CompJouleS: Energy estimate tool for Machine Learning Algorithms for multiple applications in CPU, GPU, and FPGA architectures

Murat Isik
SLAC National Laboratory
Stanford University, Stanford
Menlo Park, California, USA
mrtisik@stanford.edu

Jens E. Pedersen
SLAC National Laboratory, Menlo Park
Stanford University, Stanford
California, USA
jeped@kth.se

Vedant Karia
SLAC National Laboratory, Menlo Park
University of Texas
California, USA
vedant.karia@my.utsa.edu

Sadasivan Shankar
SLAC National Laboratory, Menlo Park
Stanford University, Stanford
California, USA
sadas.shankar@stanford.edu

ABSTRACT

We introduce CompJouleS, a multi-platform energy estimation tool designed to measure the energy cost and performance of custom machine learning algorithms across various hardware architectures, including CPU, GPU, FPGA, Hybrid, and ASICs. Current energy estimation tools lack the flexibility and precision required for accurate analysis across different layers of computing, including applications, machine learning architectures, and hardware architectures. CompJouleS addresses these limitations by combining top-down and bottom-up approaches to provide accurate and efficient energy estimates. The tool is modular, allowing for expansion to additional newer and heterogeneous architectures, and incorporates a computational complexity calculator to estimate the workload of various operations for user-defined algorithms. The first version of CompJouleS is limited to machine learning algorithms on three specific architectures, with the second version expected to extend its capabilities to user-defined algorithms including scientific computations. The paper also reviews existing energy estimation tools and methodologies, highlighting the advantages and limitations of each.

I. INTRODUCTION

As the scaling of silicon-based von Neumann architectures slows down, more specialized and unconventional hardware platforms are being introduced for targeted applications (e.g. Tensor Processing Unit). Separately, there is a large gap between present-day systems and thermodynamic and biological limits (1). With the incredible growth in energy requirements from modern-day deep learning algorithms, lowering energy consumption becomes more important than ever. To address

*The website to this tool is available at https://compjoules.slac.stanford.edu.

the increasing energy consumption by applications such as Machine Learning/Artificial Intelligence (MI/AL), there is a strong need for tools to estimate energies that are standardized across different hardware systems. The heterogeneity of the computational architectures, however, challenges our current ability to make standardized measurements. Specifically, the challenge is to define methods to fairly measure the wattage of a system in a way that can be compared across platforms with different configurations.

In this work, we introduce CompJouleS: a tool for energy estimation and analysis that targets a wide range of machine learning algorithms and hardware architectures. By providing an easily accessible tool for benchmarking architecture-hardware-algorithms-software components, we hope to bring an increased awareness to the problem and enable energy estimation and analysis by a wider community.

II. RELATED WORK

Energy estimation in computing systems is a critical area of research, and several tools have been developed to measure and analyze power consumption. Marcher and Wattch are architectural-level power analysis tools that provide insights into the energy profiles of processor architectures (2; 3). Power API offers a standardized approach to power measurement, enabling the integration of power-aware metrics in performance analysis (4). Accelergy (5) is a more recent tool that focuses on the energy estimation of accelerator designs, while ScaleSim (6) is tailored for the analysis of systolic arrays commonly used in deep learning accelerators.

III. PRELIMINARY RESULTS

CompJouleS targets energy analysis on five different types of architectures: CPU, GPU, and FPGA with planned extensions to

1

Tool	CPU	GPU	FPGA	Hybrid	ASIC
Marcher	✔	✔	✘	✘	✘
Wattch	✔	✘	✘	✘	✘
Power API	✔	✔	✔	✘	✘
Accelergy	✘	✘	✔	✔	✔
ScaleSim	✘	✘	✘	✔	✔
CompJuleS	✔	✔	✔	(✔)	(✔)

Table I: Platform support across energy measurement frameworks.

Hybrid architectures, FPGA-based servers, and ASICs (such as neuromorphic architectures) in a future version. To demonstrate our approach, we performed a preliminary analysis for various ML algorithms across different hardware architectures.

