Title: A Machine Learning Approach for Network-on-Chip Architecture Design

ABSTRACT

Escalating computational demand, specially due to machine learning and high performance computing (HPC) workloads has made the design of multiprocessor system-on-chip (MPSoC) complex and challenging. These systems necessitate parallel processing, immense computational power, and high-speed on-chip communication infrastructure. While computational needs can be taken care of by integrating large number of processing elements into a chip, effective communication between them becomes a challenge. Network-on-Chip (NoC) technology addresses the issue of energy-efficient and high-performance on-chip communication. In this direction, efficient mapping of an application onto NoC is a pertinent step that remains a challenge, demanding improved algorithms to overcome the limitations of existing heuristic approaches that often encounter local minima.

This dissertation focuses on optimizing the application mapping process with packet latency and communication cost as the key performance metrics. Traditional mapping methods need to invoke NoC simulators for latency estimation, incurring high execution time for the mapping process. To address this issue, we propose a LPNet model, a deep learning-based latency prediction model. Based on the simulation results, it has been observed that LPNet reduces errors when compared with the simulator results. LPNet has been integrated into a Discrete Particle Swarm Optimization (DPSO) algorithm to identify superior mapping solutions and reduced optimization time.

ATSR (Active Search), a machine learning-based mapping algorithm has been specifically designed for application mapping in 2D mesh NoCs. Machine Learning (ML) algorithms have gained significant importance in the field of combinatorial optimization due to their ability to effectively tackle complex problems, compared to the traditional heuristic approaches. ATSR harnesses the power of message passing neural networks and pointer networks, allowing it to navigate complex mapping spaces efficiently. To enhance its capabilities further, a new IP-core numbering scheme has been introduced that improves model accuracy and compatibility across various NoC sizes. Simulation has been performed to validate and assess the practical applicability of the ATSR algorithm in real-world scenarios.

Another scheme, named NCTPAM (Neural Congestion-aware Through-silicon via (TSV) Placement and Application Mapping), an ML based approach for application mapping and TSV placement in partially connected 3D mesh NoC has been introduced. Finding optimal mapping solutions and TSV placements is an NP-hard problem. NCTPAM incorporates graph attention networks and pointer networks in its architecture. A method has been proposed to balance the load across vertical connections, thereby improving overall reliability. Performance of these proposed strategies has been studied analytically and also via simulation.

Keywords: Network-on-chip (NoC), NoC analytical model, application mapping, machine learning, message passing neural network (MPNN), pointer networks, graph attention networks, reinforcement learning, sequence-to-sequence models.