Exploring GPGPUs Workload Characteristics and Power Consumption

Sohan Lal, Jan Lucas, Mauricio Alvarez-Mesa, Ahmed Elhossini, Ben Juurlink

* *Department of Embedded Systems Architecture,*
Technical University of Berlin,
Einsteinufer 17, D-10587 Berlin, Germany,
`http://www.aes.tu-berlin.de`

ABSTRACT

While general purpose computing on GPUs continues to enjoy higher computing performance with every new generation. The high power consumption of GPUs is an increasingly important concern. To create power-efficient GPUs, it is important to thoroughly study its power consumption. The power consumption of GPUs varies significantly with workloads. Therefore, in this work we study GPU power consumption at a detailed level and its correlation with well-known workload characteristics such as IPC. The low IPC kernels are further explored for the possible bottlenecks.

KEYWORDS: GPGPUs;Power Characteristics;Low Performing Kernels

1 Introduction

GPUs offer tremendous compute power which continues to increase with every new generation. To meet future GPGPU computing performance and power requirements, it is however important to develop new architectural and applications techniques to make GPUs more power efficient [KDK+11]. Therefore, it is necessary to accurately estimate the power consumption at the component level to find opportunities for power optimizations.

There are some research works which estimate GPU power consumption, but they do power estimation at a very coarse-grained level and do not study the change in component power with changing workload. For example, Gebhart et al. [GJT+11] used a very simple and high level power model to estimate the total core power. In our previous research [LLA+13] we developed a GPGPU power simulator and compared total simulated and measured power for different benchmarks. In contrast to above mentioned works, we analyze GPU power consumption at the component level for a diverse set of workloads and

[1]E-mail: {sohan.lal,j.lucas,alvarez,ahmed.elhossini,juurlink}@aes.tu-berlin.de

explore correlation between individual component power and workload metrics such as IPC and SIMD utilization. Moreover, the workloads are studied at the kernel level.

2 Experimental Methodology

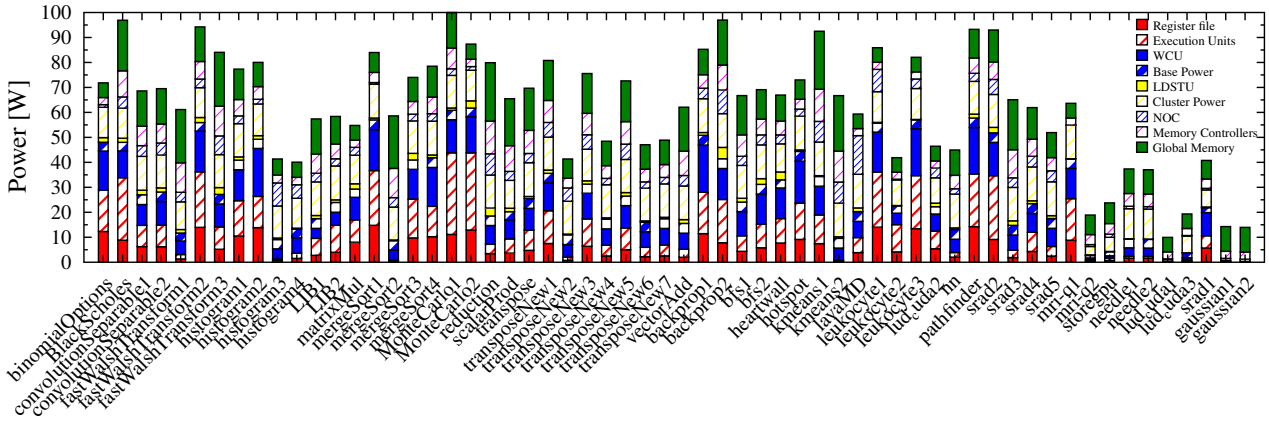
We use our GPUSimPow simulator for estimating the power of different benchmarks [LLA+13]. GPUSimPow uses GPGPU-Sim [BYF+09] as architectural simulator. GPUSimPow has an average relative error of 11.7% and 10.8% between simulated and hardware power for GT240 and GTX580 respectively.

We use benchmarks from the Rodinia benchmark suite [CBM+09], CUDA SDK. The benchmarks cover a wide variety of application domains.

3 Results

3.1 GPU Component Power Consumption

Figure 1 shows the dynamic power consumption by individual components on GTX580 GPU for different kernels. Figure 1 shows a significant change in component power consumption across different workloads.



(a) GTX580

Figure 1: GPU component power consumption. Each bar shows the total dynamic power.

3.2 Categorization of Workloads

We profiled the benchmarks and extracted metrics such as IPC, SIMD utilization etc. We found that the IPC has strongest correlation with GPU components power consumption. Hence,we choose IPC to further explore the power characteristics of various workloads. It would be interesting to study the workload characteristics that cause this change in component power consumption. The kernels with similar power characteristics can be grouped into one category. This can be useful information for the architects and programmers to prioritize the components for power optimizations as most power consuming components could change from one category of workloads to another. We classify the kernels into *high*

IPC, *medium IPC*, and *low IPC* category. The *high IPC* category contains all kernels whose IPC is equal to or higher than 60% of peak IPC. The *medium IPC* category contains all kernels whose IPC is equal to or greater than 40% and less than 60% of peak IPC. The kernels with IPC less than 40% belong to third category of *low IPC*.

