)} = 0 \tag{2}
  \]

  where \(\lambda_{loss(Norm)}^{(i)}\) depicts the probability of packet loss in a queue at the \(i^{th}\) intermediate stage of a MIN.

- **Reliability analysis:** the reliability of a fabric is the ability to overcome all unexpected circumstances. Here, the point-to-point reliability (p-t-p reliability), also known as terminal reliability, is used. With p-t-p...
p reliability, the probability of at least one fault-free path existing between an input-output pair is considered. The reliability between an arbitrary given pair (inlet-outlet) is dependent on all the switch elements involved, insofar as the failure of each element implies the failure of the current routing action [11]-[13]. Supposing there is a switching subsystem with a series of n switch components, each with a reliability of \( r_i \); then, the reliability of this subsystem can be calculated as

\[
R_n \text{Switches in Series} = \prod_{i=1}^{n} r_i
\]  

(3)

On the other hand, for a subsystem with \( n \) parallel switch components, at least one of these must be active in order for this subsystem to operate successfully. In this case, the reliability of the subsystem is calculated as

\[
R_n \text{Parallel Switches} = 1 - \prod_{i=1}^{n} (1 - r_i)
\]  

(4)

IV. RELIABILITY AND PERFORMANCE ANALYSIS

A. Reliability of the proposed fabric

For reliability analysis of the novel fabric, p-t-p reliability is used. As discussed above, in p-t-p reliability the probability of at least one fault-free path existing between an input-output pair is considered [See Figure 6]. The total p-t-p reliability \( R_{\text{fabric}} \) of an arbitrary input-output pair can be expressed by Equation (5)

\[
R_{\text{fabric}} = R_{\text{preamble}} \times R_{\text{SLMIN}} \times R_{\text{MLMIN}}
\]  

(5)

Given the probability of a multiplexer and a SE being in normal operation (i.e., \( r_{MX} \) and \( r_{SE} \) respectively are known), then the total reliability of the fabric, \( R_{\text{fabric}} \), can be calculated as

\[
R_{\text{fabric}} = r_{MX} \times r_{SE} \times \left( 1 - \prod_{i=1}^{S} (1 - r_{SE})^{L_i} \right)
\]  

(6)

where \( S \) is the length of a single-layer segment and \( G_L \) is the maximum number of layer replications. The values in Equation (6) are higher than the corresponding values of the conventional equivalents such as banyan- or delta-type MINs with the same network size \( R_{\text{MIN}} = r^{\log_2 N} \).

The p-t-p reliability for a route within this novel fabric (2SemiL and 3SemiL cases with two and three layers respectively) can be illustrated as in Figure 7.

For a given probability of a multiplexer and a SE being in normal operation (\( r_{MX} \) and \( r_{SE} \) respectively) the \( R_{\text{fabric}} \) of a SemiL network with fan-out in the final stage can be expressed as

\[
R_{\text{SemiL}} = r_{MUX} \times r_{SE} \times \left( 1 - \prod_{i=1}^{S} (1 - r_{SE})^{L_i} \right)
\]  

(7)

where \( G_L \) depicts the maximum number of layer replications, \( G_L \) has values of 2 and 3 for 2SemiL and 3SemiL networks respectively, and \( n \) in \( r_{SEn} \) is equal to 4 and 6 respectively.

The p-t-p reliability has been given in earlier studies in the literature for various types of MIN architectures (e.g., Pars, Augmented Shuffle Exchange Network (ASEN) and Augmented Baseline Network (ABN) [12] [14][15].

For Pars networks, the reliability is given in [8] as follows

\[
R_{\text{Pars}} = r_{MX} \times r_{SE} \times \left( 1 - (1 - r_{SE})^{2} \cdot r_{DEMUX}^{2} \right)
\]  

(8)

Figure 7 shows the total input-output reliability versus the reliability \( r \) of each component for the novel fabric (SemiL networks) with \( G_L = 3 \).

