EasyChair Preprint

№ 1484

A Survey on Progression of GPU Energy Efficiency

Muhammad Raza Manzoor

September 6, 2019

A Survey on Progression of GPU Energy Efficiency

Muhammad Raza Manzoor

Computer Science Department
University of Management and Technology
Lahore, Pakistan
f2018279002@umt.edu.pk[1]

***Abstract*--In this article, we present a survey of DVFS (Dynamic Voltage/Frequency) techniques for improving power efficiency of Graphic Processing Unit. We incorporate just those study that analyze Graphic Processing Unit (GPU) power consumption and methods for energy efficiency of Graphic Processing Unit. Further we only focus on the key idea of different DVFS experimental techniques and methodology**

Keywords:
GPU, CPU, Energy, Performance, DVFS

I. INTRODUCTION

The Requirement of information processing and computation are growing rapidly. Due to this demand, researchers have moved from Serial computing platform i.e. SISD to high performance platforms such as multi-core processors, field programmable gate arrays, Graphic Processing Units (GPUs), etc. Graphic Processing units (GPUs) have been increasingly used for high performance computation (HPC) due to their unmatched computational energy. All supercomputer used Graphic Processing Unit to achieve unmatched computational energy [1].

Manufacturer increases the no. of core processors for high performance which results in more power consumption of Graphics Processing Units (GPUs). They devour much power when contrasted with Central Processing Units (CPU). Power consumption of Graphic Processing

Units have remarkable influence on their reliability and productive performance. (Ashish Mishra 2015). From recent years many high-performance computing use the power of Graphic

Processing Units (GPUs). For example Titan consumes 8.2 HW of power [2]. Tianhe-IA consume electricity bill of $2.5 million [1]. Power consumption forces researchers to find methods that reduce consuming power of Graphics

Processing Units. Because of these reasons understanding of the Graphics Processing Unit (GPU) is significantly important for researchers to propose an efficient solution to overcome the power consumption challenges.

In last few years, several researchers have produced different techniques and methodology to reduce the energy utilization of Graphic Processing Unit (GPU). Based on their approach for improving energy efficiency DVFS is the best energy management approaches [3]. It is effective either in energy saving and enhancing execution. Jiao et al [4]. Studied memory/frequency scaling on NVIDIA GTX 280 and observed that power efficiency depend upon application characteristics, scaling down Graphics processing unit core frequency would save power utilization. Ma et al [5] developed online energy management framework to perform dynamic memory and frequency scaling and, Central Processing Unit and Graphics Processing Uni workload division. Results on NVIDIA GeForce8800, framework could save about 6.1% system energy (CPU+GPU) and about 14%

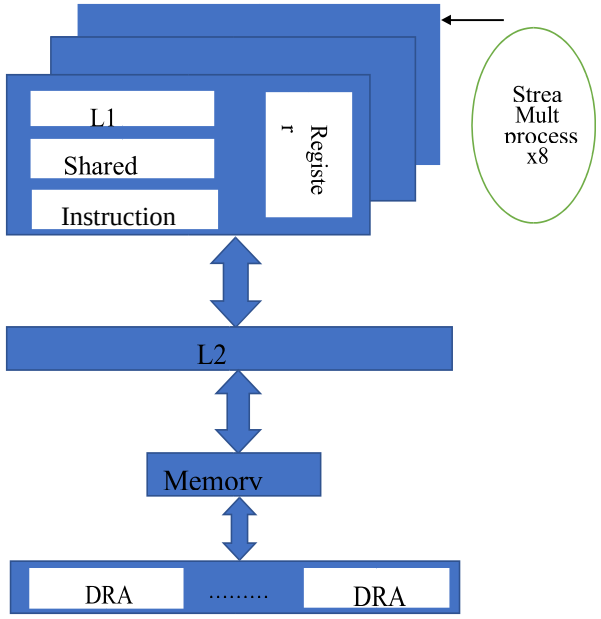
GPU power. Ge et al applied Dynamic voltage and frequency technique on both GPU and Kepler K20c CPU. They discovered that scaling Graphics Processing Unit frequency higher could not utilize greater power. Sethia et al [6] developed a runtime Graphics processing unit core and memory frequency system known as Equalizer. The equalizer either enhance the performance or conserve energy. The author categorized Equalizer into two mode; energy saving mode and high performance mode. Equalizer could save 15% energy in energy efficiency mode.

