

Studying the effect of the design parameters on the interconnection network performance in NOWs

R. Martínez, J.L. Sánchez, F.J. Alfaro, V. Chirivella, J. Flich

Abstract— With the increasing use of Network of Workstations (NOWs) as an alternative to huge parallel computers it has become essential to design high-performance interconnection networks for the communication between the nodes of these clusters.

A large number of studies have been carried out to achieve this objective. Most of them propose a new technique that affects one of the parameters that characterize the interconnection network. These techniques are completely new or inspired in the techniques previously used in multiprocessor systems. The impact of the proposal is studied (in most cases using simulation as an evaluation technique), and an analysis is made of the effect of the new technique over the system performance versus those currently in existence. In this kind of study most of the network parameters are fixed and usually only a few parameters are varied.

This paper presents a more general study of the interconnection network performance. This study consists in showing the effect of different design parameters over the network performance, and the interaction between them. This study would not be viable with the traditional techniques due to the number of simulations required. The alternative of the Experimental Design is used to carry out the study.

Keywords— Interconnection network, performance evaluation, experimental design.

I. INTRODUCTION

NETWORKS of workstations are a cost-effective alternative to huge parallel computers when executing a great variety of applications that need a lot of computing power. Some reasons for this can be found in recent developments. Personal computers have a higher performance and a lower cost than a few years ago. More important, interconnection network technology has taken important steps forward, and is now able to provide high bandwidth and low latency [1], [2], [3].

This situation has allowed many applications, which due to their requirements were limited to very specific systems, to be executed over new platforms like clusters. Some of these applications are scientific computing, web servers, videoconferences, multimedia servers, collaborative environments, distributed data bases, etc [4].

Departamento de Informática, Escuela Politécnica Superior, Universidad de Castilla-La Mancha, 02071 - Albacete (España). {raulmm, jsanchez, falfaro}@info-ab.uclm.es

Departamento de Estadística e Investigación Operativa, Universidad Politécnica de Valencia, 46071 - Valencia (España). vchirive@eio.upv.es

Departamento de Informática de Sistemas y Computadores, Universidad Politécnica de Valencia, 46071 - Valencia (España). jflich@gap.upv.es

This work was partly supported by the Spanish CICYT under Grant TIC2003-08154-C06 and JCC de Castilla La-Mancha under Grant PBC-02-008.

One of the essential design components in clusters is the interconnection network. A large number of studies have centered on this, in an attempt to obtain a higher performance [5]. Demands on the diversity of applications running over this kind of platform force networks to provide very high performance. There are applications, most of them parallel ones, which need communications with low latency, and there are many others, of a distributed nature, that also need high bandwidth.

The fast advance in high-performance networks for NOWs is partially due to the reuse of research that has been carried out for years in multiprocessor networks. Many techniques and ideas used in multiprocessor environments have been transferred to NOWs [6]. Nevertheless, this transfer is not always immediate or even possible. On some occasions, the transfer can be made but at the expense of sacrificing some particular characteristics of the clusters (scalability, flexibility in the interconnection, expansibility, etc.). This is the reason for the important research that is being carried out in different aspects of interconnection in NOWs. These aspects include the interface between nodes and the interconnection network, aspects related to the switches, and the physical links themselves. In some works, how to transfer existing techniques to the new environment is studied. In others, a study is made of how to modify them to fit the new characteristics. Finally, in many cases new strategies are proposed to obtain better results.

To summarize, in the last few years a great deal of effort has been put into research to consolidate the NOWs as distributed and parallel process platforms of high performance, making it possible to use a high number of applications at a reduced cost.

Many of the works related to this research have similar characteristics in their development and general objectives. In these studies, a particular technique considering some concrete aspect of the interconnection network is proposed. The impact of the proposal in question is studied, analyzing the effect over the system performance versus existing techniques. Usually, simulation is chosen to obtain the results. These results will show the goodness of the new technique. The evaluation process consists basically in subjecting the network model incorporating the new proposals to different tests. This model is characterized by a number of parameters (workload, network size, buffers size, routing algorithm, etc.) and the performance analysis will consist in fixing most of them and varying those that are thought to

be more significant, especially those related to the technique that is being studied. Usually, the number of different cases of study is not large, and the number of values for each parameter is not too high. This way of approaching the performance evaluation may lead to results that are too tied to the fixed conditions, and does not allow us to estimate the possible interactions between parameters.

