An Approach to Mapping Parallel Programs on Hypercube Multiprocessors

Aguilar Jose
CEMISID. Dpto. de Computación.
Facultad de Ingeniería. Universidad de los Andes.
Av. Tulio Febres. Mérida, Edo. Mérida-Venezuela
Telf: (58.74)440002   Fax:(58.74)402872
email: aguilar@ing.ula.ve

Abstract

In this work, we propose a heuristic algorithm based on Genetic Algorithm for the task-to-processor mapping problem in the context of local-memory multiprocessors with a hypercube interconnection topology. Hypercube multiprocessors have offered a cost effective and feasible approach to supercomputing through parallelism at the processor level by directly connecting a large number of low-cost processors with local memory which communicate by message passing instead of shared variables. We use concepts of the graph theory (task graph precedence to represent parallel programs, graph partitioning to solve the program decomposition problem, etc.) to model the problem. This problem is NP-complete which means heuristic approaches must be adopted. We develop a heuristic algorithm based on Genetic Algorithms to solve it.

1. Introduction

The advent of cost-effective VLSI components in the past few years has made feasible the commercial development of massively parallel computers with hundreds of processors. Because of appealing properties such as node and edge symmetry, logarithmic diameter, high fault resilience, scalability, and the ability to host popular interconnection networks, namely ring, torus, tree and linear array, *hypercube multiprocessors* has been the focus of many researchers over the past few years [2, 5, 6]. This topology has result in several commercial product (Origin2000, Intel iPSC, NCUBE/10, Caltech/JPL, etc.).

Conceptually, the hypercube interconnection network is a multidimensional binary cube with a processor or procesors cluster at each of its vertices. An n-dimmensional hypercube has 2^n processors or processor clusters and $n2^{n-1}$ links. Each processor or processors cluster has its own local memory and interprocessor communication is done by explicit message passing directly or through some intermediate processors. This type of architecture is more readily scaled up to very large numbers of processors than multiprocessors designs based on globally shared memory [4, 6].

The effective exploitation of the potential power of this type of parallel architecture requires efficient solutions to the task-to-procesor mapping problem. The problem is that of optimally allocating the tasks of parallel program among the processors in order to minimize the execution time of the program. The mapping problem is NP-complete [4, 7, 8, 9, 10] which means heuristics approaches must be adopted. The mapping of the tasks to processors may either be performed statically (before program execution) or dynamically in an adaptive manner as the parallel program executes. The appropriate approach depends on the nature of the model of the parallel programs. If the characterization of the parallel program, i.e. the dependence between tasks and their execution time can be accurately estimated a priori, then a static approach is more attractive, since the mapping computation need only be performed once. We only consider static mapping scheme in this paper.

In this work, the task-to-processor static mapping problem in the context of a local-memory multiprocessors with a hypercube interconnection topology is solved using an algorithm based on Genetic Algorithms. We will model a parallel program execution as an acyclic directed graph whose nodes represent the tasks with known (or estimated) computation times, and whose arcs represent the precedence relations between tasks, that is the explicit execution dependences [7, 8, 9, 10]. This graph is called *tasks graph*. Our approach is composed by two phases: in the first phase an initial tasks graph k-partitioning is done in a manner of minimize communication volume and load imbalance cost between the subgraphs (clusters), where *k* is the number of processors. In the second phase, task clusters are assigned among processors of the system with hypercube interconnection topology in a manner that minimize the communication distance between clusters.

This paper is organized as follows. In section 2 we formalized the mapping problem. In section 3 the theoretical basis of Genetic Algorithms are reviewed. Then, we present our algorithm. In section 4 we compare the effectiveness of our scheme with previous work [5, 7, 10]. Remarks concerning future work and conclusions are provided in section 5.

