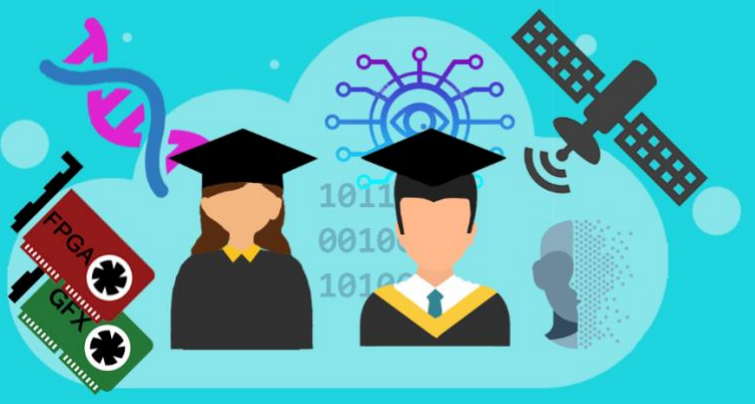
Diploma Thesis

Microproccessors and Digital Systems Laboratory



Optimizing Page Allocation over Heterogeneous Memory Systems

In modern computer systems, efficient memory management is essential for achieving high performance and optimal resource utilization. **Page fault handling** plays a critical role in this process, as it determines how the operating system responds when a program accesses a memory page that is not currently resident in physical memory. Inefficient handling of page faults can lead to significant latency, reduced throughput, and increased system overhead, especially in workloads with irregular memory access patterns or large data footprints. Dynamic allocation mechanisms, which decide where and when memory pages are placed or moved, further influence system responsiveness and memory efficiency.

Figure 1 Intel Optane DCPM Memory

As computing systems evolve to incorporate **heterogeneous memory architectures**, the need for intelligent and adaptive page fault handling and dynamic allocation strategies becomes increasingly critical. Modern systems typically combine **traditional DRAM technologies** with emerging memory technologies that complement DRAM's strengths and address its limitations. Among these, **Non-Volatile Memories (NVMs)** are becoming a mainstream component in modern high-performance computing (HPC) environments, offering significantly higher capacity at a lower cost per bit, along with data persistence and improved energy efficiency compared to DRAM. However, NVMs also introduce new challenges, such as limited write endurance and higher write latency, which must be carefully managed to ensure system reliability and performance. Consequently, heterogeneous memory systems present a complex set of trade-offs between performance, endurance, and energy consumption, emphasizing the need for effective page fault handling and dynamic data placement mechanisms to fully exploit their potential.

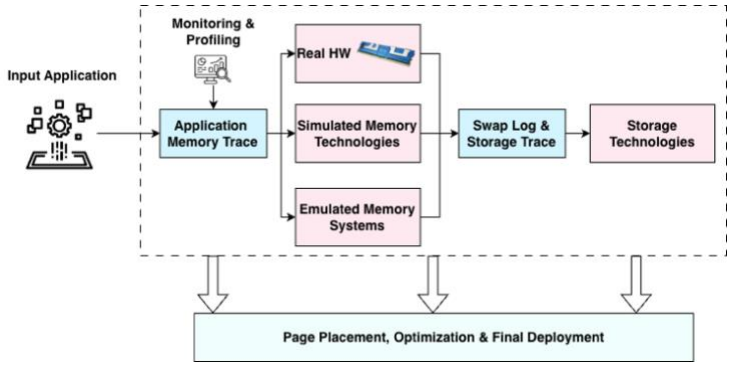


Figure 2 Thesis Projected Workflow

The proposed thesis aims to optimize the page allocation over heterogeneous memory systems. The estimated workflow begins with monitoring and profiling of the target application to collect its memory access trace. This trace captures the application's memory behavior and serves as the foundation for further analysis. The gathered application memory trace is then utilized within a controlled environment consisting of both real hardware-in-the-loop (Intel Optane DCPM) and simulated/emulated memory technologies. These simulations/emulations target various memory hierarchies and configurations, enabling detailed exploration of performance and behavior under different memory system designs. Additionally, swap logs and storage traces are generated to analyze data movement between main memory and storage layers. These traces are further correlated with storage technologies, allowing a holistic view of the system's data management and performance implications. The insights obtained from these

components are subsequently fed into the final stage of page placement, optimization, and deployment. In this stage, the collected data and analysis guide the development of strategies for optimal page placement, efficient data migration, and adaptive memory allocation. The ultimate goal of this process is to achieve an optimized balance between performance, endurance, and energy efficiency within heterogeneous memory systems.

This thesis will be performed in direct collaboration with IMEC - Leuven, Belgium.

TOPICS AVAILABLE: 1

PREREQUISITES:

- C/C++, Python, Bash/Shell Scripting
- Computer Architecture, Memory Organization
- Linux OS

CONTACT INFORMATION:

Dr. Manolis Katsaragakis, Post-Doc Researcher, NTUA: mkatsaragakis@microlab.ntua.gr

Prof. Dimitrios Soudris, NTUA: dsoudris@microlab.ntua.gr

Visiting Prof. Francky Catthor, NTUA: Catthoor@microlab.ntua.gr