Graphics processing unit performance models based on Graphic Processing Unit pipeline architecture. GPU pipeline analysis [7] [8] [9] [10] is consider as GPU frequency scaling. Nath et al [10] designed a Graphics Processing Unit performance model to evaluate the GPU performance when frequency is scaled. Statistical methods is another approach which depend on GPU DVFS performance counters. Abe et al [11] built statistical performance models which apply on real GPU hardware, their prediction errors were high. Regarding the survey on Graphics Processing Unit DVFS power models, one common approach depend on statistical methods and other one relies on empirical methods [8]. Empirical methods depends upon binary code analysis and require break up of Graphics Processing Unit (GPU) micro architecture. These approach is device-specific. Statistical method is another approach which depend on hardware performance counters. This method designed to create an energy efficiency model either by regression method [12] or machine learning approaches [13].

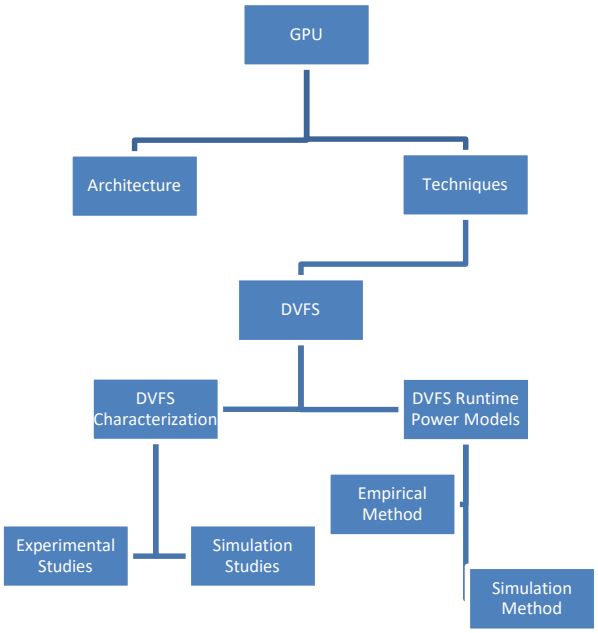
Regression methods easier to implement but for the modern GPU devices, these methods failed to capture the non-linearity, while the advance neural network(ANN) approaches suit the

complicated data dependencies results in high complexity output models, with still prediction accuracy is relatively low.

Fig.1.



TAXONOMY



The article is categorized into different sections. Section 2 provide technical framework comparing Graphic Processing Unit (GPU) energy efficiency methods with other methods. Section 3 describe comparative analysis of energy efficiency method. Section 4 provide conclusion and future research.

In this article, we present a survey of DVFS (Dynamic Voltage/Frequency) techniques for improving power efficiency of Graphic Processing Unit. We incorporate just those study that analyze Graphic Processing Unit (GPU) power consumption and methods for energy efficiency of Graphic Processing Unit. Further we only focus on the key idea of different DVFS experimental techniques and methodology.

II. GPU ARCHITECTURE

A GPU contain L2 cache and multiple stream processor as shown in block diagram of NVIDIA Maxwell GTX 980. Stream multiprocessor and L2 cache are linked with GPU memory unit. The memory unit contains multiple dynamic RAMs, via the memory controller.

NVidia introduced different generations of GPUs as shown in the table. The micro-architectures of NVIDIA GPUs from 2006 are similar except Tesla. Cache system of Tesla is different from other generations of GPU. It's essential to examine the effect of the GPU cache on the application execution and power utilization [14].

