Design optimization for HPC using FPGA “Abstract”

Field Programmable Gate Arrays (FPGAs) are increasingly recognized for their potential in high-performance computing due to their reconfigurable nature, high parallelism, and energy efficiency. However, their adoption has been limited by programming complexity, limited on-chip memory, and insufficient development tools.

This research focuses on optimizing machine learning inference performance on FPGA platforms using Tensil AI’s open-source accelerator. The authors target the ResNet20 neural network trained on the CIFAR-10 dataset and implement it on the Xilinx ZCU104 development board.

They explore and evaluate several design strategies to enhance speed, accuracy, and power efficiency. These include a baseline hardware design, a dual-clock domain approach for improved data transfer, integration of Xilinx Ultra RAM to expand on-chip memory, and an advanced compiler strategy that efficiently utilizes local memory for storing network weights and activations.

Each optimization step demonstrates a significant improvement in performance, with the best design achieving up to 293.58 frames per second and 90% classification accuracy, while maintaining low power consumption at 5.21W and a throughput of 21.12 Giga-Operations per Second (GOP/s).

Their experiments show that FPGA-based inference can reach comparable or superior efficiency compared to traditional CPU and GPU implementations. Additionally, they showed comparisons with existing state-of-the-art implementations, confirming the advantage of this approach in energy efficiency and scalability.

The results highlight FPGAs as a powerful alternative for deploying neural networks in real-time and resource-constrained environments. This work not only showcases the potential of FPGAs for machine learning tasks but also provides a structured approach for optimizing inference accelerators for high-performance applications.