On the power and performance analysis of GPU-accelerated systems

Yuki Abe
Kyushu University

Hiroshi Sasaki
Kyushu University

Koji Inoue
Kyushu University

Kazuaki Murakami
Kyushu University

Shinpei Kato
Nagoya University

1 INTRODUCTION

Graphics processing units (GPUs) have been increasingly used in general-purpose applications due to their significant benefits in performance and performance-per-watt. Figure 1 depicts the trend on the performance per watt for widely deployed NVIDIA GPU and Intel CPU architectures. Albeit energy efficient, GPUs consume significant power during operation, and commodity system software for GPUs is not well designed to control their power consumption. This is largely due to the fact that GPUs are primarily designed to accelerate computations. It is desirable that the system software can manage the power consumption of GPUs in a reliable manner. Dynamic voltage and frequency scaling (DVFS) is widely used to reduce power consumption at runtime for CPUs [2, 4], however there is not much study whether DVFS works efficiently for GPUs or not.

This paper presents our initial work on the power and performance analysis of GPU-accelerated systems. In particular, we leverage Gdev [5], a new open-source implementation of first-class GPU resource management, to demonstrate the effect of GPU power scaling on real-world hardware, whereas the evaluations provided by previous work have been limited to simulations [3]. The ultimate goal of our project is to establish the theory and practice of DVFS schemes for GPU-accelerated systems, addressing correlative power and performance optimization problems. Toward this end, we make the following contributions in this paper: (i) verify the availability of voltage and frequency scaling for NVIDIA's Fermi GPU architecture using Gdev, (ii) analyze the implication of voltage and frequency scaling with the GPU and CPU, and (iii) identify the necessity and open issues of GPU and CPU coordinated DVFS algorithms.

2 EXPERIMENTAL SETUP

We conduct a power and performance analysis of two different GPU-accelerated systems. One system uses Intel Core i5 2400 and the other uses Intel Atom D525, while they both use the same NVIDIA GeForce GTX480 graphics card. Table 1 and 2 present our experimental configurations for the frequency levels of the CPU and the GPU, respectively. Note that we have not yet implemented a DVFS algorithm, while we make a static setup for the frequency and voltage of the CPU using standard Linux interface and that of the GPU using Gdev, when running the benchmarks, in order to analyze the potential of DVFS approaches to GPU-accelerated systems. A real implementation of the DVFS algorithms, however, is left for future work.

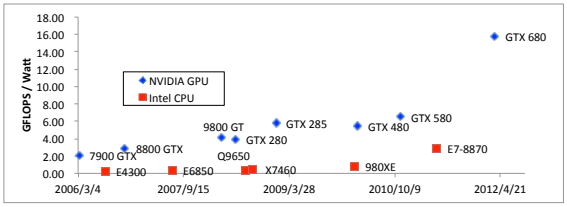


Figure 1: Performance-per-watt trends on NVIDIA GPUs and Intel CPUs

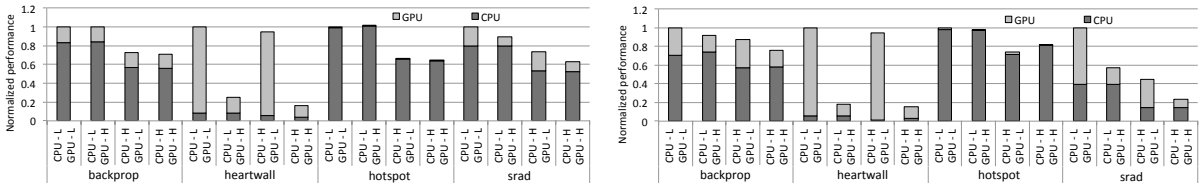
Table 1: CPU frequency levels

CPU	Low [MHz]	High [MHz]
Core i5 2400	2700	3300.1
Atom D525	1350	1800

Table 2: GPU (GTX480) frequency levels

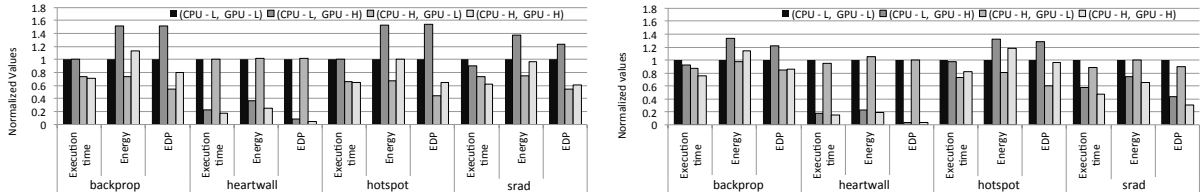
Clock Domains	Low [MHz]	High [MHz]
Core / Shader / Memory	50 / 101 / 135	700 / 1401 / 135

