A New Model of Communication Cost for Interconnection Networks with Irregular Topology.

Vicente Arnau, Juan M. Orduña, Aurelio Ruiz, Rodrigo Valero, José Duato

Abstract—**Networks of Workstations (NOWs) have become a cost-effective alternative to parallel computers. Switch-based interconnects with irregular topologies provide the wiring flexibility required in these environments. The characterization of these networks results quite difficult, since the traditional parameters used for regular topologies (node degree, diameter, average distance, etc.) do not provide information about the arrangement of the links. In this paper, we propose a new model of communication cost between network nodes. This model takes into account both the network topology and the routing algorithm, but it does not depend on the traffic pattern generated by the application running on the machine. The evaluation results show that our communication cost model is highly correlated with network performance. Since it provides a metric based on internode distance, our model can be used as the basis for both an efficient characterization of networks as well as an efficient mapping of processes to processors.**

I. Introduction

NETWORKS of Workstations (NOWs) have become a cost-effective alternative to parallel computers [1], [5]. In theses systems, workstations are connected via high-speed local area networks. Switch-based interconnects with irregular topologies [2], [6], [8] provide the wiring flexibility required in these environments, also allowing the design of scalable systems with incremental expansion capabilities.

In massively parallel computers (MPPs), interconnection networks have been characterized by their topological properties, such as number of nodes, bisection width and diameter [4]. However, the characterization of irregular networks cannot be based upon these parameters, since they do not provide information about the arrangement of the links. On other hand, the routing algorithm may seriously affect network performance by determining the traffic distribution in the network, as in the case of the Autonet networks [8]. Therefore, the routing algorithm must be considered when characterizing irregular networks.

As a result of the above considerations, a new method based on grouping network nodes into clusters has been proposed for characterizing irregular networks [7]. In this method, a link cost calculation algorithm assigns a cost depending on the load distribution to every link in the network. Based upon link cost, this method proposes to partition network into clusters in such a way that the available bandwidth within a cluster is higher than the bandwidth of the links that communicate the cluster with other clusters. This approach can easily identify the bottleneck links, helping the operating system to properly distribute the load between the correct group of workstations. However, in this approach the characterization of the network depends on the traffic pattern generated by the application running on the machine. Additionally, that traffic pattern may vary over time, thus limiting the applicability of this approach.

In this paper, another method to characterize irregular networks is proposed. The idea is to establish a model of communication cost between each node pair in the network that is independent of the traffic pattern, taking into account only the topology of the network and the routing algorithm. Therefore, this model can be used in clustering algorithms that do not depend on the traffic pattern generated by the applications running on the machine. Furthermore, this model can be used as the basis for an efficient mapping of processes to processors, since it provides a metric based on internode distance. We have studied the correlation of our model of communication costs with performance measures obtained by simulation. The results show that the proposed model is highly correlated with network performance, thus being a suitable metric for network characterization.

II. A New Model of Communication Cost

Our model proposes a simple metric, the *equivalent distance* between each pair of nodes (in what follows we will refer to a switching element as a node). This metric measures the cost of communi-

This paper is supported by the Spanish CICYT under Grant TIC97–0897–C04–01

V. Arnau, J. M. Orduña, A. Ruiz and R. Valero are with the Instituto de Robótica, Universidad de Valencia, Spain. E-mail: Juan.Orduna@uv.es. J. Duato is with DISCA, Univeridad Politécnica de Valencia, SPAIN. E-mail: jduato@gap.upv.es

cating between two nodes without explicitly considering traffic pattern. A *table of equivalent distances* can be obtained by computing the equivalent distance between each pair of nodes in the network. Unlike the metric proposed in [7], this metric does not consider different single link costs. We assume that link bandwidth is the same for all the links in the network. Under these conditions, assigning different single link costs would make the model of communication cost dependent on the traffic pattern. The equivalent distance for a pair of nodes is computed taking into account all the shortest paths between them supplied by the routing algorithm. We assumed Up/Down routing [8] because it is used in several commercial networks. However, the model proposed in this paper can be applied to any routing algorithm.. The name of the metric is derived from the analogy to the electrical equivalent resistance. Indeed, we use the same rules as for electrical circuits to compute the total communication cost between nodes, applying Kirchoff's laws.

