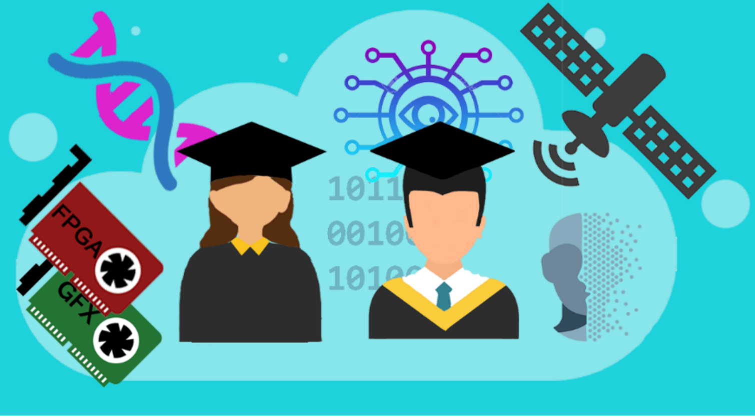
Diploma Thesis

Microproccessors and Digital Systems Laboratory



Modeling Emerging Multi-Tenant GPU Workloads

Graphics Processing Units (GPUs) have revolutionized computing by enabling high-performance acceleration for a wide range of general-purpose applications. Initially designed for rendering graphics, GPUs have become indispensable in accelerating compute-intensive tasks, especially in the field of deep neural networks (DNNs). With the explosion of artificial intelligence (AI) and machine learning (ML) applications, the demand for GPU resources has surged, pushing the boundaries of what these devices can achieve.

Recent advancements in GPU hardware and cluster management now allow multiple users to interact with a single GPU device simultaneously. This multi-tenant usage has the potential to significantly increase efficiency by sharing GPU resources across different workloads, maximizing utilization in both cloud and on-premise environments. However, this emerging use case presents new challenges, as existing simulation tools and models are not yet equipped to accurately reflect this complex multi-user interaction.

The purpose of this thesis is to address this gap by enhancing current simulation tools to support multi-tenant GPU workloads. By modifying and extending these tools, we aim to create accurate models that reflect the performance, scheduling, and resource-sharing characteristics of modern GPUs when used by multiple users concurrently. The outcome will provide insights that can help optimize resource allocation, improve GPU performance, and contribute to the evolving field of GPU workload modeling.

PREREQUISITES:

Strong knowledge of: C/C++, Multi-Processing.

Desirable: Familiarity with Linux, Bash scripting, GPU Architectures.

RELATED MATERIAL:

https://doi.org/10.1109/ISCA45697.2020.00047, https://docs.nvidia.com/deploy/mps/, https://www.usenix.org/conference/atc19/presentation/jeon

CONTACT INFORMATION:

Asst. Prof. Sotirios Xydis (sxydis@microlab.ntua.gr)

Konstantinos Iliakis (kiliakis@microlab.ntua.gr)

Panagiotis Eleftherakis (pelef@microlab.ntua.gr)