

The Effect of the Degree of Multistage Interconnection Networks on their Performance: the Case of Delta and Over-sized Delta Networks

Ahmad Chadi Aljundi, Jean-Luc Dekeyser
Universite des Sciences et Technologies de Lille
Laboratoire d'Informatique Fondamentale de Lille
59650 Villeneuve d'Ascq
{aljundi,dekeyser}@lifl.fr

Abstract

Interconnection network performance is a key factor when constructing parallel computers. Today's technological progress makes it possible to build and use crossbars of sizes up to 128. Crossbars can be used as switching elements (SEs) in parallel architectures intercommunication systems such as multistage interconnection networks (MINs). A MIN is usually defined, among others, by its topology. One of the factors defining the topology of a MIN is its degree. The degree of a MIN is the size of the SE of which it is composed. In this paper we are interested in studying the influence of the degree of two classes of MINs on their performance. The tested MINs classes are the famous Delta networks and a subclass of this family called the over-sized Delta networks. This study is to be used in future work in order to evaluate the use of MINs as an intercommunication medium in Symmetric Multiprocessors.

1. Introduction

As electronic components reach their physical speed limits, using parallelism seems to be the only known practical solution to face today's scientific applications augmenting need for computation speed. In a parallel computer the intercommunication between processors and their communication with memory modules are key factors on which the performance of the overall system depends. The important thing about an intercommunication system is that it has the capacity to route many communication tasks concurrently. A conflict occurs when more than one message try to access a same communication resource. Three types of conflicts exist while accessing the shared memory in a multiprocessor system [11]: network conflicts, bank busy conflicts, and simultaneous bank conflicts.

While a bus allows a very limited level of parallelism, a crossbar [24], which provides a full connection between

all the nodes of the system is very complex, expensive and hard to be controlled. MINs [21] are an interesting solution for intercommunication systems in parallel computers. They provide an acceptable degree of parallelism with a complexity well below that of a crossbar. Many MINs belonging to the famous Delta MINs family were studied and effectively used to build parallel computers [8]. Delta MINs form a sub-group from a bigger MINs family called banyan MINs of which networks are characterized by the existence of one and only one path between each source and destination. Non-banyan MINs are, in general, more expensive than banyan networks and more complex to control. Still, they often are fault tolerant and capable to apply rerouting strategies used to bypass routing problems such as a conflict or a faulty link or switch. Kruskal and Snir [13] studied *augmentation* techniques that might be applied on banyan networks in order to improve their performance without much loss of simplicity. Augmentation can be defined as adding links and/or switches to the basic configuration of the network.

A MIN is usually defined by, among others, its topology, routing algorithm, switching strategy, and flow control mechanism [6]. One of the factors defining the topology of a MIN is its *degree*. The degree of a MIN is the size of SEs of which it is composed. In this paper we study the effect of the MIN's degree on its performance. Networks of degrees 2, 4, and 8 were studied by Cheemalavagu and Malek in [4]. In their paper, they were limited to 8 degree MINs because of the space and time needed for the simulation. Furthermore, they used networks with and without buffers. In this paper we investigate MINs of degrees up to 64, all of them unbuffered. In order to establish this study, different degrees Delta and over-sized Delta MINs [3] are tested. The evaluation methodology is a multi-criteria simulation one that projects performance measures into multi-dimensional space. The methodology was explained in an earlier publication [3].

SMP (Symmetric MultiProcessing) machines are inter-

esting architectures used today to build parallel computers. They are either crossbar or bus based architectures. In fact, while MINs were widely studied in the literature, we think that the use of MINs as an interconnection medium for SMP architectures can improve their performance. This might be the subject of future studies in the field.

The remainder of this paper is organized as follows: In section 2 we present a topological classification of MINs before a rapid reminder of the two example architectures to be used in the paper: the Omega network and the MCRB network. This is followed by the explanation of the evaluation methodology to be used. Section 5 presents the comparison and evaluation results before ending the paper with the conclusion.

2. Topological Classification of MINs

A MIN can be defined as a network used to inter-connect a group of N inputs to a group of M outputs using several stages of small size SEs followed (or leaded) by linking stages [23].

We propose in figure 1 a topological classification of MINs. We give in the following some important definitions related to this classification.