Taking this into account, it seems that it could be useful to make a more complete study with similar characteristics, but without centering it on some concrete parameters. It would be a study that would take into account a large number of parameters, and would study the effect of each one over the system performance and the interaction between them. This would allow us to obtain a global idea of the contribution of the different design alternatives to the interconnection network performance.

In such a study, the most appropriate way of approaching the problem consists in varying a few parameters by turn while fixing the rest. At first it may appear that all the possible combinations of the different values of the parameters must be tested to achieve the required information, namely the effect of each parameter over the system performance, and the interaction between them. In this paper, it will be shown that with a proper experimental design this is not necessary. The goal of experimental design is to obtain the maximum information with the minimum number of experiments.

The rest of this paper is organized as follows. Section II presents the experimental design technique. The approach to the performance evaluation is stated in Section III, and the obtained results are shown in Section IV. Finally, some conclusions and future work are given.

II. DESIGN AND ANALYSIS OF EXPERIMENTS

In general, an *experimental design* will provide a set of experiments (number of experiments and value of parameters) for performance evaluation. As mentioned earlier, the goal of a proper experimental design is to obtain the maximum information with the minimum number of experiments. A proper experimental design also makes possible to separate out the effects of various design parameters that might affect the performance of the system being studied.

Once the experiment has been carried out, the analysis of the results determines if a design parameter has a significant effect, or if the observed difference is simply due to random variations caused by measurement errors and parameters that were not controlled.

A. Design of experiments

There are numerous varieties of experimental designs. The most frequently used are full factorial designs and fractional designs [7]. Before explaining each of them, several terms should be introduced. The **response variable** is the outcome of the experiment, generally the variable that measures the per-

formance. Each design parameter that affects the response variable and has several alternatives is called a **factor**. The values that a factor can assume are called its **levels**.

A.1 Full factorial designs

A full factorial design explores every possible combination of all levels of factors. A performance study with k factors, with the i th factor having n_i levels, requires n experiments, where

$$n = \prod_{i=1}^k n_i \quad (1)$$

The advantage of a full factorial design is that the effect of each factor and its interactions can be found. However, the main problem is the cost of the study. It may take too much time (or money) to conduct all these experiments, especially when we take into account the possibility that each of them could be repeated several times. There are three ways to reduce the number of experiments: by reducing the number of levels for each factor; by reducing the number of factors; and by using fractional factorial designs.

The first alternative presents as its main disadvantage the fact that it is not able to study all the desired levels of a quantitative factor, or the nature (lineal, quadratic) of the relationship between the response variable and a given factor. This alternative may be recommendable in an initial study where the goal is only to obtain the relative importance of each factor. A full factorial design in which each of the k factors is used at two levels requires 2^k experiments. This is a very popular design and is called **2^k design**. After reducing substantially the list of factors more levels per factor can be tried.

The second alternative would need to restate the initial objectives of the problem resolution. This option may not suppose a substantial lose of information if the factors that are going to be left out have little effect on the response variable.

A.2 Fractional factorial designs

A fractional factorial design consists in a factorial design where an important number of experiments have been removed, at the cost of not being able to study some interactions of high order. For example, a full factorial design with 6 factors and 2 levels per factor would require 64 experiments (Equation 1), and it would allow us to estimate 6 simple effects¹, 15 double interactions², 20 triple, 15 quadruple, 6 quintuple and 1 sextuple, for a total of 63 effects. Many of these estimated effects, usually those of order higher than or equal to three, will be probably non-existent; this is because fractional factorial designs are appropriate in many cases.

¹A simple effect is the mean variability produced in the response variable when a single factor varies.

²Interaction between two factors exists (double interaction) if the effect of one of them changes when the other varies its level. This concept can be generalized for superior orders.

As stated, a fractional factorial design consists in choosing only certain combinations out of all the ones possible (a full factorial design). For example, a 2^{k-1} design would be a 2^k design where a half of the experiments have been removed. In general, these are called 2^{k-p} designs, allowing us to study k factors with 2^{k-p} experiments.

To simplify the task of designing high fractional designs without mastering the theoretical basics, Genini Taguchi proposed the generalized use of some basic designs that he grouped in a set of tables and graphics [8]. These designs are called *orthogonal arrays*.

