

A Universal Performance Factor for Multi-Criteria Evaluation of Multistage Interconnection Networks

A. Chadi Aljundi ^{a,*}, Jean-Luc Dekeyser ^a, M-Tahar Kechadi ^b, Isaac D. Scherson ^c

^aLaboratoire d'Informatique Fondamentale de Lille, Université des Sciences et Technologies de Lille, 59650 Villeneuve d'Ascq, France

^bParallel Computational Research Group, University College Dublin, Belfield, Dublin 4, Ireland

^cDept. of Computer Science (Systems), School of Information and Computer Sciences, University of California, Irvine, Irvine, CA 92697, USA

Abstract

The choice of an interconnection network for a parallel computer depends on a large number of performance factors which are very often application dependent. We propose a performance evaluation and a comparison methodology. This methodology is applied to a recently introduced class of interconnection networks (Multistage Chordal Ring Based multistage interconnection network). These networks are compared to well known Omega network of comparable architectural characteristics. The methodology is expected to serve in the evaluation of the use of multistage interconnection networks as an intercommunication medium in today's Multiprocessor systems.

Key words: Interconnection Networks, Parallel Architectures, Delta Networks, Banyan Networks, MCRB networks.

1 Introduction

In a parallel computer, a medium must provide the means of communication between the different processing elements (PEs) themselves and/or with the system's memory modules. In fact, inter-connecting processors and linking them efficiently to the memory modules in a parallel computer is not an easy task. A trade-off has to be made between the hardware complexity of the network and the performance degradation caused by conflicts that might occur when more than one communication task occur simultaneously [20]. Interconnection networks [33] constitute a good communication medium for these parallel systems. Many studies were devoted to evaluate different interconnection network architectures. All these studies share

a common goal, mainly, how to design an interconnection network that provides the processors with the maximum bandwidth and fast access to data stored in memory.

Multistage Interconnection Networks (MINs) are usually used in multi-processor and multi-computer systems as an intercommunication medium between PEs and memory modules. A simple definition of MINs can be found in [35], where a MIN is a network generally used for interconnecting a set of N input terminals to M output terminals using sets of fixed-size switches arranged in stages.

Many MIN families were studied and effectively used to build parallel computers [11,1,34]. Delta MINs form a sub-group of a bigger MIN family called Banyan MINs. These MINs are characterized by the existence of one and only one path between each source and destination. Non-Banyan MINs are, in general, more expensive and

* Corresponding author.

Email address: aljundi@lifl.fr (A. Chadi Aljundi).

more complex to control. Still, they often are fault tolerant and capable to apply rerouting strategies used to bypass saturated/faulty links and conflicts.

Several MIN architectures were proposed in the literature [32,38] and a lot of work was devoted to study, evaluate and compare their performance [2,39]. Multi-criteria MIN performance evaluation is not easy to be established because of the large number of factors to be evaluated and their software and hardware dependencies [27]. Beside the application dependency of these factors, their interdependencies make the task really difficult.

The majority of the work done on the performance evaluation of MINs did not consider it as a multi-objective problem. A lot of these studies evaluated throughput, latency, or other performance metrics as separated factors.

In [13] Cheemalavagu and Malek studied the effect of the degree of Banyan MINs on their performance. The throughput and delay was measured using a packet switching strategy. After defining the throughput as the total number of packets received during a time interval, and the average delay as the average time taken by a packet to reach its destination, the authors used the throughput-delay ratio as a *figure of merit*. It is defined as the ratio of throughput to the average delay. The use of this ratio was explained by the importance of both metrics, and by the fact that it can give an *indication* of the average number of “*packets being output by the network*”. Simulation was used to measure throughput and delay.

While universality was not intended, such a metric lacks the possibility to evaluate an arbitrary number of performance metrics. As stated above, throughput and delay are not the only two important performance factors to be studied for the evaluation of MINs. On the other hand, the use of the ratio was not mathematically justified.

Another example of simulation based multi-objective approach of the performance evaluation of MINs is the work of Lakamraju *et al.* who presented an interconnection network synthesis method [24].

The goal of the authors was to be able to synthesize networks satisfying “*a set of desired properties*”. The work was applied on random regular networks, which later were successively passed through filters which discard networks that do not correspond to certain criteria. At the end of the procedure, a set of “*good*” networks of which one can be chosen. The important feature of this technique is that it can consider any number of desired performance factors and that it is useful “*when seeking to synthesize network that performs well with respect to multiple performance measures*” [SIC].

The technique is based on the specification of the performance measures of interest, the generation of a number of random regular networks which are then passed through a *bank* of filters. “*Each filter is associated with a performance requirement. The filters identify a subset of networks which have the desired performance with respect to the specified measures. This subset constitutes a short-list of networks from which the designer can choose*”.

The filter consists of two parts: an evaluation part that calculates the value of the associated performance metric, and the checking part that compares measured values with a certain threshold. The output of one filter, which is a subset of the input set of networks serves as the input of the next filter. “*Filters are arranged sequentially one after the other in decreasing priority order of the measures they represent*”.

This filtering technique is an interesting approach for multi-objective performance evaluation of interconnection networks. It falls short of a mathematical formalization. For example, the meaning of priority order is an essential point that has to be defined. On the other hand, different thresholds need different evaluation phases and the comparison does not give a global idea about the compared networks.

To date, huge efforts are still made in order to design better parallel computer communication networks. While these networks are in the best case a hybrid environment of crossbars and busses used, for example, in todays SMP (Symmetric MultiProcessors) and MPoC (Multi-Processors on-Chip)

machines, we believe that the use of MINs in such environments can be of a great benefit to their performance. Our performance methodology goals are twofold: firstly, it will help in choosing between the existing MINs, a MIN that performs better according to the architecture specifications and the design goals of the machine. Secondly, it allows performance evaluation and comparison of any new MIN.

Recently a MIN was introduced in [21] and was dubbed **Multistage Chordal Ring Based network (MCRB)**[4,6,5]. The underlying topology of its structure is a chordal ring, which is an enhanced, highly fault-tolerant, and scalable ring [17,31]. One of the main performance characteristics of the MCRB network, which is its throughput, was studied in a previous publication by the authors [4]. The evaluation and comparison methodology that we propose is validated on the MCRB and Omega networks.