Micro-architecture	Year	Compute Capability (X)
Kepler	2012	3.00
Fermi	2009	2.00
Pascal	2016	6.00
Maxwell	2014	5.00
Tesla	2006	1.00

A. GPU DVFS

DVFS is a method used for lowering power consumption of GPUs via scaling the frequency and voltage at run time. DVFS can also be used to decrease the frequency and voltage of processors during low workload or inert periods of GPU. Power consumed by GPU is given in the equation [15].

$$P \propto \lambda * C * V^{\wedge 2}$$

Where P is Energy consumed by GPU, C is Capacitance, λ is Clock frequency and V is Supply Voltage. Thus, energy consumed by decreasing voltage or frequency or both. Decreasing the frequency might also take extra time to accomplish the task. Thus almost no energy will be saved. Therefore, wise DVFS methods are required to enhance the power performance of graphics processing units.

III. DVFS CHARACTERIZATION

There are various ongoing studies on GPU DVFS, which can be carried out either by simulations or experiments. The experimental research are related to scaling the frequency and voltage of Graphics Processing unit in reality while the simulation research related to scaling the voltage/frequency in on simulators. Like GPUWattch. Both experiment and computer simulation techniques suggest that the dynamic voltage/frequency is effective in power consumption.

A. Experimental studies

Jiao et al [4]. Studied core frequency and memory scaling on 3 applications of GTX 280 GPU. The applications are the hybrid fast Fourier transform, memory-intensive dense matrix transpose and compute-intensive dense matrix multiply. They observed that power efficiency was depend upon application characteristics, scaling down GPU core frequency could save power.

Ma et al [5] developed online energy management framework to perform dynamic core and frequency/memory scaling and, Central Processing Unit and Graphics Processing Uni workload division. Results on NVIDIA GeForce8800, framework could save about 6.1% system energy (CPU+GPU) and about 14% GPU power

Ge et al applied Dynamic frequency and voltage technique on both Graphics processing unit and Kepler K20c CPU. They discovered that scaling Graphics Processing Unit frequency higher could not utilize greater power

Mei et al [16] scaled the core frequency, the core voltage and the memory of the Fermi GTX560 GPU. We concluded that the effect of Graphics processing unit DVFS relies upon the application qualities. The ideal setting to utilize the least power was a combination of proper Graphic Processing Unit memory frequency and core voltage. Results on NVIDIA Fermi GPU, saved about 20% power consumption with 4% performance loss.

Mei et al [18] presented GPU voltage and frequency scaling on Fermi as well as Maxwell GPU. They presented the influence of scaling the core voltage and frequency and also the effect of scaling the memory frequency on 24 Kernels. Mei et al observed 21% power consumption on Fermi GTX-560 GPU. They concluded that optimal setting relies on the application characteristics also Mei et al applied the same voltage and frequency approach on Maxwell GTX-980 GPU. The authors also observed that increasing the memory frequency conserve more energy in case of Maxwell GPU but not in the case of Fermi GTX GPU.

B. Simulation Studies:

Leng et al [19] devised GPUWattch, that may simulate the cycle level Graphics Processing unit core voltage and core frequency scaling, based totally at the Fermi GTX481. Leng et al configured the numerous GPU core voltage and core frequency settings in keeping with the 46 nm prediction technology version, and simulated both slow off-chip and prompt on-chip DVFS. Results on slow off-chip DVFS saved about 13% energy consumption and 14% energy consumption in prompt on-chip dynamic voltage frequency scaling with 2.9% performance loss.

Sethia et al [6] designed a runtime GPU system known as Equalizer. Equalizer either enhance the performance or conserve energy. The author categorized Equalizer into two approach; energy saving approach and high performance approach. Equalizer save 14.5% energy in energy efficiency mode.

Gopireddy et al [20]. Designed an architecture and their simulation saved about 48% power consumption as compared to conventional architecture with normal dynamic voltage frequency scaling.

IV. RUNTIME POWER MODELS

Regarding the survey on Graphics Processing Unit (GPU) DVFS power modeling, one common approach relies on statistical methods and other one relies on empirical methods [8].

Empirical methods depends upon binary code analysis and require break up of Graphics Processing Unit (GPU) micro architecture. These approach is device-specific. Another approach is statistical method which relies on hardware performance counters. This method used to create power model either by machine learning [12] approaches or regression [13].

