Hardware/Software Co-design for AI Systems

Abstract:

The rapid growth of modern neural network (NN) models' scale generates ever-increasing demands for high computing power of artificial intelligence (AI) systems. Many specialized computing devices have been also deployed in the AI systems, forming a truly application-driven heterogenous computing platform. This talk discusses the importance of hardware/software co-design in AI system designs. We first use resistive memory based NN accelerators to illustrate the design philosophy of heterogeneous AI computing systems, and then present several hardware friendly NN model compression techniques. We also extend our discussions to distributed systems and briefly introduce the automation of co-design flow, e.g., neural architecture search. A research roadmap of our group in the relevant topics is given at the end of the talk.