Improving the Efficiency of Adaptive Routing in Networks with Irregular Topology *

F. Silla and J. Duato
Facultad de Informática
Universidad Politécnica de Valencia
Camino de Vera s/n.
46071 - Valencia, SPAIN
E-mail: {fsilla,jduato}@gap.upv.es

Abstract

Networks of workstations are emerging as a cost-effective alternative to parallel computers. The interconnection between workstations usually relies on switch-based networks with irregular topologies. This irregularity makes routing and deadlock avoidance quite complicated. Current proposals avoid deadlock by removing cyclic dependencies between channels and therefore, many messages are routed along non-minimal paths, increasing latency and wasting resources.

In this paper, we propose a general methodology for the design of adaptive routing algorithms for networks with irregular topology that improves over a previously proposed one by reducing the probability of routing over non-minimal paths. The resulting routing algorithms allow messages to follow minimal paths in most cases, reducing message latency and increasing network throughput. As an example of application, we propose an improved adaptive routing algorithm for Autonet.

1 Introduction

Several switch-based interconnects like Autonet [8], Myrinet [1] and ServerNet [6] have been proposed to build networks of workstations for cost-effective parallel computing. Typically, these switches support networks with irregular topologies. Such irregularity provides the wiring flexibility required in local area networks, also allowing the design of scalable systems with incremental expansion capability. The irregularity makes routing and deadlock avoidance quite complicated. Current proposals avoid deadlock by removing cyclic dependencies between channels, routing many messages along non-minimal paths and, therefore, increasing latency and wasting resources.

A more efficient approach to deadlock avoidance consists of allowing the existence of cyclic dependencies between channels while providing some escape paths to avoid deadlock [4, 5]. The resulting routing algorithms are more flexible, usually increasing performance. Some recent router implementations like the MIT Reliable Router [3] and the Cray T3E router [9] are based on these techniques.

It is possible to apply this technique to networks with irregular topologies [11], routing most messages along minimal routes and using escape paths to avoid deadlock. In this paper, we propose a refinement of the general methodology for the design of adaptive routing algorithms proposed in [11]. This methodology is suitable for networks with irregular topologies. The new methodology reduces the probability of routing over non-minimal paths. As an example, we apply it to Autonet networks, proposing a new routing algorithm. This algorithm reduces message latency and increases network throughput over previously proposed routing schemes.

The paper is organized as follows. Section 2 introduces switch-based networks with irregular topologies. Section 3 describes how to improve the methodology for the design of adaptive routing algorithms proposed in [11]. As an example, the new methodology is applied to the Autonet routing algorithm. Section 4 describes the routing algorithm proposed in Autonet, showing how performance can be increased in Section 5. The performance of the new routing algorithm is evaluated in Section 6. Finally, some conclusions are drawn.

2 Networks of Workstations

Networks of workstations are usually arranged as switch-based networks consisting of a set of switches, each switch having several ports. Each port consists of one input and one output link. A set of ports in each switch are either connected to processors or left open, whereas the remaining ports are connected to ports of other switches to provide connectivity between the processors. Such connectivity

*This work was supported by the Spanish CICYT under Grant TIC94–0510–C02–01

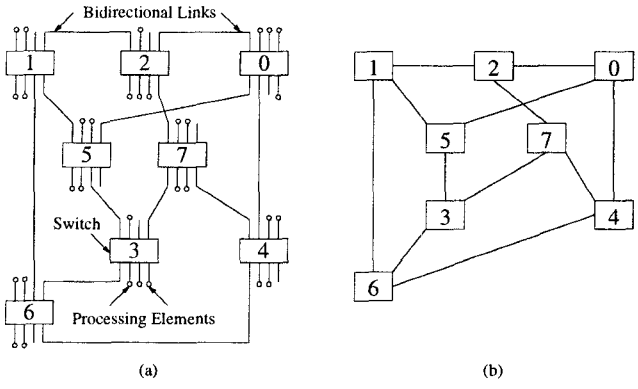

Figure 1: (a) A network with switch-based interconnect and irregular topology; (b) the corresponding graph G.

is typically irregular and the only thing that is guaranteed is that the network is connected. Typically, all links are bidirectional full-duplex and multiple links between two switches are allowed. Figure 1(a) shows a typical network of workstations using switch-based interconnect with irregular topology. In this figure, it is assumed that switches have eight ports and each processor has a single port.

