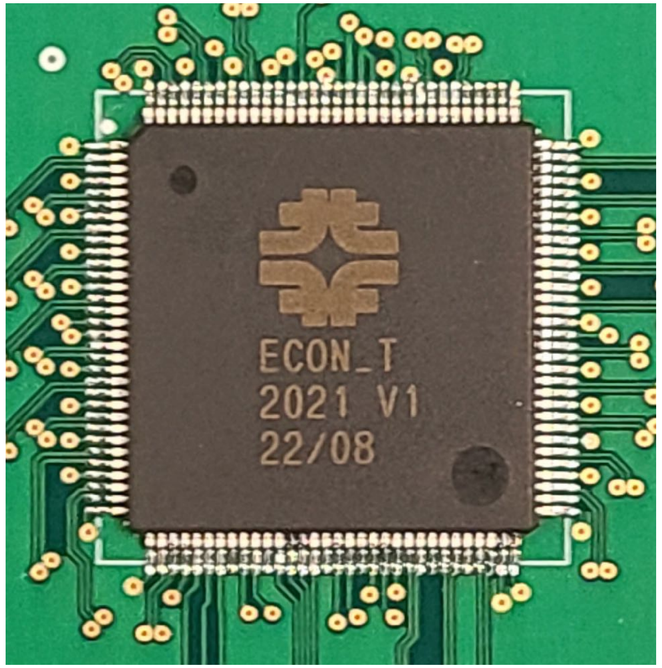
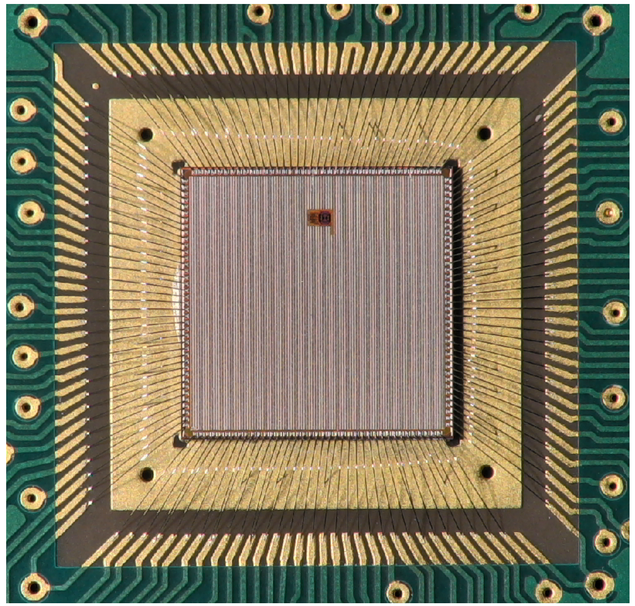
Design and Initial Test Results of a Autoencoder for data compression for ECON-T ASIC for the CMS HGCAL upgrade



Danny Noonan, for the ECON ASIC team

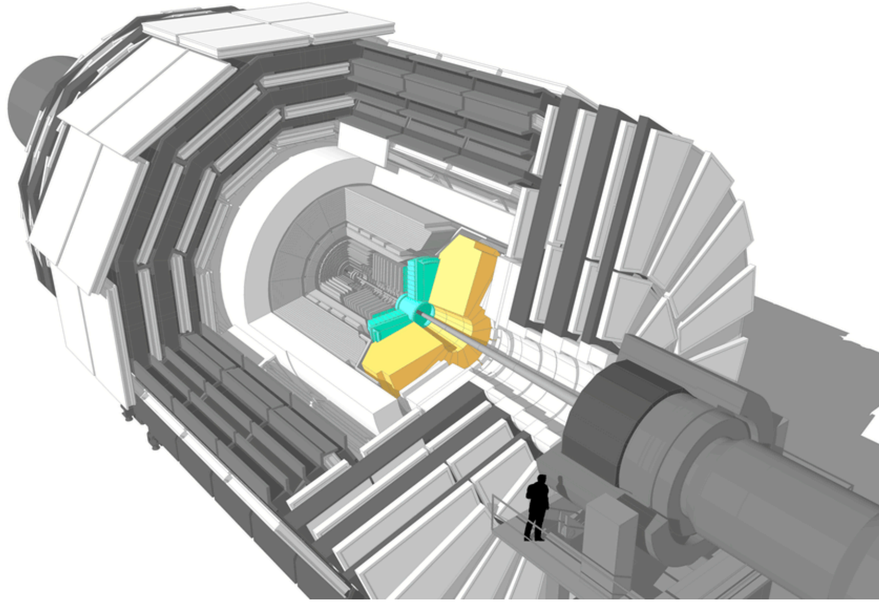
Alpha Shanai[1], Chinar Syai[1], Christian Herwig[1], Cristina Mantilla Suarez[1], Cristian Gingu[1], Danny Noonan[1], Davide Braga[1], Farah Fahim[1], Giuseppe Di Guglielmo[1], Javier Duarte[2], Jim Hoff[1], Jennifer Ngadiuba[1], Jim Hirschauer[1], Jon Wilson[3], Martin Kwok[1], Manuel Valentin[4], Matteo Lupi[5], Maurizio Pierini[5], Nhan Tran[1], Pam Klabbers[1], Paul Rubinov[1], Phil Harris[6], Ralph Wickwire[1], Rohan Shenoy[2], Seda Memik[4], Sioni Summers[5], Xiaoran Wang[1] , Yingyi Luo[4]