What makes it possible to calculate various effects and their contribution to the total variation is the concept of orthogonality. Given two factors, FI with I levels and FJ with J levels, it can be stated that *the simple effects of both factors are orthogonal if in the tests of the design each of the variants i of FI appears in the same proportion as the J variants of FJ* . In a full factorial design the orthogonality is inherent, but a fractional factorial design must be carefully designed to maintain it.

B. Analysis of results

After the experiments have been designed and carried out, the useful information must be extracted from the results. To discover what effects are significant or not over the response variable the *Analysis of the Variance technique (ANOVA)* is used. When the importance of a specific factor has been determined, it is normal to use (according to the nature of the factor) the Least Significant Difference (LSD) intervals and the orthogonal contrasts to analyze in which way those factors affect the response variable.

B.1 ANOVA

The study of the significance of the different factors and their possible interactions is carried out by elaborating the ANOVA table. The basic idea consists in splitting up the total variability observed in the results into a number of terms. Each term is associated to a different considered effect (simple effects, double interactions, etc.), plus one additional term associated with those effects that have not been considered and the possible error due to randomness, which is called the residual term. This variability is represented in the ANOVA table as the sum of the squares of differences between means. In the ANOVA table and for each factor, the fraction between the variation explained by each effect and the residual term is shown. If this fraction is high enough (taking into account an F of Snedecor distribution) the associated term is considered significant.

B.2 LSD intervals and orthogonal contrasts

After determining the significance of an effect, the next step is to determine the nature of the relationship with the response variable. In the case of a qualitative factor, means can be compared, but by taking into account the LSD interval. The LSD interval is

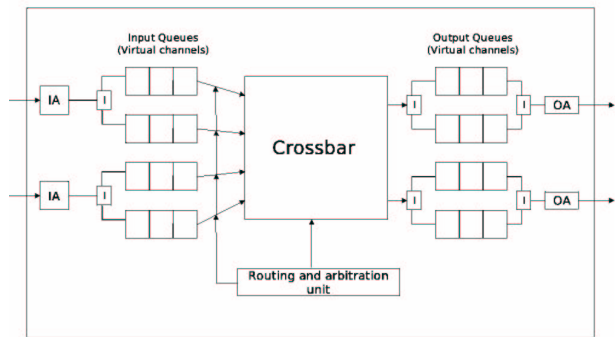


Fig. 1. Switch Model

the approximate interval where the real value of the mean can be found with a certain level of confidence. If the factor having a significant effect is quantitative, orthogonal contrasts must be used. This technique consists in studying separately the effect of the different components (lineal, quadratic, etc.). These effects are analyzed in a similar way as in ANOVA. If a factor has i levels, $i - 1$ components can be studied, but in general effects higher than quadratic are not significant.

III. A SYSTEMATIC APPROACH TO PERFORMANCE EVALUATION

To carry out this performance study, the methodology described in [7] has been used. This methodology suggests a certain number of steps that are fully developed in [9], but due to lack of space the process will be shown in a summarized form.

A. State goals and define the system

The goal of the study is to evaluate the interconnection network performance in a NOW according to different possible configurations. To simplify the study, the network will consist of only one switch attached to a certain number of hosts. The hosts will generate messages at a certain rate and messages will travel through the links to the switch, which will in turn send them to their destinations. The switch model used is typical in this kind of study.

A.1 Switch model

Figure 1 shows the switch model that we are going to use. In this model, there are input and output channels, an internal crossbar that provides connectivity between them, and a routing and arbitration unit that decides which output channel message should be sent through when it arrives at the switch. Physical channels may be split into several virtual channels.

B. List services and outcomes. Select metrics

The goal of a network interface in a NOW is to transfer messages among the nodes that form the cluster, but with the requisite of a high performance. In this work, the chosen metric has been the throughput. The throughput is the maximum traffic accepted by the network, and is usually expressed as

the number of received messages per node and time unit.

