Embedded Neural Networks on FPGAs for Real-Time Computation of the Energy Deposited in the ATLAS Liquid Argon Calorimeter

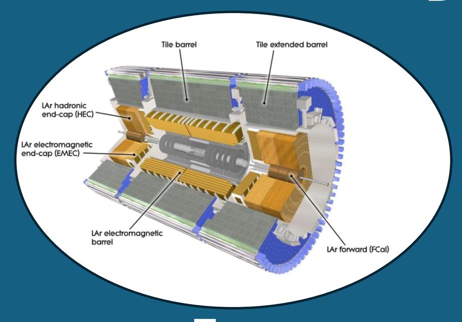
-

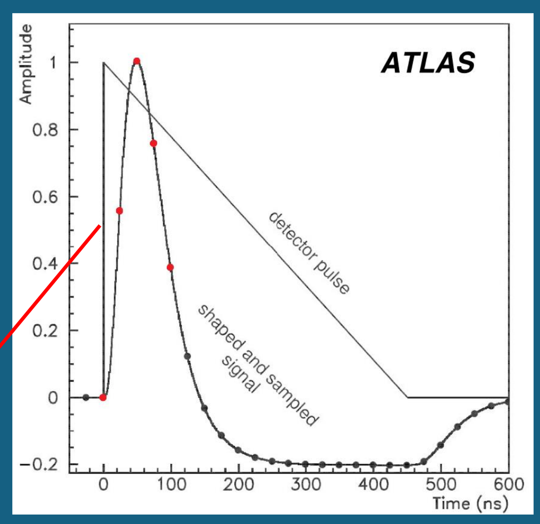
AI and Physical Sciences @AMU

Raphaël BERTRAND, Aix Marseille Univ, CNRS/IN2P3, CPPM, Marseille, France

The **Phase-II upgrade** of the LHC will **increase its instantaneous luminosity** by a factor of 5-7 leading to the HL-LHC. At the **HL-LHC**, up to **200 proton-proton collisions** are expected per bunch crossing (BC), every 25 ns. **ATLAS** is one of the **general-purpose detectors** at the LHC. The **Liquid Argon (LAr) calorimeter** of the ATLAS experiment **measures the energy of particles produced** in LHC collisions. The LAr calorimeter is formed of **182000 cells** and **provide trigger capabilities** to the ATLAS detector. Therefore, the LAr calorimeter readout electronics is required to collect and **process the calorimeter data at 40 MHz**.

