Centre de physique des particules de Marseille

AI on FPGA

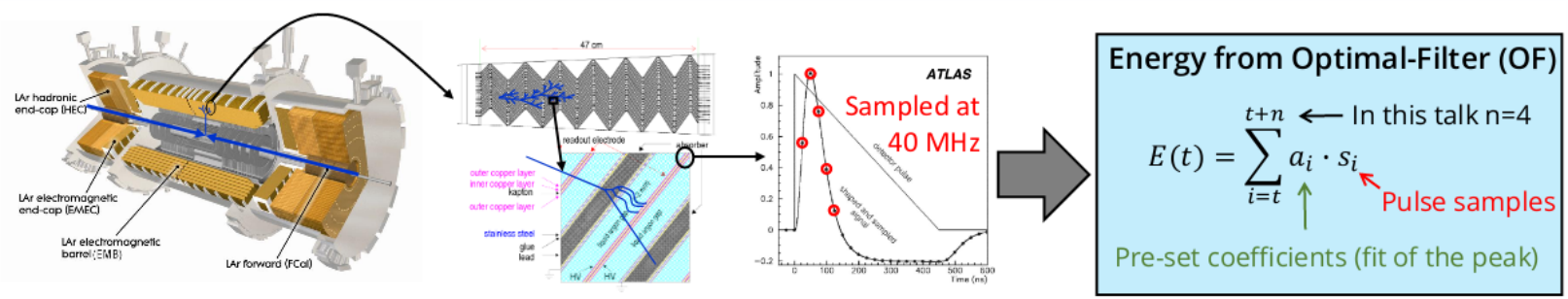
July 2, 2026

Etienne FORTIN

Outline

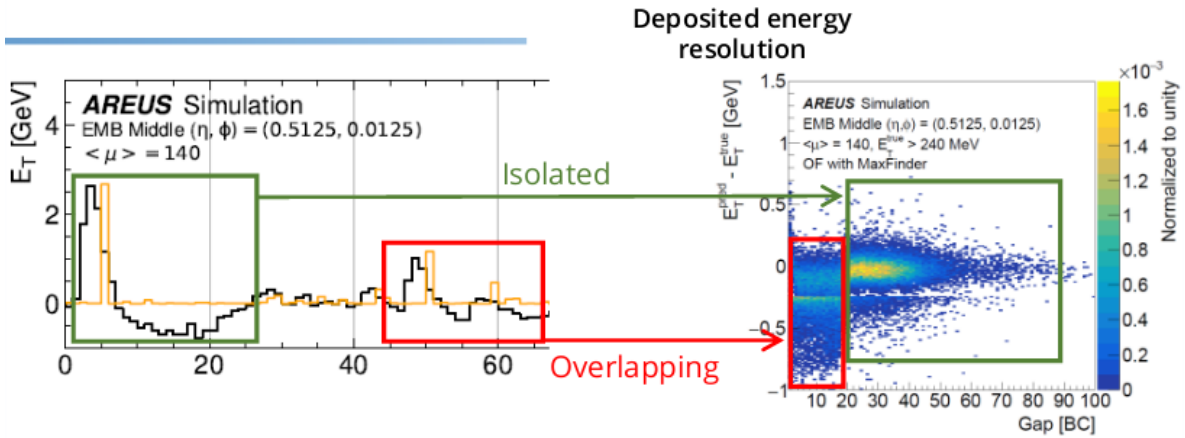
- Context
- Neural networks
- Implementation flow
- Some example of optimisations