2. Mapping Problem

The parallel program is characterized by a task graph: $Gt = (N, A, C, t)$, where N = {1, ... ,n} is the set of *n* tasks that compose the program, and *C, t* denote the times related to task execution and to communication between tasks. Thus, each task *i* has a weight *C(i)* which defines its execution time, for i=1, ..., n. t_{ij} will denote the data communication requirements between tasks *j* and *i*. $A = \{a_{ij}\}$ is the adjacency matrix representing the precedence order between the tasks. Since the graph is acyclic, we may number the tasks in a manner such that $a_{ij}=0$ if i > j [7, 8, 9, 10].

The parallel computer is represented as a hypercube graph $G_m=(P, E)$ where *m* denotes the dimension. An hypercube of dimension *m* has 2^m nodes and $(m2^{m-1})=k$ edges. That is, the nodes $P = (1, ..., k)$ represent the processors and the edges *E* represent the communication links. The system is assumed to be homogeneous, with identical processors. Hence, in contraste to the task graph, no weights are associated with the nodes or edges of the processors graph. If the nodes are labeled from *0* to 2^m-1 in binary, then an edge connects two nodes if only if their binary labels differ in exactly one bit position. The hamming distance between two nodes equals the minimal length of any path connecting these nodes.

The problem is that of assigning the *n* tasks to *k* processors. This means that we have to create task clusters (Gt_1, ..., Gt_k) in a way which optimizes performance. The problem is then characterized by the following objectives:

- The load of the different task clusters must be balanced.
- The communication between different task clusters must be kept to a minimum.
- Two task clusters with communication between them must be mapped onto nearest-neighbor processors.

That is, the task-to-processor mapping is a function *M:N->P*. *M(i)* gives the processor onto which task *i* is mapped.The tasks cluster *p (TCp)* is defined as the set of tasks assigned to cluster *p*:

$$TCp= \{j \,/\, M(j) = p\} \qquad \text{for } p = 1, ..., k$$

The load of *TCp (L_TCp)* is the total execution time of all tasks assigned onto it:

$$L\_TC_p = \sum_{i \in TC_p} C(i)$$

and the idealized average load is given by

$$L\_TC = \sum_{p=1}^{k} (L\_TCp - \sum_{p=1}^{k} L\_TCp/k)^2$$

The communication between TCp and TCq is equal to

$$C\_TCpq = \sum_{i,j \in D} t_{ij}$$

where, D={(i ∈ TCp) & (j ∈ TCq) & (p .not equal. q) & (a_{ij} = 1 or a_{ji} = 1)}

The first and second contraint can be solved as

$$\min_{M} (L\_TC + \Sigma_{p,q}\, C\_TCpq) \qquad (1)$$

If we suppose that each task cluster must be mapped to a different processor, the nearest-neighbor approach can be solved using the next cost function

$$CCP = \sum_{p,q=1 \& p \neq q}^{k} C\_TCpq\ PATHpq$$

where, PATHpq is the minimal length of any path connecting nodes *p*
and *q* (hamming distance)

and, the function to be minimized is

$$\min_{M} (CCP) \qquad (2)$$

3. Our Approach

In this section, we present Genetic Algorithms and formalize our strategy to solve the mapping problem.

3.1 Genetic Algorithms

This is an optimization algorithm based on the principles of evolution in biology. A genetic algorithm (GA) follows an "intelligent evolution" process for individuals based on the utilization of evolution operators such as mutation, inversion, selection and crossover [1, 3, 7, 10]. The idea is to find the best local optimum, starting from a set of initial solutions (individuals), by applying the evolution operators to successive solutions so as to generate a new and better local minimum. The procedure evolves until it remains trapped in a local minimum. We can represent individuals as string. The main program for a GA is the following:

```
Generation of individuals which represent potential
        solutions
Repeat until system convergence
  Evaluation of every individual
  Selection of the best individual for reproduction
  Reproduction of the individual using the evolutive
        operators
  Replace the worst old individuals by the new
        individuals
```

In this work, we used the mutation, inversion and crossover operators. The crossover used is standard, a single cutting point chosen with uniform probability over the string length (individuals representation) and a swap of the genetic material following it. The mutation operator is the standard, which modifies each string element according to probability p_m. Under inversion operator two points are chosen along the length of the individual, the individual is cut at those points, and the end points of the cut section switch places. In this method three parameters are studied: the maximum number of generations (NUMGEN), the size of the population and the probability (PM) to use the mutation operator after the crossover operator.

