ECE SEMINAR

Yang Li

Research Scientist, Meta

February 22nd, 11:00 AM to 12:00 PM

Location: SEH B1270

Resource Management: Cloud and AI

ABSTRACT

Resource management significantly impacts the performance and availability of cloud infrastructure. In the first part of my talk, I will introduce my work on managing memory, cache, and power resources to boost the performance and availability of cloud infrastructure. On memory resources, we propose UH-MEM, a utility-based data placement technique for hybrid memory. UH-MEM comprehensively accounts for various characteristics of memory pages to estimate the performance impact of moving a memory page between different memory types and uses this estimate to guide the performance optimization of hybrid memory. On cache resources, we propose dCat, a dynamic cache partitioning system that provides notable cache isolation and improves workload performance on multi-tenant clouds. On power resources, we design CapMaestro, a scalable distributed system for priority-aware power management of cloud infrastructure. This system significantly increases the server capacity of data centers while ensuring service quality for high-priority workloads. Cloud resource management, in many cases, requires predicting future resource demand, which can be viewed as forecasting spatial- and time-dependent data. In the second part of my talk, I will introduce GSA-Forecaster, our proposed machine learning model (a graph Transformer) for forecasting spatial- and time-dependent data, with applications to, but not limited to, cloud resource demand prediction. We apply GSA-Forecaster to real-world forecasting tasks and find that it predicts significantly better than the Transformer and other state-of-the-art models.

BIOGRAPHY

Yang Li is a research scientist at Meta. Before that, he was a researcher at Microsoft. He received PhD and MS degrees from Carnegie Mellon University (advisor: Prof. José M. F. Moura), an MSE degree from the University of Texas at Austin, and a BE degree from Tsinghua University. His research interests span computer architecture and systems, machine learning, and their intersection, including resource management for cloud infrastructure, machine learning models for forecasting spatial- and time-dependent data, and computer systems for large-scale AI and on-device AI. He is a recipient of the IBM Patent Application Award and the Texas Instrument Outstanding Student Designer Award.