AIDAQ



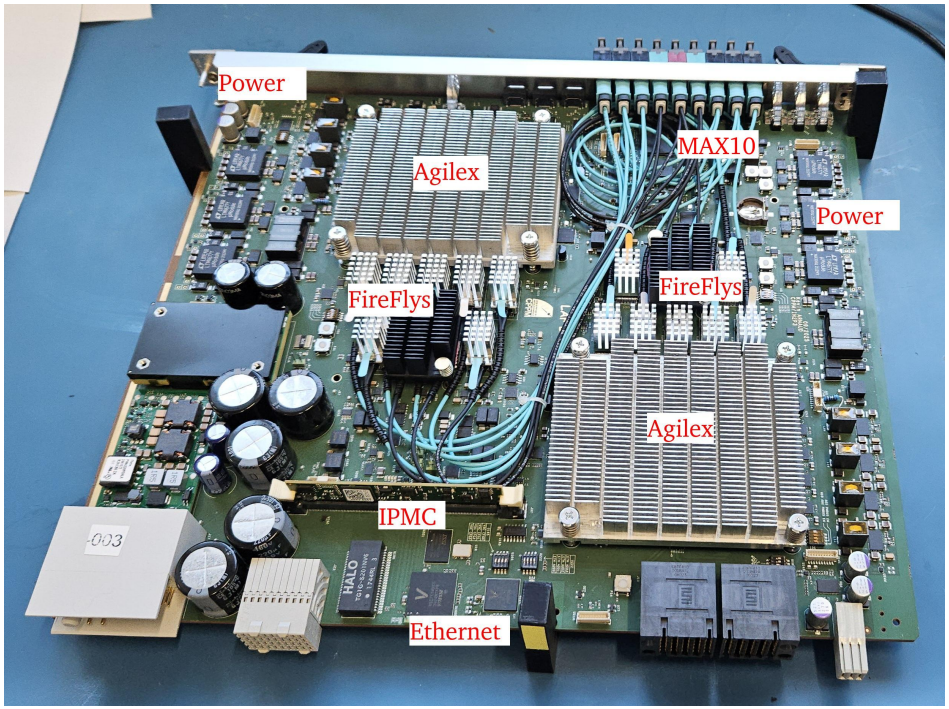
For the HL-LHC with increased luminosity probability of overlapping pulse in LAr Calorimeter will increase.

The actual algorithme for reconstruction (Optimal filtering) underestimate energies in such case.



