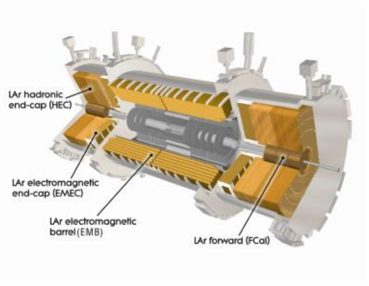
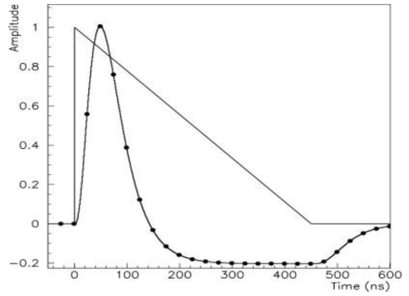
Artificial Neural Networks on FPGAs for Real-Time Energy Reconstruction of the ATLAS LAr Calorimeters

PANIC Lisbon Portugal – Particles and Nuclei International Conference

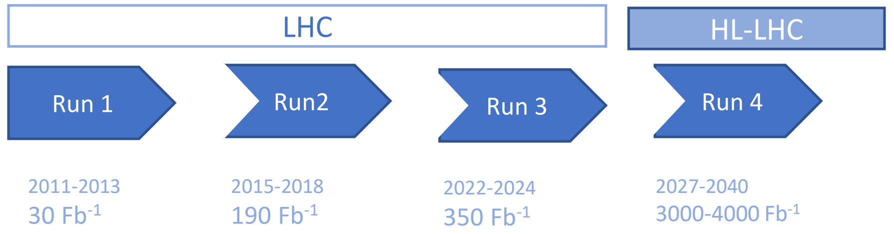
ATLAS Liquid Argon Calorimeter

- The LAr Calorimeter measure the energy deposited by electrons photons and positrons.
- The triangular pulse is shaped to be a bipolar shape. Then it is sampled and digitized at 40 MHz.
- Current Energy Reconstruction algorithm is : Optimal Filter with Maximum Finder (OFMax)
    - Linear Combination of up to 5 samples



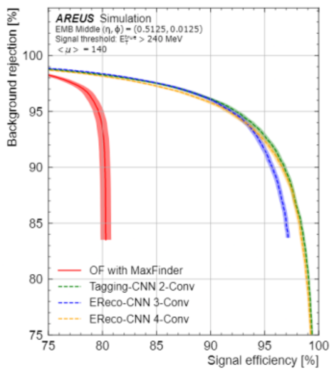


Phase-II Readout Electronic upgrade

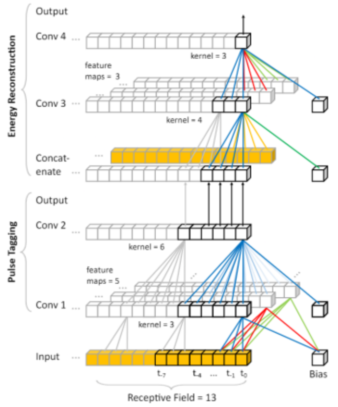


- High Luminosity LHC(HL-LHC) is expected to produce 5-7 times the instantaneous luminosity of the current system.
- 140-200 simultaneous proton-proton collisions on average
- Challenges for LAr calorimeter readout under HL-LHC conditions:
    - Overlapping of signals
    - New trigger scheme allows the selection of subsequent bunch crossing (BC), i.e. every 25ns
- OFMax decreases in performance under HL-LHC conditions
    - FPGA to implement more complex real-time energy reconstruction algorithms.
    - Very low latency required for reconstructing the energy (typically up to 150 ns)
    - One FPGA should process 384 or 512 Lar Calorimeter cells
- Machine Learning solutions under investigation to replace the OF in the future HL-LHC conditions.

Convolutional Neural Network (CNN)

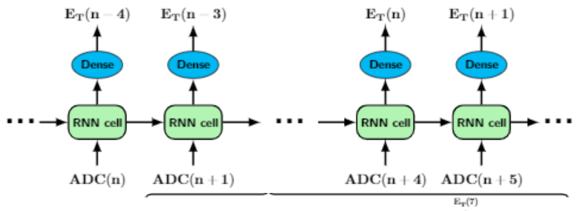


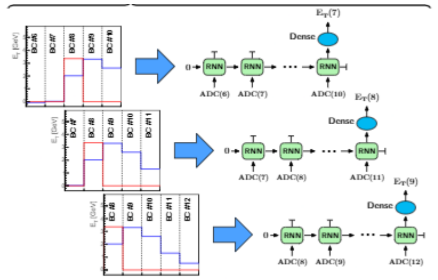
- CNN architecture is designed to compute energy deposits from samples
- Pulse tagging sub-networks (2 Layers):
    - Trained to detect energy deposits above noise threshold.
    - Sigmoid activation function
    - Receiver operating characteristic (ROC) curve shows performance better than OFmax
- Energy Reconstruction sub-network (1-2 Layers):
    - Uses results of tagging sub-network and raw ADC samples
    - Respectively one and two reconstruction layer results in 3D and 4D Conv networks
    - ReLU activation function
- Trained and evaluated on simulated samples using the AREUS software package



Recurrent Neural Network (RNN)