C. List parameters. Select factors to study

The factors that have been chosen among all the possible parameters to carry out the study are shown in the following list:

- The factors that refer to the network itself are:
 - Time of fly (TFLY): this is the time that message header takes to go from a host to a switch or vice versa.
 - Number of hosts (NHOST): this is the number of hosts connected to the switch.
- The factors that relate to the switch are:
 - Queue assignation policy (POLIC): this is the policy that decides which virtual channel corresponds with each message going into the switch.
 - Number of virtual channels (NVC): this is the number of storage queues associated with each port of the switch.
 - Crossbar organization (MUX): this indicates if the virtual channels are multiplexed or not. The crossbar can be multiplexed both at the input and at the output.
 - Routing table access (TABLE): this indicates the number of accesses that can be made to the routing table simultaneously. It ranges from one to infinite (parallel access) in the ideal case.
 - Routing time (TROUT): this is the time taken by the routing algorithm.
 - Arbitration time (TARB): this is the time taken by the arbitration algorithm.
 - Bytes per cross (FCROSS): this is the number of flits that can cross the crossbar for each connection in one cycle.
 - Size of queues (QSIZE): this is the size of the queues or virtual channels.
- The workload factors are:
 - Message size (MESSIZE): this is the number of flits of the message.
 - Destination distribution (DDEST): this indicates the policy used to generate the message destinations.

Table I shows the different levels for the factors. It also specifies if the factors are considered to be design or noise factors, and if they are quantitative or qualitative. The design factors are those that must be fixed when the network is physically made. The noise factors are those that affect the network performance but cannot be established in reality. The levels for DDEST are the uniform distribution, and the hot-spot distribution with one or two hot spots and 5%/10%/15% of messages addressed to hot spots. The levels for MUX are the four possible combinations for multiplexing the virtual channels. The value for the medium level of NVC depends on the number of hosts in the network (NHOST), two virtual channels when there are eight hosts, three when there are

sixteen and four when there are sixty-four.

D. Select evaluation technique. Select workload

Simulation has been chosen for performance evaluation because it will allow us to model a complex system with a large number of parameters to vary at a reasonable price.

Two message destination distributions have been used: uniform and hot-spots.

E. Design experiments

Taking into account the number of factors and levels considered in the study, and according to Equation 1, a full fractional design will lead to 1,469,664 simulations ($3 \times 7 \times 3 \times 4 \times 3 \times 3 \times 3 \times 4 \times 2 \times 3 \times 3 \times 3$) which is non-viable. So, the use of fractional factorial design is required.

The study has been carried out in two stages: the previous design and the final design. With the previous design the aim is to obtain the simple and double effects that are significant. In the final design only those effects worthy of interest will be studied in detail.

E.1 Initial design

At this stage, only the design factors have been considered. The noise factors and two levels per factor have been left out. The minimum and maximum levels of each factor are shown in Table I. The chosen design is a 2^{11-4} design, which allows us to study all the simple effects and double interactions between the considered factors. After making the 128 simulations the corresponding ANOVA table shows that only these effects are significant: TROUT, TARB, FCROSS, TABLE, POLIC, NVC, TFLY*POLIC³, TROUT*TARB, TROUT*TABLE, TROUT*POLIC, TROUT*NVC, TARB*FCROSS, TARB*TABLE, FCROSS*TABLE, TABLE*POLIC, TABLE*NVC, POLIC*NVC. The final design must include at least the factors that are significant by themselves, or those that are significant in a double interaction. The double interactions themselves must be included for further study.

E.2 Final design

The final design can be carried out using the information obtained from the initial design, but now taking all levels into consideration. This design will allow us to obtain the effect of the design and noise parameters over the network performance.

For this purpose it is usual to use one of the orthogonal arrays of Taguchi [8]. However, when the studied factors have a different number of levels, the selected design must be adapted. In this case, an orthogonal array has been used to accommodate those factors with three or less levels. The resulting design has been crossed with the remaining factors to obtain the final design.

The orthogonal array L_{81} is the smallest one that is able to accommodate the factors with three or less

³ A*B means a double interaction between A and B.

Factor	Levels	D/N	QN/QL
FCROSS	1, 2, 4	D	CN
DDEST	U, 1HS5%, 1HS10%, 1HS15%, 2HS5%, 2HS10%, 2HS15%	N	CL
LONGMEN	32, 164, 1024	N	CN
MUX	In&Out, In&-, -&Out, -&-	D	CL
NCV	1, 2/3/4, number of hosts	D	CN
NHOST	8, 16, 64	D	CN
POLIT	Aleatory, Most Empty Queue, VOQ	D	CL
TABLA	1, 2, 4, Parallel	D	CL
TAMCOLA	1, 4 messages	D	CN
TARB	1, 25, 50	D	CN
TFLY	8, 32, 64	D	CN
TROUT	1, 25, 50	D	CN

levels, and their interactions. Although the initial design showed that it was not necessary to include MUX and TFLY factors, they have been added without increasing the number of tests. This has been done because a quadratic relation could not be correctly observed on a two level analysis, and in some cases conclude that a significant factor is not significant.

