A Clustering Approach for Improving Network Performance in Heterogeneous Systems ⋆

V. Arnau, J. M. Orduña, S. Moreno, R. Valero, and A. Ruiz

Departamento de Informática. Universidad de Valencia.SPAIN
Juan.Orduna@uv.es

Abstract. A lot of research has focused on solving the problem of computation-aware task scheduling on heterogeneous systems. In this paper, we propose a clustering algorithm that, given a network topology, provides a network partition adapted to the communication requirements of the applications running on the machine. Also, we propose a criterion to measure the quality of each one of the possible mappings of processes to processors based on that network partition. Evaluation results show that these proposals can greatly improve network performance, providing a basis of a communication-aware scheduling technique.

1 Introduction

Networks of Workstations (NOW's) are used nowadays as parallel computers, forming heterogeneous systems. A lot of research has focused on solving the *NP-complete* problem of efficiently scheduling diverse groups of tasks to the machines that form these systems from the computational point of view. However, the increasing computational power of new processors and the growing communication requirements of the applications may cause the interconnection network in these heterogeneous systems to become the system performance bottleneck.

In a previous paper, we proposed a new model of communication cost between nodes, *the table of equivalent distances* [1]. This model consist of a table of $N \times N$ distances, where N is the number of nodes in the network. In this table, the element T_{ij} represents the communication cost for messages going from node i to node j. In this paper, we propose a clustering algorithm based on the table of distances that provides a network partition, and a criterion to measure the suitability of each allocation of network resources to the applications that the provided network partition may generate. Evaluation results show that the use of this proposals significantly improve network performance by reducing communication bottlenecks. Furthermore, this clustering technique is applicable to both regular and irregular topologies, providing a general basis for communication-based process mapping.

2 A New Clustering Approach

From a general point of view, we can consider that each application belongs to a different user. Therefore, we can assume that the processes belonging to the same application

⋆ Supported by the Spanish CICYT under Grant TIC97–0897–C04–01

may intensively communicate between them, but they will not communicate at all with processes from other applications. Therefore, we can group the processes running on the machine, forming a set of logical clusters of processes, where each cluster is formed by the processes belonging to each application. The proposed algorithm intends to provide a network partition adapted to any existing set of logical clusters.

The first step in this method is to find an Euclidean metric space in which we can represent our N nodes, in such a way that the resulting distances between them are as close as possible to the values in the table of distances (the latter one does not define a metric space). We have computed a least squares linear adjustment using the steepest gradient method [2]. The solution consists of N points in an Euclidean space whose coordinates generate a table of Euclidean distances with the least quadratic error with regard to the table of distances. It is worth mention that the table of Euclidean distances does not contain repeated values (except zero values in the diagonal).

Once a table of Euclidean distances has been computed, the furthest-neighbor algorithm is used to compute the optimal dendogram [4]. This algorithm uses a similarity measure. In each step the algorithm merges two of the existing clusters into a new one, choosing the two clusters that result in the lowest similarity measure when the step is applied. The similarity measure usually used in this algorithm is the intracluster distance, and therefore it is called the furthest-neighbor algorithm. However, we considered as the similarity measures f to be *maximized* the inverse of the Euclidean distance, defined as $f_a = \frac{1}{D_{ij}}$, where D_{ij} is the distance from cluster i to cluster j in the Euclidean table of distances. The initial network partition consists of the N nodes located at the coordinates given by the computed table of Euclidean distances. In each step a new partition is formed, decreasing the number of clusters by one. When merging two clusters, they are replaced by a new cluster, and the Euclidean table of distances must be computed again in each step.

The result of the above clustering approach is a dendogram, but not a mapping of processes to processors. The cardinal of the set of logical clusters can be used to determine when to stop the clustering algorithm, obtaining a network partition with a number of network clusters equal to the number of existing logical clusters of processes. Nevertheless, the number of nodes (switches) in each network cluster may significantly differ from the number of processes in each logical cluster of processes. Therefore, new changes in this partition are still needed in order to map all the existing processes according to the communication requirements. We have performed manually this clustering adjustment, obtaining different possible process mappings from each network partition. However, it is necessary to define a metric of the communication bandwidth achieved by each one of the possible mappings, in order to select the one who provides the best network performance.

