Performance Prediction Modeling of GPU Applications: Analytical Modeling and Machine Learning

Marcos Amarís González*, Raphael Y. de Camargo†, Alfredo Goldman*
* Institute of Mathematics and Statistics
University of São Paulo - São Paulo, Brazil
{amaris, gold}@ime.usp.br
† Center for Mathematics, Computation and Cognition
Universidade Federal do ABC - Santo André, Brazil
raphael.camargo@ufabc.edu.br

ABSTRACT

Today, most High Performance Computing (HPC) platforms have heterogeneous hardware resources (CPUs, GPUs, storage, etc.) Graphics Processing Units (GPU) are specialized coprocessor in accelerating vector operations in parallel. The prediction of application execution times over these devices is a great challenge and is essential for efficient job scheduling. There are different approaches to do this, such as analytical modeling and machine learning techniques. Analytic predictive models are useful, but require manual inclusion of interactions between architecture and software, and may not capture the complex interactions in GPU architectures. Machine learning techniques can learn to capture these interactions without manual intervention, but may require large training sets.

This document shows the summary of two main works. In the first work, we present the comparison of a developed BSP-based model to three different ML techniques, this comparison was done with 9 well-known matrix/vector applications. In this first work, we wanted to perform a fair comparison, for this reason, we decided that ML process would had the same features that the BSP-based model.

In the second work, we have compared among ML techniques. Here, a two step of extraction features are done. First a correlation analysis and after hierarchical clustering analysis. In this second work, 10 irregular CUDA kernels from 6 applications of Rodinia Benchmark suite were used. Next sections will present more details of these two works separately.