

DOCTORAL THESIS

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

Wang, Qiang

Date of Award:
2020

[Link to publication](#)

General rights

Copyright and intellectual property rights for the publications made accessible in HKBU Scholars are retained by the authors and/or other copyright owners. In addition to the restrictions prescribed by the Copyright Ordinance of Hong Kong, all users and readers must also observe the following terms of use:

- Users may download and print one copy of any publication from HKBU Scholars for the purpose of private study or research
- Users cannot further distribute the material or use it for any profit-making activity or commercial gain
- To share publications in HKBU Scholars with others, users are welcome to freely distribute the permanent URL assigned to the publication

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

Declaration	i
Abstract	ii
Acknowledgements	iv
Table of Contents	vi
List of Tables	x
List of Figures	xii
Chapter 1 Introduction	1
1.1 Energy Conservation by GPU DVFS	2
1.2 Performance Modeling for GPU DVFS	2
1.3 Power Modeling for GPU DVFS	6
1.4 High Performance DDL Training System	8
1.5 Thesis Goals and Contributions	10
1.6 Thesis Organization	12
Chapter 2 Background and Literature Review	14
2.1 Performance and Power Benchmarking	14
2.2 GPU Computing and DVFS	16
2.3 GPU Performance Modeling	18
2.4 GPU Power Modeling	21

2.5	Distributed Deep Learning Training System	23
2.5.1	Efficient Systems for A Single DDL Job	23
2.5.2	Efficient Systems for Multiple DDL Jobs	24
Chapter 3 EPPMiner		26
3.1	The EPPMiner Benchmark	26
3.1.1	Description of selected applications and the workload	26
3.1.2	Design of performance metrics	29
3.1.3	Performance and power measurements	31
3.2	Showcases	31
3.2.1	Experimental testbed	32
3.2.2	Showcase I: Comparison of different devices	33
3.2.3	Showcase II: Impact of multi-threading on performance/power/energy	36
3.2.4	Showcase III: Impact of DVFS on energy efficiency	38
3.3	Summary	42
Chapter 4 GPGPU Performance Estimation with Core and Memory Frequency Scaling		43
4.1	Memory Modeling with Frequency Scaling	45
4.1.1	Global Memory Access Latency	45
4.1.2	DRAM Latency	46
4.1.3	L2 Cache Latency	50
4.1.4	Adjustment with Frequency Scaling	51
4.1.5	High Memory Bandwidth based GPUs	52
4.2	Graphics Processing Unit Performance Modeling with Frequency Scaling	52
4.2.1	Case 1: \bar{L}_{base}^m can be hidden	53
4.2.2	Case 2: \bar{L}_{base}^m cannot be hidden	55
4.2.3	The Effects of Core and Memory Frequency Scaling	57

4.3	Experiments	60
4.3.1	Experimental Methodology	60
4.3.2	Experimental Results	61
4.3.3	Case Study of Energy Conservation	68
4.3.4	Discussion	71
4.4	Summary	73
Chapter 5 Machine-learning based GPGPU performance & Power estimation with frequency scaling		74
5.1	Cross-Benchmarking Suites	74
5.2	Modeling GPU Performance with Machine Learning Methods	78
5.3	Modeling GPU Power with Machine Learning Methods	82
5.4	Experiments	86
5.4.1	Experimental Methodology	86
5.4.2	Experimental Results	90
5.5	Summary	95
Chapter 6 Communication Contention Aware Scheduling of Multiple Deep Learning Training Jobs on GPU Clusters		96
6.1	Preliminaries	97
6.1.1	Distributed Deep Learning Training	97
6.1.2	Communication Model	97
6.2	System Modeling and Problem Formulation	98
6.2.1	System Modeling	98
6.2.2	Problem Formulation	101
6.3	Solution	104
6.3.1	Placement	104
6.3.2	Scheduling	107
6.4	Performance Evaluation	113
6.4.1	Evaluation Setup	113

6.4.2	Evaluation Results	114
6.4.3	Discussion	117
6.5	Summary	118
Chapter 7 Conclusion and Future Work		119
7.1	Future Research Directions	120
Bibliography		122
Curriculum Vitae		138