In particular, the method used to construct the table of equivalent distances, computing the equivalent distance between each pair of nodes in the network, is the following one:

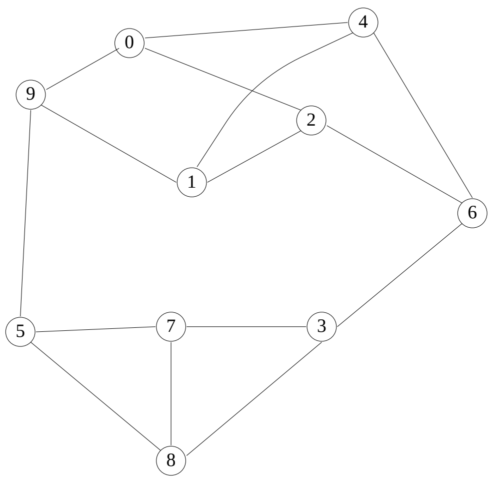

Fig. 1. Graph G for a 10-node Autonet network

1. If there exists only one shortest path between a given pair of nodes then the communication cost between those nodes will be the sum of the costs of the links that form the path. That is, this case is similar to computing the equivalent resistance of an electrical circuit consisting of serially arranged resistors. Since we have assumed that all the links in the network have unit cost, the communication cost is equal to the number of links in the path.
2. If there exists more than one shortest path between a given pair of nodes then the communication cost between them is computed similarly to the electrical equivalent resistance between two points of an electrical circuit, replacing each link in a shortest path with a unit resistor and applying Kirchoff's laws. Note that we only consider the shortest paths provided by the routing algorithm. Those paths are not necessarily minimal (as is the case for Up/Down routing). Also, the paths not supplied by the routing algorithm are not considered.

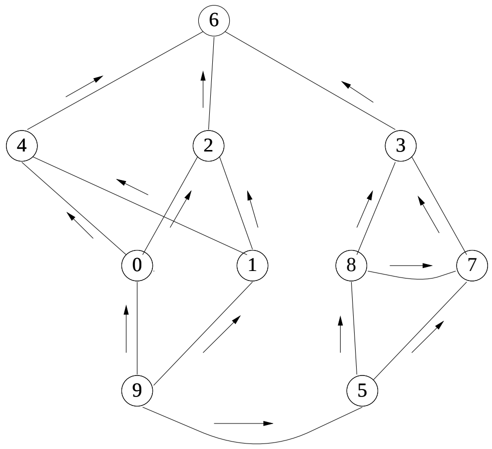

Fig. 2. Link direction assignment for the graph in Figure 1

The application of this method will produce a table T_N with $N \times N$ equivalent distances, where N is the number of nodes in the network. In this table, the element T_{ij} represents the equivalent distance between node i and node j. As an example, consider an Autonet network modeled as the multigraph $I = G(N, C)$ shown in Figure 1, where N is the set of switches, and C is the set of bidirectional links between the switches. The Autonet routing algorithm is distributed, and implemented using table-lookup [8]. In order to fill the routing tables, a breadth-first spanning tree (BFS) is computed first, using a distributed algorithm. Based on this spanning tree, the link direction assignment is computed for graph I. The "up" end of a link is connected upper level node when the link connects nodes located at different levels. When communicating nodes at the same tree level, the "up" end of a link is connected to the node with the lower label. 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. Figure 2 shows the link direction assignment for the graph in Figure 1. In this figure switches are arranged in such a way that all the switches at the same level in the spanning tree are at the same vertical position in the figure. The Up/Down routing scheme establishes that a legal route must traverse zero or more links in the "up" direction, followed by zero or more links in the "down" direction. A message cannot traverse a link along the "up" direction after having traversed a link in the "down" direction. Although such routing scheme is deadlock-free and allows some adaptivity, in some cases Up/Down routing is not able to supply any minimal path between two nodes.