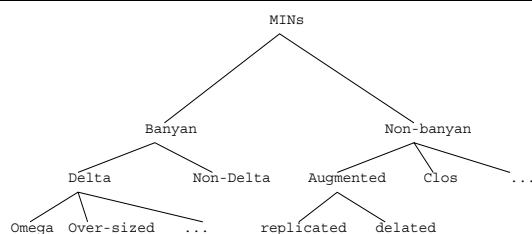


Figure 1. A topological classification of MINs

Definition 1 A banyan [7] MIN is a MIN having the property of the existence of one and only one path between each source and destination.

Banyan MINs might have the *Delta* property or not. Basing on Patel's definition of Delta networks [20], we propose the following mathematical definition of the Delta property.

Definition 2 For a Banyan MIN of size N and degree r^1 , suppose that the switch's inputs and outputs are presented to the base r , i.e. in the form d_0, d_1, \dots, d_{r-1} . Let the inputs and outputs of the SEs in the network have the same indexes then digits d_0 of all inputs of a switch must be equal. A network or a stage having this characteristic is called to be having the *Delta* property.

¹ In this paper Network(N,r) will present a MIN of size N and degree r .

SEs in Delta networks, which are banyan networks, are digit-controlled crossbars. Digit-controlled crossbars are controlled by digits of the message's control sequence.

In fact, a network having the Delta property possesses some kind of regularity so that the network's routing algorithm can be simple and well defined [12].

Definition 3 We call an over-sized MIN of size N a banyan Delta MIN composed of more than one copy of a Delta MIN gathered together by an interconnection stage having the Delta property.

Non-banyan networks can be constructed either by the *augmentation* of a banyan network [13] or by the construction of a *multipath* network such as the Clos network [5].

3. Case Studies

Our goal in this paper is to evaluate the effect of MINs degrees on their performance. The study will be applied on Delta and over-sized Delta MINs. Later in this paper, Delta networks will be presented by its sub-class Omega network for comparison purposes. The MCRB network will be the example used to study over-sized Delta networks.

3.1. Omega Network

A shuffle exchange [9] network, also called Omega network [14], is a subset of the Delta networks family proposed by Patel [20], which is a bit-controlled interconnection networks family. Figure 2 shows an Omega(16,4) network.

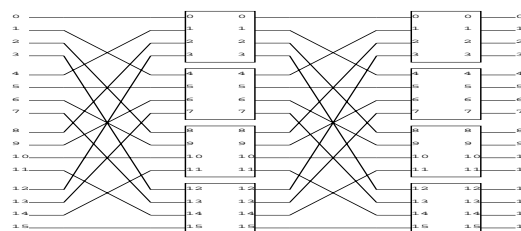


Figure 2. A block diagram of $\Omega(16, 4)$

3.2. Example of Over-sized Delta Networks: The MCRB Network

The MCRB topology, defined by Kechadi in [10], is a dynamic multistage implementation based on the static chordal ring topology [19]. For complexity reasons, described in [1], we give here a definition a bit different of the one given by Kechadi.

Proposition 1 An $MCRB(N, r)$ network is a MIN built of $r \times r$ SEs and contains n stages of $N = r^n$ (SEs) each. Let SE_{ij} be the SE j of the stage i of the $MCRB(N, r)$, then SE_{ij} is connected to SE_{i-1, k_d} such that $k_d = (j + d r^i) \bmod r^n$, for $0 \leq i \leq r^n - 1, 1 \leq j \leq r - 1$, and $0 \leq d \leq r - 1$.

As an example, the configuration of the $MCRB(8, 2)$ is shown in Figure 3.

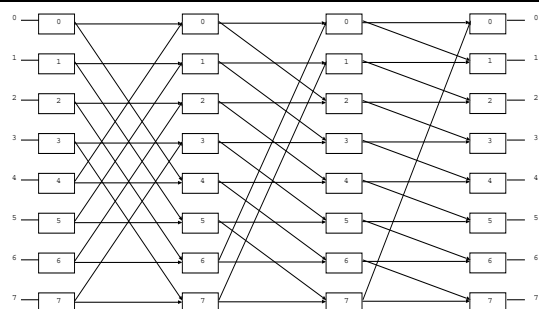


Figure 3. $MCRB(8, 2)$ Network

In [3] it is proved that the MCRB network is a special case of Delta networks. We call this special case over-sized Delta networks.

