Reliability study of U-type multistage interconnection networks

E Stergiou¹, D Liarokapis¹, E Glavas¹, C Angelis¹, S V Margariti¹
¹Department of Computer Engineering, TEI of Epirus, Arta, Greece

Corresponding E-mail: dili@teiep.gr

Abstract. Multistage Interconnection Networks (MINs) are used in wide applications of multiprocessor systems in order to set up connections between various nodes such as processors and memory modules. Reliability can be considered as one of the most challenging issues in these systems because this can degrade or cancel the network performance and consequently the effort of the whole system. Here, a type of multistage fabric called U-type is selected for study, particularly in terms of its reliability. This fabric turns out to be the most appropriate type for constructing parallel systems due to the particular features that it possesses. This study reveals that U-type fabrics are ideal devices as they present mostly high values of reliability and low values of latency compared to unidirectional fabrics. This behaviour becomes more intensive when it is used in a cluster that demands service that includes a high percentage of “local type traffic.” Thus, U-type multistage devices can considerably enhance the connection points in constructing parallel systems, improving internal data flows.

1. Introduction
Multistage interconnection networks (MINs) are usually used as interconnection networks in cases of gigabit ethernet or Asynchronous transfer mode (ATM) switches [1-3]. But beyond this, MINs are proposed for connecting a large number of processors to establish a multiprocessor system [4-6]. Interconnection networks are critical in linking a large number of processors in the effort of building an integrated multiprocessor system. This is because such fabrics in general provide linear bandwidth and logarithmic packet latency in a number of nodes. Because of this behaviour, the interconnection networks are inevitable components in the construction of modern networks on chips (NoC) [5].

In constructing parallel systems, the most suitable type of MINs are the devices that here are named U-type MINs. U-type MINs can be considered as a special subclass of MINs. The basic and obvious feature that these devices have is that the inlets and outlets are located on the same side. This feature is extremely convenient for purely technical and constructional reasons when a large multiprocessor system is being manufactured. Due to this construction, the packets travel via a U-type device following a “U” shaped path. A second feature which these devices have is that in an arbitrary inlet–outlet pair, the packets do not follow the same route in terms of the number of hops. Ascertain on the one hand the continuing need for such network interconnection systems and on the other hand the study gap that exists in the current research, we are attempting to study this subject in accordance with the latest constructional requirements. Similar approaches to clos class of interconnected networks have already taken place [7]. In brief, the novel contribution of this work can be summarized as:

- A special type of MIN called a U-type MIN was selected for study because this device is unavoidable when trying to build systems with multiple parallel processors.
A study of the proposed fabric is carried out, showing that this fabric has many benefits; the results from the proposed fabric in terms of metrics, the reliability factor and packet latency, are extremely encouraging.

The results yielded indicate that the U-type MINs outperform the corresponding (in terms of network size) unidirectional MINs in terms of the reliability, and latency.

The remainder of this paper is organized as follows: in the second section, the proposed U-type MIN is introduced, and details of its operation, internal paths and some key definitions are given. In the third section, a reliability analysis and relevant plots are shown and discussed. In the next section, numerical results relating to latency are given. Finally, in the Conclusion section, our conclusions and anticipated future work are presented.

2. Description of the U-type MIN

Figure 1 shows a typical view of a sub-system in a multiprocessor system. As can be seen in Figure 1, before the U-type MIN component a special case unit is used called an Adaptation Block (AB).

![Figure 1. Typical view of a multiprocessor sub-system.](image)

The $N \times N$ classic MIN, where $N$ is the number of inlets/outlets, consisting of $k \times k$ unidirectional switch elements (SEs) has a maximum number of stages $L = \log_k N$, where $k$ depicts the size of the SEs that are used to construct the fabric.

