Optimization of Algorithms and Parallel Applications in Heterogeneous Systems Through the Combined Use of Formal Models of Computing and Communications.

Researchers:

- Juan Antonio Rico Gallego (Principal Investigator). University of Extremadura.
- Juan Carlos Díaz Martín University of Extremadura.
- Juan Luis García Zapata. University of Extremadura.
- Javier Corral García. COMPUTAEX Foundation.
- Jesús Calle Cancho. COMPUTAEX Foundation.
- Carmen Calvo Jurado. University of Extremadura.
- Jesús Álvarez Llorente. University of Extremadura.

Idioma Inglês

Descrição:

The hardware systems used in high-performance parallel computing are heterogeneous. Data centers and big facilities have incorporated machines that combine multi-core processors and accelerators (GPUs, etc.) connected through different communication channels, mainly shared memory and high-performance networks such as Infinband. The objective of this proposal is twofold: improving the current supercomputing performance, increasingly heterogeneous, and at the same time, reducing its energy consumption.

Usually, the applications that run on these platforms use a sequence of computing and communication phases. Their codes are composed of a series of patterns or kernels, such as a matrix multiplication or a Fourier transform. These kernels are executed in parallel by processes that run on heterogeneous hardware, and therefore, with different computing capabilities. In the current state of the art, the execution time of the kernel is optimized by a non-uniform distribution of the computation load, assigning to each process the work that can be done so that the overall load is balanced, with the aim that all the processes arrive at the same time to the communication phase, preventing faster processes from waiting for the slowest ones.

In the current heterogeneous supercomputing systems, the load balancing rule is necessary but insufficient, therefore this is the objective of the project. The problem is well known and begins to be the subject of attention by the research community. The non-uniform distribution of the workload not only produces differences in the volume of data that each process transmits, but also determines the use of the different communication channels of the system.

Consequently, finding an optimal load allocation requires not only knowing the load capacity of the processors involved, but also estimating the cost of communications derived from each allocation. Currently, the search for the best configuration is made through expensive tests that make extensive use of the computational resources of the system. This project proposes the use of an analytical model, for the prediction of communication costs, developed at the University of Extremadura (UEx) and the University College of Dublin (UCD) of Ireland. The model is called τ-Lop. Evaluated in multi-core clusters with high performance networks, τ-Lop has been published in two prestigious journals (first quartile and first decile), and has been the subject of a doctoral thesis with European mention in the UEx. It is considered that the approach based on τ-Lop will make it possible to automate the process of optimal allocation of load in heterogeneous architectures, avoiding the intensive consumption of resources that the tests require, and obtaining a significant improvement of the performance during the execution of the application. This will result in significant savings in both computational and energy costs in supercomputing facilities.

The aim of the project is to develop an user tool that proposes a load distribution taking into account the computational and communication models of the application. The integration of both models is the central part of the proposal. It would suppose the first model of these characteristics that not only could be used to improve the performance and the energetic consumption, but also in simulators and optimization algorithms.

The proposal is based on the work carried out by the main researcher in the development of his doctoral thesis. It relies on a multidisciplinary team that includes experts from different fields, and that works regularly with national and international researchers to tackle a project with two main milestones, the development of a load allocation mixed formal model, and its subsequent application as a software tool usable by the end user of a supercomputing center.

Funding sources:

- Project co-financed by the Junta of Extremadura and the European Regional Development Fund (ERDF) of Extremadura to 80 percent, within the Thematic Objective 01 "Reinforcement of research, technological development and innovation", through the call for grants for the realization of research projects, oriented towards the strategic areas of the regional economy contemplated in the V Regional R+D+i Plan (2014-2017), in public R+D+i centers of the Autonomous Community of Extremadura, under Decree 68/2016 of June 6.

URL de origem:https://www.cenits.es/pt-pt/node/1607