Power, Performance and Energy Efficiency Evaluation of GPU Overclocking When Executing CUDA applications

Jun-hee LEE, Joon-ho KONG, Sung-woo CHUNG
(*Division of Computer and Communication Engineering, Korea University, Seoul 136-701, Korea*)

Abstract – Processor performance has continuously advanced since Moore's law. However, power constraints have become one of the largest concerns for processor architects to further improve processor performance. With the recent introduction of GPGPU (general purpose graphic processing unit) technology, which uses GPU for general computing, techniques to maximize GPU processor parallelism and to minimize power consumption have become popular. However, end users requiring better performance often use overclocking to improve CPU and GPU performance. This paper aims to evaluate the effect of GPU and CPU overclocking when executing CUDA applications from three perspectives: performance, power consumption, and energy efficiency. The experiment shows that when executing matrix multiplication program with CUDA, while execution time of GPU (GPU core: 625 MHz, GPU memory: 450 MHz) overclocking was similar to that of CPU overclocking, its energy consumption was 12.8% less. In terms of energy efficiency, higher GPU overclocking leads to better efficiency. We verify the positive results of GPU overclocking in execution time and energy efficiency when executing CUDA applications and the potential of GPU overclocking.

Key words – *overclocking*; *power efficiency*; *GPGPU computing*

Manuscript Number: 1674-8042(2011)supp1.-0046-03

doi: 10.3969/j.issn.1674-8042.2011.supp1.010

1 Introduction

Ever since the introduction of Moore's law, the number of transistors in processor has consistently doubled every 1.5 year. According to this trend, processor performance has continuously advanced. However, as processor performance improved, the concerns for energy consumption and temperature have increased[1]. With the increasing use of ILP (instruction level parallelism), power management has become a significant concern for processor architect. Also, with the growing temperature concerns, it has hit a wall to continue to increase clock frequency. Thus, the recent tendency is turning to integrating more cores per chip. As it is going beyond multi-core to many-core era, many diverse techniques using processor parallelism to increase performance have been introduced. However, application program to make full use of parallelism has not yet been introduced.

As concerns on power management have become important, there are attempts to use GPU (graphic processing unit) core not only for graphic computing but also for general computing. NVIDIA's CUDA[2] and ATI's STREAM[3] are some examples which provide API (application programming interface) to programmers thus enabling them to use GPU in general computing. These programming models tend to support utilizing GPU's feature of having many simple cores to increase the number of threads and maximizing use of parallelism. A number of researchers are already studying CUDA, and the number of institutes teaching CUDA is increasing.

Meanwhile, for end users, as perceptible performance has higher priorities than power consumption, an overclocking technique is widely used. Overclocking is the process of running a computer component at a higher clock frequency than it was designed but not interfering with CPU or GPU stability. The increased frequency results in higher performance. Moreover, a recent research also suggests overclocking can improve energy efficiency[4].

In this paper, we aim to compare and evaluate performance, power consumption, and energy efficiency of GPU and CPU overclocking when executing CUDA applications. The rest of our paper is organized as follows. Section 2 explains experiment environment and section 3 evaluates performance and power management, and section 4 concludes with the findings.

* **Received**: 2011-09-05
Project supported: This work was supported by the MKE (The Ministry of Knowledge Economy), Korea, under the ITRC (Information Technology Research Center) support program supervised by the NIPA (National IT Industry Promotion Agency) (NIPA-2011-C1090-1121-0010)
Corresponding author: Sung-woo CHUNG (swchung@korea.ac.kr)

2 Evaluation methodology

In order to test CUDA performance, power, and energy efficiency in real system, the evaluation environments were set according to specifications shown in Tab. 1. The application used to conduct the test was matrix multiplication program with the matrix size of 4 096×4 096, and 256 threads. GPU overclocking ranges were set at 550 MHz, 625 MHz and 700 MHz for GPU core, 400 MHz, 450 MHz and 500 MHz for GPU memory bus. These clock frequencies were selected from ranges that do not interfere with system stability. Additional evaluation was carried out with CPU overclocking from default (3.0 GHz) to 3.735 GHz for comparison.

Tab. 1 System specification used in the evaluation

Category	Specification
CPU	Intel Core2 Duo E8400
Graphic card	NVIDIA GeForce 9500GT
CUDA version	2.3
RAM	DDR2 2G
OS	MS Windows XP
Application	Matrix multiplication
Power estimator	X4-LIFE INSPECTOR II
IDE environment	MS Visual Studio 2005

3 Performance and power evaluation results

3.1 Performance and power evaluation under CUDA programming model

Fig. 1 depicts matrix multiplication program's execution time and total system power consumption with various CPU and GPU clock frequencies.