The remainder of this paper is organized as follows: after introducing our evaluation methodology and giving all necessary definitions, we define the network examples that will be tested and compared. This is followed by some evaluation and comparison results before giving some concluding notes.

2 Evaluation Methodology

The main goal of proposed methodology is the definition of a systematic decision making mechanism for choosing suitable MINs for multiprocessor systems. This methodology is based on performance measures. Unlike the existing performance evaluation studies that concentrate on individual and particular MIN parameters, the methodology that we propose has the ability to consider a number of performance factors. A main feature of this methodology is that it permits the comparison of MINs with different architectural characteristics. Given a set of parameters (architectural, application dependence, performance factors, etc.), these parameters are mapped and projected onto a multi-dimensional space representing factors to be evaluated. For

example, the network integration complexity is considered as one dimension in the parameter space. Therefore, this will allow us to compare MINs of different sizes and/or different degrees. For clarity, we limit our study to four performance evaluation factors knowing that the proposed methodology is general and that it is easy to add other factors. These four metrics are chosen because they were (and are) very largely used in the literature. This is explained of course by their representativity and wide coverage.

Complexity is a quantitative term related to the *cost*. The evaluation of a system, whether it is hardware or software or both, needs a full study of the cost to be paid to design and implement it. Thus, the cost must be calculated in space and time terms.

2.1 Integration complexity

While studying a MIN, the first evaluation to do is its hardware complexity. The hardware complexity of a MIN can be calculated in two ways: the number of connection points and the number of links needed to construct the MIN. Liu [28] defines the hardware complexity of a MIN as the maximum of the two means. The hardware complexity of a MIN in term of cross-points is equal to the total number of cross-points of all crossbars used to build it. The complexity in terms of links is the sum of links in all stages.

When the number of input and output terminals is the same, N , we say that the MIN is of the size N . The degree of the MIN is defined as the size of crossbars used to build it [22].

Definition 1 Consider a MIN of size N and degree r , that has X stages of z SEs each. The stages are connected with Y inter-stages links. The integration complexity of the MIN is defined as $C = \max(r^2 X z, Y r)$.

Because the crossbar has the highest complexity, all single path MINs of complexity higher than crossbar are generally excluded. The crossbar network is non-blocking as it can effect any permutation

of the input set onto the output set in one cycle [12–14]. However, the construction of larger size crossbars is very expensive. This is why less complex inter-communication networks, such as MINs, are studied.

2.2 Temporal complexity

We adopt Flynn’s taxonomy of functional environments for parallel architectures [16] as it is the most widely accepted one. Basic performance criteria in SIMD environments are different of those in MIMD ones [7]. When studying routing capacity of MINs, the throughput is the important performance factor in MIMD environments, while in SIMD environments, the important criterion is the network’s permutation capability. The proposed methodology allows to compare these two architectures with different performance criteria (e.g. throughput and permutation capability). In addition to the most important temporal performance factors which are throughput and permutation capability, the network latency plays also a very important role as it is a measure of input patterns which exhibit a high degree of switching setting conflicts when routed in the network. In the following we focus on these three parameters.

2.2.1 Throughput

Analytically, the throughput is defined as the number of messages delivered to their destinations per unit of time [29,32]. Many analytical studies of MIN’s throughput can be found in the literature [32,22,35]. Simulations are used frequently when more realistic results are needed and accurate analytical models are not available. They allow more flexibility in network characterization to analyze real-world and popular communication patterns. Practically, the throughput is calculated as the number of messages delivered to their destinations over a certain number of trials. In the case of unbuffered networks, when a conflict occurs between two or more messages, only one message goes through and the others are discarded.

Definition 2 *In an unbuffered MIN, the throughput is defined 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 to the same SE I/O. All other messages are discarded.*

2.2.2 Permutation capability

Permutation capability [3] refers to the MIN’s ability to route message-permutations. A message-permutation is a group of messages in which destinations are permutations of all inputs. In fact, scientific calculations for which parallel machines are used, need in general fast access to special data structures. They usually need to implement skewing schemes [37] which give processors the ability of accessing sub-structures of interest such as lines, columns, and diagonals in conflict-free way [25]. In other words, the request destinations generated to access a sub-structure of interest are permutations of all or a sub-set of the global data set [35].

The number of routable permutations on MINs is equal to $N!$, where N is the network size. In order to study the permutation capacity of a MIN, one might try to route a certain number of random permutations on the network and calculate the number of cycles needed to route all permutation messages to their destinations. Analytical studies [26,8] proved that routing random permutations is not efficient for the permutation capacity study. Therefore, frequently used permutations [26] must be used for such analysis. To establish this comparative study, a certain number of these permutations can be routed under the assumption that the accumulated number of permutations routed per cycle can be a comparison factor.

Definition 3 *The permutation capability of a MIN is the number of permutation messages that can be routed to their destinations in a certain number of interconnection cycles.*

To calculate the permutation capability of a MIN we simulate routing permu-

tations of one family of frequently used permutations, which is the BPC (Bit Permute/Complement) family.

Definition 4 *A BPC permutation [7] of a set of integers is another set of integers whose binary representations are permutations of the binary representations of the first set. Some bits of each representation of the result might then be 1's complemented.*

2.2.3 Network's Latency

Another important performance parameter is the network latency. 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 BPC simulation to measure the MIN latency and the simulation tool and approach are discussed in [4].

Definition 5 *The Latency of a MIN is defined as 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 [36].*

2.3 Universal Performance Factor

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

In order to define the UPF, suppose that the factors to be evaluated as well as their importance are a part of the designing process (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 beginning of this study of the universal factor will concentrate on factors to be

minimized only. The definition will be generalized later for the case where the two types of factors are involved. For a group p^{min} of performance metrics, the universal performance factor is defined by the Euclidean distance of the performance projection value into an n dimensions vector space, where dimensions represent different performance factors.

