Convolutional Neural Networks on FPGAs for Processing of ATLAS Liquid Argon Calorimeter Signals

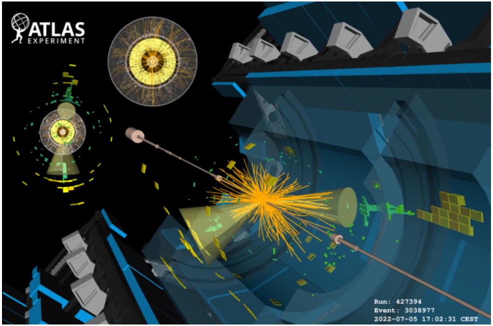
DPG SMuK 2023

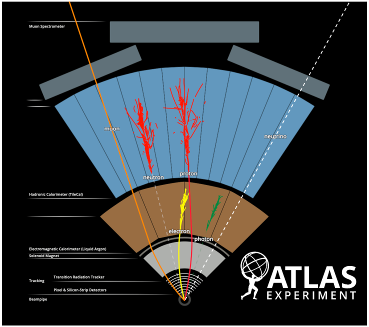
Johann C. Voigt – TU Dresden

21 March 2023

TECHNISCHE UNIVERSITÄT DRESDEN
INSTITUTE OF NUCLEAR AND PARTICLE PHYSICS
PUNCH 4NFDI
FIS ATLAS
GEFÖRDERT VOM Bundesministerium für Bildung und Forschung

LHC and ATLAS

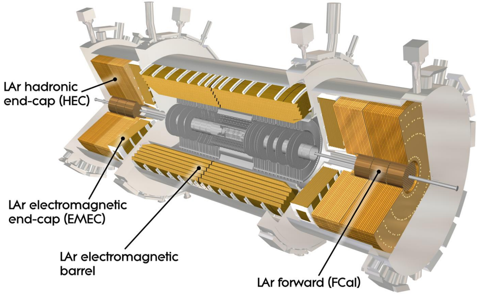


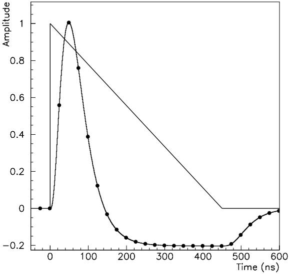


- LHC provides ≈ 50 proton-proton collisions every 25 ns $\widehat{=}$ 40 MHz
  $\rightarrow$ 140-200 simultaneous collisions after upgrade
- ATLAS Phase-II upgrade to prepare for higher load

https://cds.cern.ch/record/2814924 [2], https://cds.cern.ch/record/2770815 [5]

LAr-Calorimeter



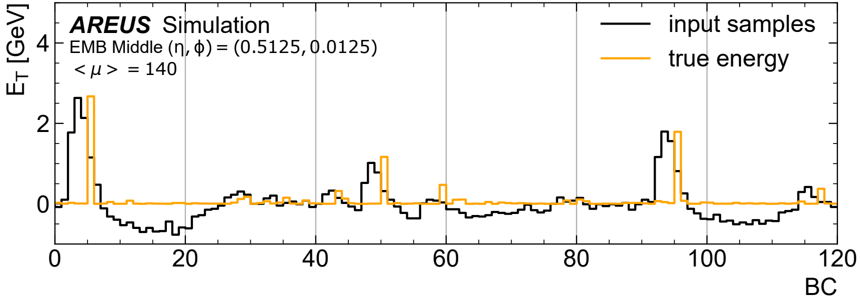


- $\approx 180\,000$ channels $\rightarrow$ data stream of $\approx 235\,\mathrm{Tbit\,s^{-1}}$
- Triangular detector pulses $\rightarrow$ Analogue pulse shaping $\rightarrow$ Digitization
- Digital energy reconstruction with Optimal Filter (OF)

$$E(t) = \sum_i c_i \cdot x(t-i)$$

https://cds.cern.ch/record/1095928 [6], http://cds.cern.ch/record/1701107 [3]