The switch may implement different switching techniques: wormhole, virtual cut-through, or ATM. However, wormhole switching [2] is used in recently proposed networks like Myrinet and ServerNet. So, we will restrict ourselves to wormhole switching in this paper. Several deadlock-free routing schemes have been proposed in the literature for irregular networks [8, 1, 6, 7]. Routing in irregular topologies can be performed by using source routing or distributed routing. In the former case, each processor has a routing table that indicates the sequence of ports to be used at intermediate switches to reach the destination node. That information is stored in the message header [1]. In the latter case, processors and switches require routing tables. However, some network mapping algorithm must be executed in order to fill those tables before routing can be performed.

Once a message reaches a switch directly connected to its destination processor, it can be delivered as soon as the corresponding link becomes free. So, we are going to focus on routing messages between switches. The interconnection network I between switches can be modeled by a multigraph $I = G(N, C)$, where N is the set of switches, and C is the set of bidirectional links between the switches. Figure 1(b) shows the graph for the irregular network in Figure 1(a).

3 An Improved Design Methodology for Adaptive Routing Algorithms

A general methodology for the design of adaptive routing algorithms for networks with irregular topology has been recently proposed in [11]. This methodology can be summarized as follows. Given an interconnection network and a deadlock-free routing function defined on it, it is possible to duplicate all the physical channels in the network, or to split them into two virtual channels. In both cases, the graph representation of the new network contains the original and the new channels. Then, the routing function is extended so that newly injected messages can use the new channels without any restriction as long as the original channels can only be used in the same way as in the original routing function. However, once a message reserves one of the original channels, it can no longer reserve any of the new channels again. This design methodology supplies deadlock-free routing algorithms, as was proved in [11].

According to the extended routing function just revisited, new channels provide more routing flexibility than original channels. Besides they can be used to route messages through minimal paths. However, once a message reserves an original channel, it is routed through the original paths, which, in most cases, are not minimal. Also, routing through original paths produces a loss of adaptivity. Following this reasoning, the general methodology proposed in [11] can be refined by restricting the transition from new channels to original channels in the following way. Newly injected messages can only leave the source switch using new channels belonging to minimal paths, and never using original channels. When a message arrives at a switch from another switch through a new channel, the routing function gives a higher priority to the new channels belonging to minimal paths. If all of them are busy, then the routing algorithm selects an original channel belonging to a minimal path (if any). To ensure that the new routing function is deadlock-free, if none of the original channels provides minimal routing, then the original channel that provides the shortest path will be used. This ensures that, at least, one escape path exists at each switch. In case that several original channels provide the shortest paths, only one of them will be provided by the routing function. Once a message reserves an original channel, it will be routed using this kind of channels according to the original routing function until it is delivered. By restricting the use of original channels in this way, we allow most of the messages to follow minimal paths, and therefore, a more efficient use of the resources is achieved.

The improved design methodology also supplies deadlock-free routing algorithms, assuming that the original routing function is deadlock-free. The proof of deadlock freedom differs from the one for the methodology pro-

posed in [11] because newly injected messages can only leave the source switch using new channels. The following theorem formally proves deadlock freedom.

Theorem 1 *The improved design methodology supplies deadlock-free routing functions.*

Proof: We proceed by contradiction. Suppose that there is a deadlocked configuration. In this configuration, each message is waiting for channels occupied by other messages in the configuration. However, messages in the configuration cannot occupy original and new channels because once a message reserves an original channel, it cannot request a new channel again. Also, messages cannot occupy only original channels because the original routing function is deadlock-free. Therefore, all the blocked messages must occupy new channels. However, those messages can escape from deadlock by using the original channels because the original routing function is able to deliver messages from any switch (or processor) to any destination. The only exception is that newly injected messages can only leave the source switch using new channels. If one of those messages is waiting for a new channel, it is occupying a channel directly connected to a processor. Therefore, no other message in the configuration can request the channel it occupies and that message is not involved in the deadlocked configuration. Thus, there is no deadlocked configuration, contrary to the initial assumption. □

This result is valid for any topology and any original deadlock-free routing function. Deadlock freedom can also be proved by using the theory proposed in [5].

4 The Autonet Routing Algorithm

The Autonet routing algorithm [8] provides deadlock-free routing and partially adaptive communication between nodes in an irregular network. It is distributed and implemented using table-lookup. When a message reaches a switch, the destination address stored in its header is concatenated with the incoming port number to index the routing table, that returns the suitable outgoing port number. When the routing table provides several output ports, one of them is randomly selected.

