The Bigger, the Better? Optimizing Neural Networks for Calorimeter Calibration in the ATLAS Detector

TRIUMF [1] THE UNIVERSITY OF BRITISH COLUMBIA [2] ATLAS EXPERIMENT

Presenter: Annabel Li
Advisors: Colin Gay[1], Kelvin Leong[1,2], Wojtek Fedorko[2], Max Swiatlowski[2]

Background & Motivations

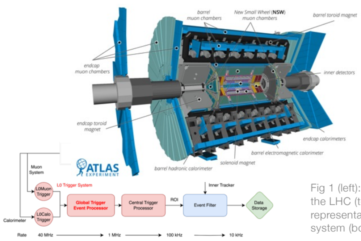


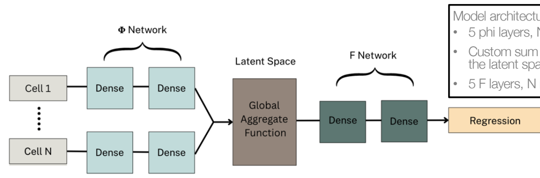
Fig 1 (left): the ATLAS detector at the LHC (top) and a visual representation of the trigger system (bottom)

- Large Hadron Collider (LHC) will be upgraded to High-Luminosity by 2030
  - **Collisions per bunch crossing will increase**
- Hardware trigger system (L0) already struggles with current data rate
  - Incorrect calibration in energy deposited → incorrect events reconstruction[1][2]
  - Low trigger rate discards potentially valuable information

We need a more accurate and efficient trigger system.

How Neural Networks Can Help

- DeepSets machine-learning model improves performance in cluster energy regression[1][2]
- 3 stages: Φ network, latent space, F network



Model architecture:
- 5 phi layers, N nodes each
- Custom sum function ("Ksum") in the latent space
- 5 F layers, N nodes each

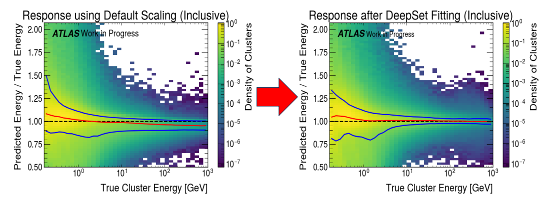


Fig1: DeepSets model visuals

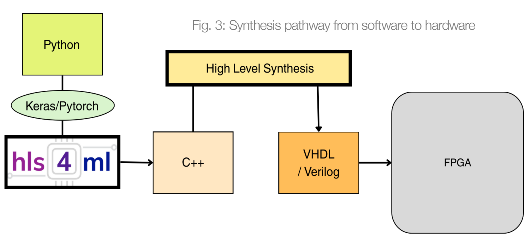
Top: Schematic of the DeepSets model

Right: Response from MC samples using default calorimeter calibration (left) vs. DeepSets model (far right). Red/blue lines represent the median and IQR responses

How is Code Implemented on Hardware?

- FPGAs are designed with hardware description languages (VHDL, Verilog)
- **hls4ml** package automatically converts python machine learning models to synthesis-ready form

Fig. 3: Synthesis pathway from software to hardware



What is Quantization?

- During HLS, floating-point numbers are **quantized** to fixed
  - "ap_fixed<M,N>" = M total bits with N integer bits

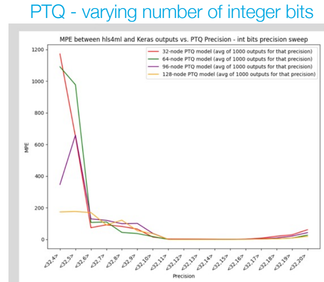
ap_fixed<16,6> → 101101.1010000000 = -18.375

- 2 methods for ML:
  - **Post-Training Quantization (PTQ)** → weights and biases quantized after training
  - **Quantization-Aware Training (QAT)** → model trained on lower-precision operations

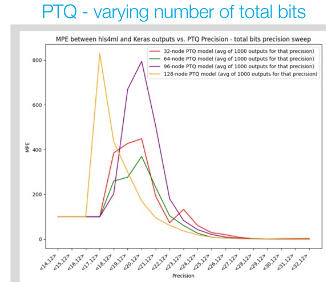
We can use hls4ml to quickly test parameterizations of the DeepSets model for optimization[4][5].

Results

PTQ - varying number of integer bits



PTQ - varying number of total bits



PTQ – setting architecture and intermediate output precisions separately

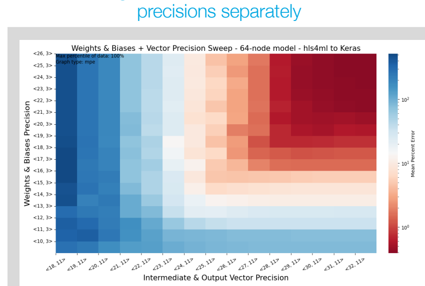


Fig4: PTQ results

Top left: MPE from Keras regression output, varying the number of integer bits for all model parameters

Top right: MPE from Keras regression output, varying the number of total bits for all model parameters

Left: MPE from Keras regression output, specifying different precisions for weights + biases and intermediate + output vectors

Conclusion & Next Steps

Problem: Current L0 trigger system at the LHC is unsuitable for the HL upgrade

Project Goal: Optimize NN size and precision for FPGA deployment

Findings:
- Larger models deviate more than smaller models from their Keras equivalents at lower precisions
- Accuracy increases with precision, but plateaus after a certain point
- Weights and biases can be represented with less bits than intermediate outputs

Next Steps:
- Further optimization strategies: QAT, pruning, High-Granularity Quantization

References:

[1] "Deep Learning for Pion Identification and Energy Calibration with the ATLAS Detector," tech. rep., CERN, Geneva, 2020.

[2] "Point Cloud Deep Learning Methods for Pion Reconstruction in the ATLAS Experiment," tech. rep., CERN, Geneva, 2022.

[3] F. Fahim, et al., "hls4ml: An open-source codesign workflow to empower scientific low-power machine learning devices," March 2021.

[4] P. Odagiu, et al., "Ultrafast jet classification at the hl-lhc," Machine Learning: Science and Technology, vol. 5, p. 035017, July 2024.

[5] C. Antel, "QDIPS: Deep Sets Network for FPGA investigated for high-speed inference on ATLAS," tech. rep., CERN, Geneva, 2025.