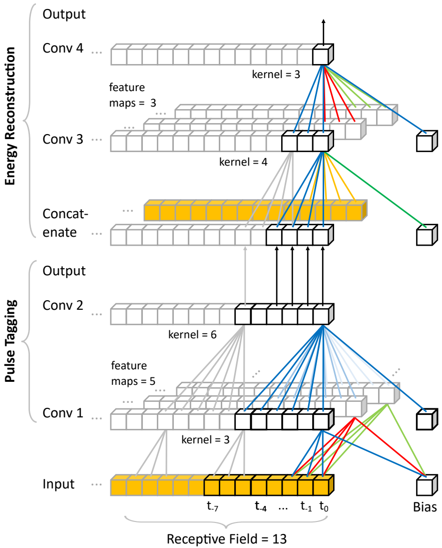
Example input sequence



→ Reconstruct true energy from digitized detector output

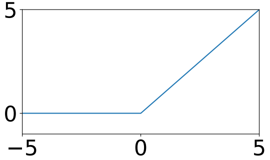
https://doi.org/10.1007/s41781-021-00066-y [1]

CNN architecture for energy reconstruction

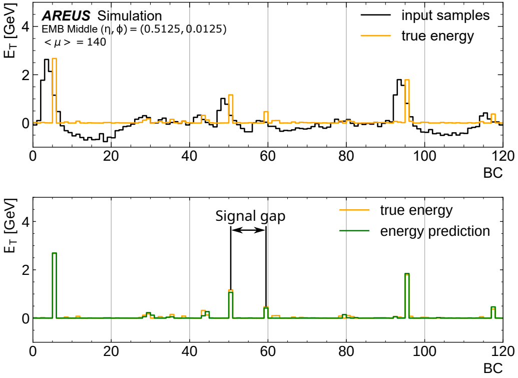


https://doi.org/10.1007/s41781-021-00066-y [1]

- Input: 1D time series of ADC samples (one detector cell)
- Output: Sequence of reconstructed energies
- CNN layer:
  - Linear combination of output of previous layers
  - Non-linear activation (i.e. ReLU)



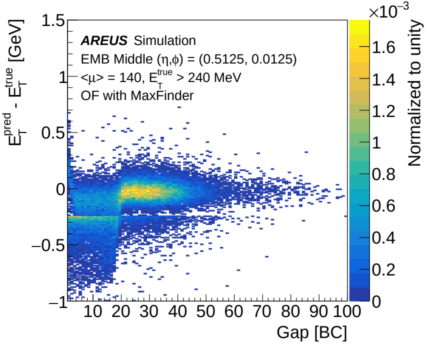
CNN example sequence



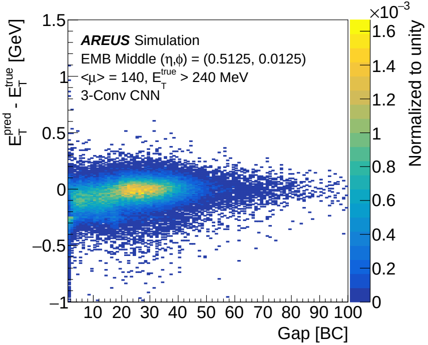
https://doi.org/10.1007/s41781-021-00066-y [1]

- Input sequence from AREUS detector simulation
- Energy reconstruction trained with true deposited energy as target
- Network performance can be evaluated by comparison with true energy

CNN energy resolution as a function of gap



Optimal Filter



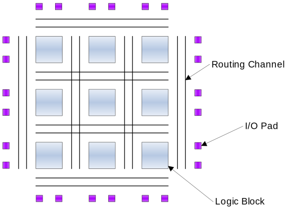
3-Conv CNN

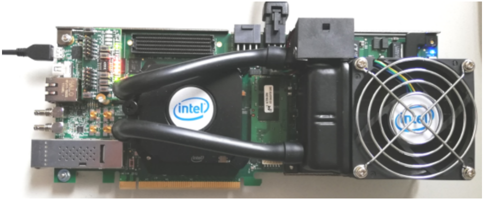
→ Significant improvement in reconstruction of overlapping pulses

https://doi.org/10.1007/s41781-021-00066-y [1]

FPGAs

- Configurable hardware chip with flexible logic cells and interconnection
- Parallelisation, Pipelining, high input/output rates
- Synthesis tool translates HDL code into configuration for FPGA
  $\rightarrow$ Combines speed of hardware solution with flexibility of software
- Main resource constraints: Number of general logic cells (ALMs) and specialized multiplication units (DSPs) vs. maximum clock frequency