Category	GT240		GTX580	
	#Kernels	Avg IPC	#Kernels	Avg IPC
High	21	387.2	20	780.7
Medium	10	274.7	12	518.4
Low	46	88.7	45	149.2

Table 1: Number of kernels and average IPC for each category for GT240 and GTX580.

Table 1 shows that over half of the kernels belong to *low IPC* category. Blem et al. [BSS11] defines benchmarks with IPC less than 40% of peak IPC as challenging benchmarks as these benchmarks under-utilize the GPU computing resources. We also simulated these benchmarks on 3 different GPU configurations. We changed the number of cores and memory controllers proportionally. We found a strong Pearson correlation of over 99% between the IPC of these benchmarks on the different configurations. This is a strong hint that these results are valid for a large class of possible GPU configurations. However, further work is needed to see if this also holds true at the component level.

	RF	EU	WCU	BP	LDSTU	CP	NOC	MC	GM
High	14.2%	26.1%	19.2%	4.6%	1.1%	15.8%	3.0%	5.4%	10.7%
Medium	11.3%	18.0%	16.1%	4.9%	2.2%	17.1%	5.3%	8.7%	16.4%
Low	5.5%	8.5%	11.2%	6.4%	1.9%	22.0%	7.9%	12.6%	24.1%

Table 2: Component power consumption for three categories of kernels for GTX580.

Table 2 shows the power consumed by Register file (RF), Execution Units (EU), Warp control unit (WCU), Base Power (BP), LDSTU, Cluster Power (CP), NOC, Memory Controllers (MC), and Global Memory (GM) for the three categories of kernels. The EU, WCU, and RF are the three most power consuming components for kernels belonging to *high IPC* category. For the medium IPC category kernels, the GPU components EU, WCU, and RF consume relatively less power as compared to *high IPC*. GM power consumption increases from 10.7% to 16.4% for *medium IPC* category of kernels. It is interesting to know that power consumed by EU, WCU, and RF is far less for *low IPC* category kernels compared to other two categories. The largest fraction of power is consumed by global memory in the *low IPC* category. The CP and the BP consumption increase from *high IPC* to *low IPC* because the overall usage of cores decreases, but activation power is still consumed. In short, we see that most power consuming components change across the three categories of workloads and thus, also change with the change in workload. We observed similar trends for GT240, but we did not include these results for reasons of space.

3.3 Low Performing Workloads

In the Section 3.2 we found that more than half of the kernels belong to *low IPC* category and hence have low performance. In order to investigate the reasons for low performance, we

simulate the benchmarks with perfect memory option enabled in GPGPU-Sim. The perfect memory option has zero memory latency with no cache misses. We found that with perfect memory approximately 60% of the benchmarks are within the *high IPC* category, but still about 30% of benchmarks are within the *low IPC* category. This clearly shows that memory bandwidth and latency are not the only important reasons for low IPC, but there are other reasons for low IPC as well.

4 Conclusion

In this work, we studied the power consumption of GPUs at the component level. We classified kernels into high, medium, and low IPC category and show a significant change in components power consumption across the three categories of kernels. This could be a vital information for the computer architects and application programmers to prioritize the components for power and performance optimizations. We further explored the bottlenecks for low performing kernels. We plan a more in-depth study of the bottlenecks of low IPC workloads and explore architectural improvements that could mitigate these bottlenecks.

References

[BSS11] E. Blem, M. Sinclair, and K. Sankaralingam. Challenge Benchmarks that must be conquered to sustain the GPU revolution. In *Proceedings of the 4th Workshop on Emerging Applications for Manycore Architecture*, June 2011.

[BYF+09] A. Bakhoda, G.L. Yuan, W.W.L. Fung, H. Wong, and T.M. Aamodt. Analyzing CUDA Workloads Using a Detailed GPU Simulator. In *Proceedings of the IEEE International Symposium on Performance Analysis of Systems and Software, ISPASS*, 2009.

[CBM+09] S. Che, M. Boyer, Jiayuan Meng, D. Tarjan, J.W. Sheaffer, Sang-Ha Lee, and K. Skadron. Rodinia: A Benchmark Suite for Heterogeneous Computing. In *Proceedings of the IEEE International Symposium on Workload Characterization*, 2009.

[GJT+11] M. Gebhart, D. R. Johnson, D. Tarjan, S. W. Keckler, W. J. Dally, E. Lindholm, and K. Skadron. Energy-Efficient Mechanisms for Managing Thread Context in Throughput Processors. In *Proceedings of the 38th Annual International Symposium on Computer Architecture*, ISCA, 2011.

[KDK+11] S. W. Keckler, W. J. Dally, B. Khailany, M. Garland, and D. Glasco. GPUs and the Future of Parallel Computing. *IEEE Micro*, 31, sept–oct. 2011.

[LLA+13] J. Lucas, S. Lal, M. Andersch, M. Alvarez Mesa, and B. Juurlink. Why a Single Chip Causes Massive Power Bills - GPUSimPow: A GPGPU Power Simulator. In *Proceedings of the IEEE International Symposium on Performance Analysis of Systems and Software*, ISPASS, April 2013.