We have defined two distinct and complementary global quality functions, the *similarity* function F_G and the *dissimilarity* function D_G. F_G measures the intracluster distances, and D_G measures the intercluster distances. The cluster quality function F_{A_i} for cluster A_i is defined as the quadratic sum of all intracluster distances. It must be noticed that for these functions we are considering the distances in the table of distances. Although the partition provided by the Euclidean approach is based on an Euclidean table of distances, the quality function must be based on the table of distances, since it measures the actual network distances. The similarity global function for the final partition F_G is computed as the sum of all the F_{A_i} values divided by the total number of intracluster distances existing in partition P and normalized by the quadratic average value of all of the distances between the network nodes. For the dissimilarity global function we define the cluster dissimilarity function D_{A_i} for a cluster A_i as the quadratic sum of all intercluster distances from nodes in cluster A_i to all the nodes in the rest of the clusters. The dissimilarity global function D_G is defined as the sum of all the D_{A_i} values divided by the total number of existing intercluster distances in partition P and normalized by the quadratic average value of all of the distances between the network nodes. F_G and D_G provide a measurement of the intracluster and intercluster communication costs, respectively. Thus, the quotient D_G/ F_G provides the relationship between the intracluster and intercluster bandwidth achieved with each process mapping. We will denote this relationship as the *clustering coefficient* C_c. This clustering coefficient can be used to measure the quality of each process mapping.

3 Performance Evaluation

We have evaluated the improvement in network performance that the proposed clustering approach can provide, as well as the correlation between the clustering coefficient and network performance. This study assumes that all the communication between processors is intracluster communication and all the processors transmit the same amount of information. 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 network is composed of a set of switches. The network topology is irregular and has been generated randomly. We assumed 8-port switches. Each switch has 4 ports available to connect to other switches. From these 4 ports, three of them are used in each switch when the topology is generated. The remaining port is left open. We have evaluated networks with a size ranging from 16 switches (64 nodes) to 24 switches (96 nodes). Several distinct topologies have been analyzed. For the sake of simplicity, we have assumed a fixed pool of N processes grouped into 4 clusters with $\frac{N}{4}$ processes, where N is the number of network nodes. Each process is assumed to send all of the generated messages to processes in the same logical cluster of processes. For each network, we have computed the clustering algorithm until it has provided a 4-cluster partition of the network, and then we have chosen several possible mappings based on this partition. Additionally, we have computed several random mappings for each considered network.

Figure 1 shows network performance for some of the mappings based on the partition provided by the Euclidean approach (E_i labels) for a 16–switch network, compared with the network performance obtained by several randomly generated mappings (R_i labels). The clustering coefficient C_c obtained for each mapping is shown on the right side of each plot label. The network throughput obtained with any of the mappings based on the proposed approach is about a 55% higher than the network throughput obtained with any of the randomly generated mappings, while the network latency is less

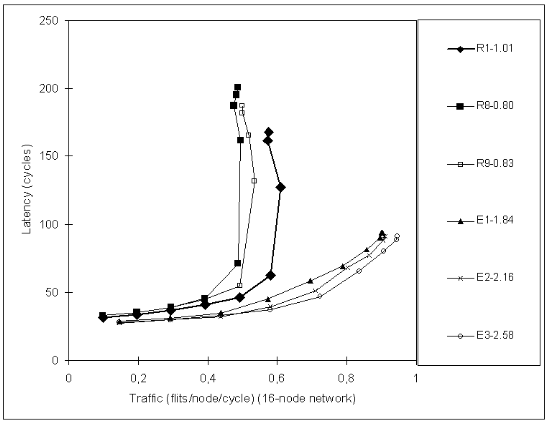

Fig. 1. Simulation results for a 16-switch network

than 63%. On the other hand, the value of C_c is clearly lower for randomly generated mappings, showing that this function is directly related to network performance.

We have also studied the correlation of the clustering coefficient C_c with network performance. We computed the correlation index between the clustering coefficient and the network performance obtained for all the mappings in all of the considered networks. In any case this index resulted higher than 80% for simulation points of both low network load and network saturation. These results validates the clustering coefficient as an "a priori" measure of relative network performance.

4 Conclusions

Network throughput and network latency are greatly improved when using the mappings based on the proposed approach, showing that it can be used as the basis for a communication-aware mapping technique. We have also studied the correlation between the proposed clustering coefficient and network performance. The results show that when only exists intracluster communication then this coefficient is highly correlated with network performance. For further information, please see technical report TR-AR99 on *http://www.gap.upv.es*

References

1. V. Arnau et al., "On the Characterization of Interconnection Networks with Irregular Topology: A New Model of Communication Cost", in *PDCS'99*, November 1999.
2. M. S. Bazaraa et al., *Nonlinear Programming: Theory and Algorithms*, *J. Wiley, 1993*.
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. R. Duda, P. Hart, *Pattern Classification and Scene Analysis*, *J. Wiley, 1973*.