Computing, Performance and Communication systems (2016) 1: 33-37
Clausius Scientific Press, Canada

A Survey of Power Consumption Modeling for GPU Architecture

Qiong Wang[1, a], Ning Li[1,b], Li Shen[1,c] and Zhiying Wang[1,d]

[1]National University of Defense Technology, Changsha, Hunan, China
[a]wangqiong@nudt.edu.cn, [b]rtyq180@gmail.com, [c]lishen@nudt.edu.cn, [d]zywang@nudt.edu.cn

Keywords: GPU architecture, Performance, Power estimation, Modeling

Abstract: GPUs are of increasing interests in the multi-core era due to their high computing power. However, the power consumption caused by the rising performance of GPUs has been a general concern. As a consequence, it is becoming an imperative demand to optimize the GPU power consumption, among which the power consumption estimation is one of the important and useful solutions. In this work, we give a survey of the power modeling for GPU. We first introduce the current development of heterogeneous architectures and then summarize the existing modeling techniques for GPU power consumption. The main two types of power modeling could be classified as simulator-based methods and real machine-based methods.

1. Introduction

Since the invention of the first computer (a.k.a, ENIAC), the development of the computer filed has made significant progress. Particularly, the high-performance computing is of increasing interests due to its powerful computation resources. However, the traditional CPUs are insufficient for building high-performance computer any more. The heterogeneous platform provides an effective solution to this problem. A heterogeneous platform contains CPU, GPU and other computing core integrated in the same platform, taking full advantages of different types of processors, and thus enhancing the overall computing platform efficiency.

The emergency of heterogeneous architecture promotes the fast development of high performance computing technology, and the heterogeneous processors have become the mainstream of high performance computing systems. However, the processor has met the bottleneck to make further development and single-core architecture cannot meet the requirements of modern applications owing to power issues. Multi-core architecture has relieved the pressure of power problems, but power is still the major concern of processors. Due to the variety of functions and on-chip resources, power issues are becoming more and more serious and put tremendous negative impacts on the advance of processors. Therefore, power is the state-of-art topic in computer architecture.

The power of the processor is mainly determined by the state of processor: when the frequency and voltage are in a lower state, the power will be smaller. According to this, some people proposed DVFS which can lower frequency and voltage to reduce the consumed power of the processor without sacrificing too much performance. The power of processor in the current and other states are the necessary parameters to finish dynamic scaling, and the scaling scheme will determine the goal state according to the power parameters.

In this paper, we aim at the GPU architecture and carry out a survey of current techniques for GPU power consumption modeling.

2. Heterogeneous Platforms

With the diversification of computing requirements and application characteristics, micro-processors have diversified development trends. As the core component of the computing platform, the micro-processor must be able to provide a variety of configurations and operation modes to meet the various requirements of the application program, and select an optimal combination of hardware and software to achieve efficient operation. Heterogeneous architecture will be able to provide a variety of hardware configuration for the processor. In a narrow sense, only when each core in the system instruction set architecture, micro-architecture, or between the core of the interconnection network structure is inconsistent, the system can be called heterogeneous multi-core Structure [1]. In a broad sense, the system is considered heterogeneous as long as each core in the system operates in different ways, such as the operating frequency, voltage, etc. of the processor. Both the broad and narrow sense of the heterogeneous computing system, can run the application to provide a variety of run-time configuration. Fig. 1 shows a typical heterogeneous processor architecture.

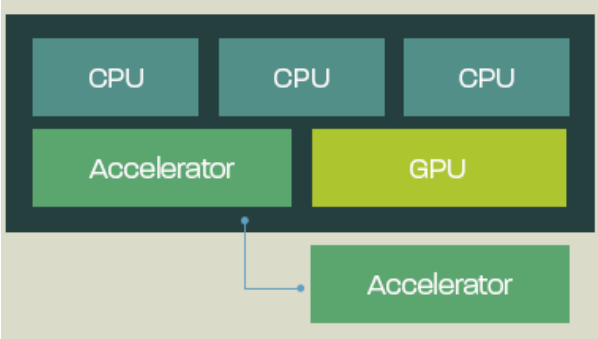

Figure 1. A typical heterogeneous processor architecture

In the field of supercomputing, the heterogeneous architecture has become one of the most popular architectures in recent years. TIANHE-1A [2], which once ranked the first in the TOP-500 supercomputers, uses the GPU as an accelerator, together with the CPU into the heterogeneous system. In November 2012, the first of TOP-500, Titan supercomputer, also used this kind of CPU-GPU heterogeneous architecture. MilkyWay-2 followed by using a more advanced coprocessor, for 7 consecutive ranking the first place in TOP-500 [3]. In June 2015, the number of supercomputers using heterogeneous architectures was 90, up from 75 of the six months ago. Fig.2 illustrates the current number of heterogeneous architecture supercomputers and the ratio of performance provided.