For the performance evaluation, we use four representative benchmark programs from the Rodinia benchmark suite 2.0.1 [1], providing the maximum size of data as inputs. The benchmark programs are compiled using CUDA. We measure the total power consumption of the entire system, since our experimental environment does not provide a way to measure the power consumption of the GPU individually apart from the CPU and other devices. We obtain the voltage and electric current from the power plug of the machine by connecting to the WT1600 digital power meter developed by Yokogawa Electric Corporation. This digital power meter is able to measure the voltage and electric current every 50ms, while it can also calculate the power consumption by multiplying them. Finally, we derive energy consumption by accumulating the power consumption measured during the executions of benchmark programs.



(a) `Atom` (b) `Core i5`

Figure 2: Breakdown of the execution time for the CPU and GPU



(a) `Atom` (b) `Core i5`

Figure 3: Normalized execution time, energy consumption, and energy-delay product (EDP) for different configurations. CPU-L and GPU-H means that the frequency level of CPU is Low and that of GPU is High.

3 EVALUATION

Figure 2 presents the normalized execution time of each benchmark program, containing the details of workload showing the breakdown with respect to times consumed by the GPU and CPU. This evaluation evinces that the frequency of the GPU can be scaled under Gdev as well as that of the CPU. One may observe that the frequency of the GPU has greater impact on GPU-intensive workload such as `heartwall`, while has less impact on CPU-intensive workload such as `backprop` and `hotspot` in terms of execution time. This observation implies possible energy savings for "CPU-intensive" workload by applying DVFS to the GPU. In future work, we conduct more precise measurement, removing errors raised from cache effect and instrument manipulation.

Figure 3 shows a comparison of the GPU and CPU with different frequency levels in terms of the execution time, energy, and energy-delay product (EDP), normalized by a combination of the Low frequency levels. It indicates that the relationship between performance and power is not identical at all but is highly variable depending on workload. First, it is important to see that both the energy consumption and EDP are minimized by different combination of frequency levels in different benchmarks. For example, in the Core i5 platform, the best configuration for `hotspot` is CPU-H and GPU-L setting, where that for `srad` is CPU-H and GPU-H setting. It is also notable that the optimal setting is not identical between two platforms. The minimum energy consumption and EDP for `srad` in Atom platform are achieved by CPU-H and GPU-L setting where the best setting for Core i5 is CPU-H and GPU-H as discussed above. This indicates that we need an in-depth investigation in order to design an optimal GPU and CPU coordinated DVFS algorithm.

4 CONCLUSION

This paper has presented a power and performance analysis of GPU-accelerated systems with different frequency and voltage levels. We demonstrated that the power and performance of GPU-accelerated systems is very sensitive to workload and platforms. Therefore, we need a sophisticated GPU and CPU coordinated DVFS algorithm that optimizes both power and performance for various kinds of workload and platforms. This is one of the greatest challenges left for our future work.

References

[1] CHE, S., SHEAFFER, J. W., BOYER, M., SZAFARYN, L. G., WANG, L., AND SKADRON, K. A characterization of the rodinia benchmark suite with comparison to contemporary cmp workloads. IISWC '10, pp. 1–11.

[2] HSU, C.-H., KREMER, U., AND HSIAO, M. Compiler-directed dynamic voltage/frequency scheduling for energy reduction in microprocessors. ISLPED '01, pp. 275–278.

[3] LEE, J., SATHISHA, V., SCHULTE, M., COMPTON, K., AND KIM, N. S. Improving throughput of power-constrained gpus using dynamic voltage/frequency and core scaling. PACT '11, pp. 111–120.

[4] MAGKLIS, G., SCOTT, M. L., SEMERARO, G., ALBONESI, D. H., AND DROPSHO, S. Profile-based dynamic voltage and frequency scaling for a multiple clock domain microprocessor. ISCA '03, pp. 14–27.

[5] SHINPEI, K., MICHAEL, M., CARLOS, M., AND SCOTT, B. Gdev: First-class gpu resource management in the operating system. USENIXATC'12.