Digitization and overlapping signals



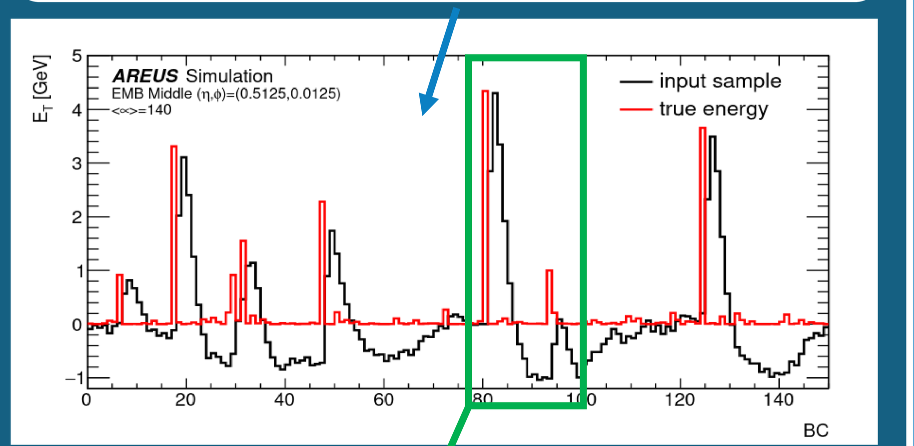


$$E_t = \sum_{i=1}^{5} c_i * x_{t+i}$$

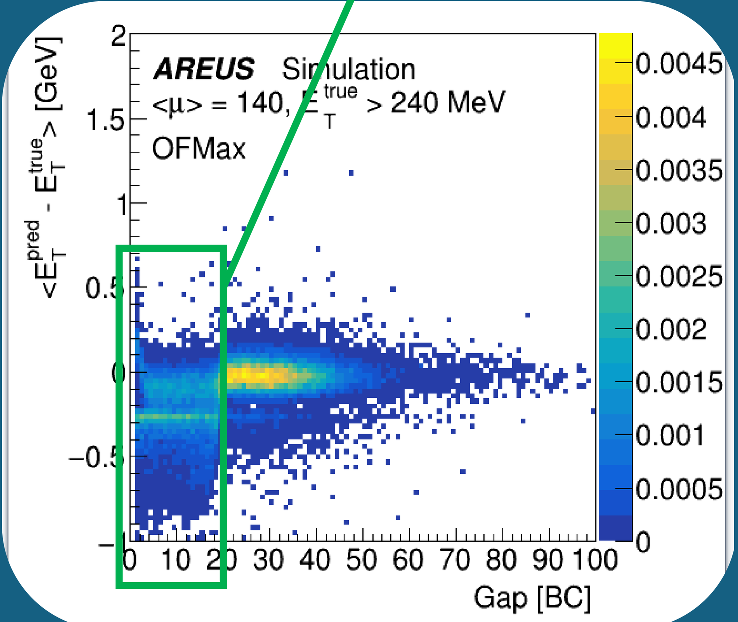
- Energy deposits in LAr calorimeter cells lead to an **electric pulse** which is **shaped**, **sampled** and **digitized at 40 MHz**.
- The pulse **spans about** 625 ns corresponding to **25 BCs**.

- The **samples around the pulse peak** are used to **reconstruct** its **amplitude** which is proportional to the **energy deposited** in the corresponding cell.
- Currently, an **optimal filtering algorithm (OF)** is used to **compute the deposited energy**.
- The computation is done using electronics boards based on the **FPGAs technology** that can efficiently **handle the high bandwidth** (~250 Tb/s).

Simulated pulse chain with a pileup of 140 and additional injected pulses at higher energy to represent hard-scattering events



