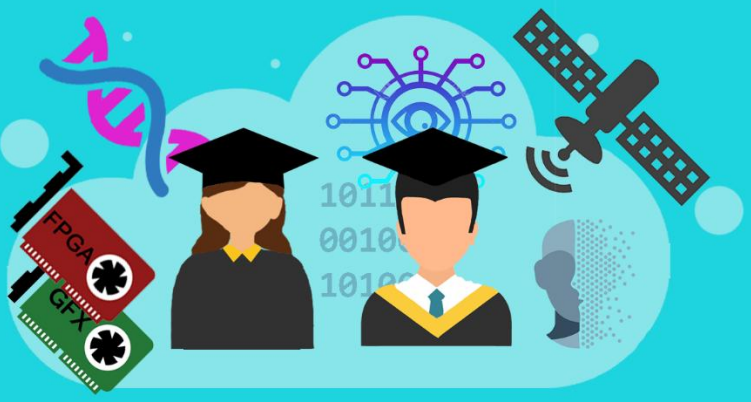
Diploma Thesis

Microproccessors and Digital Systems Laboratory

&



MAX PLANCK INSTITUTE FOR INFORMATICS

Multi-Tier Memory and Accelerator Optimization for Deep Learning Recommendation Models

Deep Learning Recommendation Models (DLRMs) have become fundamental components in modern industry applications such as advertisement systems, music and movie recommendation platforms, e-commerce product suggestions, and social media content. These models process billions of user interactions daily, making them critical infrastructure for digital services. The widespread adoption of DLRMs stems from their ability to capture complex user-item interactions and preferences through sophisticated neural architectures that combine categorical feature processing with dense neural network computations.

Specifically, DLRMs rely heavily on large embedding tables to retrieve semantic representations of items, products, users, and contextual features, often requiring large amounts of memory to store these high-dimensional embeddings. The inference process consists of two distinct computational phases: (i) *Embedding Lookup operations* that search through massive embedding tables to retrieve relevant feature vectors, and then (ii) followed by *Multi-Layer Perceptron* (MLP) inference involving dense matrix-matrix multiplications for prediction generation. The embedding lookup step is inherently memory bandwidth-bound and requires substantial memory capacity, leading to low processor utilization on GPUs that are primarily optimized for compute-intensive MLP workloads. Consequently, prior research works have investigated offloading embedding operations to disaggregated memory devices connected to GPUs, leveraging compute units within these memory nodes to perform search computations closer to the data.

However, existing approaches fail to fully exploit the multi-tier memory hierarchies and diverse compute units available in modern server architectures. Existing prior solutions either offload embedding lookup operations to remote memory devices without utilizing available CPU resources, or employ CPU cores merely for prefetching schemes to identify which embedding table entries should be loaded into GPU memory. This project aims to comprehensively leverage all available acceleration units and memory

capacity in modern servers, including disaggregated memory devices with in-memory processing capabilities (e.g., FPGA-based emulation), CPU cores and memory hierarchies, and GPU devices. Through this multi-memory-tier and multi-accelerator approach, we will design an effective recommendation system that integrates advanced load balancing algorithms, intelligent prefetching mechanisms, and access prediction optimizations to significantly improve performance by coordinating in-memory compute cores in disaggregated devices, CPU processing capabilities, and GPU parallel computation resources.

This diploma thesis has the possibility of being conducted as part of an internship at the Max Planck Institute for Software Systems (MPI-SWS) in Germany.

References:

[1] Jie Ren, Bin Ma, Shuangyan Yan, Benjamin Francis, Ehsan K. Ardestani, Min Si, Dong Li, “Machine Learning-Guided Memory Optimization for DLRM Inference on Tiered Memory”, HPCA 2025

[2] Youngeun Kwon, Yunjae Lee, Minsoo Rhu, “TensorDIMM: A Practical Near-Memory Processing Architecture for Embeddings and Tensor Operations in Deep Learning”, MICRO 2019

[3] G. Sethi, B. Acun, N. Agarwal, C. Kozyrakis, C. Trippel, and C.-J. Wu,, "RecShard: Statistical Feature-Based Memory Optimization for Industry-Scale Neural Recommendation", ASPLOS 2022

[4] Liu Ke, Udit Gupta, Carole-Jean Wu, Benjamin Youngjae Cho, Mark Hempstead, Brandon Reagen, Xuan Zhang, David Brooks, Vikas Chandra, Utku Diril, Amin Firoozshahian, Kim Hazelwood, Bill Jia, Hsien-Hsin S. Lee, Meng Li, Bert Maher, Dheevatsa Mudigere, Maxim Naumov, Martin Schatz, Mikhail Smelyanskiy, Xiaodong Wang, “RecNMP: Accelerating Personalized Recommendation with Near-Memory Processing”, ISCA 2020

Contact Information:

Prof. Christina Giannoula: (cgiannoula@mpi-sws.org)

Dimosthenis Masouros Ph.D.: (dmasouros@microlab.ntua.gr)

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