In U-type MINs, instead of the traditional unidirectional SEs with $k$ input and $k$ output ports, bidirectional $(2k \times 2k)$ SEs are used. Figure 2 depicts the typical $2k \times 2k$ SEs (where $k = 2, 4, 8, ...$) that are used in the construction of U-type MINs. Depending on the set of bidirectional SEs that are selected, a corresponding U-type MIN is constructed. In general, bidirectional SEs have the ability to implement three ways of connecting, so they can operate in three modes. The first way of operation is the classic forward direction, the second is the backward direction and the third mode of operation is the return direction. (See Figure 2).
Figure 2. A typical operation (forward, backward and return operation) of bidirectional SEs that are used in U-type MINs.

In order to denote the working connection in a bidirectional SE, special notation is introduced, particularly:

- In the case of forward direction operation, the notation $p_i(p_o_j)$ means connection between input port $p_i$ and output port $p_o_j$, where $i, j \in [0, k - 1]$.
- In the case of backward direction operation, the notation $p_o_i(p_i_j)$ is used, meaning connection between output port $p_o_j$ and input port $p_i_i$ where $i, j \in [0, k - 1]$.
- In turnaround direction operation the pair $[p_i_i, p_i_j]$ denotes the connection link between an input $p_i_i$ and output $p_i_j$, where $i, j \in [0, k - 1]$ and $i \neq j$.

In all cases the U-type MINs use the store and forward mechanism as a packet management technique.

2.1. Job cluster example

In Figure 3 an example of an arbitrary distribution of traffic load is shown. In this example, the microprocessors 0, 1 are supposed to be running a job. This means the microprocessors feed the inputs 0 and 1 and the packets are driven to corresponding outputs 6 and 7. This type of traffic is characterized as “non-local traffic,” because the packet load is going through a long path (turn level 2 in MIN), and this has the effect of resulting in high values of packet delay.

On the other hand, the microprocessor 2 is assigned to execute another job and the packets are driven to output 3. Because the distance between the input and output is the minimum (they are neighbours) the corresponding U-type path that the packets are using is the shortest. The same happens for the input/output pair 4, 5. This type of traffic is characterized as “local traffic.” The packets which use the shortest path have the minimum corresponding packet delay.

The challenge here is how to minimize the packet path distances using the shortest U-type paths in order to improve the behaviour of the system by decreasing the time for job completion.

The main idea that is behind the usage of U-type MINs is to exploit the neighbouring inlets/outlets – exploit the shortest paths – in the interconnection network. Exploiting the shortest paths in the network reduces the packet traffic population in the U-type MIN and reduces traffic congestion. This characteristic basically decreases the packet latency, which is extremely useful in the case of Network on Chip (NoC) and generally in all attempts to redistribute traffic load and increase the performance of parallel computers.
2.2 Adaptation Block (AB)

The Adaptation Block (AB) accepts jobs by processors and has the ability to rearrange those jobs, changing the feed sequence at the inputs of a U-type MIN. The main goal of the AB is to increase the ratio of ‘local type traffic’ in the multistage network by changing the topology. The AB unit has the ability to rewire the connection between inputs/outputs. Thus, the AB has the ability to rearrange the input traffic in the U-type MIN. Each application allocates an exclusive subset of processors [8]. This bounded set of processors can be seen as a processor cluster.

Communication in a parallel multiprocessor system takes place predominantly inside these clusters. Therefore, it is desirable that processors within a cluster are placed as close to each other as possible. Ordinarily, allocating applications in adjacent processors is not feasible at runtime. Moreover, this situation gets even worse when the applications have different starting times relative to each other and when a favourable allocation of clusters is mostly not possible. The moving of an application in context from one processor to the other to localize the clusters is a very costly process [9]. The AB unit reconfigures the network topology to localize application clusters and optimize the communication requirements.

Supposing we have an aforementioned sub-system of a parallel system, a significant parameter for the proper operation of a multiprocessor system is the reliability of the system and especially the reliability of the U-type multistage fabric that is used in the system. Next, some basic definitions are presented followed by a description of the reliability analysis of the U-type MIN.

2.3 Basic Definitions

Buffer size \((b)\): the maximum number of packets that an input buffer of an SE can hold.