Definition 6 *Given a MIN and a groups of performance factors p^{min} . The universal performance factor (UPF) can be broadly defined by:*

$$UPF = \sqrt{\sum_{i=1}^l (p_i^{min})^2} \quad (1)$$

Definition 7 *Given two networks μ_1 and μ_2 and their UPFs, UPF_1 and UPF_2 respectively. We say that μ_1 is more powerful than μ_2 if $UPF_1 < UPF_2$.*

By this definition, the MIN multi-criteria performance evaluation and comparison is transformed into the evaluation of a unique function, for which the choosing parameter is the distance between the value of the UPF and an ideal (non-realistic) network, for which the UPF value is equal to 0. 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 1 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 1 one can notice that μ_1 is more powerful 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.

Usually different parameters have different types of measures and scaling. In order to solve this problem, values can be normalized to a certain value, this may be the average value, or the maximum value for each factor. The normalization procedure using the maximum value gives the possibility to

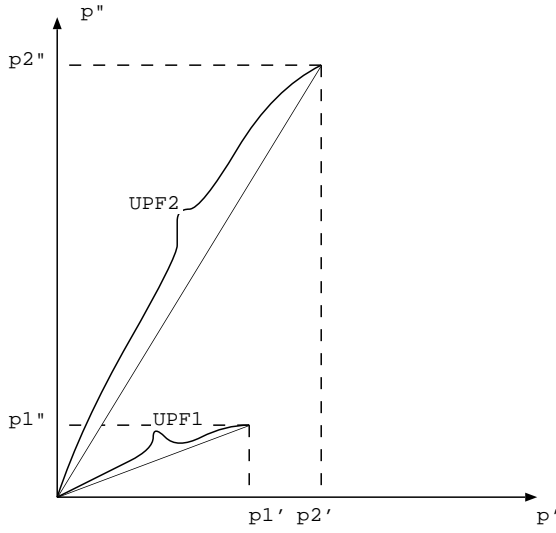


Fig. 1. An example of the use of the UPF factor.

give all factors the same importance, and thus be able to compare different scaling metrics. The equation 1 can be replaced by the following equation:

$$\text{UPF} = \sqrt{\sum_{i=1}^l \left(\frac{p_i^{\min}}{\text{MAX}(p_i^{\min})} \right)^2} \quad (2)$$

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 equation.

$$\text{UPF} = \sqrt{\sum_{i=1}^l \left(\frac{p_i^{\min}}{\text{MAX}(p_i^{\min})} \right)^2 w_i} \quad (3)$$

Now, the UPF formula (equation 3) will be generalized to the case when both factors to be maximized and factors to be minimized are to be evaluated simultaneously. To introduce this aspect it is suitable to note that normalizing the performance factors using the maximum value permits to consider the maximizing of a factor equivalent to the minimizing of $1 - \frac{p^{\max}}{\text{MAX}(p^{\max})}$. This leads to the following final formula for

the UPF:

$$\text{UPF} = \sqrt{\sum_{i=1}^l \left(\frac{p_i^{\min}}{\text{MAX}(p_i^{\min})} \right)^2 w_i + \sum_{j=1}^k \left(1 - \frac{p_j^{\max}}{\text{MAX}(p_j^{\max})} \right)^2 w_j} \quad (4)$$

As in any multi-criteria decision making problem, the importance of each criterion is a design problem. For the sake of this paper, all performance factors are considered to be of equal importance and thus all weights are equal to one.

Two conditions have to be considered to use the UPF as a performance factor. First, it is supposed that $p^{\max} \neq 0$. The second condition is the inter-independence of the measured factors.

3 Case Studies

In order to test the methodology that we propose, we will apply it on the MCRB network introduced recently and the well known Omega network, a subclass of Delta networks. In the following, definitions and some characteristics of the Delta class are presented. This is followed by a brief description of the MCRB network.

3.1 Omega Network

Before describing the Omega network, let us give a quick reminder of the delta network.

Definition 8 A *Banyan*[18] MIN is a MIN having the property of the existence of one and only one path between each source and destination. We call this attribute the *Banyan property*.

Banyan MINs may have the *delta property* or not. Delta networks, proposed by Patel [32], are built on ab crossbars. Let $o_{i,i=0,1,\dots,b-1}$ be an output of index i of a crossbar. If an input of a crossbar in stage j is connected to an output o_i of another crossbar in stage $j-1$, then all its other inputs must be connected to outputs of the same index i of crossbars in the previous

stage. We propose the following mathematical translation of the delta property.

Definition 9 For a Banyan $MIN(N,r)$ 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 switching elements (SEs) in the network have the same indexes then digits d_0 of all inputs of a switch must be equal. This characteristic is called Delta property.

The SEs in delta networks are digit-controlled crossbars. This is done by including a control sequence in the message, called *message control sequence*. The control sequence is a series of digits allocated for each stage of the network. The digit indicates which output of the SE to be connected to the input. Therefore, the control sequence represents the path to be taken by the message through the MIN. The Delta network control sequence is a representation of the destination. In a Delta $MIN(N,r)$, the control sequences can be written in r -based representation of the destinations. Such a MIN is composed of $\frac{\log N}{\log r}$ stages. This means that crossbars in stage i are controlled by the i -th digit of the 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 [23].

Omega network, a member of Delta networks, first was defined by Lawrie [25]. An Omega of degree 2 is a special case of shuffle-exchange networks [19]. For an Omega of degree $r > 2$ one may use q -shuffle [32] property to link different stages of the network. Figure 2 shows an $\Omega(16,4)$ network.

3.2 The MCRB Network

Note that for complexity reasons, which will be discussed later, the topology of the MCRB network discussed here is a bit different from the one proposed in [21].

¹ $MIN(N,r)$ stands for a MIN of size N and degree r .

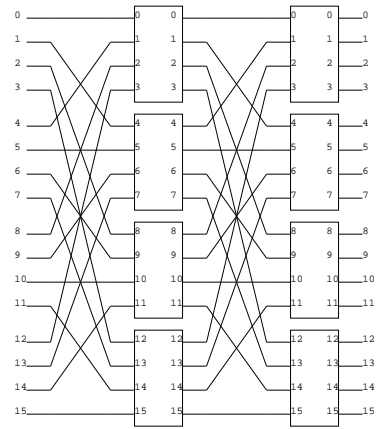


Fig. 2. A block diagram of $\Omega(16,4)$.

