Deploying ML In Hardware

FPGAs & ASICs

SLAC TID-AIR
Technology Innovation, Drectorate
Advanced Instrumentation for Research Division
R. Herbst

Matching To Common DAQ Structures

- Dual core ARM A-9 processor
- 1GB DDR3 memory
- High performance platform with 9 clustered Vector processor
- Four categories of approaches to ML in FPGAs *
  - Map digital neurons and their interconnecting weights
  - Software based processing with FPGA coprocessor
  - FPGA based solutions tend to fall into the “Streaming architecture” or
  - Dual Zynq 7045 FPGAs
  - Data processing daughter board with

Machine Learning In Hardware

- Simple to deploy FPGA framework for classification
- Leave out back propagation
- Each layer is pipelined, allowing higher frame rate
- Layers are flexible, can exist in different FPGAs
- Take advantage of bit-parallel quantization for DSP density
- First few layers have minimal overlap of data (convolution)
- Inter-FPGA connections overlap data
- Shared hybrid memory cube (HMC)
- Possible to design networks which minimize or eliminate overlaps

Proposed ML Framework

- Xilinx tool flow is geared towards co-processor based machine learning
- Possible some open-source design flows allow self contained classification system
- Proper solution allows the deployed networks to be integrated into a layered DAQ system
- Working design flow for deploying neural networks in FPGA auto generated from CellNet model

ML Framework Proof Of Concept

- VHDL record driven generation of CNN synthesized into a pipelined classification engine
- Deployed in TID-Air’s Firmware Library SURF
- https://github.com/slaclab/surf
- LetNet (simple digital written decimal recognition) Example:

Applications To ML In Hardware

- The following frameworks are being studied to see if they can be used or serve as a guide for a SLAC framework
- Xilinx: FINN
  - Open Source
  - Xilinx Vivado HLS
  - Brains of Neural Networks (BNNs)
  - CERN, Columbia, Fermilab, MIT, etc. blazed
  - Imperial College London fpgaConvNet

Existing Frameworks

- FINN: Not Available
  - Xilinx Vivado HLS
  - Source Not Available

TID-AIR ES PCI-Express Application Framework

- Provided standard application interfaces which are portable between hardware platforms
- Also provides LCLS1 and LCLS2 timing cores along with timing/data event builder blocks
- Open source version of Amazon Cloud Computing node, static support blocks with user defined partial reconfiguration core

High performance platform with 9 clustered processing elements (SOC)

- Dual core ARM A-9 processor
  - 1GB DDR3 memory
  - Large FPGA fabric with numerous DSP processing elements

The RCE Platform

- Data processing daughter board with
  - Dual Zynq 7045 FPGA
  - 2 artificial 9-10 PCIe links between each FPGA and the NREL

Commercial Xilinx Hardware

Xilinx KCU1500 co-processor
  - XCVU15 FPGA
  - 2 QSPF optical modules
  - 16GB DDR
  - Amazon AWS
  - Xilinx Virtex UltraScale+ VCU1525
  - XCVU9P FPGA
  - 2 QSPF optical modules
  - 64GB DDR
  - Xilinx Alveo U200
    - 2 QSPF optical modules
    - 64GB DDR

Numerous experiments

- LSST
- Heavy Photon Search, LDMX
- DUNE SDTn (ProtoDUNE)
- ATLAS Muon
- ITK Development
- nEXO (Baseline)