Available online at www.sciencedirect.com
SCIENCE DIRECT®
Parallel Computing 31 (2005) 711–735
PARALLEL COMPUTING
www.elsevier.com/locate/parco

Heuristics for work distribution of a homogeneous parallel dynamic programming scheme on heterogeneous systems ☆

Javier Cuenca [a], Domingo Giménez [b,*], Juan-Pedro Martínez [c]

[a] *Departamento de Ingeniería y Tecnología de Computadores, Universidad de Murcia, Spain*
[b] *Departamento de Informática y Sistemas, Universidad de Murcia, Spain*
[c] *Departamento de Estadística y Matemática Aplicada, Universidad Miguel Hernández, Alicante, Spain*

Received 31 October 2004; received in revised form 15 March 2005; accepted 15 April 2005
Available online 15 June 2005

Abstract

In this paper the possibility of including automatic optimization techniques in the design of parallel dynamic programming algorithms in heterogeneous systems is analyzed. The main idea is to automatically approach the optimum values of a number of algorithmic parameters (number of processes, number of processors, processes per processor), and thus obtain low execution times. Hence, users could be provided with routines which execute efficiently, and independently of the experience of the user in heterogeneous computing and dynamic programming, and which can adapt automatically to a new network of processors or a new network configuration.

Keywords: Auto-tuning; Heterogeneous computing; Parallel dynamic programming; Parameter optimization; Performance modelling; Scheduling; Load-balancing

☆ Expanded version of a talk given at the HeteroPar 2004 Workshop, Cork, Ireland, July 5–8, 2004.

* Corresponding author. Tel.: +34 968 36 4611; fax: +34 968 36 4151.
E-mail addresses: javiercm@ditec.um.es (J. Cuenca), domingo@dif.um.es (D. Giménez), jp.martinez@umh.es (J.-P. Martínez).

0167-8191/$ - see front matter
doi:10.1016/j.parco.2005.04.005

1. Introduction

Automatic tuning techniques have been used in the design of parallel routines in recent years. Techniques have been developed in different fields [1,2], and especially in linear algebra routines [3–5].

The selection of appropriate values of some parameters which influence the execution time of the algorithm in parallel homogeneous systems has been analyzed in previous works. Linear algebra routines [4] and a Dynamic Programming Scheme [6] have been studied.

In this paper the possibility of applying the automatic optimization techniques studied for homogeneous systems to heterogeneous networks is analyzed. In this new scenario the decisions to be taken to reduce the execution time are: the processors to use, the number of processes and their distribution in the processors system. In that way, the ideas previously presented for homogeneous systems are used in combination with a problem of assignation of tasks to processors.

Section 2 analyzes the necessary modifications to the methodology used for homogeneous systems to adapt it to heterogeneous networks. In Section 3 the use of these ideas in a Dynamic Programming Scheme is considered. Experimental results and the results of some simulations are shown in Sections 4 and 5. Finally, Section 6 presents some conclusions.

2. Automatic optimization on heterogeneous systems

To solve problems on heterogeneous systems efficiently, one possibility is to design specific algorithms for this type of systems, with a heterogeneous distribution of data between the processes. Some work is being devoted to this goal [7–9].

Another approach could be to develop some techniques to assign processes to processors, and to use homogeneous algorithms (parallel algorithms where the processes of the logical topology are considered as identical) for the solution of the problems. Existing routines could be used efficiently in heterogeneous systems with a low additional programming cost.

The first approach would lead to efficient algorithms but at a high cost of reprogramming classical algorithms. For the second approach to be successful, it is necessary to obtain a good distribution strategy, i.e. a strategy with low execution time which produces a satisfactory load balancing. A number of equal (or approximately equal) processes are generated and these processes are assigned to some particular processors in the system. Parallel systems are used to reduce the execution time needed to solve a problem, and thus the selection of the processors to use, the number of processes, and the processes to processors assignation must be made in order to reduce the execution time, and without a large execution time in the decision part. For that, theoretical models of the execution time are used. In these models appear parameters which represent the characteristics of the system. When the system is heterogeneous, distributed or load-variable, the values of these parameters vary from one processor to another, statically and dynamically, and also with the