In order to fill the routing tables, a breadth-first spanning tree (BFS) on the graph G is computed first using a distributed algorithm. Routing is based on an assignment of direction to the operational links. In particular, the "up" end of each link is defined as: (1) the end whose switch is closer to the root in the spanning tree; (2) the end whose switch has the lower ID, if both ends are at switches at the same tree level (see Figure 2). The result of this assignment is that each cycle in the network has at least one link in the "up" direction and one link in the "down" direction.

To eliminate deadlocks while still allowing all links to be used, this routing scheme uses the following up/down

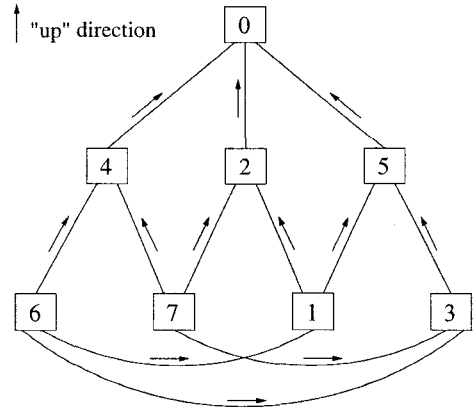

Figure 2: Link direction assignment.

rule: a legal route must traverse zero or more links in the "up" direction followed by zero or more links in the "down" direction. Thus, cyclic dependencies between channels are avoided because a message cannot traverse a link along the "up" direction after having traversed one in the "down" direction. Such routing not only allows deadlock-freedom but also adaptivity. However, in some cases, up/down routing is not able to supply any minimal path between some pairs of nodes, as shown in the following example.

Figure 2 shows the example irregular network shown in Figure 1(a). Note that every cycle has at least one link in the "up" direction and one link in the "down" direction. All the alternative minimal paths are allowed in some cases. For example, a message transmitted from switch 7 to switch 0 can be either routed through switch 4 or switch 2. In some other cases, however, only some minimal paths are allowed. For example, a message transmitted from switch 2 to switch 5 can be routed through switch 0 but it cannot be routed through switch 1. It may happen that all the minimal paths between two nodes are forbidden. This is the case for messages transmitted from switch 4 to switch 1. The shortest path (through switch 6) is not allowed. All the allowed paths (through switches 0, 2, and through switches 0, 5) are non-minimal. The amount of forbidden minimal paths increases as the network becomes larger.

5 Providing Minimal Paths

In this section we apply the design methodology presented in Section 3 to the Autonet routing algorithm, allowing the use of minimal paths to most messages in the network. Since the methodology proposed needs two channels connecting each pair of switches, we can follow two different approaches. In the first one, all the physical channels in the network are duplicated, taking advantage of spare switch ports. Thus, additional wires are required but the current

switch design is valid and the proposed routing algorithm can be implemented simply by changing the routing tables. In the second approach, physical channels are split into two virtual channels. This approach does not need more wires, but requires a new switch design that supports virtual channels. Virtual channels can be efficiently implemented by using the techniques described in [10]. In this paper we only evaluate the second approach.

New channels will be used to provide minimal routing, whereas original channels will serve as escape paths to avoid deadlock. Thus, the original up/down routing algorithm must be extended in order to use the new channels efficiently. The new routing function is a direct application of the design methodology proposed in Section 3. It is inherently different from that proposed in [11], since that methodology allows the use of original channels to leave the source switch and does not limit their use at intermediate nodes to only those original channels that provide minimal paths. Moreover, the resulting routing algorithms provide a greater adaptivity, while in the new methodology we try to restrict routing to only minimal paths, diminishing the adaptivity.

A variation of the new routing algorithm could be as follows. At intermediate switches, instead of routing through an original channel once all of the new channels belonging to minimal paths are busy, we could reduce even more the use of original channels by routing the message through a nonminimal new channel that conforms to the up/down rule. If this new channel is also busy, then try the original channel (escape path). However, this routing algorithm increments the complexity of the selection function, while the increment in performance obtained by simulation (assuming the same routing time) is negligible. This is due to the use of nonminimal paths.

