Application Driven Energy Efficiency in Embedded GPUs

Akash Sachan*, Rakesh Kumar*, Bibhas Ghoshal

Department of IT, Indian Institute Of Information Technology-Allahabad, Uttar-Pradesh, India-211015

Abstract

Improving the energy efficiency in embedded Graphical Processing Units (GPUs) through an application specific selection of frequency of operation of the GPU platform and compiler settings. Our proposed approach displays on an average 24.52% and 57.68% improvement over the baseline (lowest frequency) and DVFS approaches respectively.

Motivation

• Selecting the lowest functional frequency during execution does not guarantee energy saving.

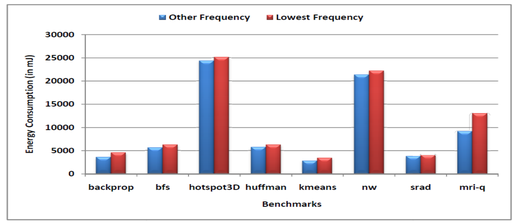


Fig. 1: Energy profile of GPU benchmarks when executed at different frequency levels along with lowest frequency available on board

• A compiler optimization selected for performance need not be suitable for power/energy optimization.

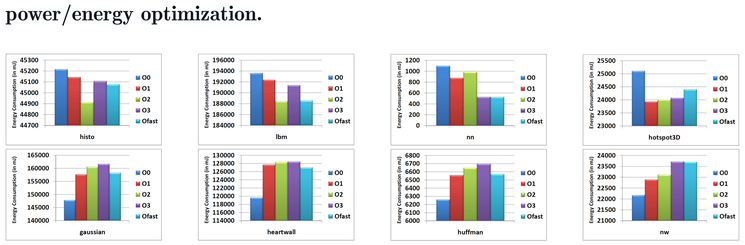


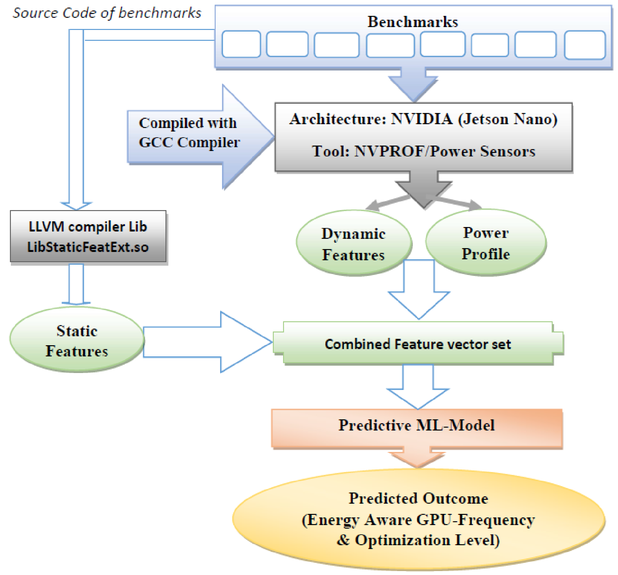
Fig. 2: Energy profile of GPU benchmarks when executed at the lowest frequency for different compiler optimization levels

• A frequency and compiler settings providing the most desired results in terms of energy consumption reduction for a certain GPU application does not provide the same for a different application.
• Making such choices requires knowledge of target architecture, code features and compiler pass independence. In order to relieve developers from making these decisions, an automated technique is required to determine the most optimal GPU frequency and compiler optimization for minimum energy consumption.

Experimental Setup

• NVIDIA Jetson Nano GPU board, an embedded system-on-module (SoM) with quad-core ARM Cortex-A57 integrated with 128-cuda cores Maxwell GPU.
• The platform supports 12 different levels of operating frequencies (921.6, 844.8, 768, 691.2, 614.4, 537.6, 460.8, 384, 307.2, 230.4, 153.6, 76.8) for GPU in MHz.
• The operating system is ubuntu-18.04 LTS with Kernel 4.9.140-tegra. The compiler used are gcc (version-7.5.0) and nvcc (version-10.2.89).
• There are 22 benchmarks which were selected from the benchmark suite, Rodinia-3.1 and Parboil-2.5 for the training and testing of the ML-model.