TABLE I

Possible paths for going from node 0 to node 1

0-2-6-4-1
0-2-1
0-4-6-2-1
0-4-1

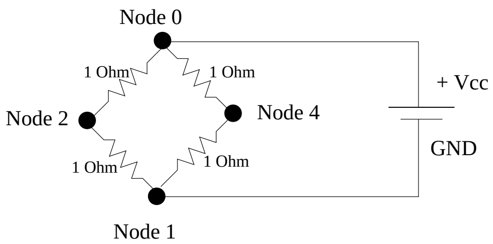

Fig. 3. Model of equivalent resistance between node 0 and node 1

Let us consider the equivalent distance between nodes 0 and 1. Taking into account the link direction assignment in Figure 2 and considering Up/Down routing, Table II shows all possible paths for going from node 0 to node 1, existing two shortest paths. Therefore, in order to compute the equivalent distance between nodes 0 and 1, the source and destination nodes must be considered as the V_{cc} and GND points of an electric circuit. Nodes 2 and 4 must be considered as two different intermediate points, and each link must be considered as a resistor with unit resistance, resulting in the circuit shown in Figure 3. Applying Kirchoff's laws to this circuit, an equivalent resistance of 1 Ohm is obtained. Thus, the equivalent distance from node 0 to node 1 is set to 1. The rest of equivalent distances in the table of distances are computed in the same way. The table of distances obtained for the network whose graph is shown in Figure 1 is shown in Table II.

An important property of the table is its symmetry. It is due to the behavior of the Up/Down routing algorithm. Effectively, there exist three cases when communicating two nodes: using only "up" paths (that is, paths that are exclusively formed by "up" channels), using only "down" paths (paths that only use "down" channels) and using "up-down" paths (paths that first use only "up" channels and then only "down" channels). When reversing communication direction, all "up" channels in a path become "down" channels, and therefore all "up" paths become "down" paths, and vice-versa. However, "up-down" paths remain unchanged. As a result, all possible paths for communicating two nodes are the same for both communication directions, and therefore the table of distances is symmetrical. As a consequence, the table of equivalent distances determines a pseudo-metric space ($T_{ij} = 0 \iff i = j$ and $T_{ij} = T_{ji}$). However, the table of distances does not satisfy the triangular inequality, and thus it does not define a metric space. This limitation prevents the application of clustering methods that use Euclidean metrics for finding and characterizing groupings of nodes. Nevertheless, the key issue for this table is its correlation with network performance. If it were well correlated, it would provide a simple method of measuring how far a node is from the rest of the network. In this case, the characterization of the network based on this table could be done following the criterion of node proximity.

TABLE II

Table of equivalent distances between nodes

Node	0	1	2	3	4	5	6	7	8	9
0	0	1	1	2	1	3	1	3	3	1
1	1	0	1	2	1	3	1	3	3	1
2	1	1	0	2	2	3	1	3	3	1
3	2	2	2	0	2	1	1	1	1	2
4	1	1	2	2	0	3	1	3	3	1
5	3	3	3	1	3	0	2	1	1	1
6	1	1	1	1	1	2	0	2	2	1.25
7	3	3	3	1	3	1	2	0	1	2
8	3	3	3	1	3	1	2	1	0	2
9	1	1	1	2	1	1	1.25	2	2	0

III. Model Validation

The *table of distances* obtained in the previous section takes into account only the topology of the network and the routing algorithm. These distances intend to measure the communication cost between each pair of nodes in the network. In order to ensure that these distances provide a good estimation of the communication cost, we must study the correlation between each distance in the table of distances and the average latency of the messages exchanged between the corresponding pair of nodes.

We have evaluated the performance of several irregular networks by simulation. The evaluation methodology used is based on the one proposed in [3]. The most important performance measures are latency and throughput. The message latency lasts since the message is injected in the network until the last flit is received at the destination node. Throughput is the maximum amount of information delivered per time unit (maximum traffic accepted by the network). Traffic is the flit reception rate. Latency is measured in clock cycles. Traffic is measured in flits per node per cycle.