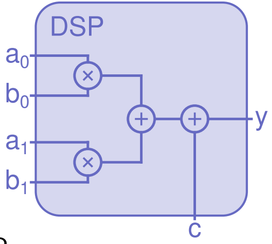


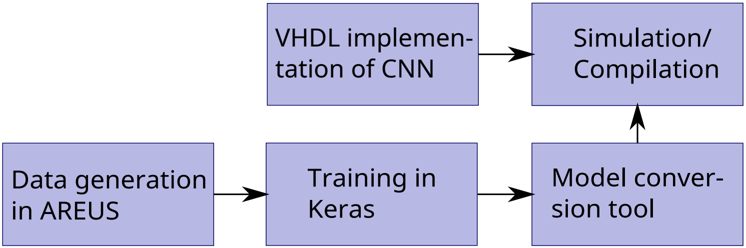


https://commons.wikimedia.org/wiki/File:Fpga_structure.svg [4]

CNN firmware implementation

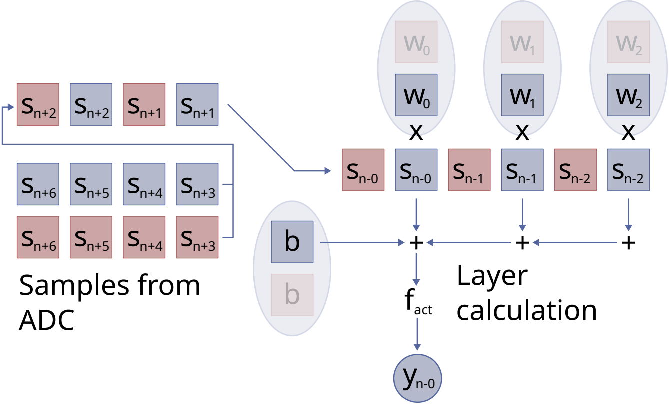
- Flexible/generic 1D-CNN model implemented directly in VHDL
- Optimized for DSP usage and latency
- DSPs can be chained for efficient multiply-add structures
- Depends on special architecture of Agilex DSPs
- Fixed point calculation with 18 bit total bit width





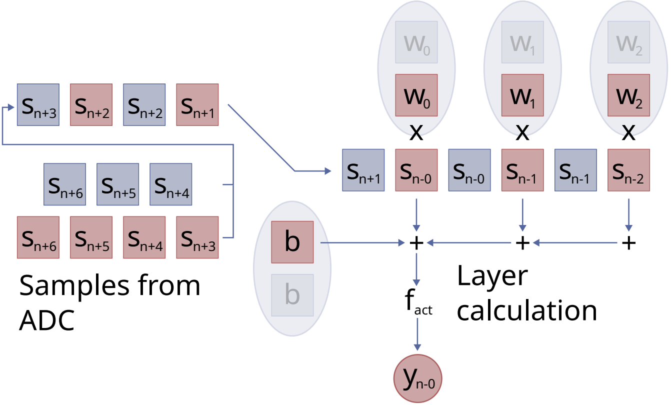
Multiplexing

- One FPGA needs to fit 33 CNN instances
  $\rightarrow$ Use less than 3 % of FPGA resources per instance
- Each instance uses 12$\times$ multiplexing
  $\rightarrow$ Design needs to run at 12$\times$ the ADC frequency: 480 MHz



Multiplexing

- One FPGA needs to fit 33 CNN instances
  $\rightarrow$ Use less than 3 % of FPGA resources per instance
- Each instance uses 12$\times$ multiplexing
  $\rightarrow$ Design needs to run at 12$\times$ the ADC frequency: 480 MHz



FPGA resource estimation

- Store weights in order required by DSP chain
  → Move complexity to pre-processing on computer
  → Previous version stored weights in logical order with very high resource impact
- 150 ns latency meets the trigger requirements

3-Conv Network	f_{max}	ALMs	DSPs
1 instance (12 channels)	570 MHz	6 k (0.4 %)	46 (0.4 %)
33 instances (384 channels)	537 MHz	186 k (14 %)	1518 (12 %)

Summary

- Flexible VHDL implementation supporting 1D CNNs for continuous input data stream
- Multiplexing support with low resource overhead
- Design fits target FPGA and runs at required clock frequency

[1] Georges Aad et al. "Artificial Neural Networks on FPGAs for Real-Time Energy Reconstruction of the ATLAS LAr Calorimeters". In: *Computing and Software for Big Science* 5.1 (Oct. 2021). DOI: 10.1007/s41781-021-00066-y. URL: https://doi.org/10.1007/s41781-021-00066-y.

[2] ATLAS Collaboration. *ATLAS First Collisions of LHC Run 3*. July 5, 2022. URL: https://cds.cern.ch/record/2814924.

