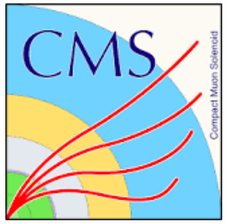
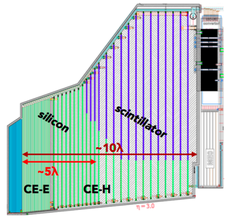
Design and Testing of a reconfigurable AI-ASIC for front-end data compression at the HL-LHC

Fermilab: F. Fahim, C. Gingu, C. Herwig, J. Hirschauer, C. Mantilla Suarez, D. Noonan, P. Rubinov, N. Tran
Baylor: J. Wilson

2022 IEEE Real Time
5th August 2022

With thanks to the CMS collaboration,
the CMS High-Granularity Calorimetry group,
hls4ml,
and the lpGBT designers (ePortRX + PLL).





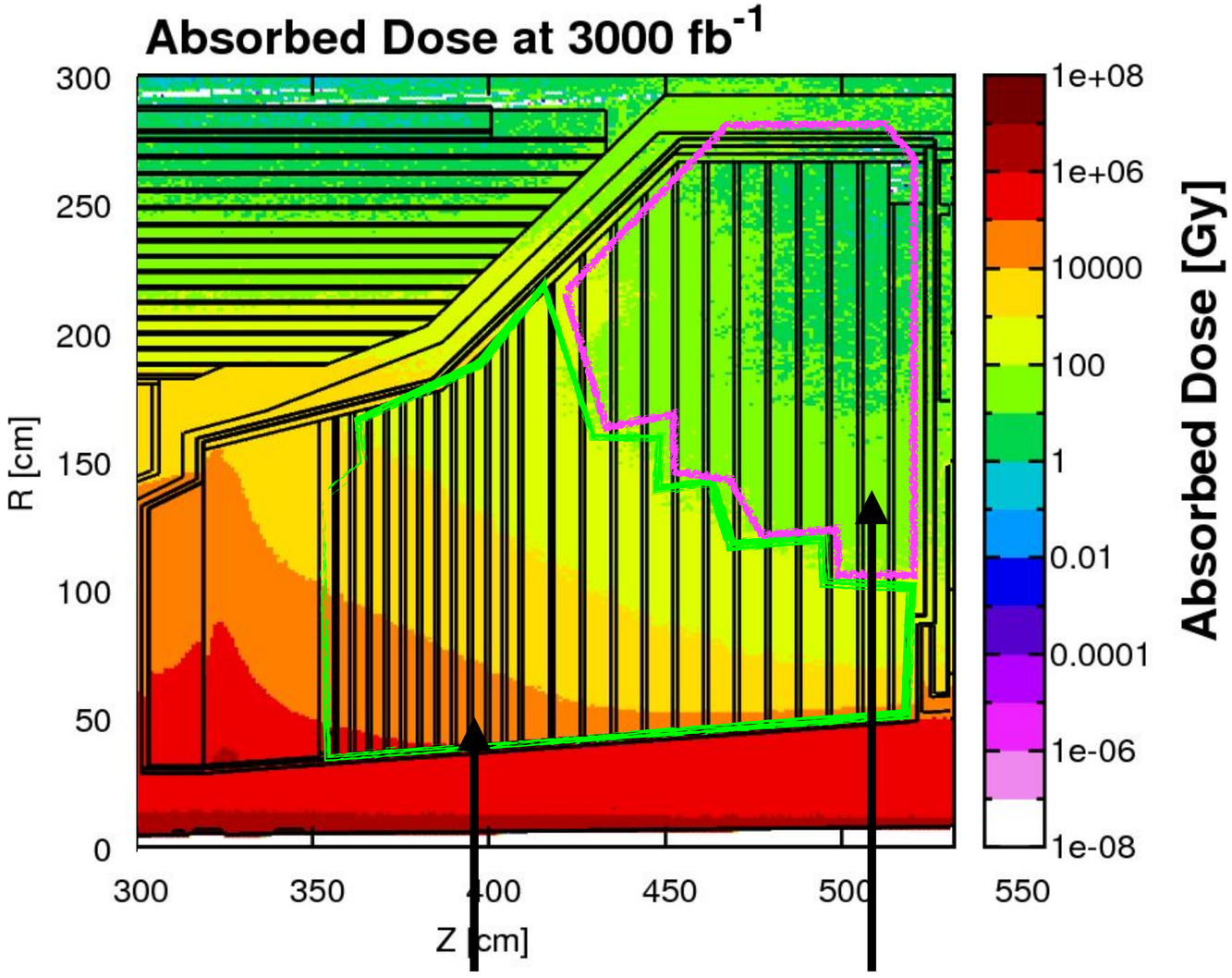
https://lpgbt-support.web.cern.ch/
https://fastmachinelearning.org/hls4ml/#
https://cms.cern/