The MCRB network structure is a ring-based topology, referred to as *chordal rings*, also called *loop networks*. Chordal rings have been widely investigated in [9,10,40,31]. and they constitute a special class of circulant graphs[30]. A chordal ring is defined by a pair (V, C) , denoted $R(V, C)$, where V is the set of N nodes labeled from 0 to $N - 1$, and $C = \{c_1, c_2, \dots, c_n\} \subset \{2, \dots, N - 2\}$ a set of integers, called *chords*. Each chord c_i connects every pair of nodes that are at distance c_i in the ring. Because $R(V, C)$ is also a ring, where any vertex is connected to its direct neighbors, we define $c_0 = 1$. We are interested in the case where the distance between two nodes is a power of a certain integer number r , such that any chord c_k can be expressed as a power of r ; $c_k = r^i$.

Definition 10 A chordal ring-based network $CRB(N, r)$ is a chordal ring $R(V, C)$, where V is the set of N nodes labeled from 0 to $N - 1$, and C a set of chords $C = \{c_0, \dots, c_n\}$ having the property: $\forall c_i \in C, c_i = r^i$. The chord c_i is called *i -type dimension*.

The MCRB network is a multistage implementation of the CRB network.

Definition 11 Let N be the number of input/output nodes labeled from 0 to $N - 1$ in a MIN. A CRB-derived MIN, denoted by $MCRB(N, r)$, has n stages defined such that for any stage $S_i, 0 \leq i \leq n$, S_i implements all paths defined only along the i -type

A $CRB(N, r)$ is called *complete* if the total number of nodes can be expressed as a power of the radix r ; $N = r^n$. The $CRB(r^n, r)$ is node-symmetric network leading to r^{n-1} groups of r nodes each.

Proposition 1 Any network of type $CRB(r^n, r)$ has an $MCRB(r^n, r)$ implementation with n stages of r^n SEs each. Let SE_{ij} be the SE j of the stage i of the $MCRB(r^n, r)$, then SE_{ij} is connected to SE_{i-1, k_d} such that $k_d = (j + d r^i) \bmod N$, for $0 \leq i \leq N - 1, 1 \leq j \leq r - 1$, and $0 \leq d \leq r - 1$.

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

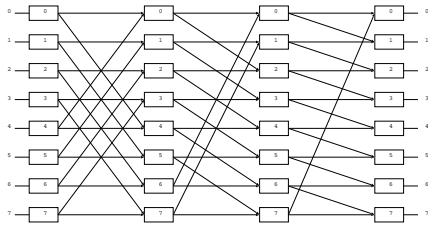


Fig. 3. $MCRB(8, 2)$ Network.

Following the definitions of the MCRB network and the Data manipulator, studied in [15], this later can be defined as the superposition of two MCRB networks. However as we focus on this paper on the study of the performance evaluation of the MCRB compared to the Omega network, a detailed study of the architectural features of the MCRB network is not the main focus of this paper.

4 Performance Evaluation Results

In this section we evaluate and compare the two MINs, which are the MCRB and the Omega networks for all the performance factors defined above. Note that a complete comparative study of the throughput of these two networks is presented in [4]. This will allow us to validate our evaluation methodology based on the universal performance factor.

4.1 Integration Complexity

It is easy to verify that the inter-stage complexity is smaller than cross-points complexity for both MCRB and Omega networks. Therefore, the study of the integration complexity is limited to the cross-points one. First of all, we calculate the complexity in terms of cross-points and compare them to the crossbar complexity. This step is important to eliminate networks of complexity greater than the crossbar of the same size.

The MCRB network is composed of $\log_r(N)$ inter-stages and $\log_r(N) + 1$ stage -switches. The two end-points stages are composed of N multiplexers/demultiplexers. Each multiplexer and demultiplexer has r cross-points. Thus the complexity of the input stage and the output stage is equal to $2Nr$. Each of the remaining $\log_r(N) - 1$ stages is composed of N crossbars of degree r^2 , which gives a total complexity of Nr^2 for each stage. Thus, the MCRB network complexity is

$$C_{MCRB} = 2Nr + \left[\frac{\log(N)}{\log(r)} - 1 \right] Nr^2 \quad (5)$$

In this paper, the interesting parameter values of $MCRB(N, r)$ are those that lead to a complexity lower than N^2 , which is the complexity of a crossbar of size N . Figure 4 shows that MCRB complexity is, for some values of N and r , greater than this of the crossbar. For $r \geq 3$, the interesting values of r for which the complexity of the MCRB network is lower than N^2 are $r = \frac{\log(N)}{2}$.

Figures 5 and 6 show different projections of Figure 4. We observe that for $r = 2$, $MCRB(4, 2)$ and $MCRB(8, 2)$ are more complex than the crossbar. When $N = 16$ complexity values of MCRB network and the crossbar are equal. Therefore, before implementing an $MCRB(N, r)$, one should make sure, from the equation 5, that its complexity is lower than N^2 .

We call the network defined in [21] Augmented MCRB (AMCRB). Augmented

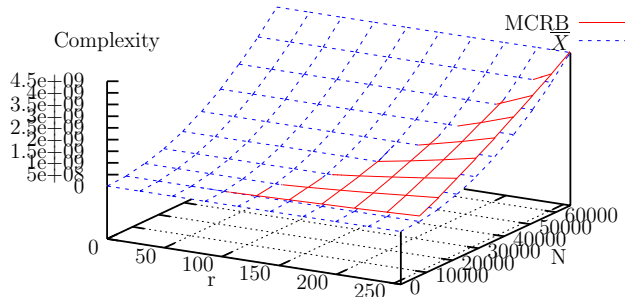


Fig. 4. Crossbar and MCRB network complexity as a function of their size N and the degree r (only MCRB). \bar{X} stands for crossbar

networks were defined in [22] to improve Banyan networks performance without great loss of simplicity. One way of augmenting a MIN is the super-position of two equivalent MINs. One possible equivalent network that can be superposed to the original architecture is the complement network.