The 81 combinations of the L_{81} have been crossed (in a full factorial way) with the four levels of TABLA, the three levels of MESSIZE and the seven levels of DDEST, for a total of 6804 ($81 \times 4 \times 3 \times 7$) simulations. This number could have been reduced through more fractional factorial design, but it allows us to study a large number of double interactions with a high precision, being simulated in a reasonable time.

IV. RESULTS

The execution of the experiments of the final designs produced a great deal of information. With the results of the 6804 simulations, the results of the combinations that have not been executed can be estimated. Furthermore, many effects and interactions have been extracted. Due to lack of space, only the more interesting results will be shown. The more interesting interactions are shown in Figure 2.

FCROSS*TABLE: With a parallel access when the number of bytes per cross increases, then the throughput also increases. This is because at a certain moment, with a parallel access to the routing table there are more packages ready to cross. This allows a larger number of packages to cross at the same time, taking advantage of a crossbar with a higher capacity.

DDEST*POLIC: VOQ decreases the negative effect of having a large number of messages whose target is the same hot spots. The problem with hot spots is the congestion of some target hosts and the HOL Blocking. This is because with VOQ and a single switch, the traffic to the different target hosts is completely isolated, and

HOL blocking is not a problem.

MESSIZE*TABLE: The longer the messages, the smaller the difference between the different options on the number of ports of the routing table. This is specially so in the case of having four and infinite ports. This is because at the same injection rate, the number of packages decreases, reducing the need to route a high number of messages at the same time.

NHOST*NVC: The higher the number of hosts (in this case, ports of the switch), the better the results with a higher number of virtual channels. With a high number of ports, the problem of HOL Blocking increases. With a high number of virtual channels this problem decreases. This is because having a high number of virtual channels is more important with a high number of ports.

TROUT*MESSIZE: The longer the messages, the smaller the significance of having a high routing time. This is because with a lower number of messages the routing must be performed a lower number of times.

TROUT*TABLE: When the routing time increases, the positive effect of increasing the number of ports in the routing table is higher. A high routing time exacerbates the problem of routing table having few ports. This is because more messages have to wait until a port is freed. More messages now arrive at the switch while the routing is in process, these messages also have a longer wait.

V. CONCLUSIONS AND FUTURE WORK

In this paper, the high dependency between the different design parameters in a switch has been shown. This means that the effect of one parameter over the performance of the switch depends on the value of some other parameters. This should be taken into account when carrying out a study of the effect of some concrete design parameter.

With the appropriate design not all the possible combinations of the parameter values require test-

