Novel GPU Micro-Architectures for General-Purpose Workload Acceleration

GPUs were originally designed for accelerating graphics workloads. Nowadays, due to their massive computing capacity, cost-effectiveness and intuitive programming model, they form the most prevalent platform for accelerating general-purpose workloads. GPUs rely on massive Thread-Level Parallelism (TLP) and fast context switching to achieve high resource utilization.

Among the diverse range of general-purpose workloads ported to GPUs, there exists a class of applications or functions (kernels), which due to limited data parallelism, they fail to hide the long-latency of memory operations and suffer from frequent stalling time. These irregular kernels can profit from more aggressive Instruction-Level Parallelism (ILP).

Emerging Out-Of-Order GPU execution schemes exploit ILP, promise improved performance across a wide spectrum of applications and pave the way for a complete GPU execution model reshaping.

GPGPU-Sim and Multi2Sim are the cycle-accurate, open-source software packages mostly used and approved by the research community to design and evaluate novel GPU micro-architectural concepts. Our lab is supporting a number of diploma theses that will aim to study, comprehend, improve and extend the boundaries of this fascinating domain of novel GPU architectures. The selected candidates will have the chance to (i) familiarize with these state-of-art simulators, (ii) dive into the internals of NVIDIA and AMD GPU architectures and (iii) work on the cutting edge of GPU architecture research.

Required Qualifications: Micro-architecture, C, Object Oriented Programming, CUDA.
Desirable Qualifications: C++, Python, Git, Bash, OpenCL, familiarity with Linux.

Academic Advisors:
Member 1: Konstantinos Iliakis, Ph. D candidate, kiliakis@microlab.ntua.gr
Member 2: Sotirios Xydis, Ph.D, sxydis@microlab.ntua.gr
Member 3: Dimitrios Soudris, Ph.D, dsoudris@microlab.ntua.gr