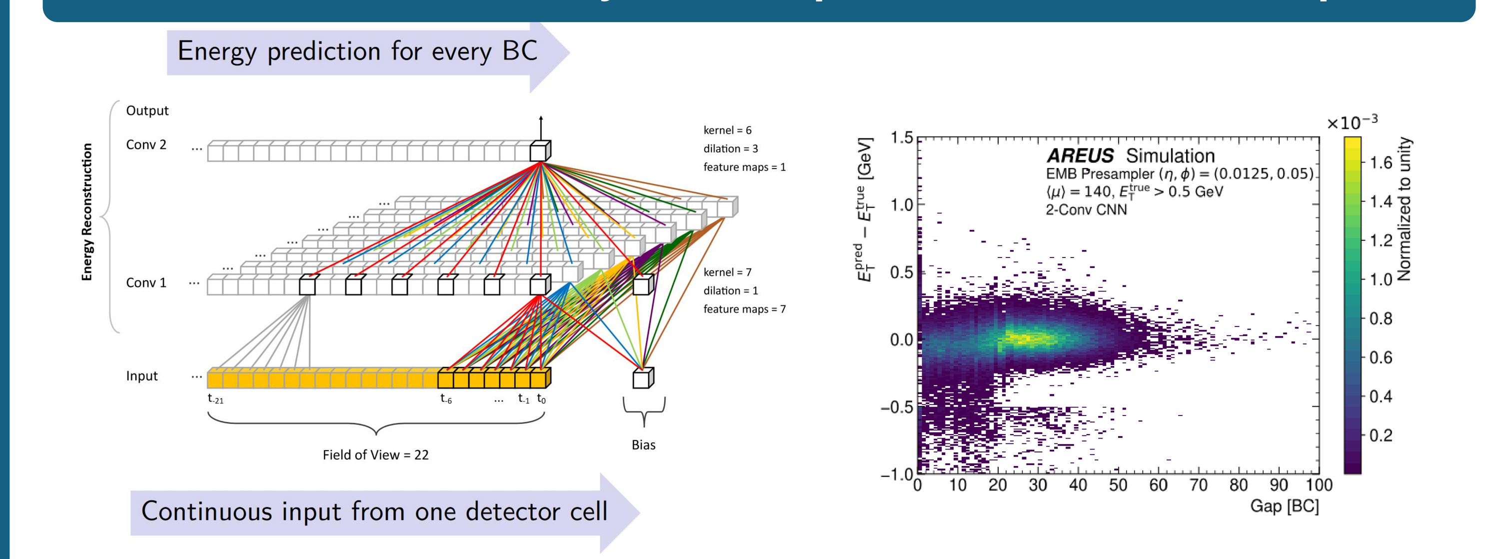
Region where pulses overlap



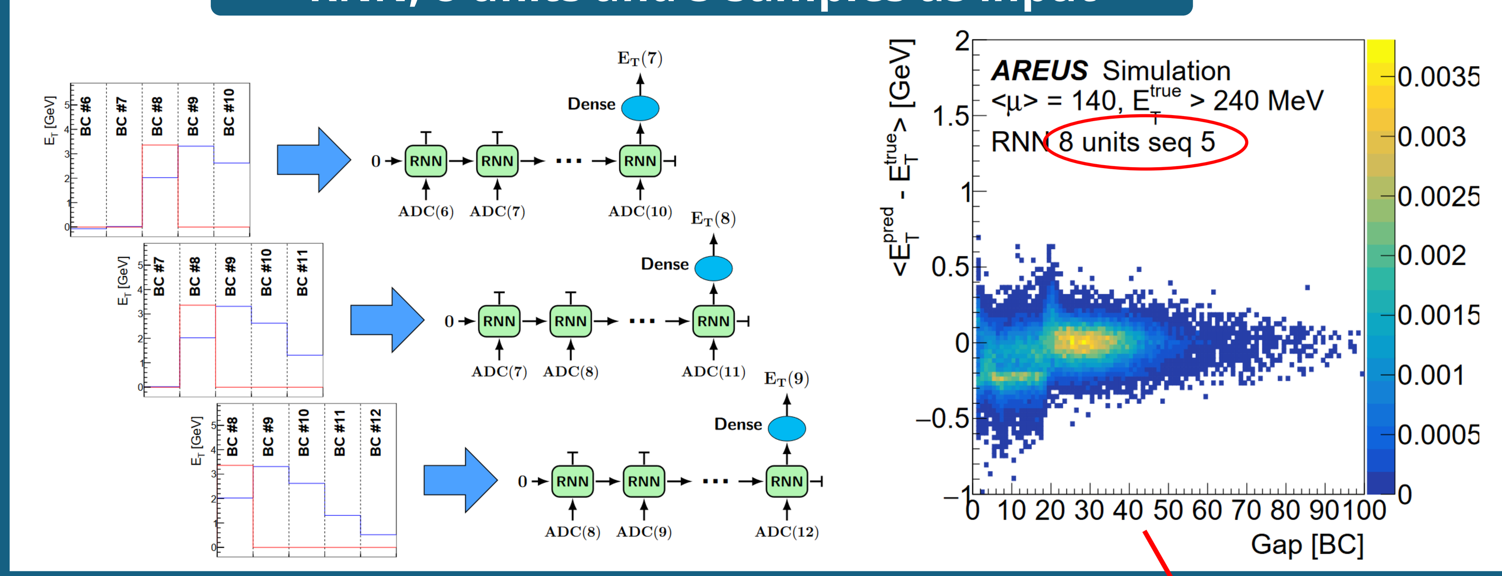
- Due to the HL-LHC **high pileup**, the **probability of overlapping pulses** will **increase**.
  - Optimal filtering **performance degrades** with overlapped pulses.
- **Neural networks** are investigated as an **alternative solution** to the optimal filtering algorithm.