[1]**Fermilab** [2]**UCSD** [3]**Baylor University** [4]**Northwestern University** [5]**CERN** [6]**MIT**



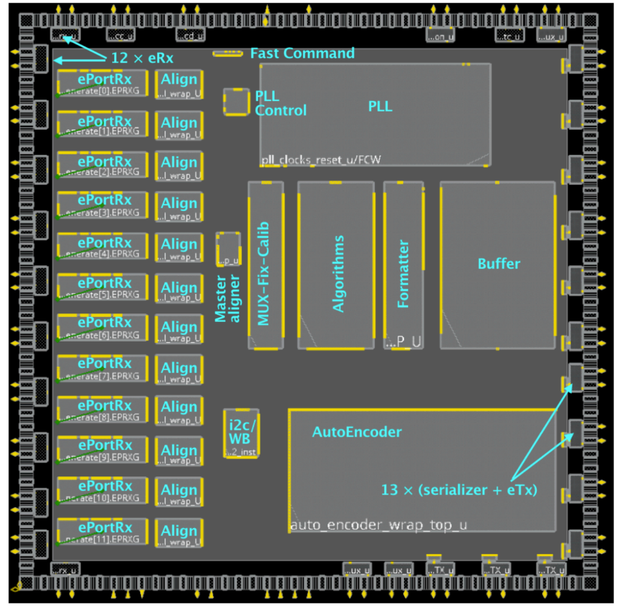
Fast Machine Learning for Science Workshop 2022
Southern Methodist University
October 3-6, 2022

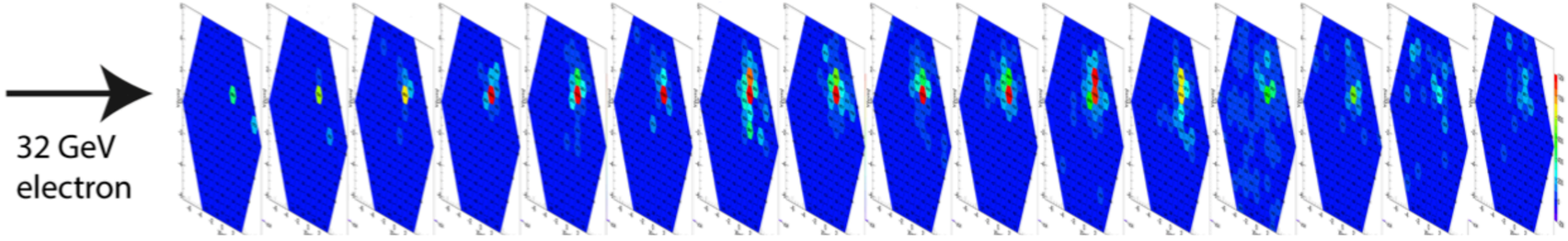
HGCAL and ECON-T ASIC



- High Granularity Calorimeter (HGCAL) is an upgrade to the CMS detector for the High Luminosity LHC
  - Replacement of end-cap calorimeter
  - Imaging calorimeter: improved position, timing, and energy resolution
  - ~6 million channels, leading to significant data volumes
- Endcap Concentrator Trigger (ECON-T)
  - ASIC designed to reduce data volume for readout of modules from HGCAL
  - Implemented in 65nm LP CMOS
  - Operates at 40 MHz, transmitting data used for making decision about whether trigger the readout of the full CMS detector
  - Reduces data volume through selection/compression of data coming from upstream readout chip
  - Prototype produced (ECON-T-P1) for testing/verifying functionality