3.2 Our Heuristic Algorithm

The mapping algorithm proceeds in two phases: An initial mapping is first generated by grouping tasks of the tasks graph into clusters in a manner that improves load balancing and communication cost. Then, clusters are assigned among processors in a manner that the nearest neighbor property is satisfied.

For the first phase (*Cluster formation)*, the tasks graph is partitioned into as many clusters as the number of processors. We define this problem as a graph partitioning problem, the objective is to split the tasks graph in several subgraphs, so as to minimize the cost of imbalance and the cost of connection (*communication cost*) between them (*cost function 1*). The GA applied in this case follows the next procedure: we define a search space of *n* vectors where everyone represents an individual, and every individual represents a possible solution (partition). Each vector has *n* elements (tasks) and every element has a value among *1...K*, according to the cluster to which it belongs. We begin with an initial population of individuals randomly defined and we choose the individuals with minimal cost for generating new individuals using the *mutation* and *crossover* operators. Since the population is constant, we substitute the worst individuals of initial solution by the best individuals generated.

For the second phase (*Processor Allocation*), the clusters generated in the first are allocated to some processor, one cluster per processor, in a manner that minimizes the intercluster communication path length. That is, an optimal mapping is generated by assigning clusters to processors in a manner to minimize the *cost function 2* (nearest-neighbor property). The GA applied in this case follows the next procedure: we define a search space of k vectors where everyone represents an individual, and every individual represents a possible solution (cluster assignment). Each vector has *k* elements (clusters) and every element has a value different among *1...K*, according to the processor to which it is assigned. So that, each cluster is assigned to a different processor. We begin with an initial population of individuals randomly defined and we choose the individuals with minimal cost for generating new individuals using the *inversion* operators. We use only this operator because it mades modifications in individuals which assure each cluster is assigned to a different processor. Since the population is constant, we substitute the worst individuals of initial solution by the best individuals generated.

4. Result Analysis

In this section, we compare the effectiveness of our algorithm (GA) with a mapping algorithm (NN) proposed in [5], using a number of sample Task Graphs. One of the sample is a finite element graph (figure 1) and one is a random graph. The first graph is representative of the kinds of graphs that result from exploiting parallelism from computation modeling physical systems by finite elements (it is not a directed graph). The last sample is completely a random acyclic directed graph. It is defined for the number of tasks in the graph (n) and the average degrees (d) of the tasks [7, 10].

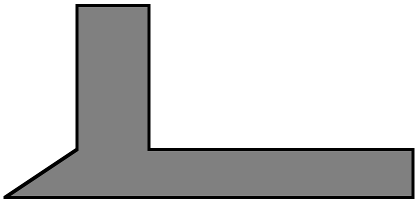

Figure 1. Sample problem graph used for performance evaluation.

We have used the parameters that give the best performances in GA according to the results of the works [7, 8, 9, 10]. NUMGEN allows to optimize the speed-up of the algorithm to reach an optimal solution. We remark than the quality of the solutions improves more rapidly in the first generations that in the following. Thus, a satisfactory quality can be obtained rapidly without to wait that the algorithm converges (NUMGEN=10). If the size of the population is large, we obtain better results, but the execution time if large. For small size is possible a rapid convergence but an optimal solution can not be found. We begin with an initial population of *n* and *k* individuals for each phase respectively. In the first phase, we used the crossover operator and then the mutation operator according to the PM probability. If PM is large we obtain good results, but it implies an execution time large. We define PM as 0,8.