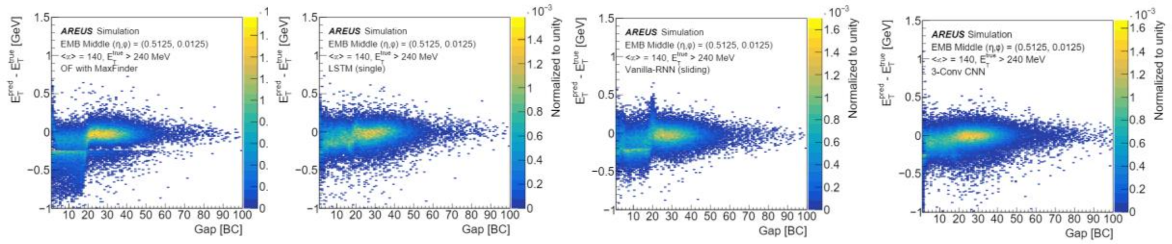
- RNN algorithms are designed for interference of time series and underlying parameters:
    - Natural candidates for the interference of deposited energies from time ordered digitalized Lar signals
    - Vanilla RNN : Simple recurrent structure with a ReLU activation function.
    - Long Short-Term Memory(LSTM) with sigmoid and tanh functions which is more suited for longer term effects
- Single Cell LSTM:
    - Operates sample per sample on entire sequence
    - Feeding data by continuous stream of digitized samples for single cells.
- Sliding-Window LSTM:
    - Feeding data in a window size of 5 including one sample before the pulse
- Sliding-Window and Single cell LSTM:
    - Dense operation which corresponds to the single neuron decoder which reads the LSTM output and calculates the energy

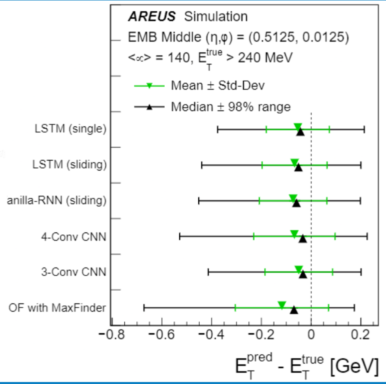




Energy Reconstruction Performance

- CNNs and RNNs outperform OF in both mean and resolution
- CNNs and RNNs better reconstruct pulses distorted by previous events
    - Reconstruction of overlapping pulses is better when additional information from past events are used (number of samples before the pulse)



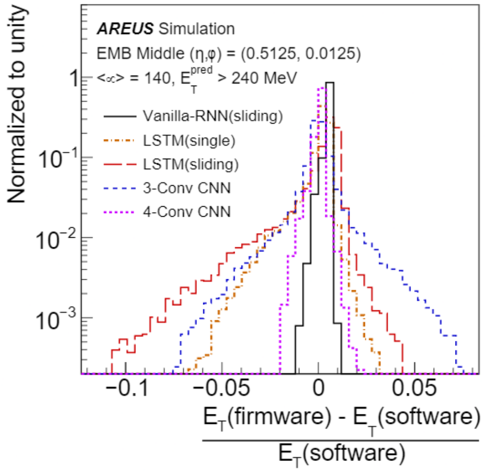


FPGA Firmware Implementation

- CNN: Direct Implementation in VHDL
    - Optimal use of DSP internal architecture in FPGA
    - Architecture automatically configured from files obtained during training
- RNN: High Level Synthesis approach
    - More flexibility in architecture, optimized for high frequency
- Good agreement between firmware simulation results and software reference model

	3-Conv CNN	4-Conv CNN	Vanilla RNN (sliding)	LSTM (single)	LSTM (sliding)
Frequency F_{max} [MHz]	493	480	641	560	517
Latency clk_{core} cycles	62	58	206	220	363
Resource Usage #DSPs	46 0.8%	42 0.7%	34 0.6%	176 3.1%	738 12.8%
#ALMs	5684 0.6%	5702 0.6%	13115 1.4%	18079 1.9%	69892 7.5%

Performance and rss usage for single instance



	3-Conv CNN	4-Conv CNN	Vanilla RNN
Multiplicity	6	6	15
Frequency F_{max} [MHz]	344	334	640
Latency clk_{core} cycles	81	62	120
Max. Channels	390	352	576
Resource Usage #DSPs	46 0.8%	42 0.7%	152 2.6%
#ALMs	14235 1.5%	15627 1.7%	5782 0.6%

Performance and rss usage for multiplexed model

- CNN shows lower latency and resource (rss) usage but needs improvement for frequency in the multiplexed model.
- Vanilla RNN meets the requirements for rss usage and Clock frequency in the multiplexed model, but has higher latency than CNN.

Energy Reconstruction using CNNs/RNNs can be implemented on LASP FPGA and show good agreement between software and firmware model and outperform OF

FPGA Hardware Implementation

- RNN implemented on Intel FPGA development Kit Stratix 10 GX
- Build a test firmware environment (with VHDL) around the RNN IP generated from the HLS code
- Tests on Hardware compatible with firmware simulation
- Validation of the design in the Hardware: The Firmware runs at the expected frequency without timing violations.
- Tests carried with one RNN implemented in the FPGA, multiple RNN instances inside one FPGA is ongoing.

Peter Matta on behalf of the ATLAS Liquid Argon Calorimeter Group
Ecole des Mines de Saint-Etienne, Centre de Physique des Particules de Marseille