It is important to note that using adaptivity does not increase the complexity of the switch. As switches become faster, they tend to be simpler, and therefore adaptivity could introduce an additional complexity, reducing the maximum performance the switch could achieve. However, since our routing algorithm is implemented using table-lookup, decisions are taken by looking up in a programmable memory. Thus, modifying the routing algorithm in order to increase adaptivity does not add any significant extra delay since the basic operation (consulting a table) is the same.

6 Performance Evaluation

In this section, we evaluate the performance of the routing algorithm proposed in Section 5. We will refer to this routing scheme as MA-2VC, since it provides minimal adaptive routing with two virtual channels. We have also evaluated the up/down scheme for the same network for comparison purposes. We will refer to this routing scheme as UD.

As the MA-2VC routing scheme requires two virtual channels per physical channel, in order to evaluate the effect of using virtual channels with the original routing algorithm, a third routing scheme was evaluated. We will refer to as UD-2VC the routing scheme in which physical channels are split into two virtual channels, and both of them use up/down routing. Note that we do not divide the irregular network into two different virtual networks, since messages at a switch can use any of the two outgoing virtual channels independently of the incoming virtual channel they arrived through.

The routing algorithm proposed in [11] focuses on adaptivity and the one we propose in this paper does on minimal routing. In order to compare the performance achieved by both algorithms, we will refer to the former as A-2VC.

Instead of analytic modeling, simulation was used to evaluate the routing algorithms. Our simulator models the network at the flit level. The evaluation methodology used is based on the one proposed in [5]. The most important performance measures are latency and throughput. The message latency lasts since the message is introduced in the network until the last flit is received at the destination node. Latency is measured in clock cycles. Traffic is the flit reception rate, measured in flits per node per cycle. Throughput is the maximum amount of information delivered per time unit (maximum traffic accepted by the network).

6.1 Network Model

The network is composed of a set of switches. Network topology is completely irregular and has been generated randomly. However, for the sake of simplicity, we imposed three restrictions to the topologies that can be generated. First, we assumed that there are exactly 4 nodes (processors) connected to each switch. Also, two neighboring switches are connected by a single link. Finally, all the switches in the network have the same size. We assumed 8-port switches, thus leaving 4 ports available to connect to other switches. We have evaluated networks with a size ranging from 16 switches (64 nodes) to 64 switches (256 nodes). For each network size, several distinct irregular topologies have been analyzed.

Each switch has a routing control unit that selects the output channel for a message as a function of its destination node, the input channel and the output channel status. Table-lookup routing is used. UD, UD-2VC, A-2VC or MA-2VC routing strategy can be chosen. The routing control unit can only process one message header at a time. It is assigned to waiting messages in a demand-slotted round-robin fashion. When a message gets the routing control unit but it cannot be routed because all the alternative output channels are busy, it must wait in the input buffer un-

