Vrije Universiteit Amsterdam

Universiteit van Amsterdam

Master Thesis

GPU energy efficiency

An analysis of energy consumption, usage patterns and energy saving strategies

Author:
Quincy Bakker
q.bakker@students.uva.nl
q.bakker@student.vu.nl

1st supervisor: Ana Lucia Varbanescu
daily supervisor: Sagar Dolas (SURF)
2nd reader: N/A

A thesis submitted in fulfillment of the requirements for the joint UvA-VU Master of Science degree in Computer Science

July 19, 2020

"I am the master of my fate, I am the captain of my soul"
from Invictus, by William Ernest Henley

Abstract

Here goes the abstract of this thesis.

To ...

Acknowledgements

TODO

Contents

List of Figures

List of Tables

1

Introduction

1.1 Context

TODO

1.2 Objective

TODO

1.3 Research Questions

To address the goal of this study some research questions were formulated which help guide the extraction of data and conclusions from the literature that is reviewed. This section outlines those research questions and their underlying motivation.

1.3.1 CPU-GPU Workload Characterization

TODO

RQ1: What are the different types of CPU-GPU workloads?

RQ2: How can CPU-GPU workloads be detected?

1.3.2 Energy Saving When Running CPU-GPU Workloads

TODO

RQ3: Can information about the CPU-GPU workload be used to save energy?

1.4 Research Method

This section describes the research method that was used.

1.4.1 Literature Study

In this section we first provide a description of the process by which literature was collected for the purpose of this study to ensure its replicability.

1.4.1.1 Search

The search for literature was conducted using Google Scholar[1]. Google Scholar works similarly to Google in that it uses a search query and presents relevant results depending on the input. Search queries on Google Scholar can be built from keywords and Boolean operators such as OR to provide constraints to the search query [5].

To build the search query, keywords are extracted from the research questions so that those keywords can be used as a base for a search query on Google Scholar. This results in the following keywords along with any potential synonyms:

- CPU
- GPU
- workload
- analysis/analyzing/characterize/characterization
- energy/power
- saving/conservation

From these keywords the following search queries were then constructed:

SQ1: CPU GPU workload analysis OR analyzing OR characterize OR characterizing OR characterization

SQ2: CPU GPU energy OR power saving OR conservation

[1] `www.scholar.google.com`

1.4.1.2 Gathering Literature

To initial round of literature gathering was performed with the assistance of a software tool used for bibliography management called Mendeley [1]. In this tool, four categories were created to organize the literature:

Unread: Literature that was gathered from a search query but that has not yet been read.

Related: Literature that has been read and is indirectly related to this literature review.

Selected: Literature that was read and that matches the inclusion criteria.

Not selected: Literature that was read but that does not match the inclusion criteria.

All papers that were found during the initial search were placed in the unread category, after which they were moved to another category depending on the contents of the paper and their applicability to the topic of this literature review.

Snowballing To gather more relevant literature the snowballing technique was used, which is the process of gathering additional literature from the references of a paper.

1.4.1.3 Application of Selection Criteria

In order to restrict the amount of papers that need to be processed and to filter out any irrelevant papers selection criteria were used. To this end, the literature review process was conducted by looking for papers that fulfil all of the specified inclusion criteria while matching none of the specified exclusion criteria. This section outlines those criteria and the reasoning behind them. Most of these criteria were sourced from the research questions and are meant to help answer them.

Inclusion Criteria At least one of these inclusion criteria must be fulfilled by each of the papers selected:

IC1: The study covers methods of CPU-GPU energy consumption measurement

IC2: The study covers methods to characterize CPU-GPU workloads

IC3: The study covers methods to save energy when running CPU-GPU workloads

[1] `www.mendeley.com`

Exclusion Criteria None of these exclusion criteria must be fulfilled by each of the papers selected:

EC1: The study does not make use of the GPU

1.4.1.4 Data Extraction

TODO

1.4.1.5 Data Synthesis

TODO

2

Background

This chapter outlines some of the research and other resources that are relevant to the topic of Graphics Processing Unit (GPU) energy conservation.

2.1 Energy Consumption

This section outlines some of the work that has been done to measure and predict energy consumption.

2.1.1 Measuring

Measuring live energy consumption is an important aspect of many power saving strategies. There exist tools that can perform these types of measurement, the most important of which are outlined in this section.