A. Empirical Method:

Isci and Margaret introduced the empirical power modeling method to measure power consumption in Pentium IV. It decomposed entire motherboard into independent hardware sub components. In every segment, they calculated maximum energy intake.

Empirical energy modeling equations is shown in equation A.

$$E = E_0 + E_1 * r_1 + \ldots\ldots\ldots + E_n * r_n$$

Where E_0 , $E_{1,\ldots\ldots\ldots}$, E_n are the maximum energy utilization of an independent hardware sub component and $r_{0,}$ $r_{1,\ldots\ldots\ldots}r_n$ are access rates and E_0 is constant.

Hong and Kim used the same strategy for NVIDIA Fermi GTX281. They determined the access rates and execution cycles of separate graphics processing unit components. The

access rate depended on binary PTX code analysis and execution cycle depended upon pipeline architecture. After that they created a set of micro-benchmarks to search for $E_0, E_1 \ldots E_n$ that provide minimal error among the measured energy. They also designed power increment model for the fact that Graphics Processing Unit power consumption rise when the chip temperature increase. The designed suite of micro-benchmarks achieved 2.5% error prediction and GPGPU kernel achieved 9.2% error prediction.

Leng et al [21] refined the Hong and Kim empirical power architecture with massive amount of micro-benchmarks to control the power error prediction of NVIDIA GTX280 GPU.

Leng et al model has exceptional performance and the model assumed by many analysts [6] [22]. Some researchers identified that Leng et al model is device-specific and difficult to scale the parameters when applying on other NVIDIA Graphics Processing Unit.

Sen and wood [22] designed an energy efficient power model that particularly depended upon the processing time of each core. Sen and Wood energy model was similar to GPU-Wattch.

B. Statistical Method:

Many researchers designed statistical power modeling method to measure power consumption. They used software program to monitor the signals of the GPU application and trained the model based on the signals. This methodology treats the Graphics processing unit micro architecture as a black box and looks for connections between the GPU runtime energy utilization and micro architecture.

Traditional regression power modeling equations is shown in equation B.

$$E = O_0 + O_1 * I_1 + \ldots\ldots\ldots + O_n * I_n$$

Where $O_0, O_1 \ldots O_n$ are output variables and $I_0, I_1 \ldots I_n$ are Input variables.

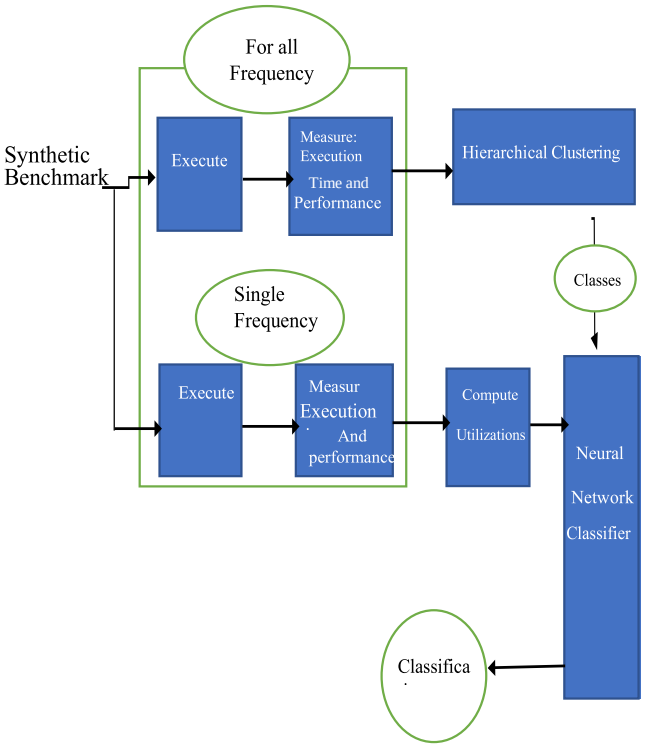
Abe et al [11] designed regression model for three different GPUs. They set 3 different core and memory frequency for Kepler GPU, Tesla GPU and Fermi GPU as a model input, additionally they select ten best performance counter. The prediction error differed from 16% to 24% relying on the GPUs. The latest models have higher prediction error.

