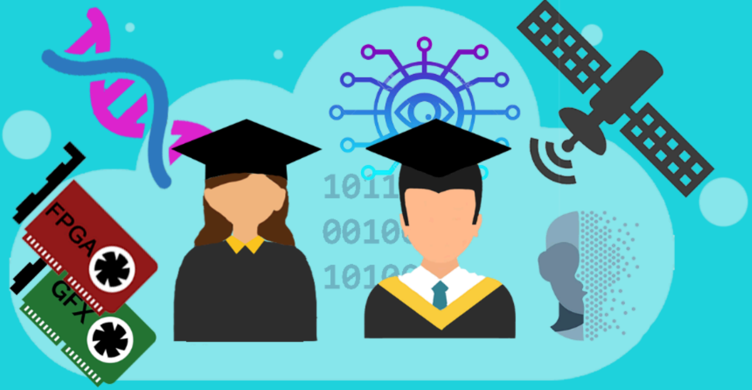
Diploma Thesis

Microproccessors and Digital Systems Laboratory



Intelligent DRAM DIMM Placement for Throughput Optimization in Homogeneous Memory Systems

Modern operating systems manage physical memory placement through generic policies such as first-touch allocation and hardware-level interleaving across memory channels. These policies treat all DRAM equally and make no distinction between which physical DIMM a given data structure resides in. As a result, when multiple data structures are accessed concurrently, they may compete for bandwidth on the same memory channel, leaving available parallelism across channels unexploited.

The research community has demonstrated that intelligent, application-aware memory placement yields significant performance benefits in heterogeneous memory systems - architectures that combine memory tiers with different characteristics, such Non-Volatile Memory (NVM) alongside volatile DRAM. In these systems, placing latency-sensitive or bandwidth-intensive data on the appropriate tier has shown measurable performance and energy improvements across database, scientific computing, and machine learning workloads (Performance, Energy and NVM Lifetime-Aware Data Structure Refinement and Placement for Heterogeneous Memory Systems, Katsaragakis et al, 2025).

However, the vast majority of deployed systems - workstations, servers, and embedded platforms - use homogeneous DRAM configurations with no memory tiers. Despite this, such systems still expose a form of hardware parallelism: multiple independent memory channels, each backed by one or more DIMMs with their own independent buses to the memory controller. This intra-tier parallelism is largely ignored by current OS policies and application allocators, which offer no mechanism for co-placement decisions at the channel or DIMM granularity. Prior work on NUMA-aware placement has demonstrated that distributing data across nodes reduces contention and improves throughput, yet no equivalent mechanism exists at the sub-node level of individual memory channels.

The aim of this thesis is to investigate whether application-aware DIMM placement can improve memory throughput in homogeneous DRAM systems. The work will first characterize the memory access patterns of representative workloads using hardware performance counters and profiling tools, identifying data structure pairs that are accessed concurrently and could benefit from placement on separate memory channels. Using Linux's existing NUMA interfaces (mbind, set_mempolicy) in a controlled QEMU-based virtual NUMA environment as well as on real hardware, we will empirically measure the throughput impact of explicit DIMM placement. Based on these findings, the thesis will explore the design of a lightweight, transparent memory placement framework.

PREREQUISITES:

Familiarity with:

- Computer Architectures
- Linux System Programming
- C/C++ Programming
- Operating Systems, Virtual Machines

DESIRABLE QUALIFICATIONS:

- Kernel Level Development

RELATED MATERIAL:

- https://dl.acm.org/doi/full/10.1145/3736174
- https://www.kernel.org/doc/html/latest/admin-guide/mm/numa_memory_policy.html
- https://www.usenix.org/legacy/event/atc11/tech/final_files/Blagodurov.pdf

CONTACT INFORMATION:

Manos Xenos: mxenos@microlab.ntua.gr

Dr. Manolis Katsaragakis: mkatsaragakis@microlab.ntua.gr

Asst. Prof. Sotirios Xydis: sxydis@microlab.ntua.gr

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