The complexity difference between AMCRB and MCRB MINs can be calculated by observing that the multiplexers of the first stage are of size $1 \times 2r$, the demultiplexers of the last stage are of size $2r \times 1$ and finally, the SEs of the other stages are of size $2r \times 2r = 4r^2$. The complexity of the AMCRB is calculated using the following equation.

$$C_{AMCRB} = 4Nr + 4 \left(\frac{\log(N)}{\log(r)} - 1 \right) Nr^2 \quad (6)$$

As in the case of the MCRB network, the crossbar forms a complexity upper limit, and networks having complexity higher than this limit are not particularly considered in this paper. Thus, interesting characteristics of AMCRB networks are $N \geq 128$ for $r = 2$ and $N \geq 256$ for $r = 4$. On the other hand, AMCRB network's complexity exceeds the complexity of a same degree crossbar for all values of $r \geq 8$.

All crossbars in square Omega networks, which are considered in this paper, are of size r^2 and are distributed on $\frac{\log(N)}{\log(r)}$ stages

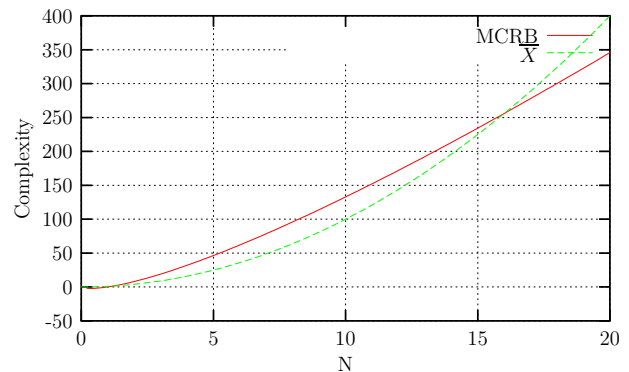


Fig. 5. Crossbar's and MCRB network's complexity as a function of the network's size and for $r=2$. \bar{X} stands for crossbar

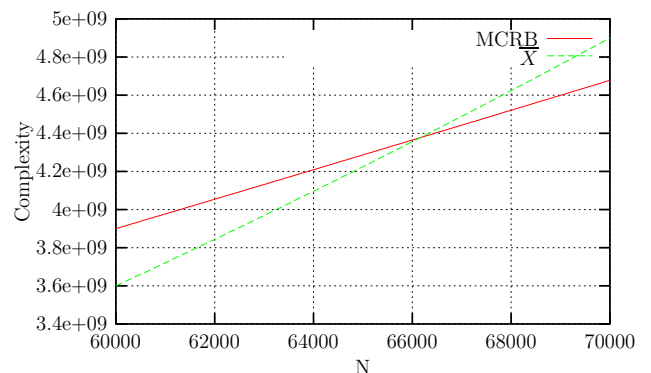


Fig. 6. Crossbar's and MCRB network's complexity as a function of the network's size and for $r=256$. \bar{X} stands for crossbar

each containing N/r SEs, which gives the Omega network's complexity:

$$C_{\Omega} = rN \frac{\log(N)}{\log(r)} \quad (7)$$

Note that from equation 7 one can notice that the Omega network complexity is always less than N^2 . The order of complexity between two MINs is defined as the ratio of their complexities, as both of them will be normalized to the complexity of the crossbar of the same size.

$$\Delta_C = \frac{C_{MCRB}}{C_{\bar{X}}} \quad (8)$$

By substituting the values of C_{MCRB} and C_{Ω} from equations 5 and 7 respectively, one

can write

$$\Delta_C = \frac{\log r}{\log N}(2 - r) + r \quad (9)$$

Note that for a certain value of $r > 2$, the term $\frac{\log r}{\log N}(2 - r)$ is negative and smaller than r . Thus, when N gets bigger, Δ_C gets bigger too and so MCRB network complexity is always greater than this of Omega. The question is, then, how to justify the increase in complexity. In other words, will the network's performance be better than this of the less complex network?

4.2 Temporal Complexity

To test the temporal aspects of the MINs that we are analysing, the same simulator explained in [4] is used to route BPC permutations. In order to calculate the number of cycles needed to route or reroute a certain number of permutations, circuit switching strategy is used. Thus, when a message is detected at input buffer of an SE of the first stage, a routing path is reserved if all the SEs describing it are free. A conflict situation occurs when a message tries to use a buffer already occupied by another message. In this case, the message stays in the input buffer of the first stage, but it will be erased from all other buffers previously allocated. When all routable messages arrive to their destinations, the simulator starts another attempt to reroute non-routable messages of the previous cycle. This procedure will be repeated until all *input* messages arrive to their destinations.

4.2.1 Permutation Capability of MCRB and Omega Networks

Figures 7, 8, and 9 show some examples of the permutation capabilities of Omega networks and MCRB networks with different values of r . These figures present the percentage of permutations that can be routed *within* a certain number of cycles. This means that for a cycle i , the presented value is the percentage of permutations that could be routed in all cycles

$i, i - 1, i - 2, \dots, 1$.

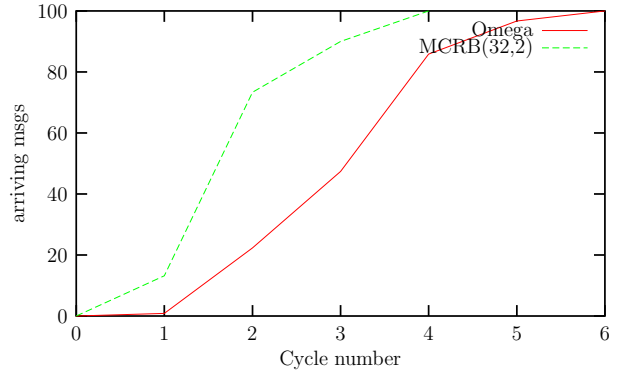


Fig. 7. Permutation capacities of MCRB and Omega networks of size 32.

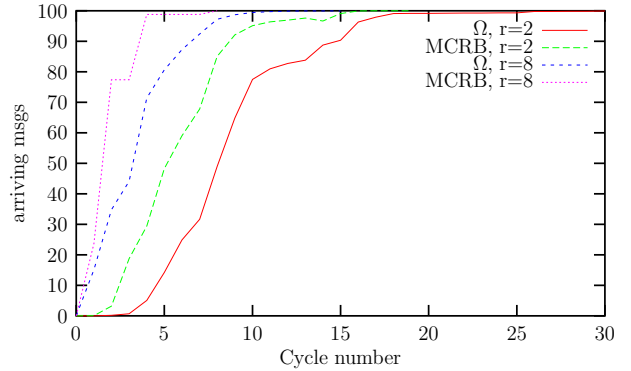


Fig. 8. Permutation capacities of MCRB and Omega networks of size 512.