A. Network Model and Message Generation

The network is composed of a set of interconnected switches. The network topology is irregular and has been generated randomly. However, for the sake of simplicity we imposed three restrictions. First, we assumed that there are exactly 4 nodes connected to each switch. Second, two neighboring switches are connected by a single bidirectional link. Finally, all the switches in the network have the same size. We assumed 8-port switches. Therefore, each switch has 4 ports available to connect to other switches. From these four ports, only three of them have been used at each switch in our simulation experiments. The remaining port is left open. We have evaluated networks with size ranging from 8 switches (32 nodes) to 24 switches (96 nodes). For some network sizes, several distinct topologies have been analyzed.

In order to study the correlation between the table of distances and the performance of the network, we have used a uniform distribution for message generation. Nevertheless, in order to ensure that the performance results do not depend on the traffic pattern, we have also used other nonuniform traffic patterns for message generation when evaluating some networks. In particular, we have considered perfect-shuffle, bit reversal, and butterfly distributions. The message generation rate is constant and the same for all the nodes in the network. We have considered a fixed message size of 16 flits, and we have evaluated a wide range of traffic, ranging from low load to saturation.

B. Simulation and Correlation Results

Our simulator models the network at the flit level, and produces both global and per-node results. Global average latency shows the average latency of all messages transmitted through the network for a given traffic load. It provides an estimation of the degree of saturation of the network. On the other hand, the average latency for each source-destination pair in the network allowed us to study the correlation of the table of distances with network performance.

Figure 4 shows performance evaluation results for the 10-switch network modeled as the graph in Figure 1. This figure shows the global average message latency. Also, average message latencies for each pair of nodes were computed. However, they are not shown here due to space limitation.

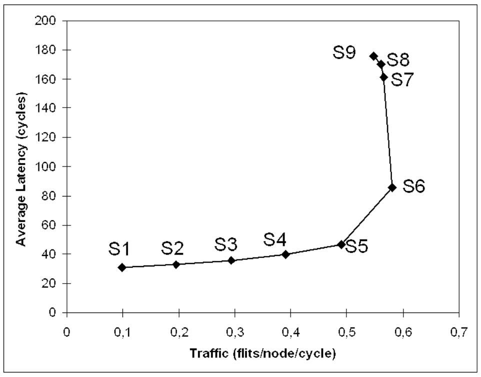

Fig. 4. Global performance results for the 10-node network modeled as the graph in Figure 1

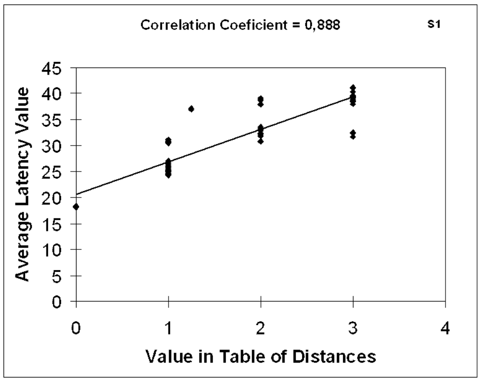

Fig. 5. Least square linear adjustment for simulation point S1

In order to establish the correlation between the table of distances and these performance evaluation results, we first computed the least square linear adjustment for each point in Figure 4. The results for the first point S1 (traffic equal to 0.097 flits/node/cycle and global average latency equal to 30'93 cycles) are shown in Figure 5. In this figure, the values in the table of distances are shown in the X-axis, while the specific latencies obtained by the simulator for each pair of nodes are shown in the Y-axis. The correlation index obtained in this case was 88.8%. The worst results were obtained when the network was under deep saturation (point S8 in Figure 4, with traffic load equal to 0.56 flits/node/cycle and average latency equal to 170.17 cycles). The correlation index in this case was 74.4%. These results show that there is a strong correlation between equivalent distances and the corresponding average latencies for all load conditions of the network. However, correlation decreases as the network approaches saturation.