Probability of packet arrivals \((p)\) at an input (offered load): in our experiment the probability of packet arrivals is ranked from 0.1 to 1.0, with steps equal to 0.1.

Arrival process at the input of the U-type MIN: the generation of succeeding packets is assumed to be a mutually independent process. The arrival process for packets at the output queues of the first stage of the network \(x_{k,n}^{(0)}\) is given by a binomial distribution \(bin\left(k,\frac{p}{k}\right)\), where \(p\) is the fixed probability of a packet generated by a processor at each cycle. Therefore:
3. Essential reliability analysis of U-type MINs

The reliability of a MIN is defined as its ability to overcome all unexpected circumstances. In this study, the point-to-point reliability (p-t-p reliability) is considered. This type of reliability is also known as the terminal reliability. With p-t-p reliability, the probability of at least one fault-free path existing between an inlet–outlet pair is calculated. The reliability between an arbitrary given pair (inlet–outlet) is dependent on all of the switch elements involved, insofar as the failure of each element implies the failure of the current routing action [10]-[13].

3.1 Reliability of the unidirectional N×N MINs

In the case of unidirectional N×N MINs, \( L = \log kN \) is the number of stages, where \( k \) depicts the size of the unidirectional SEs that are used to construct the fabric. In such a case, it is supposed that all of the packets traverse a series of \( L \) SE components, so all the packets travel equal paths in terms of the node (or stage) hops. Assuming that each unidirectional SE has a reliability of \( r_i \), then the reliability (\( R \)) of this subsystem can be calculated as

\[
R_{\text{UnidirectionalMIN}} = \prod_{i=1}^{L} r_i \tag{2}
\]

Supposing the reliability of all SEs are the same due to the homogeneity of the MINs, the total terminal reliability is \( r^{L \log kN} \).

3.2 Reliability of the U-type MIN

In the case of N×N U-type MINs, with a total depth \( L = \log kN \) number of stages, the length of the paths that the packets use is not fixed in general. The path lengths (in number of node hops) of various packets can vary from 2 to \( n = (2 \times \log kN) - 1 \). Two hops is the shortest U-type path in a U-type MIN. This route is taken when the input and the output are adjacent and there is no other node involved between this pair. On the other hand, \( n = (2 \times \log kN) - 1 \) is the longest path. The longest path happens when the distance between the input and the output is the maximum. So, the corresponding maximum value of the terminal reliability can be calculated as

\[
R_{\text{U-typeMIN}} = \prod_{i=1}^{(2 \times \log kN) - 1} r_i \tag{3}
\]

where \( r_i \) is the reliability of an arbitrary bidirectional SE acting in one of the three aforementioned operations. This reliability factor (\( r_i \)) is a manufacturer characteristic of the construction.

Consequently, the total terminal reliability value is in the range

\[
R_{\text{U-typeMIN}} \in \left[ \prod_{i=1}^{2} r_i, \prod_{i=1}^{(2 \times \log kN) - 1} r_i \right] \tag{4}
\]

so, supposing the fabric has homogeneity, without harming generality the total terminal reliability value varies as

\[
R_{\text{U-typeMIN}} \in \left[ r^2 \cdot r^{[(2 \times \log kN) - 1]} \right], \text{ where } r = r_i \ \forall i \in [1 - ((2 \times \log kN) - 1)] \tag{5}
\]
In Figure 4, the minimum values of the normalized terminal reliability for U-type MINs (8x8, 64x64, 256x256 and 1024x1024 network size) are portrayed for varying values of the bidirectional SE’s reliability \( (r_i) \), acting in one way. The minimum values of the total normalized terminal reliability occur when the packets travel the maximum route (in hops) inside the fabric. Thus, for U-type MINs with a network size of 1024x1024 (MIN depth equal to 10) and SEs with bidirectional SE reliability of 90%, the total terminal reliability drops below 20%. Strictly speaking, that means this network size is prohibitive for such uses. According to this scenario, in all of the U-type MINs under study, all packets of traffic follow the longest “U” type route, in other words, the maximum route.