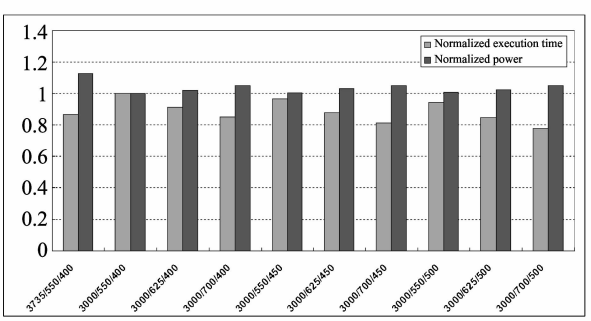

Fig. 1 Normalized execution time and power consumption per clock frequency (MHz) of CPU/GPU core/GPU memory

Fig. 1 shows respective normalized execution time and power consumption per each frequency in comparison with baseline configuration of 3 000/550/400. The result shows that when the clock frequency of GPU core and GPU memory increases, the execution time is reduced. In case that overclocking is applied on CPU only, system-wide power consumption is increased by 12.8%. However, differences of execution performance were insignificant compared to configuration of 3 000/625/450. As the increased power consumption of configuration of 3 000/625/450 is only about 2% compared to configuration of 3 000/550/400, it shows that using GPU is more efficient in CUDA programming model. Also increasing overclocking rate of GPU core than that of GPU memory turns out to be seemingly more efficient with regard to execution time. For the reference, when the same matrix multiplication program is executed in CPU without use of GPU (in case of 1 024×1 024 matrix calculation), GPU calculation show about 300 times better performance.

3.2 Energy efficiency evaluation under cuda programming model

To verify trade-off between energy consumption and execution time, this paper uses EDP (energy-delay product) metric.

Fig. 2 shows trade-off of energy consumption and execution time. Taking into account that the power consumption of GPU overclocking is not so high, the results show that efficiency of energy and execution time is best when only the graphic card overclocking is applied to the highest level (3 000/700/500). EDP is highest when overclocking is not applied to graphic card or CPU. Therefore, with CUDA programming model we can know that GPU overclocking brings considerable benefits in both aspects of energy and execution time. Moreover, compared to CPU overclocking, the EDP shows 24.1% decrease. Therefore, with CUDA programming model we have verified that GPU overclocking has much better energy efficiency over CPU overclocking.

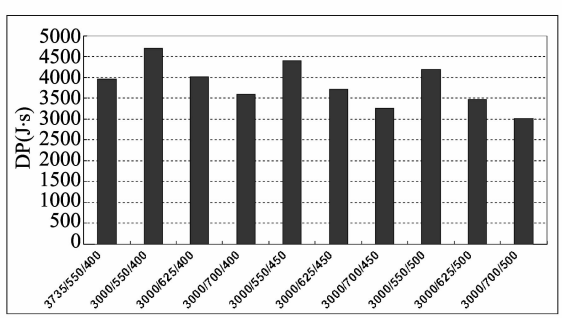

Fig. 2 EDP comparison per clock frequency (MHz) of CPU/GPU core/GPU memory

4 Conclusions

This paper evaluates power, performance, energy efficiency, and trade-off between performance and energy consumption of GPU and CPU overclocking through matrix multiplication program with CUDA. Through our evaluation outcomes we were able to conclude that use of overclocking can increase considerable amount of energy efficiency when executing CUDA applications. Therefore, GPU overclocking technique is expected to become prominent in executing CUDA applications. In future research we plan to diversify CUDA applications and evaluate efficiency of GPU overclocking in areas beyond matrix multiplication program.

References

[1] Borkar S. Low power design challenges for the decade. Proceedings of the 2001 Conference on Asia South Pacific Design Automation, IEEE Press, Januaury 2001.

[2] NVIDIA Corporation. Cuda project http://www.nvidia.com/cuda.

[3] Advanced Micro Devices Inc. ATI stream technology. http://www.amd.com/stream.

[4] Lee J H, Kong J, Suh T, et al. An effective CPU overclocking scheme considering energy efficiency. *Journal of KSCI*, 2009, 14(12): 17-24.