Department of Electrical Engineering
香港城市大學
City University of Hong Kong

CityU Architecture Lab for Arithmetic and Security (CALAS) Seminar Series

UniNDP: A Unified Compilation and Simulation Tool for Near DRAM Processing Architectures

Tongxin Xie, PhD Candidate
Department of Electronic Engineering, Tsinghua University, China

Abstract:

Near DRAM Processing (NDP) architectures have emerged to be a promising solution for commercializing in-memory computing and addressing the "memory wall" problem, especially for the memory-intensive machine learning (ML) workloads. In NDP architectures, the Processing Units (PUs) are distributed next to different memory units to exploit the high internal bandwidth. Therefore, in order to fully utilize the bandwidth advantage of NDP architectures for ML applications, meticulous evaluations and optimizations of data placement in DRAM and workload scheduling among different PUs are required. However, existing simulation and compilation tools face two insuperable obstacles to achieving these targets. On the one hand, tools for traditional von Neumann architectures only focus on the data access behaviors between the host and DRAM and treat DRAM as a whole part, which cannot support NDP architectures with multiple independent processing and memory units working simultaneously. On the other hand, existing NDP simulators and compilers are designed for specific DRAM technology and NDP architecture, lacking compatibility for various NDP architectures. In order to overcome these challenges and optimize data mapping and workload scheduling for different NDP architectures, we propose UniNDP, a unified NDP compilation and simulation tool for ML applications. Firstly, we propose a unified tree-based NDP hardware abstraction and the corresponding instruction set, enabling the support for various NDP architectures based on different DRAM technologies. Secondly, we design a cycle-accurate and instruction-driven NDP simulator to evaluate hardware performance by accurately tracking the working status of memory elements and PUs. The accurate simulation can provide effective guidance for compilation. Thirdly, we design an NDP compiler that optimizes data partition, mapping, and workload scheduling in different DRAM hierarchies. Furthermore, to enhance the compilation efficiency, we propose a hardware status-guided search space pruning strategy and a fast performance predictor using DRAM timing parameters. Extensive experimental results show that, compared to existing mapping and compilation methods, UniNDP can achieve 1.08-3.37x speedup across multiple NDP architectures and different ML workloads. Furthermore, based on the results of UniNDP, we provide insights for the future NDP architecture design and deployment in ML applications.

Biography:

Tongxin Xie received his B.S. degree from the Department of Electronic Engineering, Tsinghua University, China, in 2023. Currently, he is a 3nd year Ph.D student in the Department of EE, Tsinghua University. His research interests include Processing-In-Memory, Near-Memory-Processing, Computer Architecture, Compiler, etc.

16 January 2026 (Friday); 10am – 12nn; G6315; Zoom link (https://cityu.zoom.us/j/96742093029)