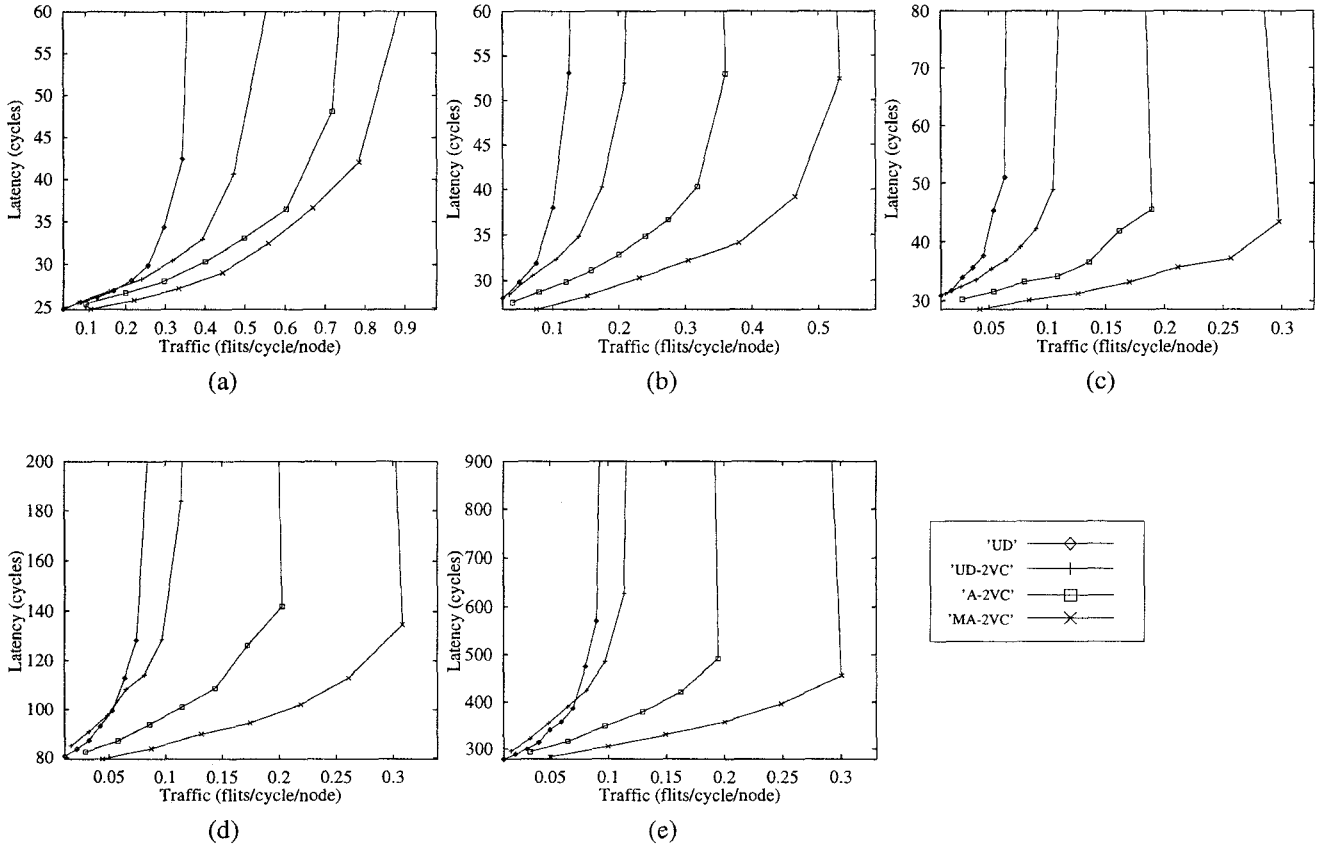

Figure 3: Average message latency versus accepted traffic.

til its next turn. A crossbar inside the switch allows multiple messages traversing it simultaneously without interference. It is configured by the routing control unit each time a successful routing is made. We assumed that it takes one clock cycle to compute the routing algorithm, or to transmit one flit across a crossbar or a physical channel. When physical channels are split into two virtual channels, we assumed flit-by-flit multiplexing. Although this is not a realistic assumption for networks of workstations, performance results do not differ significantly from those obtained with more realistic flow-control protocols [10]. See [10] for a description of flow-control protocols that support long wires and different wire lengths as well as the corresponding performance evaluation.

6.2 Message Generation

For each simulation run, we considered that message generation rate is constant and the same for all the nodes. Once the network has reached a steady state, the flit generation rate is equal to the flit reception rate (traffic). We have evaluated the full range of traffic, from low load to saturation. On the other hand, we have considered that message destination is randomly chosen among all the nodes. For message length, 16-flit, 64-flit and 256-flit messages were considered.

6.3 Simulation Results

For each network size, we analyzed several distinct randomly generated irregular topologies. However, the average latency values achieved by each topology for each traffic rate are almost the same. The only differences arise when the networks are heavily loaded, close to saturation. Additionally, the throughput achieved by all the topologies is almost the same. Hence, we will show the results obtained by one of those topologies, chosen randomly.

Figure 3(a) shows the average message latency versus accepted traffic for each routing scheme on a randomly generated irregular network with 16 switches. Message size is 16 flits. As can be seen, when virtual channels are used in the original routing scheme (UD-2VC), throughput increases by a factor of 1.5, while with the A-2VC routing scheme throughput is doubled. When the MA-2VC routing scheme is used, throughput is almost tripled. Moreover, the latency achieved by MA-2VC is lower than the one for

the rest of routing strategies for the whole range of traffic. Therefore, most of the improvement achieved by MA-2VC is due to the use of shorter paths, besides the increment of adaptivity with respect to UD scheme.