With adjacent groups of bars corresponding to various values of a component’s reliability \( r \); the graph shows a gradual increase in the total p-t-p reliability for networks with numbers of inlets/outlets \( N \) gradually increasing (where \( N = 2^k \), \( k = 3,...8 \)).

Figure 6: Typical point-to-point paths in 2 and 3SemiL networks

For low values of component factor \( r \) (e.g., \( r=0.9 \)) it can clearly be seen that the total p-t-p reliability for networks with high network diameter (e.g., size \( N=256 \)) is greatly reduced.

When the factor \( r \) tends to 1, all the fabrics, regardless of network diameter (size), tend to have the maximum p-t-p reliability of value 1.
Fig. 7 shows the total input-output reliability for various types of MIN. With adjacent groups of bars corresponding to given values of the component reliability (r), the two left-hand bars of each group represent the total p-t-p reliability of MLMINs (SeLMINs) that do not use internal paths and have replication factors of 3 and 2 respectively. The two right-hand bars of each group depict the total p-t-p reliability of MLMINs (SemiLs) that include internal paths and have replication factors of 3 and 2 respectively. The central bar represents the p-t-p reliability of a Pars network.

The two upper solid curves depict the normalized latency of a single-layer MIN (SiLMIN) with buffer sizes of 2 and 3 respectively. The dotted curves represent the same quantity for semi-layer MINs (SeLMINs) without internal path redundancy, with buffer sizes of 2 and 3 respectively and a replication factor of 3. The lowest solid curve depicts the packet latency of the novel fabric (with internal paths redundancy, also referred to here as SemiL) with buffer size equal to 1 and replication factor of 3.

Finally, the lowest solid curve depicts the packet latency of the novel fabric (with internal paths redundancy, also referred to here as SemiL) with buffer size equal to 1 and replication factor of 3. This diagram demonstrates that when the offered load exceeds a value of 70%, the novel fabric presents the smallest packet latency in comparison with other devices which are equivalent in terms of network size. This improvement in latency can be explained as follows. When a heavy load is required to be transferred, this leads to a large uniform traffic. It operates with various input parameters such as buffer size, number of inlets/outlets, offered load, number of stages, and number of layers in the last segment of the fabric. Each SE was modelled using an array of non-shared buffer pairs of queues. Each queue operates on a FIFO (first input, first output) principle and each buffer is considered to be empty initially. In the same way, the simulation was carried out for single-layer MINs, and the results were used for comparison with the corresponding fabrics with redundant paths.

All simulation experiments were performed at packet level, assuming fixed-size packets transmitted in fixed-size timeslots, where the timeslot cycle is defined as the time required by a packet to be transferred from one stage to the next. All packet contentions, which occur when two packets claim the same next point, are resolved randomly.
number of blocks (and particularly in the last stages). As a consequence, this leads the fabric to use a greater number of alternative routes, thus relieving the blocking phenomenon. The conclusion which can clearly be drawn from this plot is that the novel fabric is the most efficient choice in terms of packet delay, in comparison with other similar architectures that do not include internal paths.

V. CONCLUSION

An innovative MLMIN is presented in this paper. The proposed fabric is composed of a preamble containing MUXs, a single-layer MIN with internal path redundancy implementation, and a fan-out in the final segment. This structure is composed of $2 \times 2$ and $2 \times 2^m$ switches, depending on the replication factor. Multiple paths for each input-output pair are provided by the fabric. This allows network connections to handle the traffic more efficiently. The current preliminary but careful study indicates that the proposed MLMIN architecture outperforms conventional MLMINs in terms of performance metrics and offers acceptable values of reliability and fault tolerance.

This work has many possibilities for extension. For example, this system requires study under conditions of hotspot traffic or a multicast service, and should also be evaluated in terms of performance metrics, priorities and cost metrics when operating with a backpressure or relaxing mechanism.

REFERENCES