2.1.1.1 NVIDIA System Management Interface

NVIDIA's System Management Interface (SMI) tool is a command line utility that is able to query the GPU device state [4]. Support is limited to NVIDIA GPUs. What makes this tool useful to this research is the fact that it can retrieve the current power consumption from the GPU as it is running and that it can output this information to the console, which makes it possible to easily integrate the output programmatically.

2.1.2 Workload Analysis

An important component in any energy saving strategy is to perform a workload analysis, since the decisions that are made often depend on the type of workload that is running [2].

2.1.2.1 GPGPUSim

GPGPUSim is a tool that can be used to simulate a GPU and run synthetic workloads. It offers a lot of detailed insights that can be used for workload analysis [1].

2.1.2.2 Usage Patterns

TODO

2.1.3 Statistical Analysis and Prediction

Ma and Zhong [3] developed a method to statistically analyze and model the power consumption of a mainstream GPU. To achieve this they make use of the fact that there exists an innate coupling among the power consumption characteristics, runtime performance and dynamic workloads. They found that their model is capable of robustly and accurately predicting the dynamic power consumption estimation of a target GPU at runtime, especially for graphics applications.

Ma and Zhong [3] state that due to the relatively simpler cache hierarchy, higher level of parallelism, less complex control requirements, and more computation units, GPU power modeling differs from general-purpose processing units. Some limitations of their approach they state are that micro architectural knowledge of the GPU is needed to provide more complex and accurate modeling approaches, and that quantitative analysis of GPU workloads and statistical selection of the power consumption correlated workloads are necessary in the data preprocessing step.

Chen et al. [2] also developed a method to statistically analyze GPU power consumption. They designed a high-level GPU power consumption model using sophisticated tree-based random forest methods which can correlate the power consumption with a set of independent performance variables. Their model is able to accurately predict GPU runtime power consumption and provides insights for understanding the dependence between the GPU runtime power consumption and the individual performance metrics. To gain detailed insights they used a GPU simulator, *GPGPUSim* [1].

2.2 Energy Saving

TODO

2.2.1 Dynamic Voltage and Frequency Scaling

Dynamic Voltage and Frequency Scaling (DVFS) is a technique that

3

Usage Patterns

TODO

4

Energy Saving Strategies

TODO

5

Dynamic Energy Saving

TODO

6

Discussions

TODO

7

Conclusion

TODO

Appendix A

TODO

References

[1] Ali Bakhoda et al. "Analyzing CUDA workloads using a detailed GPU simulator". In: *ISPASS 2009 - International Symposium on Performance Analysis of Systems and Software*. 2009, pp. 163–174. ISBN: 9781424441846. DOI: 10.1109/ISPASS.2009.4919648. URL: https://www.microsoft.com/en-us/research/wp-content/uploads/2017/02/gpgpusim.ispass09-2.pdf.

[2] Jianmin Chen et al. "Statistical GPU power analysis using tree-based methods". In: *2011 International Green Computing Conference and Workshops, IGCC 2011*. 2011. ISBN: 9781457712203. DOI: 10.1109/IGCC.2011.6008582.

[3] Xiaohan Ma and Lin Zhong. "Statistical Power Consumption Analysis and Modeling for GPU-based Computing". In: *Proceedings of the SOSP Workshop on Power Aware Computing and Systems (HotPower '09)* (2009), None. URL: https://www.yecl.org/publications/ma09hotpower.pdf%20http://www.sigops.org/sosp/sosp09/hotpower.html.

[4] NVIDIA. *NVIDIA System Management Interface | NVIDIA Developer*. URL: https://developer.nvidia.com/nvidia-system-management-interface (visited on 07/18/2020).

[5] Daniel M. Russel. *Google Advanced Search Operators*. URL: https://docs.google.com/document/d/1ydVaJJeL1EYbWtlfj9TPfBTE5IBADkQfZrQaBZxqXGs (visited on 07/19/2020).

Statement of Originality

This document is written by Student Quincy Bakker who declares to take full responsibility for the contents of this document.

I declare that the text and the work presented in this document are original and that no sources other than those mentioned in the text and its references have been used in creating it.

The Faculty of Science is responsible solely for the supervision of completion of the work, not for the contents.