Song et al [13] used Advance Neural Network techniques of two hidden layers [13]. They trained their model with this technique and their version executed better prediction accuracy than linear regression. Wu et al [23] trained the runtime power model with Advance Neural Network and k-means algorithm. They applied k-means technique to cluster the identical scaling behavior kernels. Then applied Advance Neural Network technique of two hidden layers on each cluster. The prediction error was 11%.

Regression methods easier to implement but for the modern GPU devices these methods failed to capture the non-linearity, while the advance neural network approaches better suit the complicated data dependencies, with still prediction accuracy is relatively low.

Apart from the above studies, in 2018 J. Guerreiro et al (Guerreiro, et al. 2018) proposed the methodology using 35 applications from different benchmarks Rodinia, CUDA SDK, SHOC etc. on Maxwell Titan X and Titan XP GPU. Experimental results showed that the proposed methodology attain 20% energy saving and as high as 36%, also predict the optimal operating frequency of memory subsystems and graphics. The methodology is shown in the fig. The fig shows procedure to classify the dynamic voltage and frequency impact on the performance of GPU applications. The proposed methodology based on synthetic benchmarks, followed by a training set of Classifier. J. Guerreiro et al applied hierarchical clustering to define the class of benchmarks. Then applied Neural Network technique on each classifier. After the classifier is trained it is possible to classify any application into specific class that specify the dissimilarities in performance.

Fig.2.



V. COMPARATIVE ANALYSIS

The following table shows that energy efficiency and performance depend upon benchmarks and GPU device. Dynamic Voltage frequency techniques designed either to increase performance or energy efficiency or both.

Author	Method	Benchmarks	Energy improvement	Performance improvement	GPU
Ma et al	Experiment	Online management system	6%(CPU+GPU) 11% GPU	Not specified	NVIDIA GeForce8800
Ge et al	Experiment (Liner relation)	Traveling salesman problem Matrix multiplication inite state machine	Not specified	Not specified	Kepler K20c GPU
Mei et al(2013)	Experiment (Application Dependent)	37 applications	20%	4% degradation	Fermi GTX560Ti GPU
Abe et al	Experiment	dense matrix multiply (various matrix sizes)	28% (small matrix size)	Not specified	NVIDIA Fermi GTX480
Abe et al	Experiment	33 popular applications	75% for recent kepler GTX-680 Not specified for other GPU	30% degradation	GTX460/GTX480 Tesla GTX285, Fermi, and Kepler GTX680
Jiao et al	Experiment (Application Dependent)	27 kernels from Rodinia benchmark and the CUDA SDK	34.5%	Not specified	Kepler GTX640 GPU
Mei et al(2017)	Experiment (Application Dependent)	24 Kernels	21%	Not Specified	Fermi GTX-560 Maxwell GTX-980
Leng et a	Simulation GPUWattch	GPUWattch	14.4%	3% degradation	Fermi GTX480
Sethia et al	Simulation GPUWattch	compute-intensive, memory-intensive, and cache sensitive,	15%	Not specified	Not specified
Gopireddy et al	Simulation Scal Core	/	48%	Not specified	Not specified

Leng et al	Empirical method power modeling	large amount of micro-benchmarks	14.4%	3% performance loss	Fermi GTX480
Hong and Kim	Empirical method power modeling	5 memory bandwidth-limited benchmarks.	25.85%	Not specified	GTX280 GPU
Abe et al	Statistical method regression model	10 benchmarks	Fermi 40% Tesla 13%	Prediction error 15% to 23.5%	Kepler GPU, Tesla GPU and Fermi GPU
Song et al	Statistical method ANN of two hidden layers	49 kernels in the SDK Rodinia and the CUDA benchmark suite	Not Specified	4.7%	NVIDIA Fermi C2075
Wu et al	Statistical method ANN+K-mean Clustering	12+ applications	Not Specified	10% prediction error	Not Specified
Guerreiro et al	Statistical method Hierarchical clustering+ ANN	35 applications from Rodinia, CUDA SDK, SHOC	36%	Not specified	Maxwell Titan X and Titan Xp