Neural networks and performance

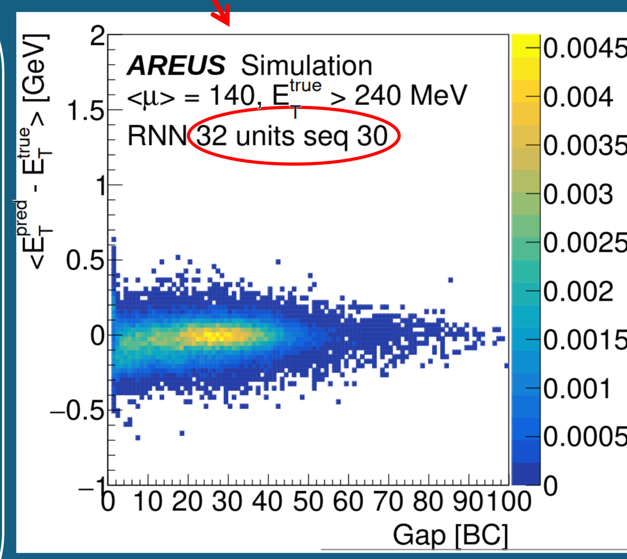
CNN, 2 convolutional layers, ~100 parameteres and 22 samples



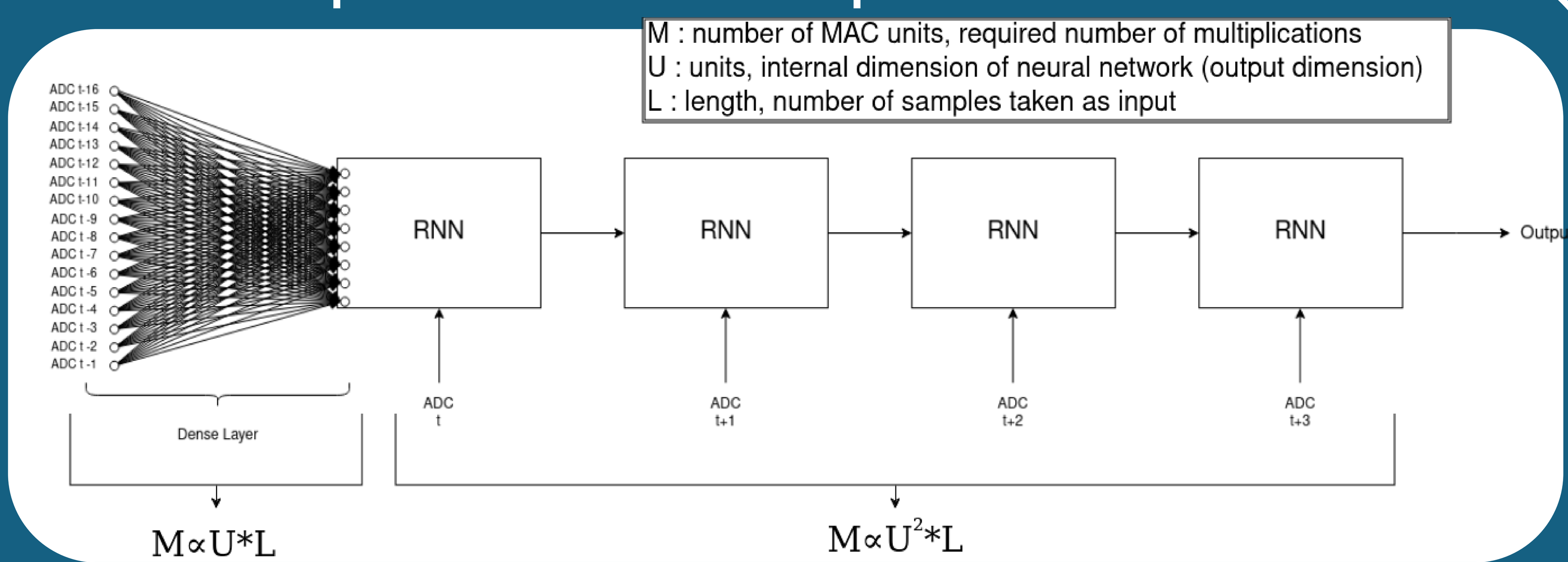
RNN, 8 units and 5 samples as input



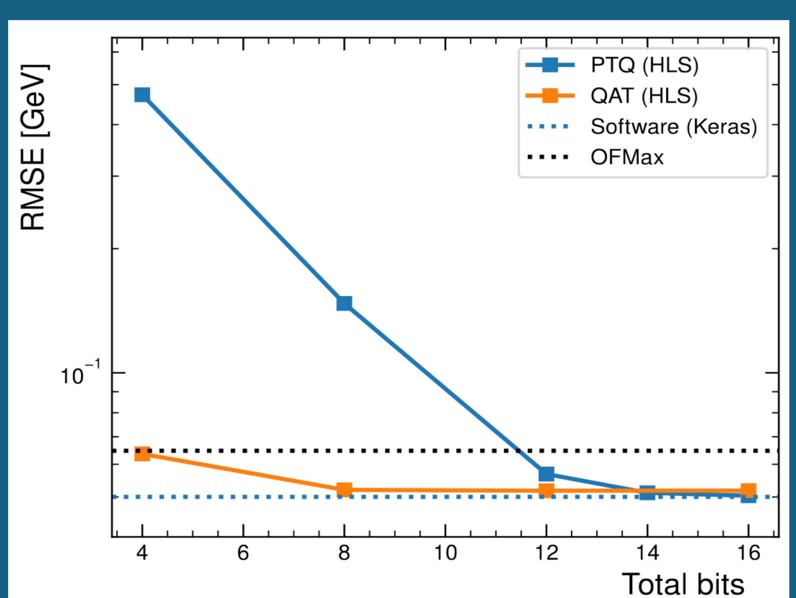
- **CNNs** and **RNNs** are designed to **compute deposited energy**.
  - **4 samples around the peak** are sufficient to **reconstruct the amplitude.**
  - **Samples prior the peak correct** the effect of the **previous deposit**.
- NNs can correct the degradation of the energy resolution.
- **Better performance** is observed :
  - With **increased number of samples** prior to the energy deposit.
  - With **increased internal dimension** (units).



