Title: Energy-efficient Communication Architecture for beyond von-Neumann DNN Accelerators

Speaker: Mr. Sumit K Mandal, University of Wisconsin, Madison

Date and Time: 30th November 2021 (Tuesday), 4 PM.

Abstract

Data communication plays a significant role in overall performance for hardware accelerators of Deep Neural Networks (DNNs). For example, crossbar-based in-memory computing significantly increases on-chip communication volume since the weights and activations are on-chip. State-of-the-art interconnect methodologies for in-memory computing deploy a bus-based network or mesh-based network-on-chip (NoC). Our experiments show that up to 90% of the total inference latency of a DNN hardware is spent in on-chip communication when the bus-based network is used. To reduce communication latency, we propose a methodology to generate an NoC architecture and a scheduling technique customized for different DNNs. We prove mathematically that the developed NoC architecture and corresponding schedules achieve the minimum possible communication latency for a given DNN. Experimental evaluations on a wide range of DNNs show that the proposed NoC architecture enables 20%-80% reduction in communication latency with respect to state-of-the-art interconnect solutions.

Graph convolutional networks (GCNs) have shown remarkable learning capabilities when processing data in the form of graph which is found inherently in many application areas. To take advantage of the relations captured by the underlying graphs, GCNs distribute the outputs of neural networks embedded in each vertex over multiple iterations. Consequently, they incur a significant amount of computation and irregular communication overheads, which call for GCN-specific hardware accelerators. We propose a communication-aware in-memory computing architecture (COIN) for GCN hardware acceleration. Besides accelerating the computation using custom compute elements (CE) and in-memory computing, COIN aims at minimizing the intra- and inter-CE communication in GCN operations to optimize the performance and energy efficiency. Experimental evaluations with various datasets show up to 174x improvement in energy-delay product with respect to Nvidia Quadro RTX 8000 and edge GPUs for the same data precision.

Speaker Bio

Sumit K. Mandal received his dual (B.Tech + M.Tech) degree in Electronics and Electrical Communication Engineering from IIT Kharagpur in 2015. After that, he was a Research & Development Engineer in Synopsys, Bangalore (2015-2017). Currently, he is pursuing Ph.D. in University of Wisconsin-Madison. He is expected to graduate in June, 2022. Details of his research work can be accessed at https://sumitkmandal.ece.wisc.edu/.