Also, the upper solid line in Figure 4 depicts the maximum values of the normalized terminal reliability of all the corresponding U-type MINs. The maximum values of the terminal reliability occur when the packets use the shortest path, equal to 2 hops, use only one bidirectional SE and operate exclusively in the return operation. The maximum set of terminal reliability values is common for all of the U-type MINs regardless of their size.

3.3 Traffic patterns of the cluster
Nonetheless, when we have a given cluster that includes various jobs, the packages of those jobs follow different paths in general according to the needs of each job (see Figure 3). This in turn changes the reliability factor of a U-type MIN. So, the results that are illustrated in Figure 4 show the boundaries of the reliability factor. However, for a given cluster with a specific traffic distribution, the reliability of its fabric needs calculation. In order to monitor the changes that the reliability factor undergoes, three types of typical traffic patterns are defined for study reasons.

Traffic pattern A: 90% non-local traffic and 10% uniform type traffic. Thus, according to this scenario 90% of the packet population turns in the last stage of the U-type MINs.

Traffic pattern B: Uniform distribution of “U” type paths. This means that we have a gradual return of packages at each stage.

Traffic pattern C: 90% local traffic and 10% uniform type traffic. Thus, according to this scenario 90% of the packet population turns in the first stage of the U-type MINs.

Figure 5 depicts the probability of turn for the three types of traffic patterns (A, B and C type) in a 5-stage depth U-type MIN.
3.4 Total terminal reliability for various traffic patterns

Figure 6 depicts the total normalized reliability for U-type MINs with depths varying from 3 to 10 stages and for the aforementioned three types of traffic patterns (A, B and C) in a quantitative manner. The upper solid line in Figure 6 shows the total normalized reliability when the traffic pattern is 90% local traffic for network depths varying from 3 to 10 stages. It is clear that the total reliability for this case remains at high values compared to the lower solid red line depicting the corresponding reliability of the opposite case (90% non-local traffic). In the case of non-local traffic, the reliability values drop dramatically.

It’s clear that the network size affects the values of the total normalized reliability of a U-type MIN. The prevalence of non-local traffic of non-local traffic in large-scale networks is the worst case, which should be avoided. Moreover, in Figure 6, the solid line with triangular triangular points illustrating the uniform distribution of turn traffic reveals that the values range over intermediate levels. Finally, the dashed line portrays the normalized reliability of a corresponding unidirectional MIN with an equivalent network size (5 stages, 32 × 32 MIN).
4. Latency metric of U-type MINs

Most of the studies of MIN performance for simplicity use simulation as an evaluation tool [14]. So, here for the same reasons, a special-purpose simulator was developed to evaluate the overall network performance of the U-type MINs. Also, another equivalent simulator was created for the unidirectional MINs. These tools were developed in C++ and are capable of operating under various configuration scenarios and handling various traffic patterns.

Figure 7 depicts the normalized packet latency of U-type and unidirectional (UD) MINs with the number of total stages equal to 6, 8 and 10 plotted versus the total offered load (varying from 10-100%). The solid lines depict the latency of unidirectional MINs while the dashed lines show the latency of U-type MINs. All of the MINs under study have a buffer size equal to 1, and the pattern applied to the U-type MIN inputs was 80% non-local traffic (Long Distance paths – LD) while the rest (20%) was uniform traffic.

![Figure 7. Normalized packet latency versus the total offered load for U-type and unidirectional MINs with maximum network depths of L=6, 8 and 10.](image)

The plot reveals that the U-type MIN provides significantly less delay in the flow of packets in comparison to the corresponding (in terms of network size) unidirectional (UD) MIN. For example, an 8-stage (256x256) U-type MIN with full traffic (100% offered load of which 80% is non-local traffic (LD paths) and 20% is uniform traffic) presents a normalized packet latency equal to 1.33, whereas it is 1.53 for the corresponding unidirectional MIN (reduction of ~13%). This behaviour (latency reduction) is particularly useful in constructing parallel systems.