Also, we have studied the correlation between the equivalent distances and the average values of the average latencies supplied by our network simulator, weighted by the number of occurrences of the corresponding distance in the table of distances. Figure 6 shows the corresponding regression curve for simulation point S8. In this figure, there is a point for each different value in the table of distances. These values are represented in the X-axis. The number of occurrences for each value is shown on the right side of the figure. For example, point P2 represents value 1.25 in the table of distances and appears twice in the table of distances (see Table II). Similarly, point P1 represents value 1 in the table of distances, and appears 42 times in Table II. The weighted average values of average latencies show a very high correlation with the table of distances (99.2%).

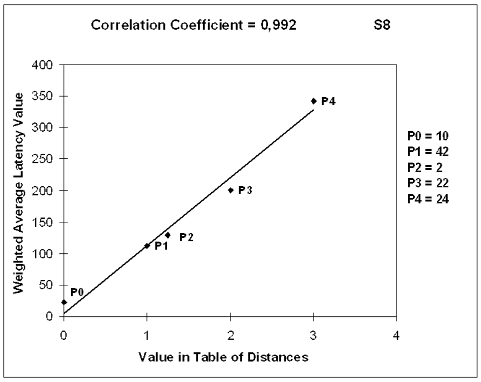

Fig. 6. Least square linear adjustment for weighted average values of average latencies

Figure 7 shows the correlation between the values in the table of distances and both latency values and the weighted average values of average latencies. Effectively, it can be clearly seen in this figure that the weighted average latencies correlate better than the latency values. Both measurements show a high correlation coefficient, exceeding 70% for all simulation points. However, the correlation of latency values with the table of distances decreases when the network enters saturation, while the weighted average latencies even improve correlation when the network reaches saturation.

Figure 8 shows the correlation between the values in the table of distances and both latency values and weighted average values of average latencies for a network with 16 switches (64 nodes). The correlation coefficient for low load is about 80% (points

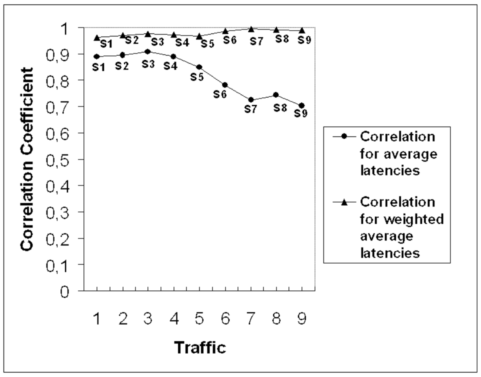

Fig. 7. Correlation of average latency values and weighted average values of average latencies

S1 to S4). However, when the network approaches saturation the correlation of latency values begins to decrease (starting from point S5), reaching 40% when the network is deeply saturated (points S8 and S9). That is, there is a poor correlation between the table of distances and latency values when the network is heavy loaded. On the contrary, the correlation of the weighted average latencies remains about 90% for all load conditions.

In order to ensure that the topology does not affect correlation results, 6 different randomly generated 16-node topologies have been evaluated. We have computed the average value of the correlation coefficients both when the network is under low load, as well as when the network is deeply saturated. The average correlation coefficient for average latencies is 81.9%, with a standard deviation of 2%. When the network is deeply saturated, the average correlation coefficient falls to 46%, with a standard deviation of 12%. However, the average correlation coefficient for weighted average latencies reaches 89.9% with a standard deviation of 2% when the network is under low load. When the network is deeply saturated, the average value of this correlation coefficient decreases to 83.2%, with a standard deviation of 7%. Therefore, the computed standard deviations show that the correlation between the table of distances and network performance is not significantly affected neither by the network topology nor by the traffic load, particularly when the weighted average latencies are used as the performance measurement.

