Abstract

Using Graphics Processing Units (GPUs) to solve general purpose problems has received significant attention both in academia and industry. Harnessing the power of these devices however requires knowledge of the underlying architecture and the programming model. In this paper, we develop analytical models to predict the performance of GPUs for computationally intensive tasks. Our models are based on varying the relevant parameters - including total number of threads, number of blocks, and number of streaming multi-processors - and predicting the performance of a program for a specified instance of these parameters. The approach can be used in the context of heterogeneous environments where distinct types of GPU devices with different hardware configurations are employed.