4. Evaluation Methodology

In [3] a multi-criteria methodology based on performance measures which defines a systematic decision making mechanism for choosing more suitable MINs for a multiprocessor system was presented. This methodology is used in this paper to evaluate the effect of the network's degree on its performance. We will limit our study to three performance evaluation factors knowing that the proposed methodology is general and that it is easy to add other factors chosen to evaluate the performance of a MIN. Here we give small reminders of the definitions of the used metrics.

4.1. Integration complexity

While studying a MIN, the first evaluation to do is its hardware complexity. The hardware complexity of a MIN can be calculated by two means: the number of cross-points and the number of connections or wires needed to construct the MIN. Liu [16] defines the hardware complexity of a MIN as the maximum of the two means. The hardware complexity of a MIN in term of crosspoints is equal to the total number of crosspoints of all crossbars used to build it. The complexity in terms of connections is the sum of links or wires in all stages.

Definition 4 consider a MIN of size N and degree r , that has X stages of x SEs each. The stages are connected with Y inter-stages links. The integration complexity of the MIN will be defined as $C = \max(r^2 X x, Y r)$.

4.2. Throughput

Throughput is defined as the number of messages delivered to their destinations per unit of time [17, 20]. Many analytical studies of MIN's throughput can be found in the literature [20, 13, 22]. Simulation is used frequently when more realistic results are needed. It allows more flexibility in network characterization in order to make it possible to analyze real-world and popular communication patterns. To study the throughput of an unbuffered network, messages leave sources to their destinations. In the case of a conflict, only one message goes through and the others are discarded. The throughput is calculated as the number of messages that arrive to their destinations per unit of time over a certain number of trials.

Definition 5 In an unbuffered MIN, We define the throughput as the number of messages delivered to their destination per unit of time knowing that only one message goes through when more than one message assigned the same interconnection resource. All other messages are discarded.

4.3. Network's Latency

Another important performance parameter is the network's latency, which is defined below. The network latency analysis depends directly on the maximum number of cycles needed to route a certain number of permutations to their destinations. We use the same previously explained simulation to measure the analyzed MINs' latency.

Definition 6 The latency of a MIN is defined by the number of network cycles needed for all messages of a permutation to arrive to their destinations. This is referred to as the network's universality [23].

4.4. Universal Performance Factor

Here we explain how performance factors can be combined in order to get a universal performance evaluation factor. The above defined factors will serve as examples to validate our proposed evaluation methodology.

We suppose that the importance of the factors is a design choice (i.e. the performance factors to be evaluated are chosen). In general, performance evaluation factors can be divided into two major groups: factors to be maximized and factors to be minimized. We call the group of factors to be maximized $p^{max} = \{p_1^{max}, p_2^{max}, \dots, p_k^{max}\}$ and the factors to be minimized $p^{min} = \{p_1^{min}, p_2^{min}, \dots, p_l^{min}\}$,

where k is the number of factors to be maximized and l is the number of factors to be minimized. The universal performance factor is defined by the following:

Definition 7 Given a MIN and groups of performance factors p^{max} and p^{min} . The universal performance factor (UPF) can be broadly defined by:

$$UPF = \sqrt{\sum_{i=1}^k (p_i^{min})^2 + \sum_{j=1}^l \frac{1}{(p_j^{max})^2}} \quad (1)$$

Definition 8 Given two networks μ_1 and μ_2 and their UPFs, UPF_1 and UPF_2 respectively. We say that μ_1 is more performant than μ_2 if $UPF_1 < UPF_2$.

In order to clarify the idea behind this definition consider an example of two MINs performance evaluation with two performance factors p' and p'' . We assume that these two factors are both to be minimized. Figure 4 presents in two dimensional space the performance of the two MINs μ_1 and μ_2 . Let $p1'$ (resp. $p2'$) and $p1''$ (resp. $p2''$) be the calculated values of these factors for μ_1 (resp. μ_2). From Figure 4 one can notice that μ_1 is more performant than μ_2 as μ_1 gives smaller values than those of μ_2 . Note that the UPF is the length of the vector having for coordinates (p'_i, p''_i) . One can notice that smaller value for UPF, better is the network performance.

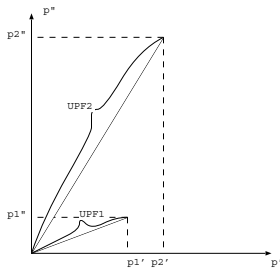


Figure 4. An example of the use of the UPF factor.