Energy estimation methodologies generally fall into either top-down or bottom-up categories. While the bottom-up method relies on a statistical model that integrates workload analysis and power consumption, top-down methods zero in on power dissipation during algorithm simulations. To conduct our evaluation, we utilized both methods (7).
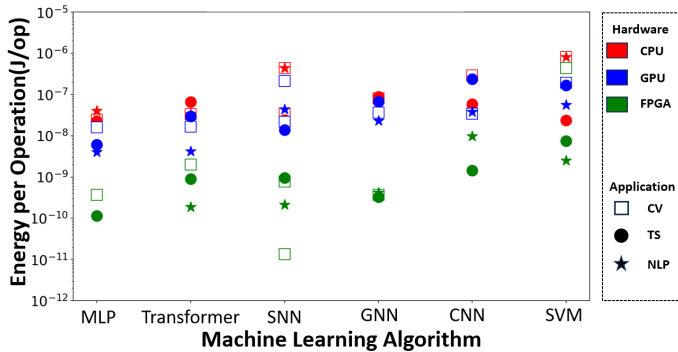

Figure 1: Energy consumption to compute one MAC operation in an ML algorithm, on three different types of hardware.

Figure 1 shows the energy consumption measured for the five algorithms (multi-layer perceptron (MLP), transformer, spiking neural network (SNN), graph neural network (GNN), and support vector machine (SVM)) measured for three different problem domains (computer vision (CV), time-series (TS), and natural language processing (NLP)). Our findings illustrate that FPGA architectures need lower energies when compared to CPUs and GPUs within the realm of hardware platforms, while also highlighting that transformers require the highest energy among all machine learning algorithms. It is important to note that as the models are scaled to larger problems with complex tasks, these trends may change. Although these findings are preliminary, they offer insights to address the energy aspects with finer granularity in the future.

IV. DISCUSSION

We presented CompJuleS: a novel cross-platform energy-measuring tool. By measuring top-down requirements on a systems level, we provide preliminary metrics for the energy consumption of algorithms across different levels. Additionally, we explored the algorithms under various workloads, to emulate real-life simulations. We are presently extending CompJuleS to hybrid and ASIC architectures, enabling a more complete comparison.

Furthermore, we are also planning to supplement the top-down numbers with bottom up estimations. Using detailed and platform-dependent energy estimates, we are presently studying the ability to generalize these estimates to future platforms and algorithms.

REFERENCES

[1] S. Shankar, "Energy estimates across layers of computing: From devices to large-scale applications in machine learning for natural language processing, scientific computing, and cryptocurrency mining," in *2023 IEEE High Performance Extreme Computing Conference (HPEC)*, Sep. 2023, p. 1–6. [Online]. Available: https://ieeexplore.ieee.org/document/10363573

[2] Z. Zong, R. Ge, and Q. Gu, "Marcher: A heterogeneous system supporting energy-aware high performance computing and big data analytics," *Big data research*, vol. 8, pp. 27–38, 2017.

[3] D. Brooks, V. Tiwari, and M. Martonosi, "Wattch: A framework for architectural-level power analysis and optimizations," *ACM SIGARCH Computer Architecture News*, vol. 28, no. 2, pp. 83–94, 2000.

[4] D. DeBonis, R. Grant, S. L. Olivier, M. J. Levenhagen, S. M. Kelly, and K. Pedretti, "A power api for the hpc community." Sandia National Lab.(SNL-NM), Albuquerque, NM (United States), Tech. Rep., 2014.

[5] Y. N. Wu, J. S. Emer, and V. Sze, "Accelergy: An architecture-level energy estimation methodology for accelerator designs," in *2019 IEEE/ACM International Conference on Computer-Aided Design (ICCAD)*. IEEE, 2019, pp. 1–8.

[6] A. Samajdar, Y. Zhu, P. Whatmough, M. Mattina, and T. Krishna, "Scale-sim: Systolic cnn accelerator simulator," *arXiv preprint arXiv:1811.02883*, 2018.

[7] M. Isik, V. Karia, A. Daram, C. E. Kayan, B. Taskin, D. Kudithipudi, and S. Shankar, "Energy and temperature analysis of ai/machine learning algorithms on different hardware systems," 2023, manuscript in preparation.