VI. Discussion:

In the above table we present the comparison of different DVFS technique that we described in section 4. The power consumption in GPU is considered as a hard issue to resolve. The power consumption in GPU faces a lot of challenges. We highlighted various DVFS techniques such as experimental research, simulation research, GPU runtime power models, etc. that is used to lower the power consumption of different GPU devices. The researchers suggested that DVFS techniques is product specific. So, different researchers work on different FPU models. For Maxwell Titan X and Titan XP Guerreiro et al model is best because the experimental results showed that the proposed methodology attain 20% energy saving and as high as 36%, also predict the optimal operating frequency of memory subsystems and graphics. If we use Fermi GTX-560 and Maxwell GTX-980 then Mei et al models is best. As the proposed methodology saved up to 21% power consumptions. Similarly Hong and Kim proposed the methodology and their experimental results for Fermi GTX280 GPU showed that model could save 25.85% power consumption. For Kepler GPU and Tesla GPU devices, Abe et al statistical model can save 40% and 13% power consumptions respectively. Abe et al also scaled the GPU core and memory frequency on Kepler GTX680, Tesla GTX285, Fermi GTX460 and Fermi GTX480 with multiple applications noticeably, they determined that, for the Kepler GTX680, the default frequency configuration was not ideal, as for the Tesla GTX285. They could safe 75% system power with 30% performance loss. We also noted that many researchers perform experiment but their models have different prediction errors. For example Wu et al experimental results showed 10% prediction error and Song et al experimental results showed .7% prediction error.

VII. Conclusion:

In this paper we discuss the various GPU dynamic voltage and frequency techniques for power efficiency. We discuss the up to date GPU DVFS approaches and their performance and effect on energy consumption. We classify the research on DVFS into different methodology such as experimental and simulation methodology. We also discuss nonlinear power modeling approaches, like Advance Neural Networks and linear regression techniques. Applying proper voltage and frequency we can conserve energy.

VIII. References

[1] N. K. Ashish Mishra, "Analysis of DVFS Techniques for Improving the GPU Energy Efficiency," 2015.

[2] J. S. V. SPARSH MITTAL, "A Survey of Methods for Analyzing and Improving GPU Energy Efficiency," pp. 2-2, 2015.

[3] B. G. M. H. R. Gonzalez, "Supply and threshold voltage scaling for low power CMOS, IEEE J. Solid-State Circuits," 1997.

[4] H. L. P. B. W. F. Y. Jiao, "Power and Performance Characterization of Computational Kernels on the GPU," *GREENCOM-CPSCOM '10 Proceedings of the 2010 IEEE/ACM Int'l Conference on Green Computing and Communications & Int'l Conference on Cyber, Physical and Social Computing* , 2010.

[5] X. L. W. C. C. Z. X. W. K. Ma, " GreenGPU: a holistic approach to energy efficiency in GPU-CPU heterogeneous architecture," *Proceedings of the 41st IEEE International Conference on Parallel Processing*, 2012.

[6] S. M. A. Sethia, " Equalizer: dynamic tuning of GPU resources for efficient execution," *IEEE Computer Society*, 2014.

[7] H. K. S. Hong, " An analytical model for a GPU architecture with memory-level and thread-level parallelism awareness," *ACM SIGARCH Computer Architecture News*, 2010.

[8] H. K. S. Hong, "An integrated GPU power and performance model," *ACM SIGARCH Computer Architecture News*, 2010.

[9] S. Hong, "Modeling Performance and Power for Energy-efficient GPGPU Computing," *Georgia Institute of Technology*, 2012.

[10] D. T. R. Nath, "The CRISP performance model for dynamic voltage and frequency scaling in a GPGPU," *Proceedings of the 48th ACM International Symposium on Microarchitecture*, 2015.

[11] H. S. S. K. K. I. M. E. M. P. Y. Abe, " Power and performance characterization and modeling of GPU-accelerated systems," *Proceedings of the 28th IEEE International Parallel and Distributed Processing Symposium*, 2014.

