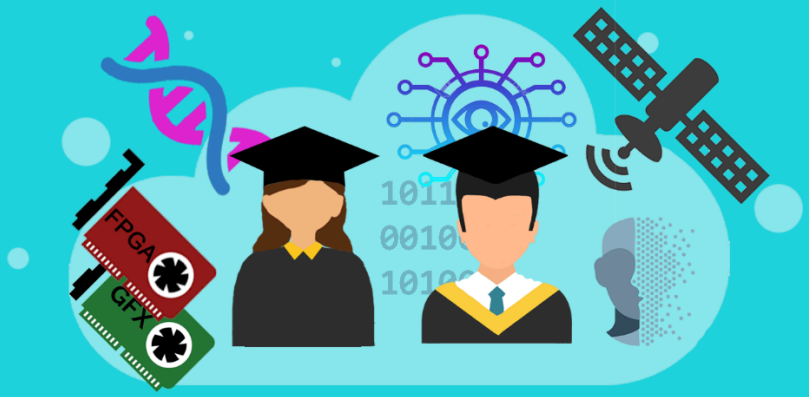


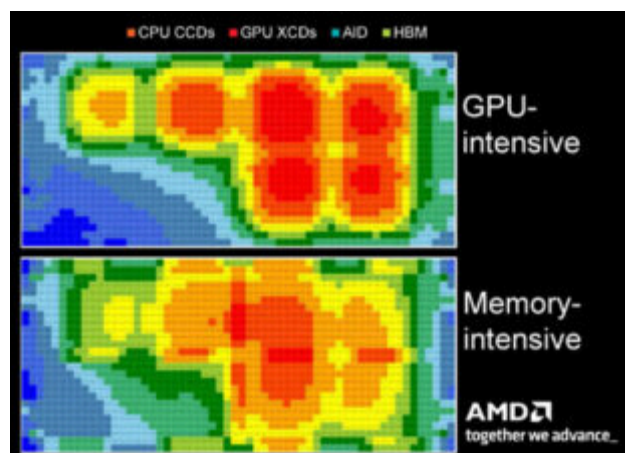


Workload-Aware Power and Energy Modeling for GPU Architectures

Motivation

Energy efficiency has become a primary design constraint for modern processors and accelerators. However, conventional power analysis is often performed either at a high architectural level, using simplified analytical models [1], or at the physical-design level [2], using activity assumptions that do not fully capture the behavior of realistic applications. This separation makes it difficult to understand **where, when, and why energy is consumed during the execution of different workloads**, and consequently limits the ability to identify effective architectural and circuit-level optimization opportunities.

Bridging **real workload behavior with detailed physical-level power characterization** can provide a significantly more accurate view of the energy consumption of complex architectures. Such information can expose spatial and temporal power variations, identify workload-dependent energy bottlenecks, and guide optimizations spanning multiple abstraction levels.



Scope

This diploma thesis aims to develop a **workload-aware power and energy modeling methodology** that combines detailed physical-design power characterization with workload activity obtained through FPGA-based hardware emulation. The resulting framework will enable power and energy consumption to be analyzed not only at the level of the complete GPU, but also across individual architectural components and execution phases. This information will be correlated with detailed power measurements of the corresponding hardware components, obtained through **logic synthesis and, where possible, complete physical implementation (place-and-route)**. Particular emphasis will be placed on understanding how different **workload characteristics, computational kernels, memory-access patterns, and utilization levels** translate into physical power consumption.

Is it just a power/energy simulation framework?

Beyond power estimation, the workload can enable several research directions. The extracted spatial and temporal power information can support **thermal and hotspot analysis**, while component-level workload characterization can guide **architectural design-space exploration**. Finally, the methodology can also investigate cross-layer power-management opportunities, including **fine-grained voltage/frequency domains, multi-VTh cell assignment, body-biasing strategies, and workload-aware power optimization**. The ultimate goal is to establish a reusable framework that connects **application behavior, architectural activity, and physical implementation characteristics**, enabling accurate workload-dependent energy analysis and providing the information required to guide future architectural and circuit-level optimization of GPUs and other accelerator architectures.

RELATED WORKS:

[1] [VIJAY KANDIAH, "ACCELWATCH: A POWER MODELING FRAMEWORK FOR MODERN GPUS". 2021 IEEE/ACM INTERNATIONAL SYMPOSIUM ON MICROARCHITECTURE \(MICRO\).](#)

[2] [M. ZERVA, P. -E. ELEFTHERAKIS, A. MARAS, K. ILIAKIS, A. MOIRAS AND S. XYDIS, "SCROOGE: CIRCUIT-LEVEL DESIGN AND OPTIMIZATION FRAMEWORK FOR RISC-V OUT-OF-ORDER GPUS," 2026 ACM/IEEE 53RD ANNUAL INTERNATIONAL SYMPOSIUM ON COMPUTER ARCHITECTURE \(ISCA\).](#)

CONTACT INFORMATION:

- Alexis Maras, Ph.D. Student, NTUA: (amaras@microlab.ntua.gr)
- Maria Zerva, Research Engineer, Barcelona Supercomputer Cluster: (maria.zerva@bsc.es)
- Panagiotis Eleftherakis, Ph.D. Student, NTUA: (pelef@microlab.ntua.gr)
- Associate Professor Sotirios Xydis, NTUA: (sxydis@microlab.ntua.gr)