Optimization of NN computational needs

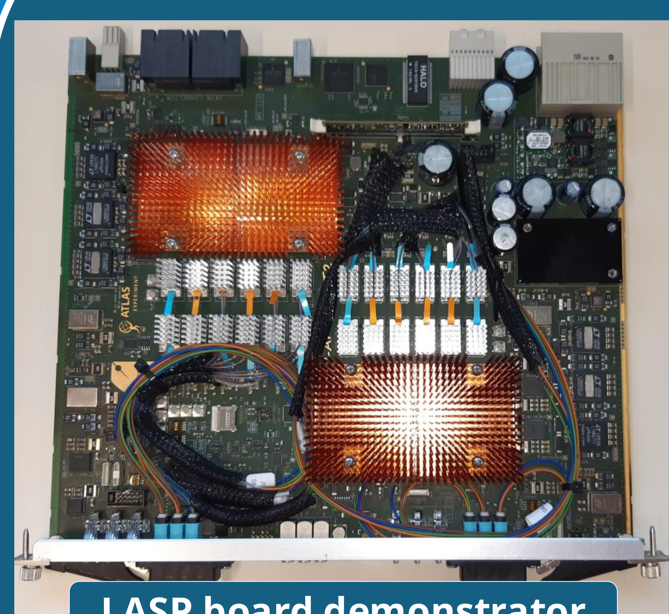


- Long sequence length requires many **RNN cells** that are **computational heavy.**
- Replace the RNN cells **prior to the pulse** by a Dense layer
  - **RNN** to compute the **amplitude** on the peak
  - **Dense** to correct for **pileup**
- **No performance drop** with this optimization.



- **Fixed-point** representation of arithmetic operations to **reduce the resources required** in FGPAs
- Possible **degradation** of the resolution in firmware **due to quantization**
- The number of **bits can be reduced by a factor of 2** by using **Quantization Aware Training (QAT)** rather than posterior quantization of weights trained with float precision.