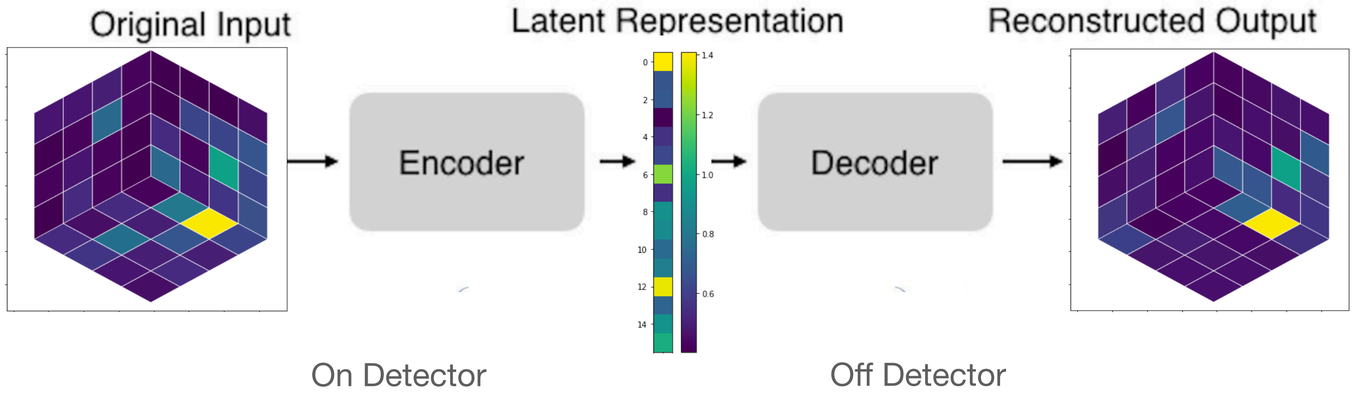
ECON-T ASIC floorpan

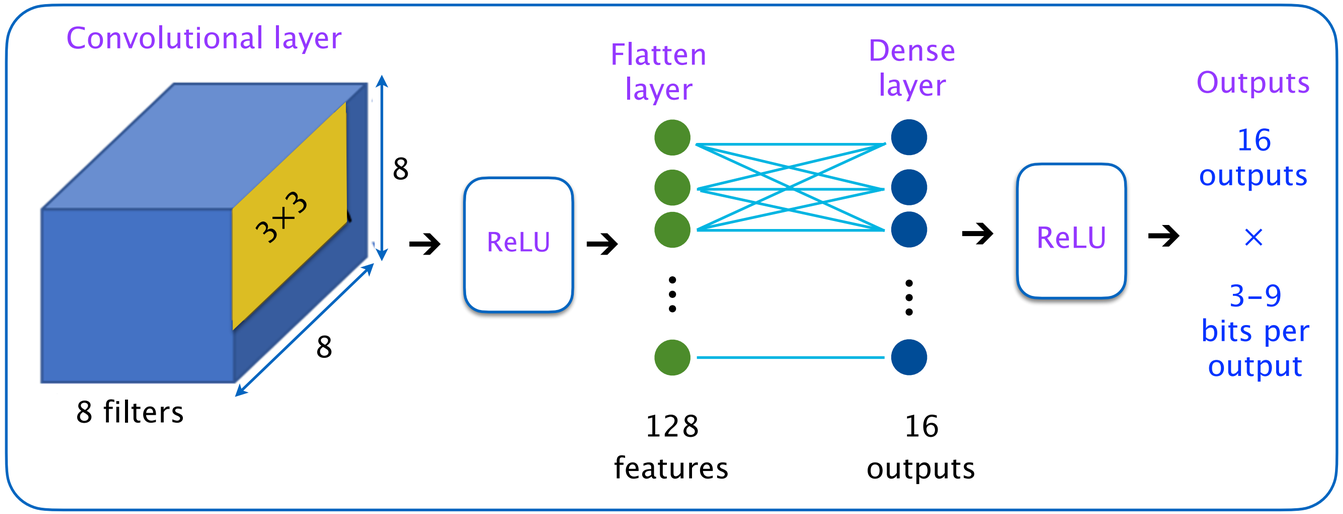




Autoencoder Algorithm

- Achieve data compression through an auto encoder implemented on the ASIC
  - Encode data to smaller latent space on ASIC, transmit smaller latent space off detector rather than full data
  - Up to 7x data compression
- Algorithm trained QKeras, implementation for the chip using hls4ml + Catapult HLS
- Encoder **network is reconfigurable** on ASIC
  - Network architecture is fixed, but weights and biases are reprogrammable on the chip over I2C
- All weight/bias registers are **triplicated and voted to improve radiation/SEU tolerance**

