Abstract

In recent years, the growing need for computing capacity has become a challenge that has led the industry to look for alternative architectures to conventional out-of-order superscalar processors, with the goal of enabling an increase of computing power while achieving higher energy efficiency.

GPU architectures, which just a decade ago were applied to accelerate computer graphics exclusively, have been one of the most employed alternatives for several years to reach the mentioned goal. A particular characteristic of GPUs is their high main memory bandwidth, which allows executing a large number of threads in a very efficient way. This feature, as well as their high computational power regarding floating-point operations, have caused the emergence of the *GPGPU computing* paradigm, where GPU architectures perform general purpose computations. The aforementioned characteristics make GPU devices very appropriate for the execution of massively parallel applications that have been traditionally executed in conventional high-performance processors.

The work performed in this thesis aims to help improve the performance of GPUs in the execution of GPGPU applications. To this end, as a first step, a characterization study is carried out. In this study, the most important features of GPGPU applications, with respect to the memory hierarchy and its impact on performance, are identified. For this purpose, a detailed cycle-accurate simulator is used to model the architecture of a recent GPU. The study reveals that it is necessary to model with more detail some critical components of the GPU memory hierarchy in order to obtain accurate results. In addition, it shows that the achieved benefits can vary up to a factor of 3× depending on how these critical components are modeled.

Due to this reason, as a second step before realizing a novel proposal, the work in this thesis focuses on determining which components of the GPU memory hierarchy must be modeled with more detail to increase the accuracy of simulator results and improving the existing simulator models of these components. Moreover, a validation study is performed comparing the results obtained with the improved GPU models against those from a real commercial GPU. The implemented simulator improvements reduce the deviation of the results obtained with the simulator from results obtained with the real GPU by about 96%.

Finally, once simulation accuracy is increased, this thesis proposes a novel approach, called FRC (*Fetch and Replacement Cache*), which highly improves the GPU computational power by enhancing main memory-level parallelism. The proposal increases the number of parallel accesses to main memory by accelerating the management of fetch and replacement actions corresponding to those cache accesses that miss in the cache. The FRC approach is based on a small auxiliary cache structure that efficiently unclogs the memory subsystem, enhancing the GPU performance up to 118% on average compared to the studied baseline. In addition, the FRC approach reduces the energy consumption of the memory hierarchy by a 57%.