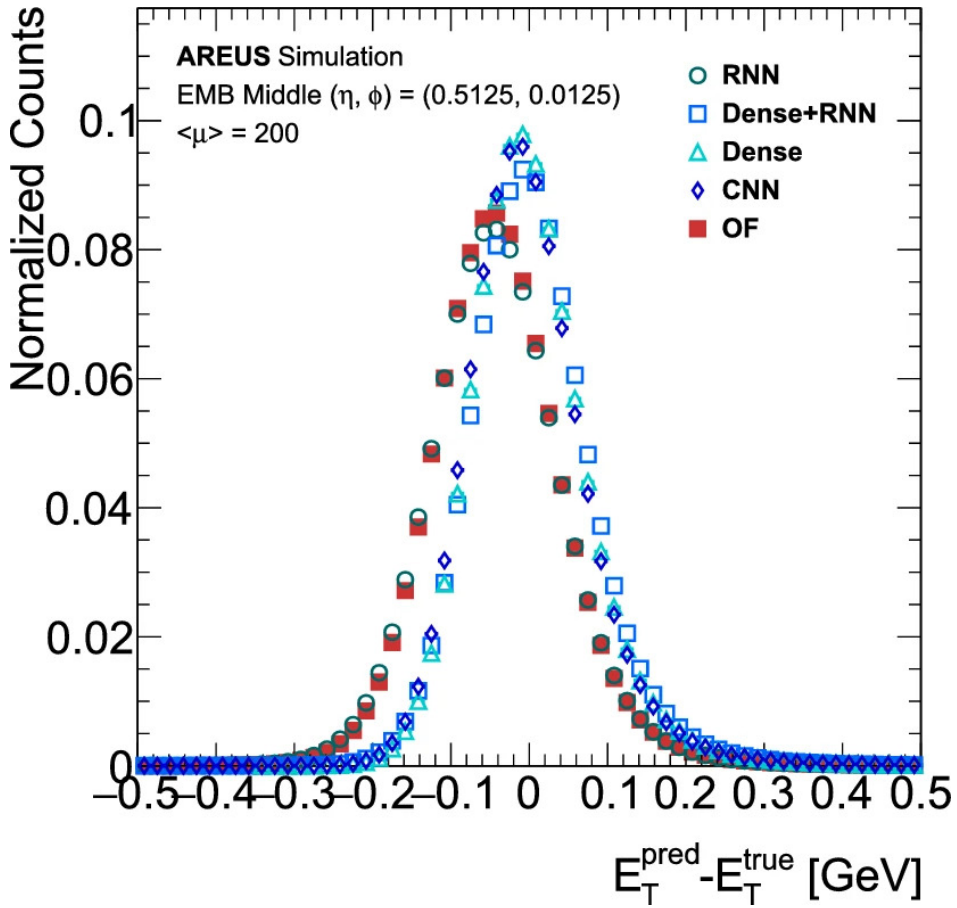
LASP

- A new back-end board will be in charge of energy reconstruction
- The LASP board (CPPM) with 2 Altera Agilex FPGA
  - Demonstrator board with Stratix
- Big enough to think about implementing neural networks.



Neural network performances

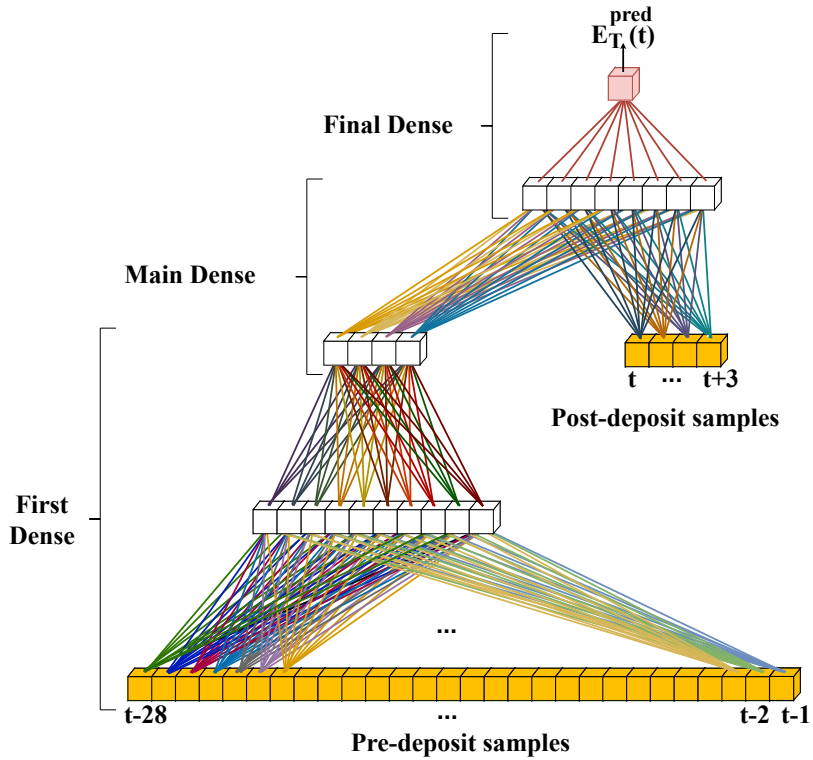
- Several NN algorithms studied in LAr
  - CNN : Dresden
  - RNN : CPPM
  - Dense : CPPM
- Outperforme the optimal filtering



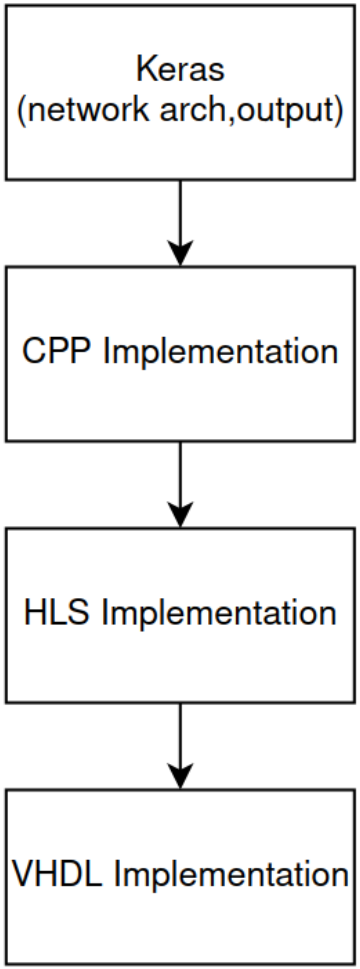
Aad, G., Bertrand, R., Laatu, L. et al. Optimised neural networks for online processing of ATLAS calorimeter data on FPGAs. Eur. Phys. J. C 86, 111 (2026). https://doi.org/10.1140/epjc/s10052-026-15321-y

