USC Viterbi School of Engineering
Ming Hsieh Department of Electrical and Computer Engineering

Seminar

Domain Specialized Architectures and Systems for AI/ML

Dr. Divya Mahajan
Senior Researcher
Cloud Accelerated Systems & Technologies
Microsoft

Thursday, February 10, 2022
10:00am – 11:00am

Zoom link: https://usc.zoom.us/j/96503892197?pwd=Nk13S1RZb25tMlN1QnUxRWZXN2lNZz09
Meeting ID: 965 0389 2197 **Passcode:** 959384

Abstract: Advances in Artificial Intelligence (AI) and Machine Learning (ML) are beginning to revolutionize medicine, manufacturing, commerce, transportation, and other key aspects of our lives. However, such transformative effects are predicated on providing high-performance compute capabilities to enable these learning algorithms. Domain specific accelerators are an efficient and performant means to meet the compute requirements of these large-scale AI/ML. As the new age data-centers become heterogeneous with these emerging domain specific hardware, we must rethink both the architecture and the corresponding system stack.

In this talk, I will provide an overview of my contributions to design, deploy, and utilize accelerators for a wide class of AI/ML applications. I will first discuss pioneering works TABLA and DaNA, which are comprehensive full-stack solutions for machine learning accelerators that integrate with data management systems. These solutions expose a high-level programming interface to programmers that have limited knowledge about hardware design, nevertheless, can benefit from performance and efficiency gains through acceleration. Then, I will describe FAE, a novel framework that leverages statistical properties of data to best utilize the heterogeneous compute and memory resources for recommender model training. Finally, I will conclude with my future research vision towards devising architectures and systems for sustainable massive-scale distributed AI/ML by exploring the challenges which arise from the cross-pollination of different components in the data processing pipeline.

Bio: Divya Mahajan is a Senior Researcher in the Cloud Accelerated Systems & Technologies group at Microsoft. She leads the research, design, and deployment of communication primitives for massive-scale distributed deep learning. She obtained her PhD in Computer Science from Georgia Institute of Technology. She obtained her Masters from The University of Texas Austin, Texas and Bachelors from Indian Institute of Technology Ropar. Her research interests lie in designing novel architectures and building robust systems to address the needs of new and emerging applications. She is passionate about continuing innovative research to have a broad impact on computing and society in general.

Divya is the recipient of National Council for Women and Information Technology Collegiate Award, President of India Gold Medal at IIT, and has been a Finalist in the Qualcomm Innovation Fellowships. Her work has been recognized with the College of Computing Dissertation Award, HPCA Distinguished Paper Award, and has appeared in top architecture, database, systems, and machine learning venues like ISCA, MICRO, HPCA, ASPLOS, VLDB, NeurIPS and high impact journals like IEEE Micro.

https://www.microsoft.com/en-us/research/people/divyam/

Host: Dr. Murali Annavaram, annavara@usc.edu