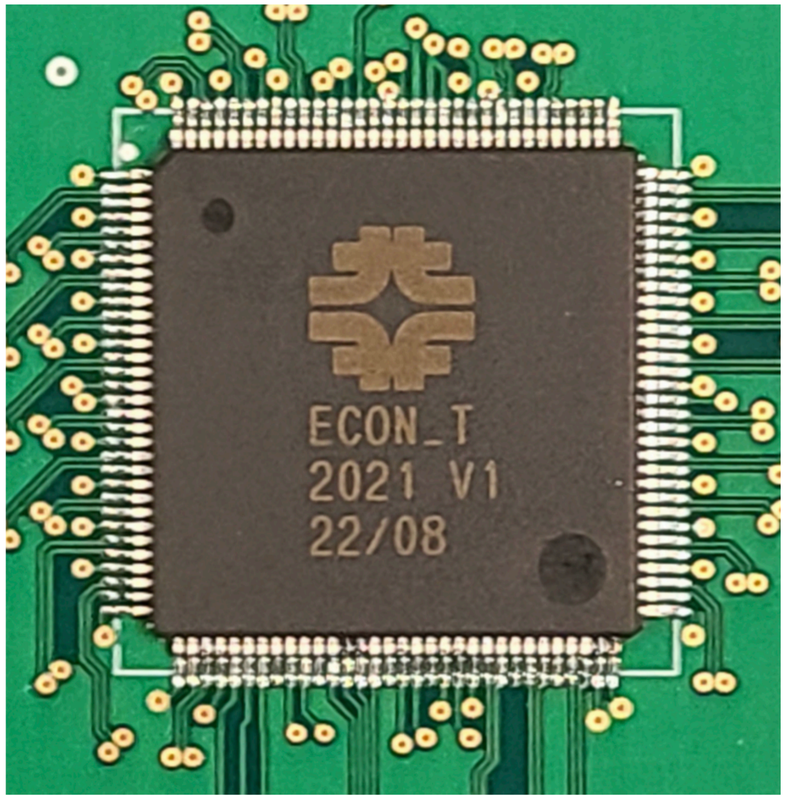
lpgbt

hls4ml

CMS High Granularity Calorimeter HGCAL

- A high granularity detector to deal with high occupancy.
- Harsh radiation environment: full volume operated at -30C.
- ~50 layers of active material (Si, scintillator) + absorber:
  - Each front layer is tiled with 300-500 8" hexagonal silicon modules.
- Spatial granularity: 6M channels in ~40 m^3.





ECON-T is an on-detector data concentrator ASIC for the trigger path.

It aggregates, selects and compresses charge data @ 40 MHz.

It runs an encoder neural network as one of the data compression algorithms.

The HGCAL trigger data challenge

Raw-data

5 Pb/s, 6M channels

HGCROCv3:

Sends sum of 4 (9) channels (7-bit floating point format) @ 1.28 Gpbs/s.

300 Tb/s, 1M channels

ECON:

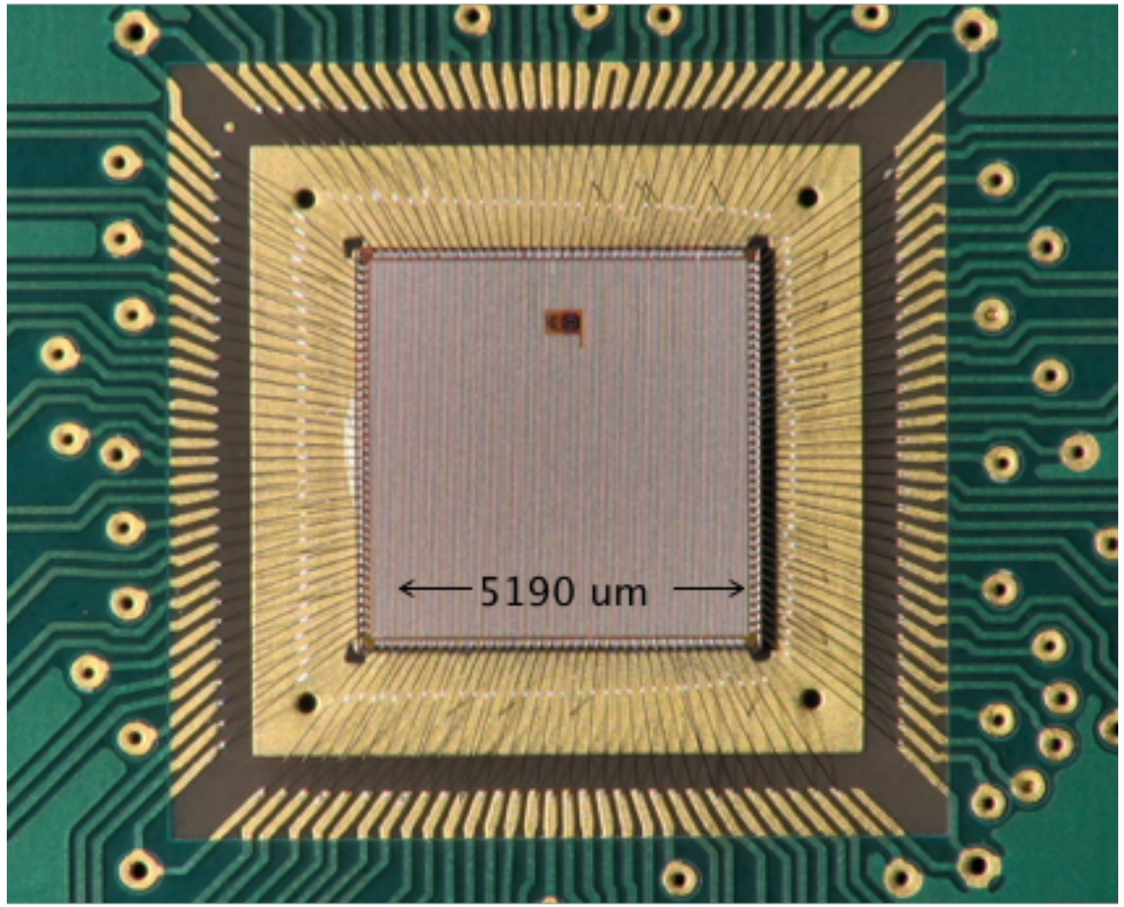
Traditional algorithm selects trigger charge data.

