DOCTORAL THESIS

Performance and power modeling of GPU systems with dynamic voltage and frequency scaling

Wang, Qiang

Date of Award:
2020

Abstract

To address the ever-increasing demand for computing capacities, more and more heterogeneous systems have been designed to use both general-purpose and special-purpose processors. The huge energy consumption of them raises new environmental concerns and challenges. Besides performance, energy efficiency is another key factor to be considered by system designers and consumers. In particular, contemporary graphics processing units (GPUs) support dynamic voltage and frequency scaling (DVFS) to balance computational performance and energy consumption. However, accurate and straightforward performance and power estimation for a given GPU kernel under different frequency settings is still lacking for real hardware, which is essential to determine the best frequency configuration for energy saving.

In this thesis, we investigate how to improve the energy efficiency of GPU systems by accurately modeling the effects of GPU DVFS on the target GPU kernel. We also propose efficient algorithms to solve the communication contention problem in scheduling multiple distributed deep learning (DDL) jobs on GPU clusters. We introduce our studies as follows.

First, we present a benchmark suite EPPMiner for evaluating the performance, power, and energy of different heterogeneous systems. EPPMiner consists of 16 benchmark programs that cover a broad range of application domains, and it shows a great variety in the intensity of utilizing the processors. We have implemented a prototype of EPPMiner that supports OpenMP, CUDA, and OpenCL, and demonstrated its usage by three showcases. The showcases justify that GPUs provide much better energy efficiency than other types of computing systems, and especially illustrate the effectiveness of GPU Dynamic Voltage and Frequency Scaling (DVFS) on the energy efficiency of GPU applications.

Second, we reveal a fine-grained analytical model to estimate the execution time of GPU kernels with both core and memory frequency scaling. Compared to the cycle-level simulators, which are too slow to apply on real hardware, our model only needs one-off micro-benchmarks to extract a set of hardware parameters and kernel performance counters without any source code analysis. Our experimental results show that the proposed performance model can capture the kernel performance scaling behaviors under different frequency settings and achieve decent accuracy.

Third, we design a cross-benchmarking suite, which simulates kernels with a wide range of instruction distributions. The synthetic kernels generated by this suite can be used for model pre-training or as supplementary training samples. We then build machine learning models to predict the execution time and runtime power of a GPU kernel under different voltage and frequency settings. Validated on three modern GPUs with a wide frequency scaling range, by using a collection of 24 real application kernels, the model trained only with our cross-benchmarking suite is able to achieve considerably accurate results.

At last, we establish a new DDL job scheduling framework which organizes DDL jobs as Directed Acyclic Graphs (DAGs) and considers communication contention between nodes. We then propose an efficient job placement algorithm, Least-Workload-First-κ (LWF-κ), to balance the GPU utilization and consolidate the allocated GPUs for each job. When scheduling the communication tasks, we propose Ada-SRSF for the DDL job scheduling problem to address the communication contention issue. Our simulation results show that LWF-κ achieves up to 1.59$\times$ improvement over the classical first-fit algorithms. More importantly, Ada-SRSF reduces the average job completion time by up to 36.7%, as compared to the solutions of either avoiding all the communication contention or accepting all of it.

Keywords: Benchmark Suite, GPU Dynamic Voltage and Frequency Scaling, GPU Performance and Power Modeling, Job Scheduling

Table of Contents