CSDS 500 Fall 2022 Colloquium

11:30 AM to 12:30 PM
Thursday, October 20, 2022

(Zoom Meeting ID: 935 2774 6217, Passcode: 946606)

"Survey and Trends of Machine Learning Accelerators"

Abstract: Certain aspects of Moore's law arguably have ended, as have a number of related laws and trends including Denard's scaling (power density), clock frequency, core counts, instructions per clock cycle, and instructions per Joule (Koomey's law). Taking a page from the system-on-chip (SoC) trends first seen in automotive systems and smartphones, advancements and innovations are still progressing by developing and integrating accelerators for often-used operational kernels, methods, or functions. Over the past several years, startups and established technology companies have been announcing, releasing, and deploying a wide variety of machine learning accelerators. The focus of these accelerators has been on accelerating deep neural network (DNN) models, and the application space spans from very low power embedded voice recognition to data center scale training. Understanding the relative benefits of these technologies is of particular importance to applying AI to domains under significant constraints such as size, weight, and power, both in embedded applications and in data centers.
This talk will share the results of an on-going, five-year study that has been surveying ML accelerators (and accelerators, in general), including their architectures, their capabilities, and their applicability to various embedded and data center applications. The survey has grown to include over 100 ML accelerators, and they provide a basis with which we will discuss the trends of the accelerators and what to expect in the coming years.

Albert Reuther
MIT

Bio: Dr. Albert Reuther is a senior technical staff member of the MIT Lincoln Laboratory Supercomputing Center (LLSC). He brought supercomputing to Lincoln Laboratory through the establishment of LLGrid, founded the LLSC, and leads the LLSC Computational Science and Engineering team. He developed the gridMatlab high-performance computing (HPC) cluster toolbox for pMatlab and is the computer system architect of the MIT SuperCloud and numerous interactive supercomputing clusters based on SuperCloud, including those in the LLSC.
As a computational engineer, he has worked with many teams within the Laboratory and beyond to develop efficient parallel and distributed algorithms to solve a wide array of computational problems. The SuperCloud architecture earned him an Eaton Award for Design Excellence and his computational engineering work earned him a 2017 R&D 100 Award. He is the technical chair of the IEEE High Performance Extreme Computing Conference and has organized numerous workshops on interactive HPC, cloud HPC, economics of HPC, and HPC security. His areas of research include interactive HPC; computer architectures for machine learning, graph analytics, parallel signal processing; and computational engineering. Dr. Reuther earned a dual BS degree in computer and electrical engineering in 1994, an MS degree in electrical engineering in 1996, and a PhD degree in electrical and computer engineering in 2000, all from Purdue University. In 2001, he earned an MBA degree from the Collège des Ingénieurs in Paris, France, and Stuttgart, Germany.

This is to certify that _____attended this seminar. Certified by _____.
Certificates of attendance and other evidence of CPD activity should be retained by the attendee for auditing purposes.

CASE SCHOOL OF ENGINEERING
CASE WESTERN RESERVE UNIVERSITY