Dense network

- Limiting ressources is the MAC(multiply and accumulate)
- 392 MAC in the network, 382 channels per LASP fpga (at 40MHz)
- 25582 MAC in the FPGA DSP
- 585% of FPGA
- We need to multiplex and increase frequency
  - Several cell in one network implementation
  - Target is 320MHz, 8 cells/network , 70% of FPGA MAC used



Implementation flow



- Check that the output if the same as keras
- Validate the understanding of the network architecture

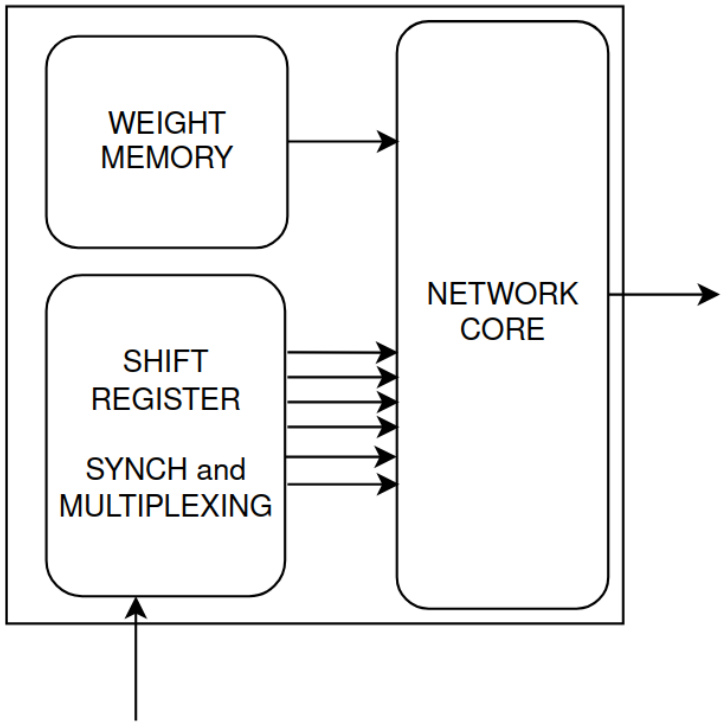
- Study quantization
- Architecture optimisation
- First insight on the ressource usage

- Ressource usage
- Fmax
- Placement optimisation

We made a python generator that generates all the codes from the keras file.

Architecture

- 3 blocks
  - Weight memory : Storing the weights and send them to the network
  - Core : network implementation
  - Shift register: Handle the time series to parrallel

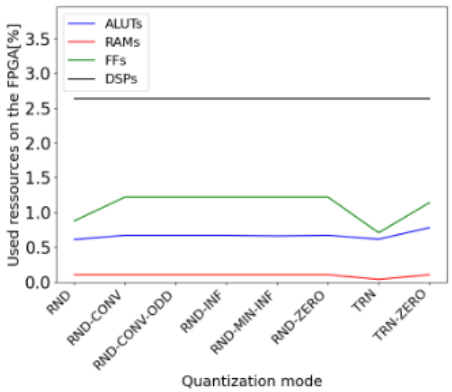


The optimisations of these blocks depends on the network itself.

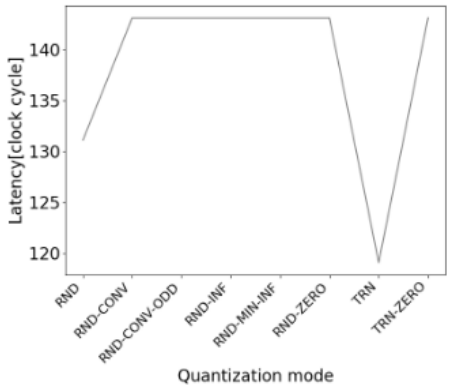
HLS optimisation: Quantisation

Fixed point library provide 8 way of rounding when reducing number of bits of a signal. Usefull during keras values(float) conversion and output of DSP(37bits).