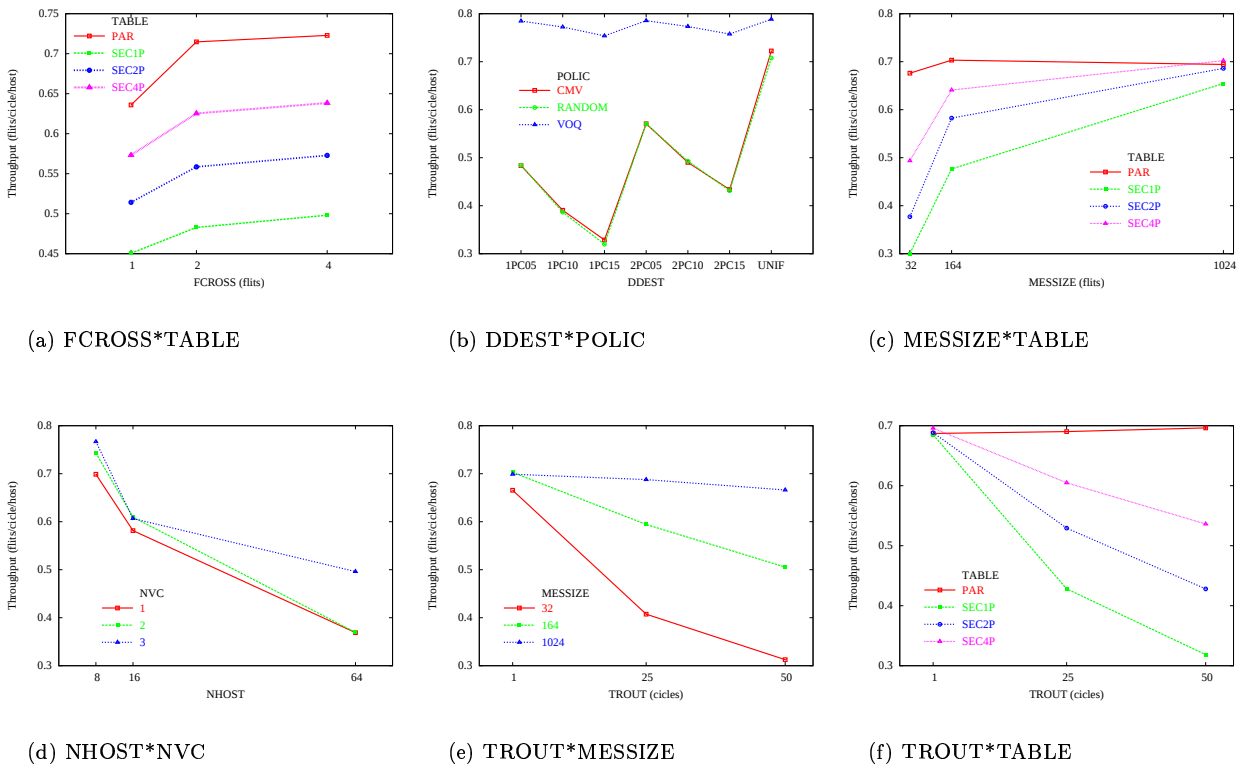


Fig. 2. Some double interactions between factors

ing. With a small number of simulations a high amount of information can be obtained. In this case, it would be necessary to simulate more than one million cases to consider all the possible combinations. With the technique used only 6804 simulations have been needed for the final design. Another relevant aspect is that it is not necessary to consider a high amount of different values for each parameter. For the quantitative parameters only three values are sufficient to study quadratic tendencies. This points to the conclusion that a good experimental design is necessary in order to obtain the maximum information with the minimum number of experiments.

The first conclusion is that a study of this kind can produce a very large amount of information that requires further study to obtain concrete results.

As regards to the specific results obtained, it must be said that some of them were totally in line with expectatives both logical reasons and in the light of previous knowledge. However, other results are more difficult to explain and would require additional studies in order to obtain more information.

The methodology used and the characteristics of the results obtained pave the way for further studies. These studies could be more specific and more contained the present study or more ambitious. Thus, some possible future work could focus on the study of some of the obtained results. One of these studies could be centered on the TABLA factor. The goal of this study could be to determine how many ports would be necessary at the routing table to obtain the

advantages of parallel access. Other studies could be variations of this one, as for example, the study of a different set of factors or an evaluation with other performance metrics. Our study has been carried out with a single switch. In this sense, an interesting study could be to focus on a network with multiple switches. This would also allow us to consider other design parameters, as for example the routing algorithm or the topology.

REFERENCES

- [1] N.J. Boden, D. Cohen, and R.E. Felderman, "Myrinet – A Gigabit per Second Local Area Network," *IEEE Micro*, pp. 29–36, Feb. 1995.
- [2] "Gigabit-ethernet overview," Gigabit-Ethernet Alliance, 1997, <http://www.gigabit-ethernet.org>.
- [3] InfiniBand Trade Association, *InfiniBand Architecture Specification Volume 1. Release 1.0*, Oct. 2000.
- [4] A.C. Arpacı-Dusseu, et al., "High performance sorting on networks of workstations," *Proceedings of the ACD SIGMOD*, May 1997.
- [5] R. Buyya, *High performance cluster computing*, Prentice Hall, 1999.
- [6] J. Duato, S. Yalamanchili, and L. Ni, *Interconnection networks. An engineering approach*, IEEE Computer Society, 1997.
- [7] R. Jain, *The art of computer systems performance analysis: techniques for experimental design, measurement, simulation and modeling*, John Wiley & Sons, Inc., 1991.
- [8] G. Taguchi, *System of Experimental Design*. White Plains, New York. Ed. UNIPUB/Kraus International Publications; Dearborn, Michigan, Ed. American Supplier Institute, Inc, 1985.
- [9] R. Martínez, "Studying the Effect of the Design Parameters on the Interconnection Network Performance in NOWs," M.S. thesis, Dpto. Informatica, Universidad de Castilla-La Mancha, 2003.