Usually different parameters have different types of measures and scaling. In order to solve this problem, values can be normalized to a certain value, which may be the maximum value, or the average value for each factor. We call \bar{p}_i the average value of factor p_i . The equation 1 can be replaced by the following equation:

$$UPF = \sqrt{\sum_{i=1}^k \left(\frac{p_i^{min}}{\bar{p}_i^{min}}\right)^2 + \sum_{j=1}^l \left(\frac{\bar{p}_j^{max}}{p_j^{max}}\right)^2} \quad (2)$$

The equation 2 can be further improved by including the importance aspect of each factor in the design process of a MIN. This can be done by multiplying each term by a factor called the *weight* (w). The weight w_i expresses the importance of the performance parameter p_i . This leads to the following final equation.

$$UPF = \sqrt{\sum_{i=1}^k w_i \left(\frac{p_i^{min}}{\bar{p}_i^{min}}\right)^2 + \sum_{j=1}^l w_j \left(\frac{\bar{p}_j^{max}}{p_j^{max}}\right)^2} \quad (3)$$

From now on all performance factors are normalized and all factors' weights are of equal importance.

5. Performance Evaluation

In this section, we evaluate and compare the effect of the degree of two MINs on their performance. Tested MINs are the Omega and the MCRB networks. Two stages MCRB networks will not be tested as they are non-blocking and more complex than the crossbar [2]. Simulation mechanism and the way results were obtained can be found in a previous publication [1]. Note that BPC (Bit Permute Complement) permutations are used as work loads [15, 18].

5.1. One dimension evaluations

We mean by one dimension evaluations those considering only one performance factor.

5.1.1. Latency Concerning latency, results that we got correspond to those found by Cheemalavagu and Malek [4]. Figure 5 shows that latency of networks of degree 4 is better than those of degree 2 and 8 for 64 size SW-banyans. In fact, this is not true for another banyan special case Delta network, which is the over-sized network. Figure 6 shows that, always, for over-sized Delta network less latency is obtained with crossbars of bigger degrees. The same figure shows that there is an optimal value of the Omega network's degree so that it has the least latency. This can be explained as follows: It is evident that bigger size crossbars can pass a larger number of permutations of the same size. However, the permutation capacity of a crossbar might be limited by the great number of requests coming to its inputs. The probability of a conflict to occur in a 2 degree Delta network is greater than that in a bigger degree one. This can be demonstrated by simple calculation using, for example, Patel's formula [20] to calculate the probability of message arrival in a Delta network. In fact, for a certain degree value, this conflict probability reaches an important level so that less messages can pass, which implies a larger latency. This explains the existence of the optimal value for Delta networks. On the other hand, as the number of crossbars in an over-sized Delta network is r times the number of crossbars

in a Delta network, this limit is not reached and the universality of the network is always better when using crossbars of bigger sizes.