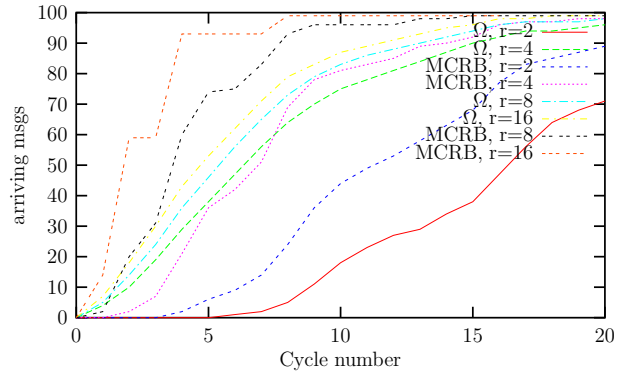


Fig. 9. Permutation capacities of MCRB and Omega networks of size 4096.

These figures show that there is a considerable gain in the permutation capability of the MCRB network relatively to Omega network. Although the MCRB network is very powerful in terms of permutation capability, which is expected, it still has higher integration complexity. A detailed

discussion about the relation between the improvement of MCRB network's performance and its complexity is presented in the following.

4.2.2 Comparing MCRB and Omega Networks' Latencies

The same conditions of simulations presented on the previous section were used to obtain results on the network latency. Figure 10 shows that for $r = 2$ the maximum number of cycles needed to route permutations in an MCRB is considerably less than Omega network. The gap increases with the network size. This is good for the MCRB, specially that for $r = 2$ it is only two times more complex than Omega (see equation 9).

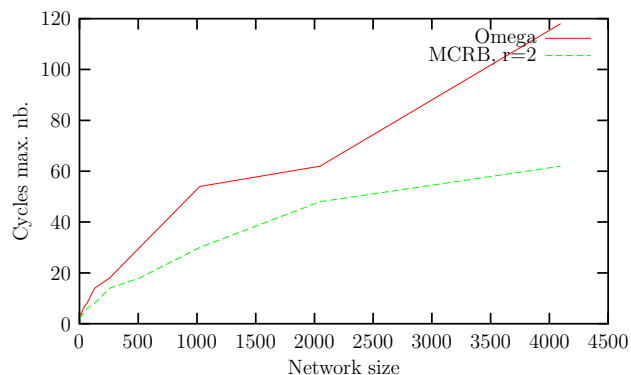


Fig. 10. The maximum number of cycles needed to route a certain number of BPC permutation on MCRB and Omega networks. The y axis is the maximum number of cycles needed to route tested permutations

In the next section we will gather all calculation and simulation results in order to give an example result of our comparison methodology between MCRB and Omega networks. Note that every composition of factors is not always available. For example, the use of the complexity with the latency does not represent an acceptable test as MINs of small size are not complex and do not need a lot of cycles to route simple permutations.

Two and three dimensions evaluations are to be presented in this section. Networks of different sizes and degrees are to be tested. Note that not every composition of factors is available to be used to compare MINs.

By example, the use of the complexity with the universality does not represent an acceptable test as MINs of small size are not complex and do not need a lot of cycles to rout the simple permutations that they can do.

4.3 Complexity and Throughput UPF

Comparing the UPFs of several MINs regarding only the complexity and the throughput means that the universality of the networks is not an important factor to be evaluated. On the other hand, the betterness of the network is judged by a small complexity and a high throughput. Figure 11 shows the normalized values of throughput and complexity of considered MINs.

Considering the comparison as a distance function, one can note that MCRB(256,4) is the best among the tested MINs and MCRB(1024,4) and MCRB(512,8) have high UPFs and thus they are considered to be unacceptable networks. The figure shows also the penalizing reason for these two network, which is their complexities. In fact, these two MINs are among the most complex networks of the studied networks sample.

One attracting point considering the application of the UPF as a distance function for the comparison of MINs performance is the possibility to compare different MINs of different architectural characteristics. One can note on figure 11 that using this logic MCRB(256,4), MCRB(256,2), MCRB(512,2) are better than Omega(256,2) and Omega(256,4).

4.4 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 universality AND throughput at the same time regardless of the complexity. Figure 12 shows the normalized values of throughput and universality of considered MINs.

The best way to study this figure and to

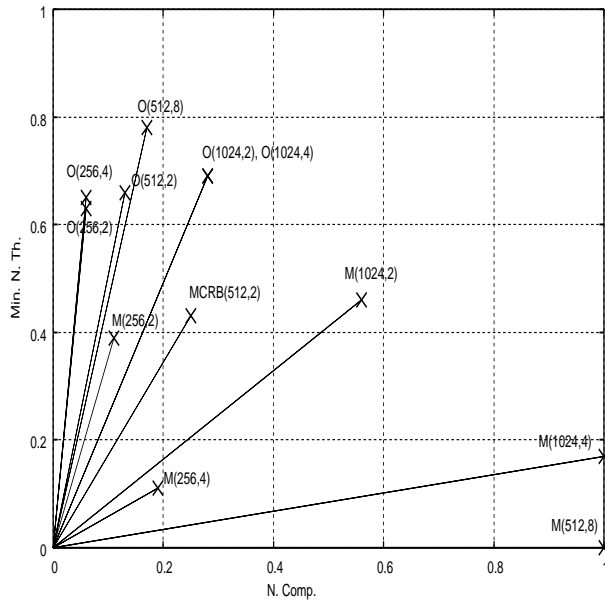


Fig. 11. Comparing some MCRB and Omega MINs regarding their throughputs and complexities

understand the use of the UPF to evaluate and compare MINs performance would be the comparison of figures 11 and 12.

