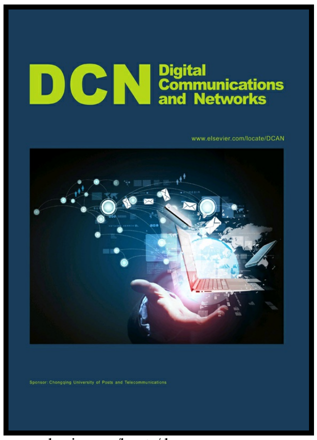
Author's Accepted Manuscript

A Survey and Measurement Study of GPU DVFS on Energy Conservation

Xinxin Mei, Qiang Wang, Xiaowen Chu



www.elsevier.com/locate/dcan

PII: S2352-8648(16)30073-6
DOI: http://dx.doi.org/10.1016/j.dcan.2016.10.001
Reference: DCAN55

To appear in: *Digital Communications and Networks*

Cite this article as: Xinxin Mei, Qiang Wang and Xiaowen Chu, A Survey and Measurement Study of GPU DVFS on Energy Conservation, *Digita Communications and Networks,* http://dx.doi.org/10.1016/j.dcan.2016.10.001

Digital Communications and Networks(DCN)

HOSTED BY
ELSEVIER
Available online at www.sciencedirect.com
ScienceDirect
journal homepage: www.elsevier.com/locate/dcan


A Survey and Measurement Study of GPU DVFS on Energy Conservation

Xinxin Mei[a], Qiang Wang[a], Xiaowen Chu*[ab]

[a]Department of Computer Science, Hong Kong Baptist University, Hong Kong S.A.R., China
[b]Institute of Research and Continuing Education, Hong Kong Baptist University, Shenzhen, 518057, China

Abstract

Energy efficiency has become one of the top design criteria for current computing systems. The dynamic voltage and frequency scaling (DVFS) has been widely adopted by laptop computers, servers, and mobile devices to conserve energy, while the GPU DVFS is still at a certain early age. This paper aims at exploring the impact of GPU DVFS on the application performance and power consumption, and furthermore, on energy conservation. We survey the state-of-the-art GPU DVFS characterizations, and then summarize recent research works on GPU power and performance models. We also conduct real GPU DVFS experiments on NVIDIA Fermi and Maxwell GPUs. According to our experimental results, GPU DVFS has significant potential for energy saving. The effect of scaling core voltage/frequency and memory voltage/frequency depends on not only the GPU architectures, but also the characteristic of GPU applications.

KEYWORDS:
Graphics processing unit, dynamic voltage and frequency scaling, energy efficiency

1. Introduction

The graphics processing units (GPUs) have become prevalent accelerators in current high performance clusters. They substantially boost the performance of a great number of applications in many commercial and scientific fields, such as bioinformatics [1, 2], computer communications [3, 4, 5], machine learning [6, 7, 8], especially the emerging deep learning [9, 10, 11]. In the TOP500 supercomputer list [12] as of June. 2016, 94 systems are equipped with accelerators and 69 out of them are equipped with GPUs [13]. The CPU-GPU hybrid computing is more energy efficient than traditional many-core parallel computing [13, 14]. However, this kind of high performance clusters still consume a lot of energy. To power the clusters remains a great expense. For example, the Titan supercomputer, 3rd in the TOP500 list as of this writing, is accelerated by 18,688 NVIDIA Tesla K20X with a power supply of 8.21 million Watts, which cost about 23 million dollars per year [15]. Given the fact that saving even a few percent of energy can reduce a large amount of electricity cost, efficient GPU power management becomes indispensable for GPU-accelerated data centers and supercomputers.

One of the promising power management strategies is the dynamic voltage and frequency scaling (DVFS) [16, 17], which refers to changing the processor voltage/frequency during task processing. It is effective in either saving energy or improving performance. The CPU DVFS technology is well developed and has been adopted in both personal computing devices and large scale clusters [18]. Despite the maturity of CPU DVFS, the GPU DVFS study started only a few years ago. According to existing studies, simply transplanting the CPU DVFS strategy to GPU platforms could be ineffective [19, 20]. For example, scaling up the processor frequency (described as "racing" in [21]) is proved to be energy efficient for the CPUs but not always for the GPUs [21, 22]. We summarize some challenges of the GPU DVFS study as below. First, the GPU hardware and power management information is

*Corresponding author. Email: chxw@comp.hkbu.edu.hk.