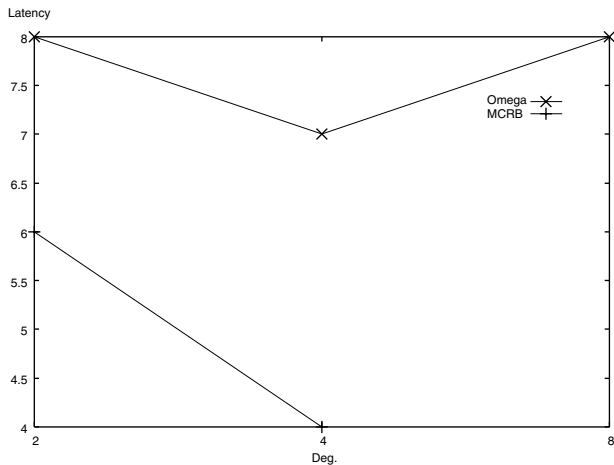


Figure 5. Latency of 64 size networks as a function of the degree

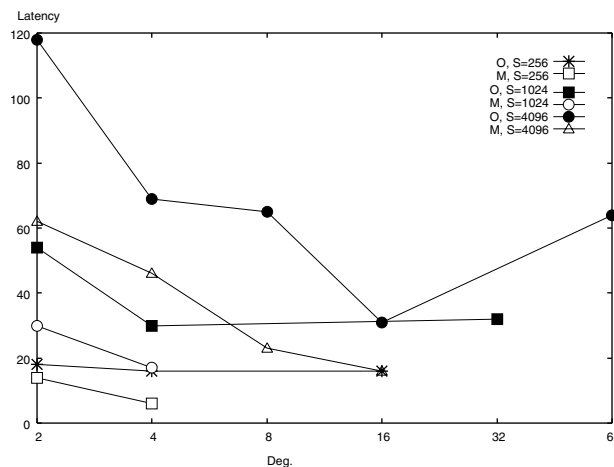


Figure 6. Latency as a function of the degree

5.1.2. Throughput Figure 7 shows the throughput of different sizes networks as a function of the degree. The figure shows that for over-sized Delta networks, throughput is bigger when larger size crossbars are used. On the other hand, Delta networks have, once again, an optimal degree value, which is 2 for size 64 networks, 2 and 4 for size 1024 and 4 for 4096 size networks.

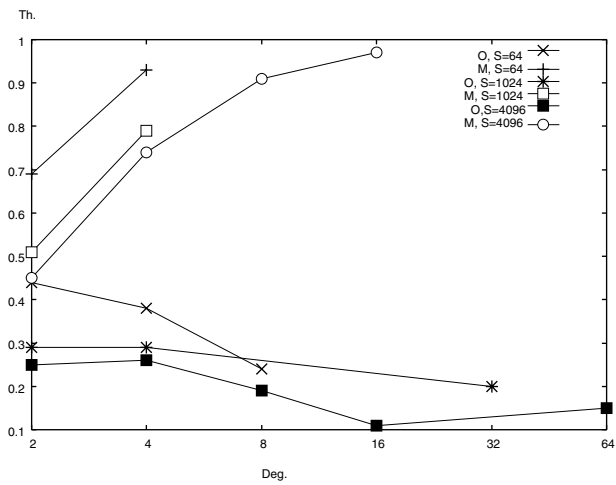


Figure 7. Throughput as a function of the degree

Note that when a very big number of messages arrive to the first stage of a network, using our approach to calculate the throughput, a considerable number of messages are discarded. No conflicts can occur on the last stage as the destinations are groups of permutations. This means that conflicts can occur only on a limited number of stages in large degree MINs which might explain the slight augmentation of the the throughput of Omega(4096,64) as related to Omega(4096,16).

5.2. Two dimensions evaluations

Here, two performance factors are evaluated simultaneously.

5.2.1. Complexity and throughput UPF Comparing the UPFs of several MINs regarding only the complexity and the throughput means that latency is not an important factor to be evaluated. In other words, the betterness of the network is judged by a small complexity and a high throughput only, which are to be evaluated at the same time. What is really important to note in figure 8 is the complexity effect on the performance of the over-sized Delta networks. This will be further cleared in the next section where we study the universality-throughput UPF.

Note also the approach of the performance values of the Delta network to the optimal degree value. For small size networks, small degrees give the best results, while for networks of size 1024, crossbars of size 2 and 4 give the same best performance. However, with 4096 size network the optimal degree value is 4.

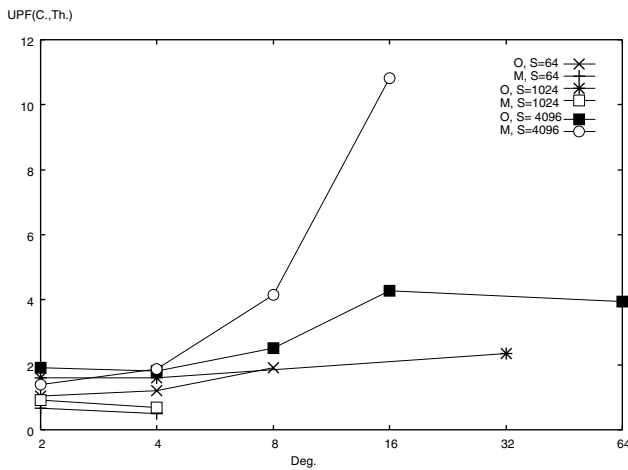


Figure 8. Complexity-throughput UPF as a function of the degree

5.2.2. Universality and throughput UPF Here, the cost of the MIN is not an important factor to be evaluated. So, we are looking for the MIN that gives the best latency AND throughput at the same time regardless to the complexity. In fact, MCRB networks show a very good performance when complexity is not to be considered. This is why their universality-throughput UPF is considerably smaller when using bigger sizes crossbars. Remember that a smaller UPF means better performance. Note that complexity plays, when considered, an important role in degrading the network's performance as related to the complexity-throughput UPF case. This is because over-sized Delta networks' complexities augment very rapidly with their degrees.