[12] H. S. S. K. K. I. M. E. M. P. Y. Abe, " Power and performance characterization and modeling of GPU-accelerated systems," *Proceedings of the 28th IEEE International Parallel and Distributed Processing Symposium*, 2014.

[13] C. S. B. R. K. C. S. Song, " A simplified and accurate model of power-performance efficiency on emergent GPU architectures," *Proceedings of the 27th IEEE International Symposium on Parallel and Distributed Processing*, 2013.

[14] K. S. M. M. W. Jia, "Characterizing and improving the use of demandfetched caches in GPUs," New York, 2012.

[15] S. a. V. J. Mittal, "A Survey of Methods for Analyzing and Improving GPU Energy Efficiency," *ACM Computing Surveys,* 2014.

[16] L. S. Y. K. Z. X. C. Xinxin Mei, "A measurement study of GPU DVFS on energy conservation," *Proceedings of the Workshop on Power-Aware Computing and Systems ,* 2013.

[17] H. S. M. P. K. I. K. M. S. K. Y. Abe, "Power and performance analysis of GPU-accelerated systems," Berkeley, 2012.

[18] X. Meia, Q. Wanga and X. Chua, "A survey and measurement study of GPU DVFS on energy conservation," *Elsevier,* 2017.

[19] Y. C. W. Zhao, "hniques, 2011, pp. 111New generation of predictive technology model for sub-45 nm early design exploration," *IEEE Trans Electron device,* 2006.

[20] C. S. J. T. N. K. A. A. A. M. B. Gopireddy, " ScalCore: designing a core for voltage scalability," *Proceedings of the 22nd IEEE International Symposium on High Performance Computer Architecture,* 2016.

[21] T. H. A. E. S. G. N. K. T. A. V. R. J. Leng, " GPUWattch: enabling energy optimizations in GPGPUs," *Proceedings of the 40th ACM Annual International Symposium on Computer Architecture,* 2013.

[22] D. W. R. Sen, " GPGPU footprint models to estimate per-core power," *IEEE Comput. Archit,* 2015.

[23] J. G. A. L. N. J. D. C. G. Wu, "GPGPU performance and power estimation using machine learning," *Proceedings of the 21st IEEE International Symposium on High Performance Computer Architecture,* 2015.

[24] J. Guerreiro, A. Ilic, N. Roma and P. Tomas, "DVFS-aware application Classification to improves GPGPUs energy efficiency," *Elsevier ,* 2018.

[25] A. I. N. P. João Guerreiro, "DVFS-aware application classification to improve GPGPUs energy efficiency," 2018.

[26] X. X. N. X. L. T. Y. W. Z. Zhuowei Wan, " Analysis of Parallel Algorithms for Energy Conservation with GPU," 2010.

[27] R. V. J. M. A. A. M. B. Z. Z. R. Ge, "Effects of dynamic voltage and frequency scaling on a K20 GPU," in *Proceedings of the 42nd IEEE International Conference on Parallel Processing,* 2013.

[28] M. L. H. H. T. M. Q. Jiao, "Improving GPGPU energy-efficiency through concurrent kernel execution and DVFS,," *Proceedings of the 13th Annual IEEE/ ACM International Symposium on Code Generation and Optimization,* 2015.

[29] F. B. F. F. M. S. Nicola Bombieri, "MIPP: A Microbenchmark Suite for Performance, Power, and Energy Consumption Characterization of GPU architectures," 2016.

[30] H. H. Muhammad Husni Santriaji, "MERLOT: Architectural Support for Energy-Efficient Real-time Processing in GPUs," 2018.

[31] F. Guo, Y. Mu, W. Susilo, D. Wong and V. Varadharajan, " Cp-abe with constant Size keys for lightweight devices,," *IEEE,* 2014.

[32] D. Evans and D. Eyers, "EEfficient data tagging for managing privacy in the internet of things," 2012.

[33] I. G. E.-E. t. C. K. E. a. DVFS, *Qing Jiao, Mian Lu, Huynh Phung Huynh, Tulika Mitra,,* 2015.

[34] W. J. Y. W. F. S. Akrem Benatia, "Energy Evaluation of Sparse Matrix-Vector Multiplication on GPU," 2016.