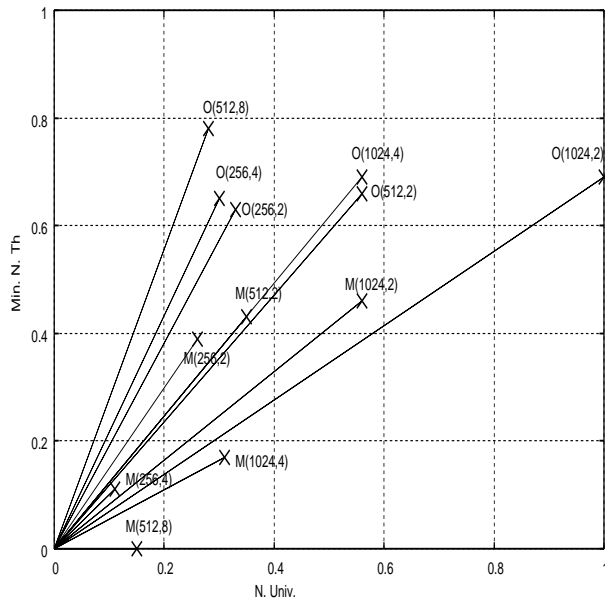


Fig. 12. Comparing some MCRB and Omega MINs regarding their throughputs and universality

It is expected that MCRB networks show a very good performance when complexity is not considered. This claim is clearly presented on figure 12. All Omega networks are less powerful than all MCRB-MINs even

with size difference between the two families networks.

MCRB(512,8) (which has one of the biggest complexities among studied MINs) becomes one of the most powerful networks for the universality-throughput UPF case. On the other hand, considering distance, both MCRB(512,8) and MCRB(256,4) are equal with respect to the universality-throughput UPF.

4.5 3-dimensions evaluations

Here all the three metrics previously studied either separately or on two dimensions are to be analyzed together. Figure 13 plots the values of considered MIN UPFs.

The analysis of this figure will be given here also as a comparison of some of the networks with their performance given on figures 11 and 12. The analysis of the figure shows that MCRB(256,4) has the smallest UPF, and thus can be considered as the best among the studied MINs. This result has been already noted with the previous evaluations.

Because of their high complexities, the MCRB(512,8) and MCRB(1024,4) which seemed to be very powerful in the case of the universality-throughput UPF is among the less powerful MINs for this inclusive UPF case. Note that the MCRB(1024,4) was not as bad as it seems in this case for the case of the complexity-throughput UPF case.

5 Conclusion

In this paper a multi-criteria evaluation and comparison methodology for MINs was presented. This methodology is based on different performance measures. Its main feature is to be able to incorporate many performance factors at the same time. The methodology was applied on a recently introduced MIN, the MCRB network, which was compared to a well known MIN, Omega, on a certain number of most desirable performance parameters.

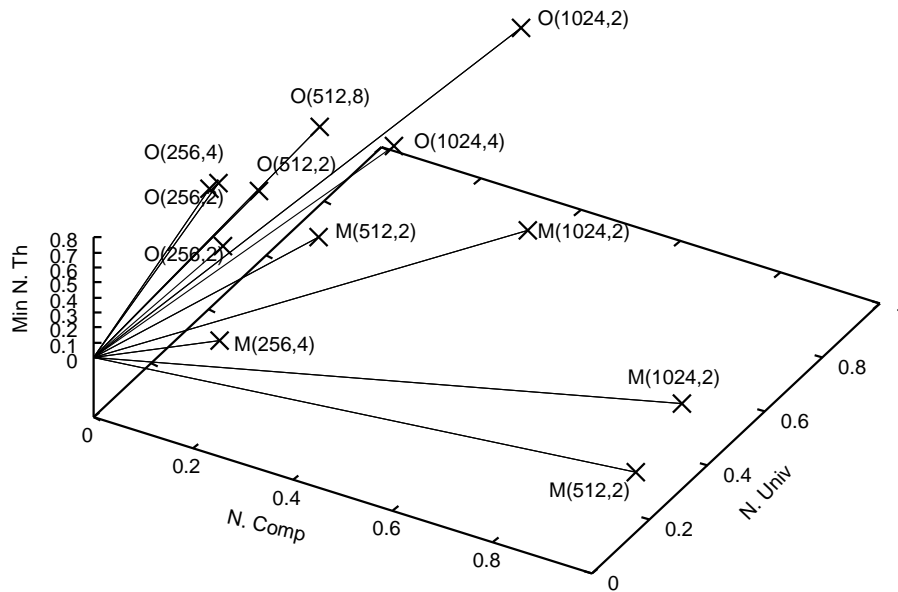


Fig. 13. 3-dimensions representation of the UPF

The proposed methodology is based on a distance function, that we call UPF. Many practical considerations were taken into account in the definition of the function. One interesting issue included in the definition is the flexibility. While only a certain number of performance metrics were studied in this paper, the distance proposed function is able to deal with other performance metrics that might be essential for special cases communication system design.

On the other hand, the association of each metric with a certain weight in the calculation of the UPF enables the treatment of a non uniform importance distribution of the studied metrics. In other words, the weights are used when some metrics are more important than others.

Results obtained from the application of this methodology, considering as well the above mentioned points, are very useful in the evaluation of modern networks for parallel computers and to study some so far difficult issues, such as the scalability of interconnection networks as well as the incorporation of standard benchmarks.