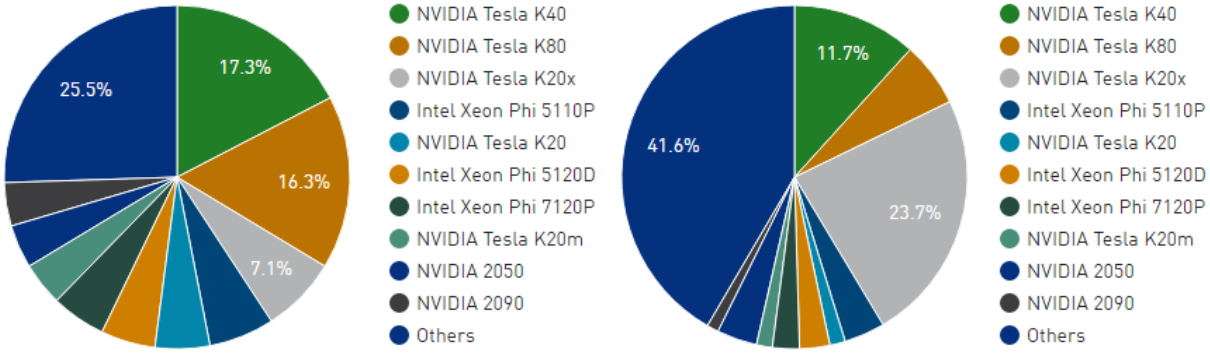

Figure 2. The number of supercomputers based on heterogeneous architecture

3. Study of Power Modeling for GPU

As the rapid development of heterogeneous platforms in recent years, GPU, as a heterogeneous platform computing core, once again becomes a hot research topic. In order to be able to fine-tune heterogeneous processors, GPU power consumption becomes an indispensable reference. Therefore, one can find a variety of GPU power consumption building methods.

3.1 Simulator-Based Modeling.

Compared to the CPU, due to the structure and some historical reasons, the internal interface of GPU is far less open. Many brands of GPU cannot even get performance counter information, but only as a vendor internal debugging. GPGPU-Sim [4] is a general-purpose GPU simulator designed by the Tor Aamodt Research Group of the University of British Columbia in 2012. As shown in Fig.3, this simulator can be used to implement clock-level GPU operating states and process simulations. The simulator integrates the GPUWattch [5] power module, which assumes that each hardware event will consume the same energy. By collecting the performance counter information of each component during the running process, the SM, Cache, Memory, File, execution unit, NoC, DRAM and other components of the power consumption and the sum of the GPU to calculate the total power consumption. The simulator compared with the GTX 480 and Quardo FX 5600, with the error of the two were 9.7% and 13.6%.

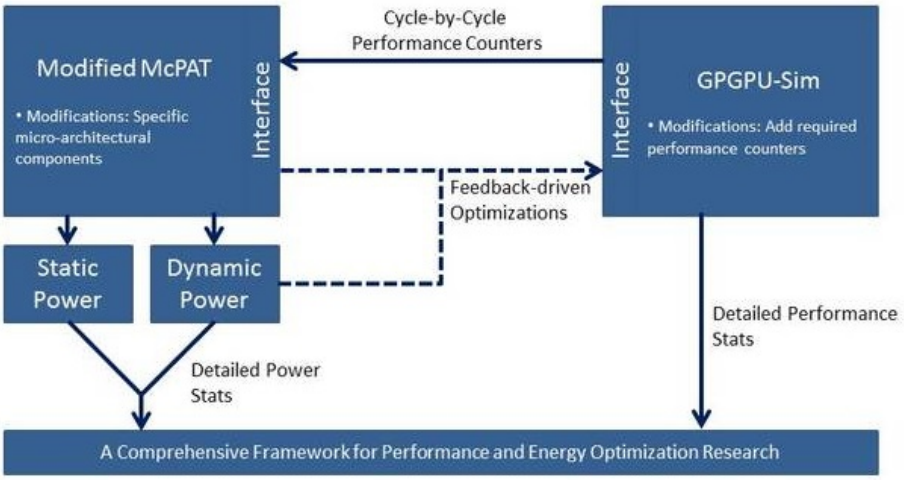

Figure. 3 Details of GPGPU-Sim [4]

Tahir [6] simulates the total power dissipation of the APU and uses the externally connected power measurement unit to measure the power consumption of the processor. We use the Multi2Sim simulator to simulate AMD's EverGreen APUs and obtain performance counter information, including Request to L2 Cache, L2 Cache Miss and DRAM Access, to estimate the power consumption of the processor using the regression model. The final error of 20% or more. Although it is the first time in the AMD platform to achieve power model, but the accuracy is poor, and Multi2Sim follow-up did not update the support of the new platform, so the study of the platform reference is not significant.

