Creating Flexible, High Fidelity Energy Modeling for Future HPC Systems

Matthew D. Sinclair[*], Bobby Bruce[†], William Godoy[+],
Oscar Hernandez[+], Jason Lowe-Power[†], and Shivaram Venkataraman[*]
[*]UW-Madison, Computer Sciences Department, {sinclair, shivaram}@cs.wisc.edu
[†]University of California, Davis, {bbruce, lowepower}@ucdavis.edu
[+]Oak Ridge National Laboratory, {godoywf, oscar}@ornl.gov

Topics: Architectures, Applications, Modeling & Simulation

I. Challenges

Computing's tremendous, transformative effect has been enabled by a virtuous synergy of (1) better hardware systems, (2) larger datasets, and (3) improved structures and algorithms that further benefit from more efficient hardware and larger datasets. However, the slowing of Moore's Law and end of Dennard's Scaling threatens this virtuous cycle. Consequently, systems are increasingly turning towards heterogeneity to continue scaling performance and energy efficiency, especially for workloads like AI and High Performance Computing (HPC). Thus, HPC systems must balance performance and energy through specialization, generalization, and extreme co-design across computing stack layers.

Energy efficient, large scale, co-designed systems will be necessary in the future. As we increasingly adopt AI, workloads are increasingly bounded by new scaling laws, where the loss scales as a power-law with model size, dataset size, and the amount of compute used for training [1]. Thus, to improve accuracy, scaling requires exponentially increasing resources. This shift means that AI growth requires more energy at a faster pace in a manner distinct from prior patterns, making energy efficiency even more critical. Accordingly, future HPC system designs need to optimize increasingly heterogeneous hardware and applications for energy consumption, without significantly compromising performance.

To drive these efforts, there is a need for credible, open-source infrastructure that can study the energy effects of new systems early in the design process, before silicon is produced. Such infrastructure will help study concrete changes to the existing state-of-the-art and evaluate the potential delta of more radical changes. Traditionally, developers rely on simulation and modeling techniques to model and prototype energy consumption. Unfortunately, the tools needed to co-design systems for performance and energy are lacking for both existing and future systems. Broadly, the current energy-aware approaches are: first principle measurements [2], [3], projecting empirical energy measurements on existing systems to future systems [4], machine learning models to predict energy consumption [5], specific components' energy estimates from design tape outs, or low-level (e.g., Spice) models to get specific components' energy estimates. Each of these approaches has issues. CACTI and McPAT have not been updated in 8 years and are no longer representative. Likewise, design tape outs are time consuming, expensive, and prevent co-design from happening early in the design process. And low-level Spice models are accurate, but scale poorly to increasingly large, complex systems. Finally, state-of-the-art tools like AccelWattch often give inaccurate results under even minor perturbations. Thus, these techniques are too inaccurate, inflexible, or outdated to enable effective, early stage co-design for increasingly complex and heterogeneous systems.

II. Opportunity

In order to perform the early-stage, deeply co-designed design space exploration, necessary to efficiently trade off performance and energy, the community urgently needs flexible, higher fidelity energy models. Specifically, we see the opportunity to ask: **"How should energy models be designed to scalably represent modern systems, pre-tapeout, with high fidelity?"** We believe there is an opportunity to design new energy models which can improve accuracy for post-tapeout hardware, first principle methods for estimating energy use in future systems by leveraging open-source hardware, and finally integration of energy models into tools such as simulators and development environments.

Accurate, Post-Tapeout Energy Measurements: There is an opportunity to develop novel, accurate methods to model energy consumption on hardware that has been taped out. This will help developers understand how their applications are consuming energy and is especially useful when energy consumption information for the hardware is not available. Although we focus on developing energy models, we envision integration with tools that will provide developers with fine-grained information about their application's energy consumption. Moreover, this integration should make it easy for domain scientists to use in an architecture-neutral, portable way. Thus, this work will enable energy-aware optimization of workloads that are of particular interest to DOE.