On the other hand, we see once again the optimal value aspect when studying the Delta network. Also note the improvement of the UPF for the Omega(4096,64) relatively to the Omega(4096,16). This is due to the improvement of the throughput of this network, previously explained.

5.3. Inclusive UPF

Figure 10 presents the comparison and evaluation of a number of networks as related to the three factors we are considering in this paper as examples, i.e. complexity, universality and throughput.

Once more, we can observe the important effect of complexity on the overall system performance. This can be easily seen by comparing figures 10 and 8, in which curves have almost identical shapes (the shapes are totally different when complexity is not considered, see figure 9). What is remarkable in figure 10 is that one of the over-sized Delta cases, which is when the size of the network is 4096, has

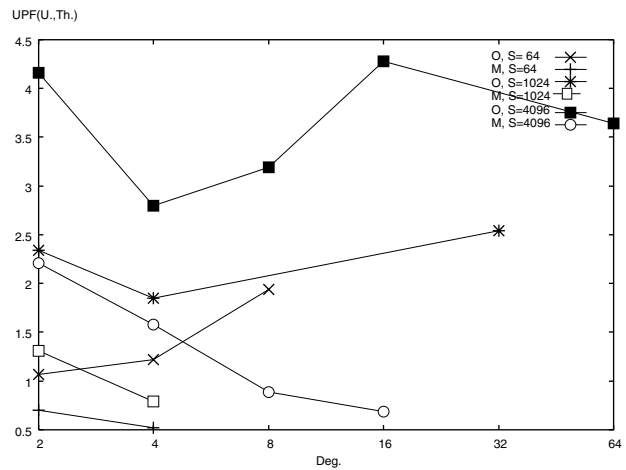


Figure 9. Latency-throughput UPF as a function of the degree

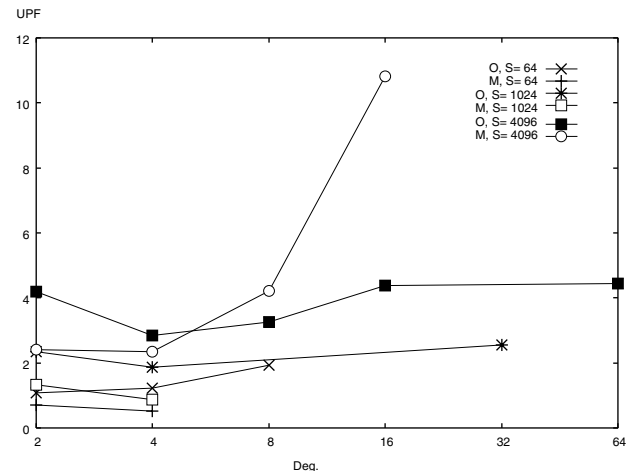


Figure 10. Inclusive UPF as a function of the degree

an optimal degree value which does not correspond to the biggest one. The best case for 4096 size over-sized Delta networks, when considering all three performance factors, is when the degree is equal to 4.

6. Conclusion

MINs have been used as intercommunication systems in parallel machines. Their performance is considered as a key factor on which the system's overall performance de-

pend. The need of a multi-criteria performance evaluation methodology is due to the large number of performance metrics that one may want to test to evaluate a MIN. We proposed a measures based evaluation and comparison methodology in previous publications. New proposed architectures can be tested using this methodology and they can be compared to other known MINs. In this paper, we used this comparison and evaluation methodology that applies simple heuristics to compare networks of different degrees. The main interest of the paper was to evaluate the network's degree effect on its performance. Our observations showed that an optimal degree value exists for Delta MINs. However, in general, bigger degrees MCRB networks lead to better performance. An exception is noted to this rule when network's complexity becomes really important and all performance factors are considered. In this case the over-sized Delta network can even approach the Delta networks performance network optimal value aspect.

This study of MINs is aimed to be a basis for future study of the use of MINs in SMP machines.