Proposed Methodology



Phase-1: Energy Profiling

• Involves the execution of different benchmark applications on the target embedded platform with different compiler settings on all supported frequencies.
• The power consumed by GPU is then recorded as benchmarks are executed. The power values were obtained through the INA3221 monitors available on the Jetson Nano experimental board.
• Finding out all mappings of each benchmark applications to its best frequency and optimization level pair defined as a class (Refer Table) in terms of energy consumption which will be used for the training of ML-model in Phase-3.

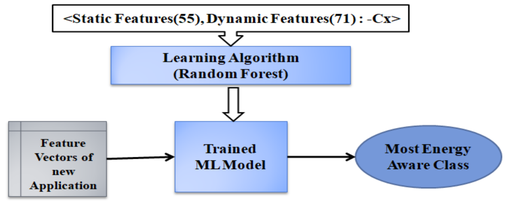
C1: 76.8-O0	C2: 76.8-O3	C3: 153.6-O0	C4: 153.6-O1
C5: 153.6-Ofast	C6: 230.4-O0	C7: 230.4-O1	C8: 230.4-O3
C9: 230.4-Ofast	C10: 307.2-Ofast	C11: 384-Ofast	C12: 460.8-O1
C13: 460.8-O2	C14: 460.8-Ofast	C15: 537.6-O0	C16: 537.6-O2

Phase-2: Feature Engineering

• *Static Features:* Static Features are attributes which are mined out of the source code of an application. It gives information about basic blocks, total no. of instructions, information extracted from control flow graph etc. There are 55 features considered extracted through LLVM compiler pass "LibStaticFeatExt.so".
• *Dynamic Features:* Dynamic Features reveal information of a code only at code execution time such as memory access patterns, cache behavior, CPU utilization etc. There are 71 dynamic features of GPU considered extracted through Nvprof (Version-9.0.252), which is a CUDA command line profiler.

Phase-3: ML-model

• The prediction accuracy of the ML model (Random Forest) was done through prediction of applications not belonging to the training set. The accuracy achieved is around 72%.



Experimental Results

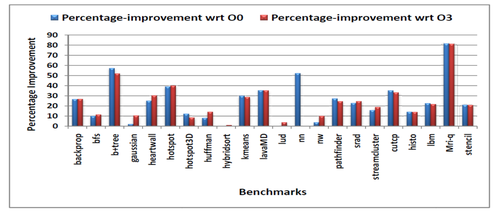


Fig. 5: Percentage Improvement in energy efficiency (24.52% at -O0 and 23.11% at -O3) over baseline

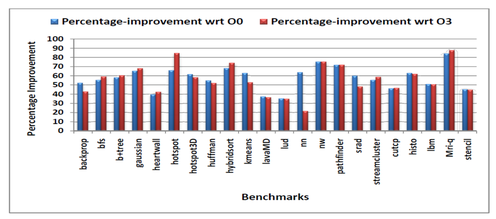


Fig. 6: Percentage Improvement in energy efficiency (57.68% at -O0 and 55.91% at -O3) over DVFS

Acknowledgement

This research is supported by Ministry of Human Resource and Development, Government of India under the scheme "Impacting Research Innovation and Technology (IMPRINT INDIA)", Project No.7482.
https://imprint-india.org/knowledge-portal-project-pages-list

References

•Fan, Kaijie and Cosenza, Biagio and Juurlink, Ben,"Predictable GPUs Frequency Scaling for Energy and Performance", Proceedings of the 48th International Conference on Parallel Processing, ICPP 2019.
•J. Guerreiro and A. Ilic and N. Roma and P. Tom´as,"Modeling and Decoupling the GPU Power Consumption for Cross-Domain DVFS",IEEE Transactions on Parallel and Distributed Systems,Nov-2019.
•Ilager Shashikant and Muralidhar Rajeev and Rammohanrao Kotagiri and Buyya Rajkumar,"A Data-Driven Frequency Scaling Approach for Deadline-aware Energy Efficient Scheduling on GPUs",2020.