5. Conclusion

U-type MINs consisting of bidirectional SEs have been selected for study in terms of their reliability and packet latency. This class of MINs is particularly useful in constructing parallel systems. The current study shows that the reliability metric of U-type MINs is in general higher compared to unidirectional MINs. This metric is extremely good when a U-type MIN is fed by a large percentage of local traffic. This means that the jobs submitted in the system have a small distance between the input/output pair (e.g., they are neighbours). This preliminary study reveals that increased values of reliability combined with low packet latency values make U-type MINs an absolutely essential component in the construction of parallel systems.

U-type MINs are adroit multi-connection systems that have the ability to efficiently handle different types of traffic (e.g., hot spot, multicast, etc.) thus improving network performance. Therefore, this construction seems to have significant merit for use in the information industry. This work could be extended in many ways. For example, the system could be studied under hot spot traffic or multicast
service and also in terms of performance metrics, priorities or cost metrics when operating with a backpressure mechanism.

6. References

[1] Soumiya T Nakamichi K, Kakuma S Hatano T & Hakata A (1999). The large capacity ATM backbone switch “FETEX-150 ESP”. Computer Networks, 31(6), 603-615.

[2] Garofalakis J & Stergiou E (2011). Mechanisms and analysis for supporting multicast traffic by using multilayer multistage interconnection networks. International Journal of Network Management, 21(2), 130-146.

[3] Ra'ed Y A & Mouftah H T (1995). Survey of ATM switch architectures. Computer networks and ISDN systems, 27(12), 1567-1613.

[4] Abandah G A & Davidson E S (1996, April). Modeling the communication performance of the IBM SP2. In Parallel Processing Symposium, 1996., Proceedings of IPPS’96, The 10th International (pp. 249-257). IEEE.

[5] Dabiri B Rezaei S H S & Modarressi M (2017, August). Low-power Parallel Data Processing Using Computation Reuse. In Proceedings of the 7th International Conference on Information Communication and Management (pp. 1-4). ACM.

[6] Sadawarti H Bansal S & Kaur N (2015). A Survey on Multi-Stage Interconnection Networks. Int. J. Comput. Sci. Trends Technol, 3(1), 143-151.

[7] Zhao Z Li B Hu Z Guan Y Wu B Dong X & Zhu G (2017, November). Minimizing delay by multi-hop routing and scheduling in clos networks for DCNs. In Systems and Informatics (ICSAI), 2017 4th International Conference on (pp. 1018-1023). IEEE.

[8] Liu J C Shin K G & Chang C C (1995). Prevention of congestion in packet-switched multistage interconnection networks. IEEE Transactions on Parallel and Distributed Systems, 6(5), 535-541.

[9] Lu C & Lau S M (1994, August). A performance study on load balancing algorithms with task migration. In TENCON'94, IEEE Region 10's Ninth Annual International Conference. Theme: Frontiers of Computer Technology. Proceedings of 1994 (pp. 357-364). IEEE.

[10] Jahanshahi M & Biestouni F (2015). Improving the reliability of the Benes network for use in large-scale systems. Microelectronics Reliability, 55(3-4), 679-695.

[11] Rajkumar S & Goyal N K (2016). Review of multistage interconnection networks reliability and fault-tolerance. IETE Technical Review, 33(3), 223-230.

[12] Md Yunus N A Othman M Mohd Hanapi Z & Lun K Y (2016). Reliability review of interconnection networks. IETE Technical Review, 33(6), 596-606.

[13] Arya S & Singh N (2017). Reliability Analysis of Modified Irregular Augmented Shuffle Exchange Network (MIASEN). International Journal of Computer Science Issues (IJCSI), 14(3), 59.

[14] Stergiou E & Garofalakis J D (2011). A Simulation Study for Optimizing the Performance of Semi-layer Delta Networks. In SIMULTECH (pp. 257-265).