Contribution ID: **129** Type: **Standard Talk**

da4ml: Distributed Arithmetic for Real-time Neural Networks on FPGAs

Tuesday 2 September 2025 14:20 (20 minutes)

Neural networks with a latency requirement on the order of microseconds, like the ones used at the CERN Large Hadron Collider, are typically deployed on FPGAs pipelined with II=1. A bottleneck for the deployment of such neural networks is area utilization, which is directly related to the required constant matrix-vector multiplication (CMVM) operations. In this work, we propose an efficient algorithm for implementing CMVM operations with distributed arithmetic (DA) on FPGAs that simultaneously optimizes for area consumption and latency. The algorithm achieves resource reduction similar to state-of-the-art algorithms while being significantly faster to compute.

We release da4ml, a free and open source package that enables end-to-end, bit-exact neural network to Verilog or HLS design conversion, optimized with the proposed algorithm. For easy adoption into existing workflows, we also integrate da4ml into the hls4ml library. The results show that da4ml can reduce on-chip resources by up to a third for realistic, highly quantized neural networks while simultaneously reducing latency compared to the native implementation hls4ml, enabling the implementation of previously infeasible networks.

Author: SUN, Chang (California Institute of Technology (US))

Co-authors: QUE, Zhiqiang (Walkie) (Imperial College London); LONCAR, Vladimir (CERN); LUK, Wayne; SPIROPULU, Maria (California Institute of Technology (US))

Presenter: SUN, Chang (California Institute of Technology (US))

Session Classification: Contributed talks