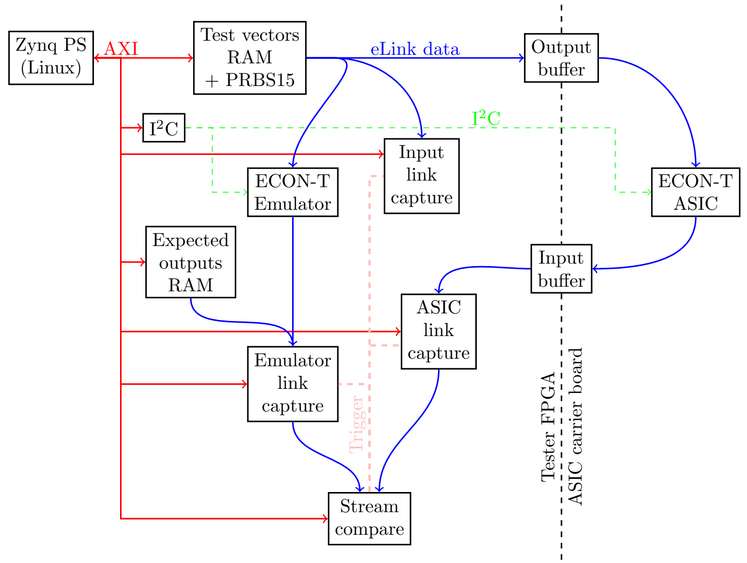




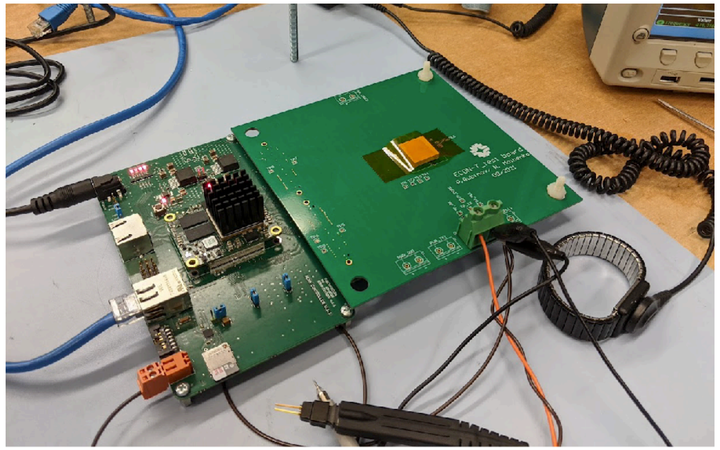
Testing Setup

- Test setup for feeding input data, clock, and control signals to ECON-T ASIC
  - Running on Zynq FPGA
- Provides **real-time check** of captured data, monitoring output stream and comparing to expected outputs
- Allows for **testing of full functionality** of the ECON-T ASIC
  - Performance of all components
  - Measuring power consumption
- Same setup can be used for bench testing as well as measuring ASIC performance in proton and X-ray beams to study radiation tolerance

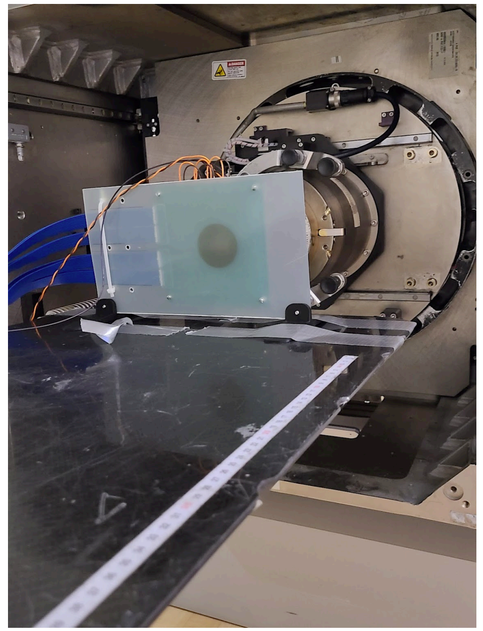
FPGA Test Firmware structure



ECON-T ASIC in test setup



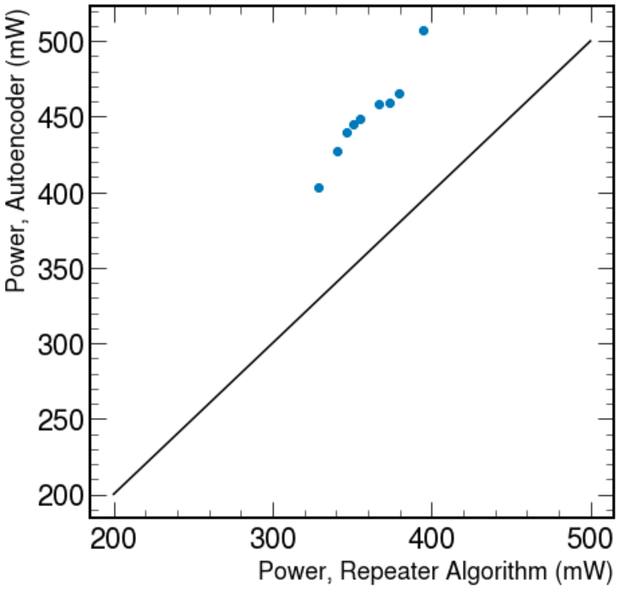
ECON-T ASIC in proton beam

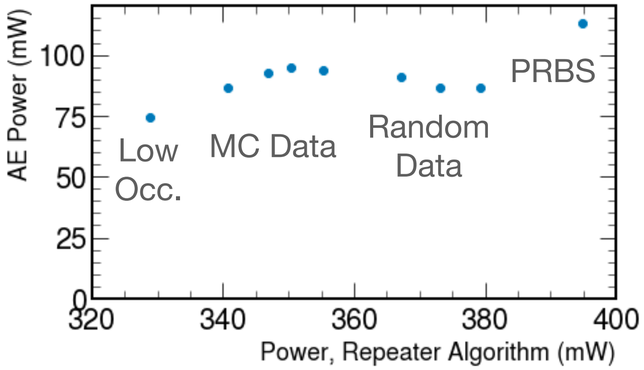


Testing Results & Summary

- Prototype testing results:
  - Functionality:
    - **Autoencoder on ECON-T performs exactly as expected**
    - Inference on chip matches predictions when same test vectors are run through software implementations (hls4ml and QKeras predictions)
  - Power:
    - Consumption varies depending on configuration
    - Autoencoder algorithm consumes ~75-100 mW
    - Total power within the 500 mW power budget of ECON-T
  - Radiation:
    - Rate of single event upset bit-flips in line with expectation
- Summary
  - Reconfigurable Autoencoder algorithm implemented in radiation tolerant way on ASIC
    - Trained with QKeras, implemented with hls4ml + Catapult HLS
  - Tests of prototype ASIC show network is performing exactly as expected