The performance criteria studied are: execution time of the algorithms (in seconds) and cost function 2 value of the solutions. Mappings were generated for a target 16 and 32 processor hypercube system. The number of simulations per a given set of parameters is either (depending of which occurs first): 30 simulations or the number of simulation required to obtain a given standard deviation of the cost function 2 (σ). Due to space limitations, the results presented in this section were chosen because they are representative of the phenomena studied. We fix $\sigma = 0.1$. We have used a Ultra I SPARCstation with 32K RAM.

Both approaches result in mappings requiring many interprocessor communication messages (cost function2), that is due to load balancing constraint. In the first case (figure 2), NN gives the best results, and the execution time are similar. NN make a nearest-neighbor mapping that permits good solutions for high degree of locality and planar nature graphs (sample 1). Our approach explicitly minimizes total communication (length and volume), and NN approach minimizes total number of messages. This is the reason of the best results in the second case (figure 3) when we use our GA based heuristic (due to a low degree of random task graphs). In the second case (figure 3), NN execution times are smaller. That is due, GA need a large time to reach a good suboptimal solution.

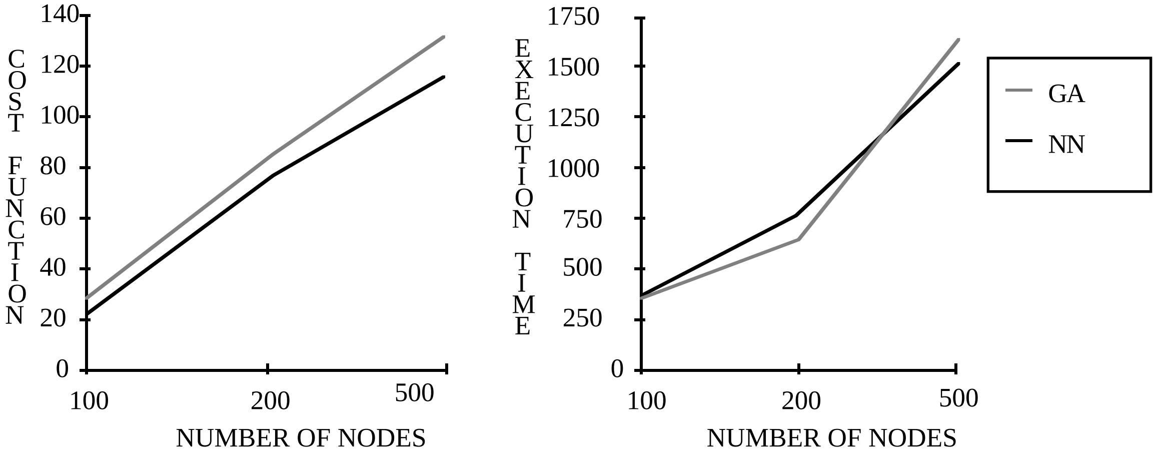

Figure 2. Results and Execution Time of the simulation for 16 processors and sample 1.