References

- [1] Ksr1 technical summary. Technical report, Kendall Square Research Corp., 1992.
- [2] G.B. Adams III and H.J. Siegel. On the number of permutations performable by the augmented data manipulator network. *IEEE Trans. on computers*, C-32(4):270–277, Apr. 1982.
- [3] D. P. Agrawal. Graph theoretical analysis and design of multistage interconnection networks. *IEEE. Trans. Comp.*, C-32(7):637–648, Jul. 1983.
- [4] A. Ch. 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.
- [5] A. Ch. Aljundi, J. L. Dekeyser, and I. D. Scherson. An interconnection networks comparative performance evaluation methodology: Delta and oversized delta networks. In *Proc. of the 16th International Conference on Parallel and Distributed Computing Systems*, pages 1–8, Reno NV., USA, Aug. 2003.
- [6] A. Chadi Aljundi, Jean-Luc Dekeyser, M-Tahar Kechadi, and Isaac 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.
- [7] B. D. Alleyne. *Methodologies for Analysis and Design of Data Routers in Large SIMD Computers*. PhD thesis, Princeton Univ., June 1994.
- [8] B.D. Alleyne and I.D. Scherson. On evil twin networks and the value of limited randomized routing. *IEEE Trans. on Parallel and Distributed Systems*, 11(9), Sep. 2000.
- [9] B.W. Arden and H. Lee. Analysis of chordal ring networks. *IEEE Transactions on Computers*, 30(4):291–295, April 1981.
- [10] B.W. Arden and K.-T. Tang. Routing for generalized chordal rings. In *Proc. ACM 18th Computer Science Conf.*, pages 271–275, February 1990.
- [11] L.N. Bhuyan, R.R. Iyer, T. Askar, A.K. Nanda, and M. Kumar. Performance of multistage bus network for a distributed shared memory multiprocessor. *IEEE Transactions on Parallel and Distributed Systems*, 8(1):82–95, Jan. 1997.
- [12] D. G. Cantor. On non-blocking switching networks. *Networks*, 1(4):367–377, 1971.
- [13] 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.
- [14] C. Clos. A study of non-blocking switching networks. *Bell system tech. journal*, 32(2):406–424, Mar. 1953.
- [15] T.Y. Feng. Data manipulation functions in parallel processors and their implementations. *IEEE Transactions on computers*, C-23(3):309–318, Mar. 1974.
- [16] M.J. Flynn. Some computer organizations and their effectiveness. *IEEE Trans. Comput.*, C-21(9):948–960, Sep. 1972.
- [17] S.A. Ghozati. Multidimensional token ring networks: Routing and operation. *Journal of Computers and Electrical Engineering*, 23(3):151–164, 1997.
- [18] G.R. Goke and G.J. Lipovski. Banyan networks for partitioning multiprocessor systems. In *Proc. 1st Annu. Symp. Comput. Arch.*, pages 21–28, 1973.
- [19] K. Hwang and F.A. Briggs. *Computer Architecture and Parallel Processing, 5th printing*. McGraw-Hill series in computer organization and architecture. McGraw-Hill International Editions, 1989.
- [20] M. T. Kechadi. *Un Modele de Fonctionnement Desordonne Pour les Systemes Multiprocesseurs Pipelines Vectoriels a Mmoire partagees (Definition, Modelisation et Proposition d'Architecture)*. PhD thesis, Universite des Sciences et Technologie de Lille, Laboratoire d'Informatique Fondamental de Lille, Mar. 1993.
- [21] M. T. Kechadi. Mrcb: 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.
- [22] C. P. Kruskal and M. Snir. The performance of multistage interconnection networks for mutliprocessors. *IEEE Trans. Comput.*, C-32(12):1091–1098, Dec. 1983.
- [23] C.P. Kruskal and M. Snir. A unified theory of interconnection network structure. *Theoretical Computer Science*, 48:75–94, 1986.
- [24] V. Lakamraju, I. Koren, and C.M. Krishna. A synthesis of interconnection networks: a novel approach. In *Proc. of the Int'l. Conf. on dependable systems and networks (DSN2000)*, pages 501–509, NY. USA, June 2000.
- [25] D. A. Lawrie. Access and alignment of data in an array processor. *IEEE Trans. Comput.*, C-24(12):1145–1155, Dec. 1975.
- [26] J Lenfant. Parallel permutations of data: A benes network control algorithm for frequently used permutations. *IEEE Trans. Comp.*, C-27:637–647, July 1978.
- [27] K.J. Liszka, J.K. Antonio, and H.J. Siegle. Problems with comparing interconnection networks, is an alligator better than an armadillo? *IEEE Concurrency*, 5(4):18–28, October-December 1997.
- [28] Y.-S. Liu. *Architecture and performance of processor-memory interconnection networks for MIMD shared memory parallel processing systems*. PhD thesis, New York University, 1990.

- [29] A. Merchant. *Analytical Models for the Performance Analysis of Banyan Networks*. PhD thesis, Stanford University, 1991.
- [30] K. Mukhopadhyaya and B. Sinha. Fault-tolerant routing in distributed loop networks. *IEEE Transactions on Computers*, 44(12):1452–1456, December 1995.
- [31] B. Parhami and D.-M. Kwai. Periodically regular chordal rings. *IEEE Transactions on parallel and Distributed Systems*, 10(6):658–767, Jun. 1999.
- [32] J. H. Patel. Performance of processor-memory interconnections for multiprocessors. *IEEE. Trans. Comput.*, C-30(10):771–780, Oct. 1981.
- [33] I. D. Scherson and A. S. Youssef. *Interconnection networks for high-performance parallel computers*. IEEE computer society press, 1994.
- [34] C.B. Stunkel, D.G. Shea, D.G. Grice, P.H. Hochschild, and M. Tsao. The sp1 high-performance switch. In *Proc. 1994 Scalable High-Performance Computing Cong.*, pages 150–157, May 1994.
- [35] T.H. Szymanski and V.C. Hamacher. On the permutation capability of multistage interconnection networks. *IEEE Trans. Comp.*, C-36(7):810–822, Jul. 1987.
- [36] T.H. Szymanski and V.C. Hamacher. On the universality of multistage interconnection networks. In *Interconnection Networks for High-Performance Parallel Computers*, pages 73–101. IEEE Computer Society Press, 1994.
- [37] H.A.G. Wijshoff and J. Van Leeuwen. On linear skewing schemes and d-ordered vectors. *IEEE Trans. Comp.*, C-36(2):233–239, Feb. 1987.
- [38] C. Wu and T. Feng. On a class of multistage interconnection networks. *IEEE Trans. on Computers*, C-29(8):694–702, Aug. 1980.
- [39] Y. Yang and J. Wang. Optimal all-to-all personalized echange in self-routable multistage networks. *IEEE Trans. on Patallel and Distributed Systems*, 11(3):261–274, Mar. 2000.
- [40] G.W. Zimmerman and A.H. Esfahanian. Chordal rings as fault-tolerant loops. *Discrete Applied Mathematics*, 37/38:563–573, July 1992.