Additionally, we have evaluated a 20-node network under the uniform traffic distribution as well as under three different nonuniform traffic distributions. The goal is to analyze whether the proposed approach (that does not take into account the traffic pattern when constructing the table of distances) is

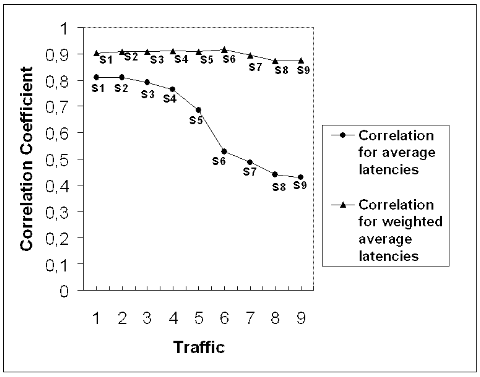

Fig. 8. Correlation of average latency values and weighted average values of average latencies for a 16-switch network

valid and accurate enough for any traffic distribution. In particular, we have evaluated the network under bit reversal, perfect shuffle and butterfly distributions. Although the results are not shown here due to space limitations, they do not significantly differ from the correlation results for the same network under uniform traffic distribution. As when analyzing the effect of network topology on the correlation results, we have computed the average correlation coefficient for different traffic patterns both when the network is under low load as well as when the network is deeply saturated. The computed average correlation coefficient for average latencies is 96.4% with a standard deviation of 2% when the network is under low load, and decreases to 58.5% with a standard deviation of 13% when the network is deeply saturated. However, the average correlation coefficient for weighted average latencies is 99.1% with a standard deviation of 0.4% when the network is under low load, and only falls to 86.2% with a standard deviation of 1% when the network is deeply saturated. Therefore, the computed standard deviations allow us to conclude that the correlation of the table of distances with network performance does not significantly depend on the traffic pattern, particularly when the weighted average latencies are used as the network performance measurement.

IV. Conclusions and Future Work

In this paper, a new model of communication cost is proposed in order to provide a single method for characterizing networks with irregular topology. This model introduces a new metric, the equivalent distance between each pair of nodes. This metric is independent of the traffic pattern, and assumes that all the links have the same cost. By computing the equivalent distances between all the nodes in the network, the table of equivalent distances is formed. This table is symmetrical but does not define a metric space. We have evaluated the correlation of the table of equivalent distances with network performance.

The results show that the proposed metric is strongly correlated with the average latency obtained by simulations. However, for all the network sizes the degree of correlation decreases drastically when the network reaches saturation. Nevertheless, the proposed metric is highly correlated with the weighted average values of average latencies obtained by simulation. Furthermore, the correlation of the table of distances with these weighted average values is not significantly affected when the network enters saturation. Additionally, these weighted average values are independent of both traffic pattern and network topology. Therefore, this table provides a good metric for the communication cost under any load condition.

As a result, we conclude that although the table of equivalent distances does not define a metric space, it can be used by clustering algorithms based on internode distances as the main communication cost index. Furthermore, the resulting clusters can be used as a basis for an efficient mapping of processes to processors. As for future work, we plan to develop efficient clustering algorithms based on the table of equivalent distances, as well as efficient process mapping algorithms.

References

[1] T. E. Anderson, T. E. Culler and D. A. Patterson, "A Case for NOW (Networks of Workstations)," *IEEE Micro*, pp. 54–64, February 1995.

[2] 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.

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

[4] J. Duato, S. Yalamanchili, L. Ni, *Interconnection Networks: An Engineering Approach, IEEE Computer Society Press, 1997.*

[5] R. Felderman et al., "Atomic: A High Speed Local Communication Architecture," *J. High Speed Networks*, vol. 3, no. 1, pp.1-29, 1994

[6] R. Horst, "TNet: A Reliable System Area Network," *IEEE Micro*, vol. 15, no. 1, pp. 37–45, February 1995.

[7] T.M. Pinkston and W. H. Ho, "A Clustering Approach in Characterizing Interconnection Networks", in *Proceedings of Fifth International Conference on High Performance Computing*, pp. 277-284, December 1998.

[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.