ECON-T Power consumption for Autoencoder

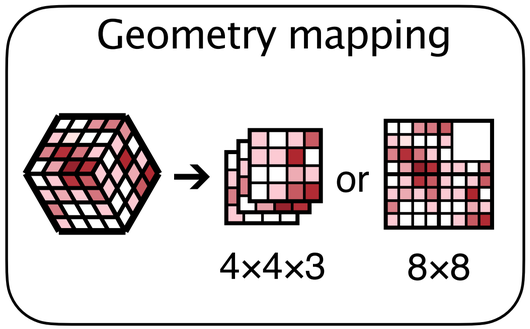


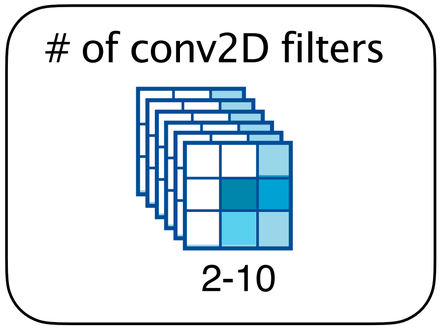


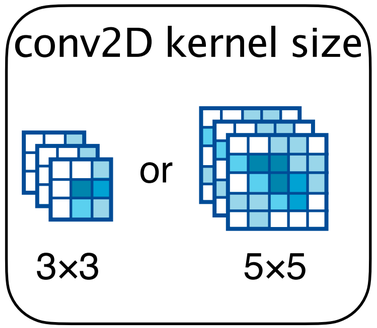
Backup

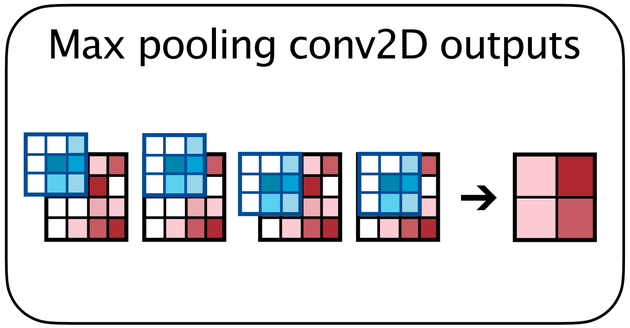
Encoder Architecture Optimization

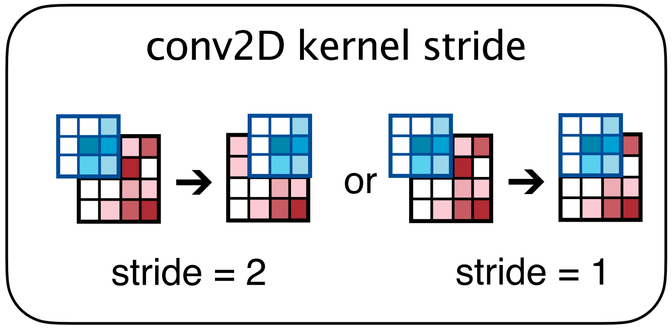
- Network architecture chosen to balance physics performance with power and area constraints of the ASIC
  - Reduce number of operations and parameters
  - Minimize Energy Mover's Distance (metric for how well we reconstruct original image) as proxy for physics performance
- Optimize architecture choices like
  - Number and types of layers
  - Size of kernels
  - Etc.







