USC Viterbi School of Engineering

PhD Defense

Ming Hsieh Department of Electrical and Computer Engineering

Machine Learning for Memory Access Prediction and Data Prefetching

Pengmiao Zhang

Friday, November 22, 2024
2 P.M. EEB 248

Abstract: Modern applications often experience performance bottlenecks due to memory system limitations. Data prefetching can hide memory access latency by predicting and loading data before it is needed. Machine Learning (ML) algorithms present a promising opportunity to enhance prefetching strategies. However, developing a high-performance ML-based prefetcher presents the following challenges: 1. ML modeling for memory access prediction, including extracting features from historical patterns, identifying future access targets, and designing models to capture their correlations. 2. Domain specific irregular memory access patterns due to multi-core execution and processing phases. 3. Balancing ML model complexity with hardware constraints, ensuring low-latency predictions while maintaining high performance. 4. Coordinated management of multiple prefetchers for ensemble prefetching. In this dissertation, we develop highly optimized ML models for data prefetching. First, to efficiently predict memory accesses for prefetching, we propose TransFetch, a novel attention-based approach that models prefetching as a multi-label classification problem. Second, we introduce a Domain Specific Machine Learning approach for prefetching, utilizing the context of architecture and computation to build high-performance ML-based prefetchers. Using this approach, we develop MPGraph and GraFetch to accelerate the execution of graph applications. Third, towards practical hardware deployment of ML-based prefetchers, we propose a novel tabularization approach that uses table hierarchies to approximate neural networks. We introduce DART, a table-based neural network prefetcher, and Net2Tab, a flexible tabularization framework. Lastly, we present ReSemble, an adaptive framework that uses reinforcement learning to optimize the coordination of multiple prefetchers. Our ML-based prefetchers show significant IPC improvements, demonstrating their performance advantages.

Bio: Pengmiao Zhang is a sixth-year PhD candidate in Computer Engineering, advised by Professor Viktor K. Prasanna. He received his BS degree in Electrical Engineering from Northeastern University (China) and MEng degree in Electrical Engineering from Harbin Institute of Technology. His research interests include machine learning for computer systems, memory system optimizations, and efficient machine learning.

Defense Committee: Prof. Murali Annavaram, Prof. Rajgopal Kannan, Prof. Viktor K. Prasanna (Chair), Prof. Cauligi Raghavendra, Prof. Vatsal Sharan

Zoom:
https://usc.zoom.us/j/9379439223