Hardware implementation

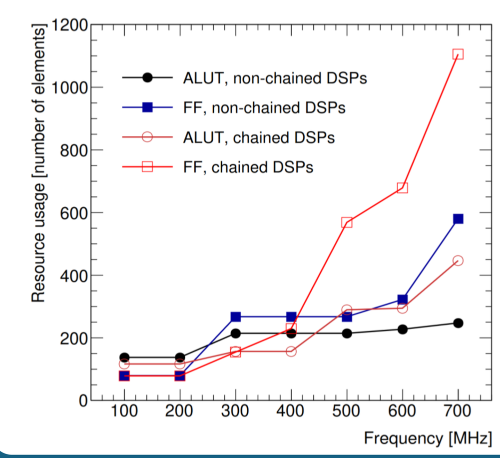


LASP board demonstrator

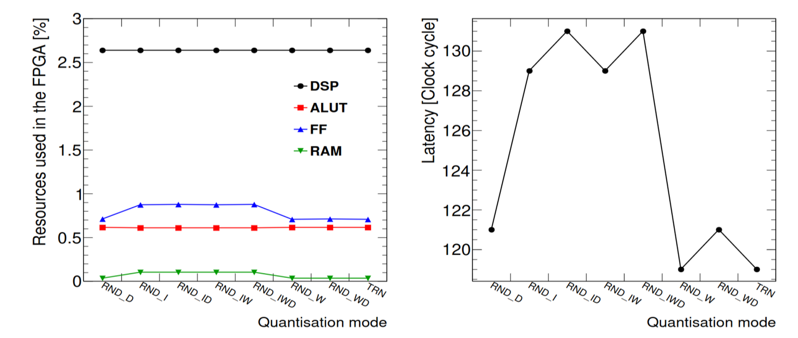
LASP demonstrator built with FPGA Stratix-10
- Prototype with Agilex 7 ongoing

Each FPGA needs to reconstruct the energy for **384 channels** :
- Impossible to implement 384 NNs on the FPGA
  - Need **multiplexing**
  - Need to go to **higher frequency** (multiple of 40 MHz)

Firmware implementation



Optimization of matrix multiplication :
- **Basic building block of NNs**
- **Scalar multiplications implemented with DSP**
- **Multiple solutions for additions implementation :**
  - **Chained DSPs design better at low frequency**
  - **Usage of general logic (ALUT) better at high frequency**



Quantization mode impact on resources usage and the latency

RNN and CNN Implemented on Stratix-10
- **CNN implemented on Agilex, RNN still in progress**
- **CNN directly implemented in VHDL, RNN Implemented first in HLS for fast prototyping and then optimized in VHDL**
- **Fits LAr requirements for both**

FPGA	Network	Multiplex.	Detector cells	$f_{\max}$	ALMs	DSPs
Stratix-10	RNN (HLS)	10	370	393 MHz	90 %	100 %
	RNN (VHDL)	14	392	561 MHz	18 %	66 %
	CNN (100 param.)	12	396	415 MHz	8 %	28 %
Agilex	CNN (100 param.)	12	396	539 MHz	4 %	13 %
	CNN (400 param.)	12	396	510 MHz	19 %	50 %

CNN and RNN outperform Optimal Filtering algorithm, especially for overlapping signals. NNs are optimized to get the best performance while reducing the complexity to be able to implement them on FPGAs. **Both CNN and RNN** have been **implemented in FPGAs** and **match requirements.**

References :
- Georges Aad et al. “Firmware implementation of a recurrent neural network for the computation of the energy deposited in the liquid argon calorimeter of the ATLAS experiment”. JINST 18 (2023) P05017
- Georges Aad et al. “Artificial Neural Network on FPGAs for Real-Time Energy Reconstruction of the ATLAS LAr Calorimeters”. Comput Softw Big Sci 5, 19 (2021)
- Lauri Antti Olavi Laatu. PhD thesis. “Development of artificial intelligence algorithms adapted to big data processing in embedded (FPGAs) trigger and data acquisition systems at the LHC.” (2023) CERN-THESIS-2023-198
- Nemer Chiedde. PhD thesis. “Implementation of embedded artificial intelligence algorithms in the readout system of the ATLAS liquid argon calorimeter.“ (2023) CERN-THESIS-2023-294
- Etienne Fortin. PhD thesis. "Commissioning and performance of the trigger and readout system of the liquid argon calorimeter of the ATLAS experiment." (2022) CERN-THESIS-2022-134.
- Anne-Sophie Berthold. “Simulation Studies of Convolutional Neural Networks for the Real-Time Energy Reconstruction of ATLAS Liquid-Argon Calorimeter Signals at the High-Luminosity LHC”. TU Dresden, IKTP, Dec. 21, 2023.