[3] ATLAS Collaboration. "Monitoring and data quality assessment of the ATLAS liquid argon calorimeter". In: *JINST* 9.arXiv:1405.3768. CERN-PH-EP-2014-045 (May 2014). Plot available separately: http://atlas.web.cern.ch/Atlas/GROUPS/PHYSICS/PAPERS/LARG-2013-01/, P07024. 39 p. URL: http://cds.cern.ch/record/1701107 (visited on 05/28/2017).

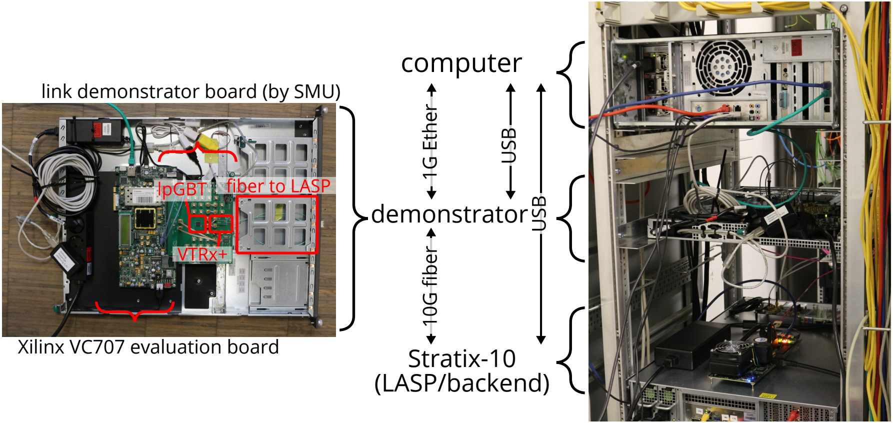
Sources II

[4] Johnteslade commonswiki. *Structure of an FPGA*. Feb. 22, 2006. URL: https://commons.wikimedia.org/wiki/File:Fpga_structure.svg (visited on 03/19/2023).

[5] Sascha Mehlhase. *ATLAS detector slice (and particle visualisations)*. 2021. URL: https://cds.cern.ch/record/2770815.

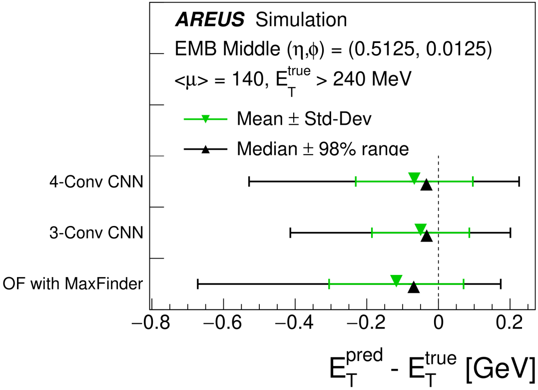
[6] Joao Pequenao. *Computer generated image of the ATLAS Liquid Argon*. CERN. Mar. 27, 2008. URL: https://cds.cern.ch/record/1095928 (visited on 03/29/2021).

Test setup



link demonstrator board (by SMU)
lpGBT
fiber to LASP
VTRx+
Xilinx VC707 evaluation board
computer
1G Ether
USB
USB
demonstrator
10G fiber
Stratix-10
(LASP/backend)

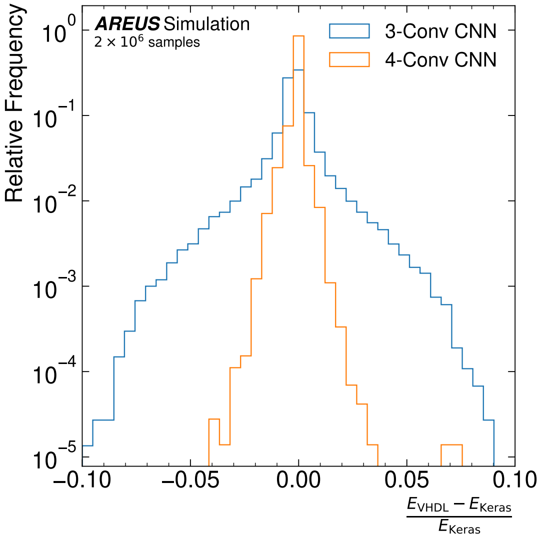
Energy resolution



- CNNs show better energy resolution and less bias than optimal filter (OF)

https://doi.org/10.1007/s41781-021-00066-y [1]

Relative deviation between firmware and software



- Good agreement between firmware and software (for samples with pred. energy above 240 MeV)

https://doi.org/10.1007/s41781-021-00066-y [1]