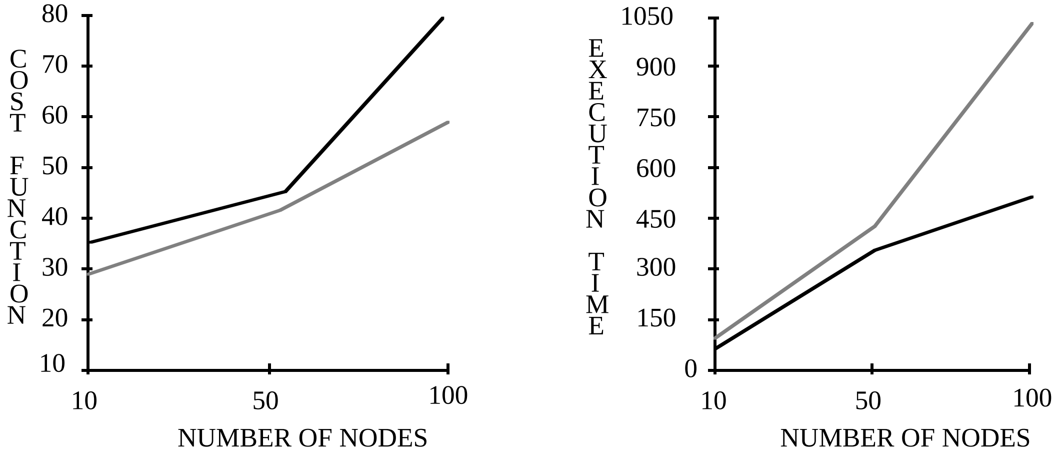

Figure 3. Results and Execution time of the simulation for 16 processors and random graphs with d = 3.

5. Conclusions

In this paper, we introduced a method to solve the task-to-processor mapping problem in the context of a local-memory multiprocessors with a hypercube interconnection topology. Our method explicitly minimizes total communication (length and volume) and load imbalance costs. The method is based on two phases: in the first phase an initial tasks graph k-partitioning is done in a manner of minimize communication (volume) and load imbalance costs between the subgraphs. In the second phase, task clusters are assigned among processors of the system with hypercube interconnection topology in a manner that minimizes communication length .

To evaluate the quality of the results obtained, we compared results obtained with a heuristic proposed in [5] for the same problem. The experiments we have run show that the results obtained by our method vary widely depending on the type and size of the graphs considered. Overall, GA appears to be preferable for random task graph with low messages between tasks. That is due to our approach uses a procedure that explicitly attempts to improve load balance. For de case of the sample 1 there is a lot of communication messages, and NN gives the best results because it minimizes the total number of messages.The Genetic Algorithm is easy to implement on a parallel machine, and this can considerably improve the speed obtained with our approach.

Acknowledgments

This work was supported by CONICIT grant S1-95000884, CDCHT-ULA grant I-503-95-A05 and CeCalCULA (High Performance Computing Center of Venezuela).

References

[1] M. Muhlenbein, G. Schleutter and D. Kramm. Evolution algorithms in combinatorial optimization. *Parallel Computing*, 7(2): 65-93, 1988.

[2] V. Lo. Heuristic Algorithms for task assignment in Distributed Systems. *IEEE Transaction on Computer*, 37:1384-1397, 1988.

[3] D. Golberg. *Genetic algorithms in search, optimization and machine learning*, Addison-Wesley, NY, 1989.

[4] K. Shin and M. Chen. On the number of acceptable task assignment in Distributed Computer Systems. *IEEE transaction on Computer*, 39(1), 1990.

[5] P. Sadayappan, F. Ercal and J. Ramanujam. Cluster partitioning approaches to mapping parallel programs onto a hypercube. *Parallel Computing*, 13:1-16, 1990.

[6] N. Bowen and C. Nikolau. On the assignment problem of arbitrary process systems to heterogeneous Distributed Computer Systems. *IEEE Transaction on Computers*, 41(3), March, 1992.

[7] J. Aguilar. Heuristics to optimize the Speed-up of Parallel Programs. *Lecture Notes in Computer Science*, 1127:174-183, 1996.

[8] J. Aguilar and T. Jimenez. A Processor Management System for PVM. *Lecture Notes in Computer Science*, 1300:158-161, 1997.

[9] J. Aguilar. Estudio del Problema de Asignación de Tareas en los Sistemas Distribuidos: funciones de costo y métodos de resolución", *Revista Técnica de Ingeniería*, Universidad del Zulia, 20(3): 203-213, 1997.

[10] J. Aguilar. and E. Gelenbe. Task Assignment and Transaction Clustering Heuristics for Distributed Systems. *Information Sciences*, 97(2):199-219, 1997.