References

- [1] A. C. Aljundi, J.-L. Dekeyser, M.-T. Kechadi, and I. D. Scherson. Comparative simulations and performance evaluation of mcrb networks using multidimensional queue management. In *Proc. Int'l Symp. on Performance Evaluation of Computer and Telecommunication Systems*, pages 288–296, San Diego, USA, July 2002.
- [2] A. C. Aljundi, J.-L. Dekeyser, M.-T. Kechadi, and I. D. Scherson. A study of an evaluation methodology for unbuffered multistage interconnection networks. In *Proceedings of 17th International Parallel and Distributed Processing Symposium, IPDPS'03*, Nice, France, April 2003.
- [3] A. C. Aljundi, J. L. Dekeyser, and I. D. Scherson. An interconnection networks comparative performance evaluation methodology: Delta and over-sized delta networks. In *Proc. of the 16th International Conference on Parallel and Distributed Computing Systems*, pages 1–8, Reno NV., USA, Aug. 2003.
- [4] S. Cheemalavagu and M. Malek. Analysis and simulation of banyan interconnection networks with 2x2, 4x4 and 8x8 switching elements. In *Proc. Real-Time Syst. Symp.*, pages 83–89, 1982.
- [5] C. Clos. A study of non-blocking switching networks. *Bell system tech. journal*, 32(2):406–424, Mar. 1953.
- [6] D. E. Culler, J. P. Singh, and A. Gupta. *Parallel Computer Architecture (A Hardware/Software Approach)*, chapter Interconnection Network Design. Morgan Kaufmann Publishers, 1999.
- [7] G. Goke and G. Lipovski. Banyan networks for partitioning multiprocessor systems. In *Proc. 1st Annu. Symp. Comput. Arch.*, pages 21–28, 1973.
- [8] A. Gottlieb. An overview of the nyu ultracomputer project. Technical report, The NYU Ultracomputer project, Oct. 1987.
- [9] K. Hwang and F. Briggs. *Computer Architecture and Parallel Processing, 5th printing*. McGraw-Hill series in computer organization and architecture. McGraw-Hill International Editions, 1989.
- [10] M. T. Kechadi. Mcrb: A new interconnection network for multiprocessor systems. In *Misc. Papers, CD-ROM of the 2002 International Conference on Parallel and Distributed Processing Techniques and Applications (PDPTA'02)*, ISBN: 1-892512-39-4, Las Vegas, USA, June 2002.
- [11] M. T. Kechadi and J.-L. Dekeyser. Analysis and simulation of an out-of-order execution model in vector multiprocessor systems. *Parallel Computing*, 23:1963–1986, 1997.
- [12] C. Kruskal and M. Snir. A unified theory of interconnection network structure. *Theoretical Computer Science*, 48:75–94, 1986.
- [13] C. P. Kruskal and M. Snir. The performance of multistage interconnection networks for multiprocessors. *IEEE Trans. Comput.*, C-32(12):1091–1098, Dec. 1983.
- [14] D. A. Lawrie. Access and alignment of data in an array processor. *IEEE Trans. Comput.*, C-24(12):1145–1155, Dec. 1975.
- [15] J. Lenfant. Parallel permutations of data: A benes network control algorithm for frequently used permutations. *IEEE Trans. Comp.*, C-27:637–647, July 1978.
- [16] Y.-S. Liu. *Architecture and performance of processor-memory interconnection networks for MIMD shared memory parallel processing systems*. PhD thesis, New York University, 1990.
- [17] A. Merchant. *Analytical Models for the Performance Analysis of Banyan Networks*. PhD thesis, Stanford University, 1991.
- [18] D. Nassimi and S. Sahni. A self-routing benes network and parallel permutation algorithms. *IEEE Trans. Comput.*, C-30:332–340, May 1981.
- [19] B. Parhami and D.-M. Kwai. Periodically regular chordal rings. *IEEE Transactions on parallel and Distributed Systems*, 10(6):658–767, Jun. 1999.
- [20] J. H. Patel. Performance of processor-memory interconnections for multiprocessors. *IEEE. Trans. Comput.*, C-30(10):771–780, Oct. 1981.
- [21] I. D. Scherson and A. S. Youssef. *Interconnection networks for high-performance parallel computers*. IEEE computer society press, 1994.
- [22] T. Szymanski and V. Hamacher. On the permutation capability of multistage interconnection networks. *IEEE Trans. Comp.*, C-36(7):810–822, Jul. 1987.
- [23] T. Szymanski and V. Hamacher. On the universality of multistage interconnection networks. In *Interconnection Networks for High-Performance Parallel Computers*, pages 73–101. IEEE Computer Society Press, 1994.
- [24] W. Wulf and C. Bell. C.mmp- a multi-mini-processor. In *AFIPS Proc. Fall Joint Computer Conf.*, pages 765–777, 1972. NL.