The MA-2VC routing scheme scales very well with network size. Figures 3(b) and 3(c) show the results obtained for networks with 32 and 64 switches, respectively, when message size is 16 flits. For 64 switches, throughput increases by factors of 4.2 and 2.7 with respect to the UD and UD-2VC schemes, respectively, when using the MA-2VC scheme. Latency is also reduced for the whole range of traffic. However, the factor of improvement in performance achieved by the A-2VC scheme with respect to UD is only around 2.7. When network size increases, the performance improvement achieved by the MA-2VC scheme increases because there are larger differences among the real distance between any two switches and the routing distance imposed by the up/down scheme.

Figures 3(d) and 3(e) show the influence of message size on the behavior of the routing schemes. They show the average message latency versus traffic for a network with 64 switches when message length is 64 and 256 flits, respectively. These results show the robustness of the MA-2VC routing scheme against message size variation. The MA-2VC routing scheme achieves the maximum throughput and lowest latency for all message sizes.

Finally, it should be noted that the improvement achieved by using the theory proposed in [4, 5] for the design of adaptive routing algorithms is much higher in irregular topologies than in regular ones. This is mainly due to the fact that most paths in irregular networks are non-minimal if those techniques are not used.

7 Conclusions

In this paper we have refined the general methodology for the design of adaptive routing algorithms for switch-based interconnects with irregular topologies proposed in [11]. The main difference with respect to the basic design methodology is that the new one severely restricts the use of the original channels. When a message is injected into the network, it can only leave the source switch through the new channels that provide minimal paths. At intermediate switches, messages arriving from new channels can only use those channels that provide minimal routing. In order to ensure deadlock-freedom, at least one of the original channels is provided as escape path. Restricting the use of original channels in this way, most messages follow minimal paths, and therefore latency is reduced and throughput increased.

We have evaluated the performance of the new methodology when applied to the Autonet network. The results have been compared with the performance of the original Autonet algorithm (with and without virtual channels). They have also been compared with the performance achieved by a routing algorithm designed according to the methodology proposed in [11] when applied to the Autonet routing scheme. The results show that the new routing algorithm reduces latency in all the cases for the whole range of network load, also increasing considerably the maximum throughput achievable.

References

[1] N. J. Boden, D. Cohen, R. E. Felderman, A. E. Kulawik, C. L. Seitz, J. Seizovic and W. Su, "Myrinet - A gigabit per second local area network," *IEEE Micro*, pp. 29–36, February 1995.

[2] W. J. Dally and C. L. Seitz, "Deadlock-free message routing in multiprocessor interconnection networks," *IEEE Transactions on Computers*, vol. C–36, no. 5, pp. 547–553, May 1987.

[3] W. J. Dally, L. R. Dennison, D. Harris, K. Kan and T. Xanthopoulus, "The Reliable Router: A reliable and high-performance communication substrate for parallel computers," in *Proceedings of the Workshop on Parallel Computer Routing and Communication*, pp. 241–255, May 1994.

[4] J. Duato, "On the design of deadlock-free adaptive routing algorithms for multicomputers: Design methodologies," in *Proceedings of Parallel Architectures and Languages Europe 91*, June 1991.

[5] J. Duato, "A new theory of deadlock-free adaptive routing in wormhole networks," *IEEE Transactions on Parallel and Distributed Systems*, vol. 4, no. 12, pp. 1320–1331, December 1993.

[6] R. Horst, "ServerNet deadlock avoidance and fractahedral topologies," in *Proceedings of the 10th International Parallel Processing Symposium*, Honolulu, Hawaii, pp. 274–280, April 1996.

[7] W. Qiao and L. M. Ni, "Adaptive routing in irregular networks using cut-through switches," in *Proceedings of the 1996 International Conference on Parallel Processing*, August 1996.

[8] M. D. Schroeder et al., "Autonet: A high-speed, self-configuring local area network using point-to-point links," Technical Report SRC research report 59, DEC, April 1990.

[9] S. L. Scott and G. Thorson, "The Cray T3E networks: adaptive routing in a high performance 3D torus," in *Proceedings of Hot Interconnects IV*, August 1996.

[10] F. Silla and J. Duato, "On the Use of Virtual Channels in Networks of Workstations with Irregular Topology," in *Proceedings of the 1997 Parallel Computing, Routing, and Communication Workshop*, June 1997.

[11] F. Silla, M. P. Malumbres, A. Robles, P. López and J. Duato, "Efficient Adaptive Routing in Networks of Workstations with Irregular Topology," in *Proceedings of the Workshop on Communications and Architectural Support for Network-based Parallel Computing* , February 1997.