3.2 Real Machine-Based Modeling.

Subsequently, many scholars focus on the real machine to study the GPU power modeling. Ying

Zhang [7] aims at AMD's HD5870 and relied on the regression method to build a power consumption and performance prediction model. They figured out the factors that affect the highest weight, so as to find the program bottleneck.

Ali Karami [8] used the same method. They used GPGPU-Sim for NVIDIA's GPU modeling, through prediction and evaluation of OpenCL kernels performance and power consumption information.

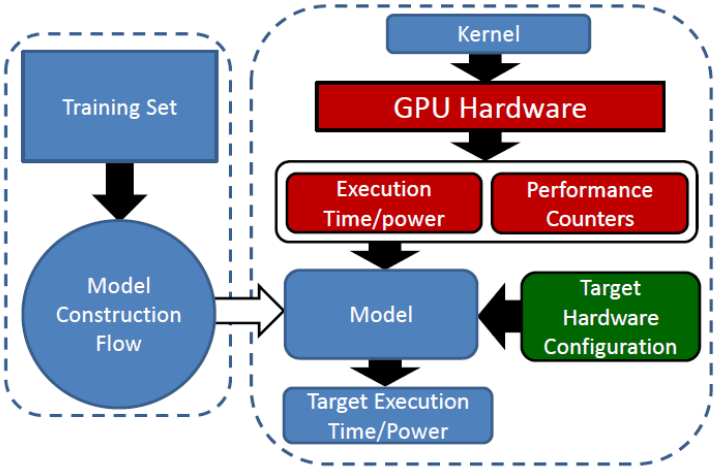

Figure. 4 Power modeling framework based on machine leanring

Gene Wu [9] proposed a new solution based on the machine learning method to predict the GPU's power consumption. Fig.4 illustrates the construction process of this model. One of the assumptions of this approach is that some programs perform a configuration conversion in a different processor configuration. The performance and power consumption are close to the value of the previous configuration state. The authors classify these programs into the same class based on the real experimental results. First of all, this method needs to run the test program in the test set to get the run time information of the program, including the number of CPU, frequency, performance counter information and power consumption of the processor, etc, and use this information as the input of training model, A power consumption prediction model is constructed by machine learning method. After the training model is obtained, it is necessary to calculate the scheduling transformation parameters of each program subclass in the model. For other programs that need to be predicted, we only need to run a specific configuration and get the run-time information, which is submitted to the previous training model. We can get the final program belongs to the class, and the procedures under the regulation conversion parameters to predict the final configuration of the target power consumption and performance information.

4. Conclusions

In this paper, we give a survey of power modeling for GPU. We first show the current development of heterogeneous architectures and then summarize the existing modeling techniques for GPU power consumption. The main two types of power modeling could be classified as simulator-based methods and real machine-based methods.

References

[1] Hennessy J L, Patterson D A. *Computer Architecture: A Quantitative Approach*. Elsevier, 2011.

[2] Yang X J, Liao X K, Lu K, et al. *The TianHe-1A Supercomputer: Its Hardware and Software*. Journal of Computer Science and Technology, 2011, 26(3): 344-351.

[3] Top 500 Supercomputer Ranking List 2015, https://www.top500.org/lists/2015/11/.

[4] Wang Y, Roy S, Ranganathan N. *Run-Time Power-Gating in Caches of GPUs for Leakage Energy Savings*. In Proceedings of the Conference on Design, Automation and Test in Europe. EDA Consortium, 2012: 300-303.

[5] Leng J, Hetherington T, ElTantawy A, et al. *GPUWattch: Enabling Energy Optimizations in GPGPUs [C]*. In ACM SIGARCH Computer Architecture News. ACM, 2013, 41(3): 487-498.

[6] Diop T, Jerger N E, Anderson J. *Power Modeling for Heterogeneous Processors*. In Proceedings of Workshop on General Purpose Processing Using GPUs. ACM, 2014: 90.

[7] Zhang Y, Hu Y, Li B, et al. *Performance and Power Analysis of ATI GPU: A Statistical Approach*. In Networking, Architecture and Storage (NAS), 2011 6th IEEE International Conference on. IEEE, 2011: 149-158.

[8] Karami A, Mirsoleimani S A, Khunjush F. *A Statistical Performance Prediction Model for OpenCL Kernels on NVIDIA GPUs*. In The 17th CSI International Symposium on Computer Architecture and Digital Systems (CADS 2013). IEEE, 2013: 15-22.

[9] Wu G, Greathouse J L, Lyashevsky A, et al. *GPGPU Performance and Power Estimation Using Machine Learning*. In 2015 IEEE 21st International Symposium on High Performance Computer Architecture (HPCA). IEEE, 2015: 564-576.