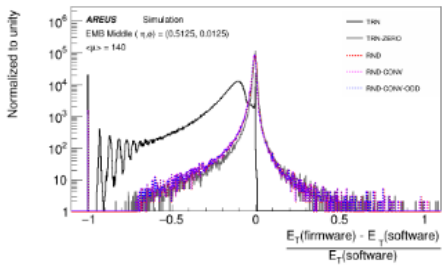
For a 5 cell vanilla network:

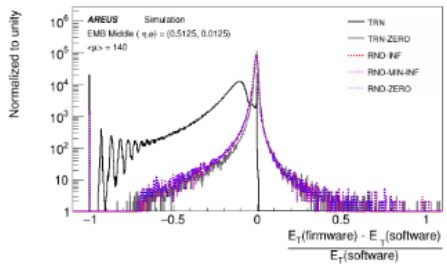


(a) Resource



(b) Latency



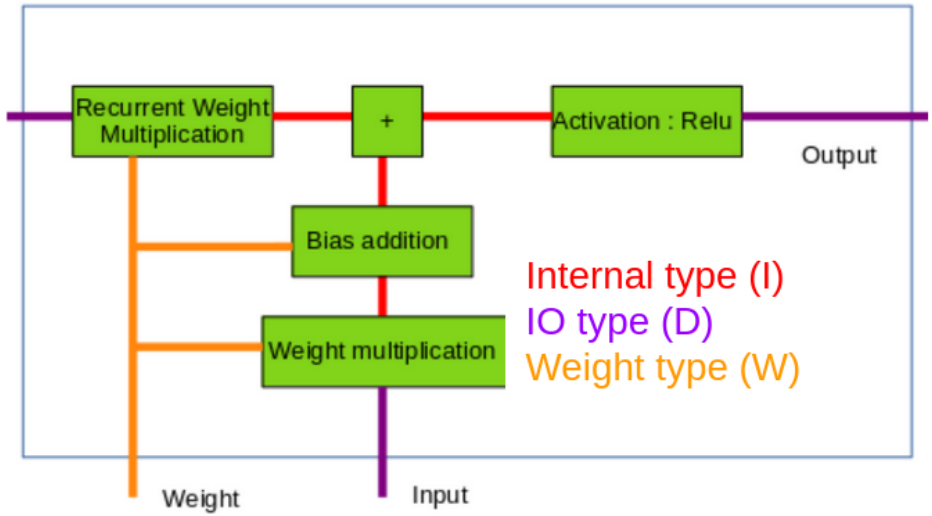


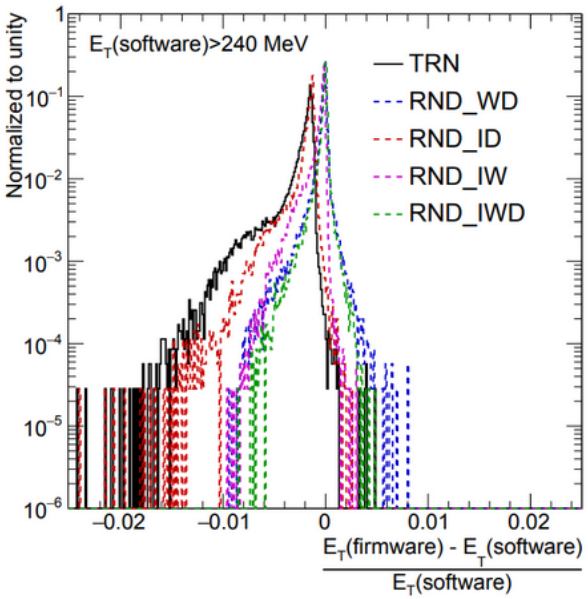
Study in details between RND and TRN

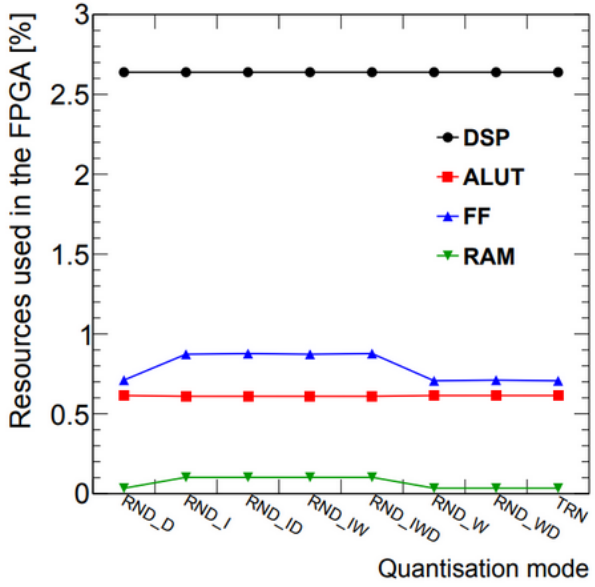
Studied in the RNN network, results used in Dense