First Principle Models for Early-Stage, Pre-Tapeout Future Systems: Empirically isolating and measuring energy works well when the underlying architecture does not change and hardware has been taped out. However, future systems, even those with similar components to existing systems, will behave differently from an energy perspective. Thus, simply scaling prior systems' empirical measurements is insufficient. Moreover, since future systems are likely to become even more heterogeneous, potentially including non-traditional or approximate architectures, estimating their energy consumption is difficult – especially for components that may not exist yet. However, relying on costly tapeouts, vendor info, or low-level Spice models for these future systems is either costly, vague (for business reasons), or scales poorly – and inhibits early-stage design space exploration. Instead, we suggest utilizing burgeoning open-source hardware [6], [7] to obtain relatively accurate, synthesizable models and energy information about most components. For components where open-source hardware is insufficient, we suggest supplementing it with analytical models. Collectively, this will create flexible, first principle energy models that can adapt to changes in technology size, system size, connectivity, and hardware heterogeneity – enabling DOE researchers and system designers to prototype energy-aware optimizations across the computing stack. However, exhaustively applying these approaches to every combination of cooling source and accelerator(s), for each generation of hardware, is impractical. Thus, we suggest developing methods to learn energy consumption patterns and predict energy consumption based on validated energy models.

Making Energy Models Easy To Use: Finally, we suggest developing best-in-class supporting infrastructure to enable turnkey ease of use. Although we do not focus on detailed simulation here, we envision integration with popular tools like gem5, a state-of-the-art computer architecture simulator which models CPUs, GPUs, and accelerators [8], [9] the full computing stack, and large-scale HPC systems.

III. Timeliness: Why Now?

A deeper understanding of energy requirements is required in order to achieve ASCR's vision for exascale and beyond HPC systems for the advancement of DOE applications as science drivers into new architectures (e.g., Discovery) [10]. Thus, our suggestions are both timely and necessary to enable early stage design exploration, advances in energy-efficient algorithms, and designing future large-scale systems. We envision our suggested energy models being part of new set of resources that combines these energy models with easy-to-use, accurate developer tools, and detailed simulator support. Overall, this work can enable users ranging from application developers to researchers to perform **energy-aware co-design** to better optimize HPC systems at multiple points in the design process.

References

[1] J. Kaplan *et al.*, "Scaling Laws for Neural Language Models," 2020.

[2] S. Li *et al.*, "The McPAT Framework for Multicore and Manycore Architectures: Simultaneously Modeling Power, Area, and Timing," *ACM TACO*, vol. 10, no. 1, pp. 5:1–5:29, Apr. 2013.

[3] N. Muralimanohar *et al.*, "CACTI 6.0: A tool to model large caches," *HP Laboratories*, vol. 27, p. 28, 2009.

[4] V. Kandiah *et al.*, "AccelWattch: A Power Modeling Framework for Modern GPUs," in *MICRO*, 2021.

[5] G. Wu *et al.*, "GPGPU Performance and Power Estimation Using Machine Learning," in *HPCA*, 2015.

[6] NVIDIA, "NVIDIA RISC-V Story," *4th RISC-V Workshop*, 2016.

[7] B. Tine *et al.*, "Vortex: Extending the RISC-V ISA for GPGPU and 3D-Graphics," in *MICRO*, 2021.

[8] A. Gutierrez *et al.*, "Lost in Abstraction: Pitfalls of Analyzing GPUs at the Intermediate Language Level," in *HPCA*, 2018.

[9] J. Lowe-Power *et al.*, "The gem5 Simulator: Version 20.0+," *CoRR*, vol. abs/2007.03152, 2020.

[10] J. Ang *et al.*, "Reimagining Codesign for Advanced Scientific Computing: Report for the ASCR Workshop on Reimagining Codesign," *DOE ASCR Workshop on Reimagining Codesign*, 4 2022.