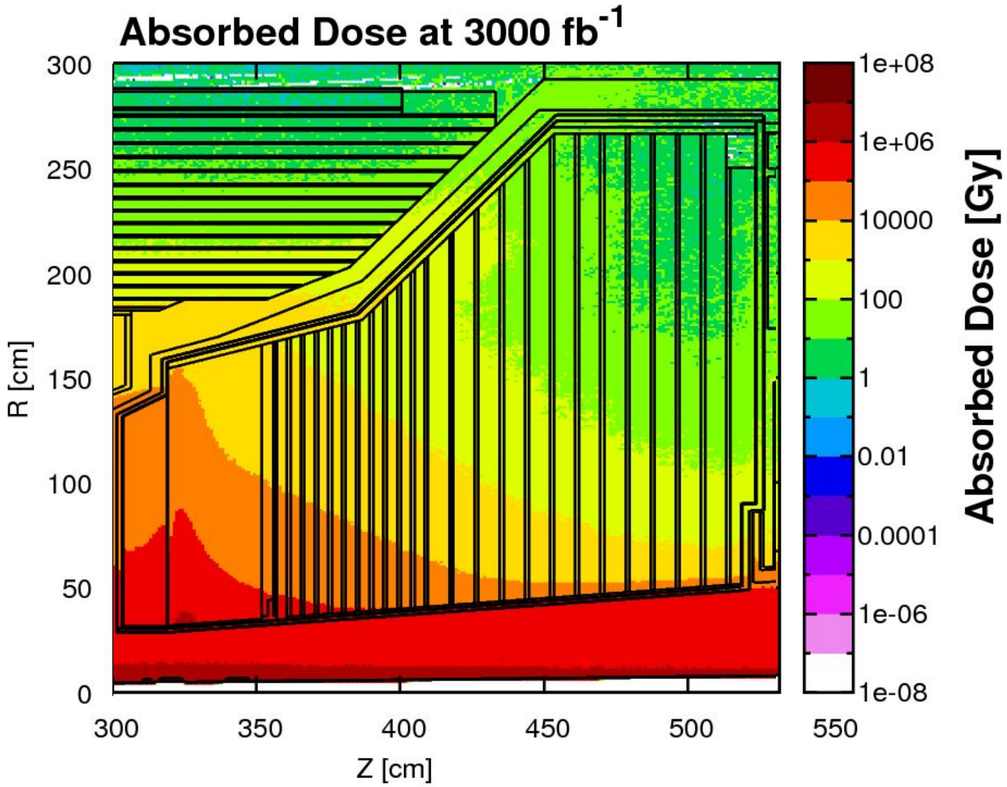


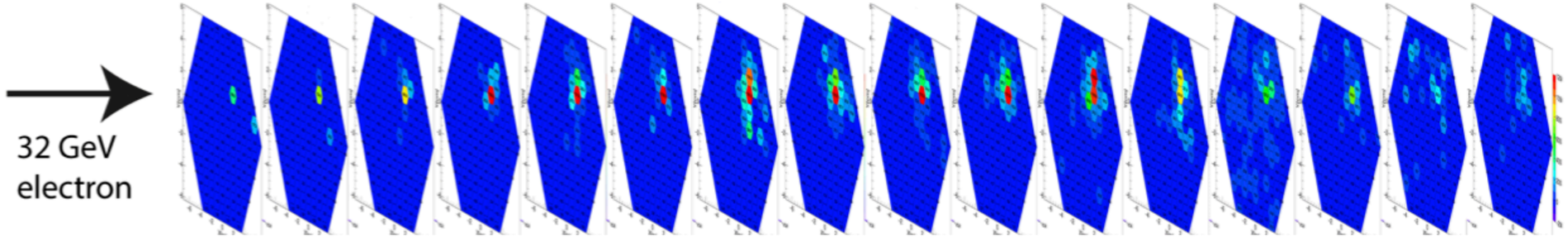


	Network Architecture					Relative Power & Area		Relative Performance	
Test feature	**Geometry**	**# filter**	**kernel**	**stride**	**pooling**	**# params**	**# operations**	**EMD Mean**	**EMD RMS**
Reference	4x4x3	8	3x3	1	none	1.00	1.00	1.00	1.00
4x4x3 -> 8x8	**8x8**	8	3x3	1	none	2.73	1.76	0.64	0.41
max pooling	8x8	8	3x3	1	**2x2**	0.71	0.97	0.59	0.33
3x3 -> 5x5 kernel	8x8	8	**5x5**	1	2x2	0.99	2.76	0.64	0.35
pooling -> stride=2	8x8	8	3x3	**2**	**none**	**0.94**	**0.59**	**0.76**	**0.46**
8 -> 10 filters	8x8	**10**	3x3	2	none	1.17	0.73	0.73	0.43
8 -> 6 filters	8x8	**6**	3x3	2	none	0.70	0.44	0.85	0.57

CMS High Granularity Calorimeter (HGCAL)

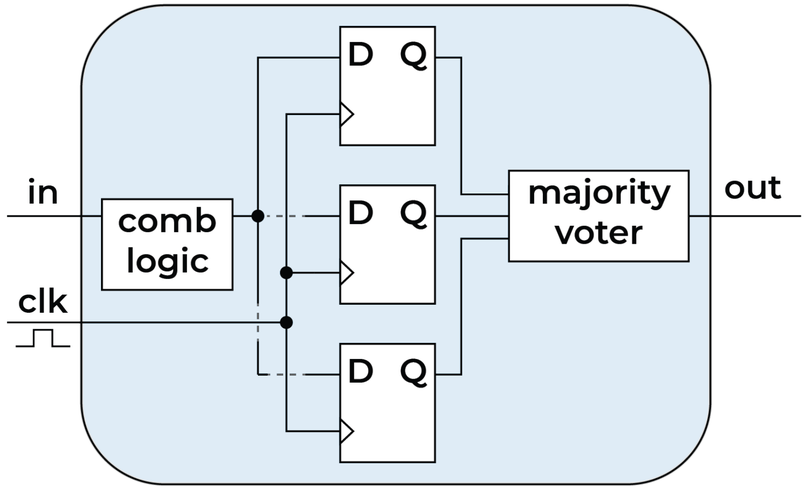
- "**Imaging calorimeter**" with ~6M readout channels.
- 50 layers of active material + absorber.
  - Front layers tiled with 300–500 8" hexagonal silicon modules.
- **HGCROC ASIC** : digitizes charge and arrival time and provides charge data for trigger path.
- **ECON ASIC** selects/compresses digital trigger data for transmission off-detector.
  - **NN Encoder to be included in ECON**.





