Status of ML on FPGAs

TA5 meeting

Johann C. Voigt

15 February 2024

TECHNISCHE UNIVERSITÄT DRESDEN

INSTITUTE OF NUCLEAR AND PARTICLE PHYSICS

PUNCH 4NFDI

General

- Last meeting on January 29
  - Opened up as part of Unplugged week
- Next meeting planned for March 1 9:30
  - Gather topics of interest for hands-on workshop

Bonn status

- Finalizing deliverable D-TA5-WP2-3: Test environment for identifying highly complex (multi-parametric) signals in huge data streams using MeerKAT data.
- hls4ml for conversion of ML model to firmware
- Vivado and Vitis HLS to for wrapper project
- Input/output data from FPGA over shared high bandwidth memory (HBM)
- Targeting Xilinx Alveo
  → See Rameshs presentation

Mainz status

- Studying Vitis/C++ programming for Xilinx Versal AI engines
- Most tutorials use HBM, not direct serial input
- Succeeded in feeding serial data to AI engines
- Working on implementing ANN in AI engines, then combine with serial input
- Manual Vitis/C++ implementation results in lower latency
- AI engines organized in layers
  - Higher latency deeper in array due to transmission between layers
  - 39 engines in first layer
  - $\approx 29\,\text{ns}$ latency for connection to next layer
- Targeting custom board with Xilinx FPGA

Dresden status

- Successfully tested custom 1D CNN firmware with 400 instead of 100 parameter networks (33 networks per FPGA with 12× multiplexing each)
- Working on QKeras support for automatic conversion from trained network to configuration file (architecture description) for firmware
  - Investigating differences in overflow handling
- Tested 1D CNNs in hls4ml for Intel FPGA
  - hls4ml from version 0.7 (released last year) should have better Intel support
  - Includes causal padding
  - hls4ml runs without error, but Intel HLS compiler fails
  - Potential workaround exists, but not followed up on
- Intel is deprecating HLS compiler in favour of oneAPI
  - Common API between CPU, GPU, FPGA
  - Supported by many vendors
  - hls4ml started implementing new backend for oneAPI, but early development stage
- Received new Intel Agilex devkit, porting transceiver code from Stratix harder than expected

Deliverables 2024

- D-TA5-WP2-3 (30 Sep 2023): Test environment for identifying highly complex (multi-para-metric) signals in huge data streams using MeerKAT data.
- D-TA5-WP2-4 (30 Sep 2024): Generic tool to convert trained neural networks into efficient HLS/VHDL FPGA firmware optimised for a real-time, low-latency environment.
- D-TA5-WP5-1 (30 Sep 2024): Development and implementation of machine learning proto-types for anomaly detection, predictive maintenance and process control.
- D-TA5-WP5-2 (30 Sep 2024): Interference recognition and mitigation schemes for transient discovery leading to a “robust” triggering system for multi-messenger follow-up.