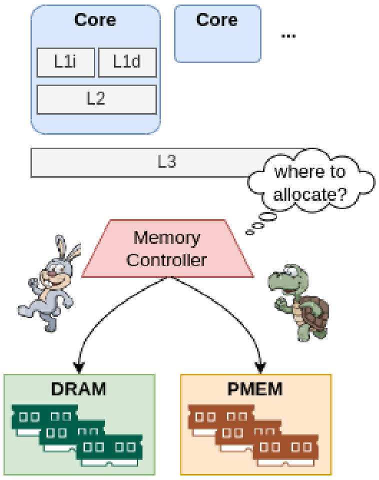
Diploma Thesis

Microproccessors and
Digital Systems
Laboratory

Machine Learning for Operating System-level Page Management

Optimizing data center infrastructure to access and analyze large data sets is crucial for gaining business insights. However, as the density of DRAM reaches its limit, this presents a challenge for infrastructure providers to find a cost-effective solution for increasing memory capacity. In order to efficiently meet memory requirements, providers are shifting towards a strategy called memory tiering. Similar to storage tiering (e.g., HDD/SSD), this approach involves using various memory technologies that are tailored to different data types, usage scenarios, technical requirements, and budget limitations. The ultimate goal is to find the optimal balance between cost, capacity, and performance. With current workloads requiring fast access to data, and businesses relying on quick, actionable insights, solutions that provide both scalability and improved performance are necessary.

Even though memory tiering can assist towards achieving a balance between the above aspects, it also introduces a plethora of new challenges on how to efficiently leverage this new memory architecture. Efficient page scheduling techniques are required, that can identify (and predict) access patterns of applications running on the system and allocate pages in the most efficient manner on heterogeneous memory tiers, taking into account optimization trade-offs, e.g., performance, cost, energy and others. This becomes even more challenging if we consider interference effects in shared computing resources caused by multi-tenant application execution. Towards this direction, machine learning and deep learning techniques can be leveraged, so as to predict future access patterns of running applications on a system and efficiently allocate and migrate pages across the various memory tiers.



In this diploma thesis, we will explore the efficacy of different machine learning (ML) and deep learning (DL) approaches for efficiently predicting the memory access patterns of running applications on a server system. Focus will be given on novel deep learning models (e.g., autoencoders/transformers) and time-series clustering approaches. Depending on the background and interests of the student, this thesis can span across two directions:

- In-depth exploration of various ML/DL models and different modeling approaches for predicting access patterns of running applications under interference.
- Actual implementation and integration of a “naive” memory controller on the Linux kernel.

RELATED MATERIAL

- Doudali, Thaleia Dimitra, et al. "Kleio: A hybrid memory page scheduler with machine intelligence." Proceedings of the 28th International Symposium on High-Performance Parallel and Distributed Computing. 2019.
- Liu, Lei, et al. "Hierarchical hybrid memory management in OS for tiered memory systems." IEEE Transactions on Parallel and Distributed Systems 30.10 (2019): 2223-2236.

PREREQUISITES:

- Linux, Bash/Shell scripting, C/C++
- Python, Machine Learning frameworks and libraries, such as tensorflow, pytorch etc.

CONTACT

- Dimosthenis Masouros, PhD candidate Microlab NTUA (dmasouros@microlab.ntua.gr)
- Thaleia Dimitra Doudali, Assistant Professor, IMDEA Software Institute (thaleia.doudali@imdea.org)
- Sotirios Xydis, Assistant Professor Microlab NTUA (sxydis@microlab.ntua.gr)
- Dimitrios Soudris, Professor Microlab NTUA (dsoudris@microlab.ntua.gr)