Title: Network-on-Chip (NoC) Performance Analysis and Optimization for Deep Learning Applications

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

Date and Time: 28th September 2021 (Tuesday), 5 PM.

Venue: Online

Abstract

Networks-on-chip (NoCs) have become the standard for interconnect solutions in industrial designs ranging from client CPUs to many-core chip-multiprocessors. Since NoCs play a vital role in system performance and power consumption, pre-silicon evaluation environments include cycle-accurate NoC simulators. Long simulations increase the execution time of evaluation frameworks, which are already notoriously slow, and prohibit design-space exploration. Existing analytical NoC models, which assume fair arbitration, cannot replace these simulations since industrial NoCs typically employ priority schedulers and multiple priority classes. Moreover, NoCs used in commercial many-core processors typically experience bursty traffic due to application workloads. Furthermore, these NoCs incorporate deflection routing to minimize queuing resources within routers and achieve low latency during low traffic load. There exists no NoC performance model which can handle all these properties of industrial NoCs. To address this limitation, we propose a systematic approach to construct priority-aware analytical performance models considering bursty traffic and deflection routing using micro-architecture specifications and input traffic. We introduce novel transformations along with an algorithm that iteratively applies these transformations to decompose the queuing system. Experimental evaluations using real architectures and applications show high accuracy of 97% and up to 2.5x speed-up in full-system simulation.

Apart from Systems-on-Chip (SoCs), 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 NoC. Our experiments show that up to 90% of the total inference latency of a DNN hardware is spent on 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.

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 found at the **URL** https://sumitkmandal.ece.wisc.edu/.