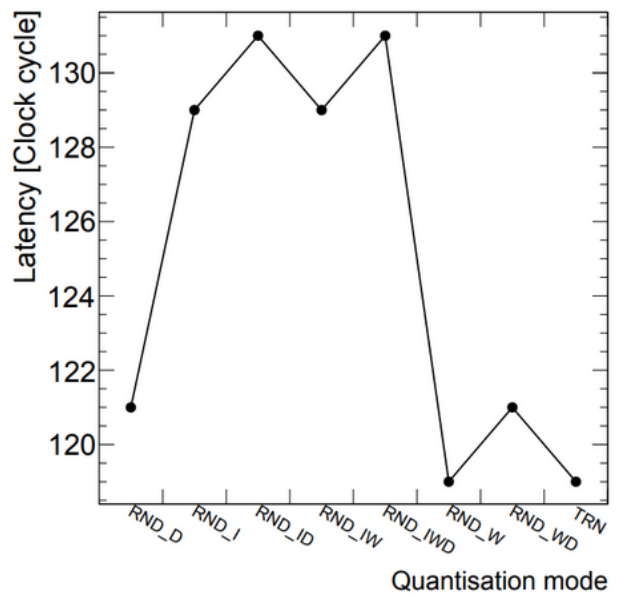
HLS optimisation: Truncation vs Rounding in detail

- Compromise between resolution and resource usage and latency
  - Truncation of I/O and Internal types leads to important reduction of latency with small impact on energy resolution
  - Weight type rounded in software: no impact on latency
- Truncation used in firmware



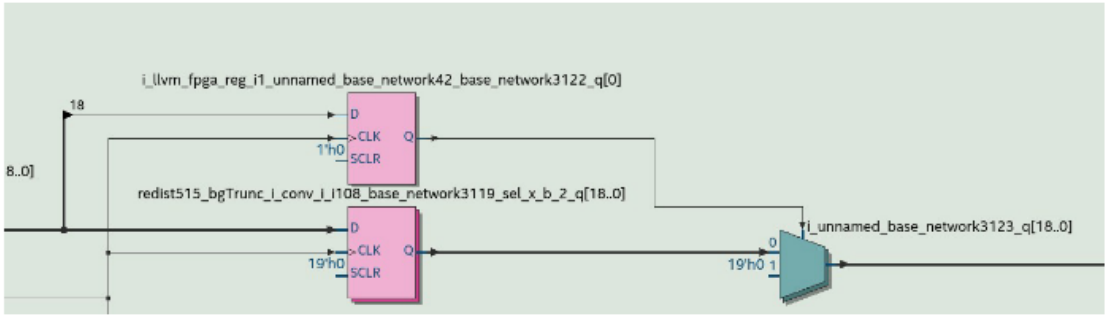






Optimisation: Activation function: RELU

- Activation functions are complex mathematical operation
- Not possible into firmare: use of Look up table
- But for each activation function we need:RAM for it....
- There is one exception the RELU: $f(x) = x$ if $x > 0$ else $f(x) = 0$



- Only logic element, no mathematical function as signed bit is used
- With a leaky relu $f(x) = x$ if $x > 0$ else $f(x) = \alpha x$ with α a power of 2 is the same

Conclusion

- Ressource usage of our network (For all channels)

	DSP	**Logic units**	**Memory**
Weights in memory	73%	43%	3.92%
Weights in logic	73%	0%	15.66%

- A proof of concept to demonstrate that we can implement neural networks in the HL-LHC readout of the LAr Calorimeter
- Still a lot of challenges in performance, calibration, clustering

Questions ?