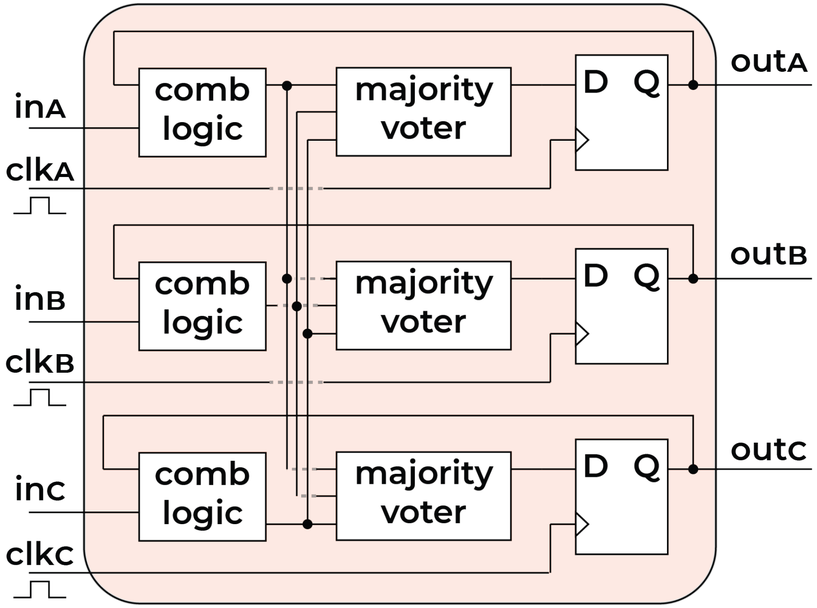
SEU Mitigation

Data path :
Encoder & Convertor



- New data every 25ns
- Triplicate registers
- No auto-correction

Configuration : I²C secondary

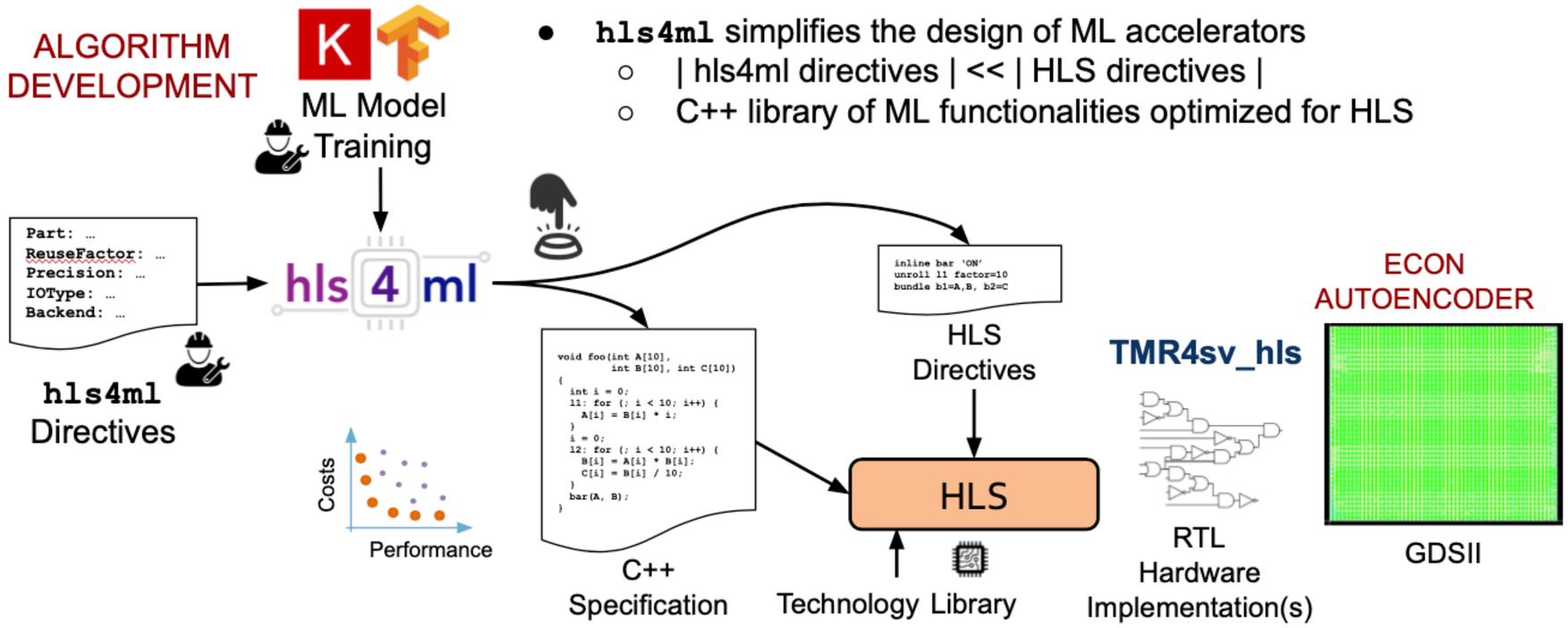


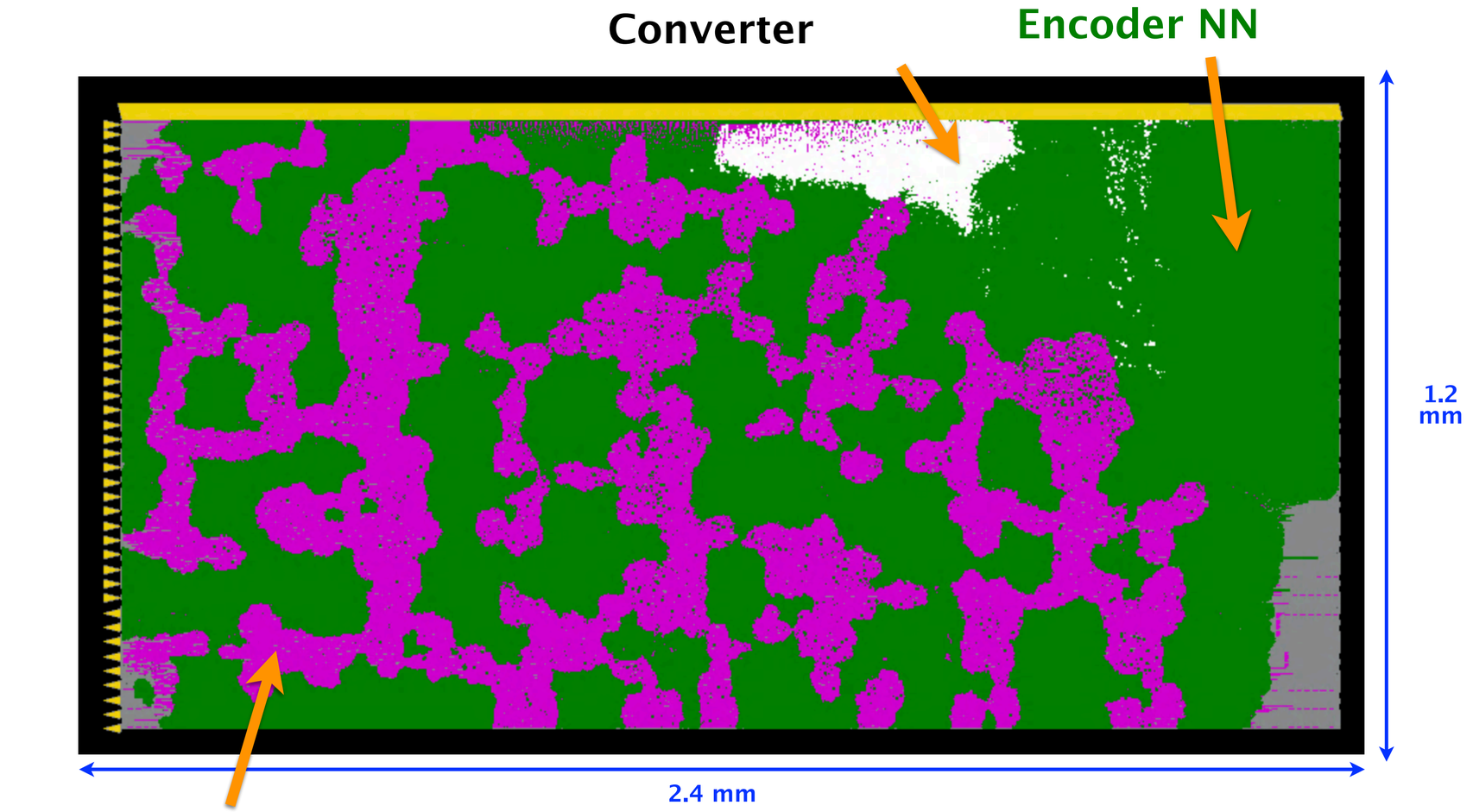
- Long term weights storage
- Triplicate registers, logic, and clocks
- Auto-correction included

Design Flow

Rapid prototyping and optimization of network achieved through

- **QKeras** : network development with **quantization-aware training** and physics simulation
- **hls4ml** : neural network description (h5 file e.g.) ➔ HLS-compliant C++ format
- **Catapult HLS** : C++ ➔ RTL
- **TMR4sv_hls** : Automated TMR for System Verilog




Converter
Encoder NN
1.2 mm
2.4 mm
Distributed i2c weights