40 Tb/s, 1M channels

~2 x 1.28 Gbps links per module

~ 9k 10.24 Gbps links in total

The ECON ASIC overview



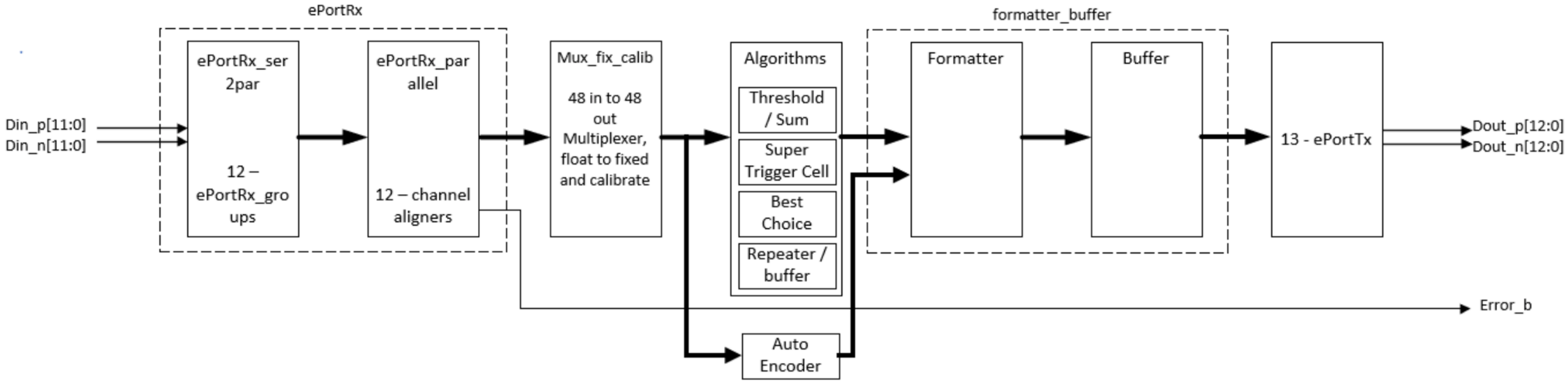
- Latency: 400 ns = 16 clock cycles
  - Encoder NN: 50 ns
- Radiation tolerance: 200 Mrad, 1×10^{16} Neq/cm^2
  - Using 65nm CMOS with standard cells characterized for radiation performance.
- Low power: ≤5 mW/channel
- 1.28 Gbps links: 12 inputs and 13 outputs (most of the modules use only 2 outputs)
- Packaging: 128-pin Low Profile Quad Flat Pack
  - 200 ASICs have been packaged from 300 produced parts in P1.

ASIC design (including AI on chip)

ASIC blocks

12 input receivers

13 output transmitters



Word Aligner

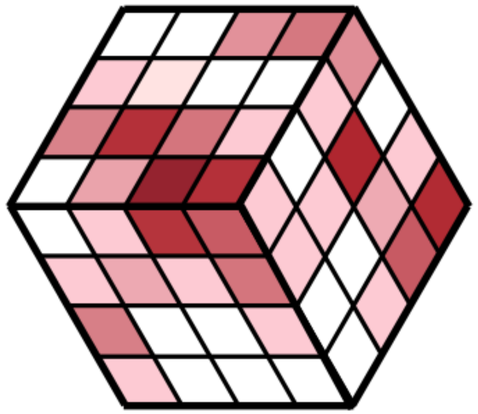
Multiplexer and calibration

Algorithms

Formatter and Buffer

The data compression algorithms

Starting from 48 Trigger Cells (TC)



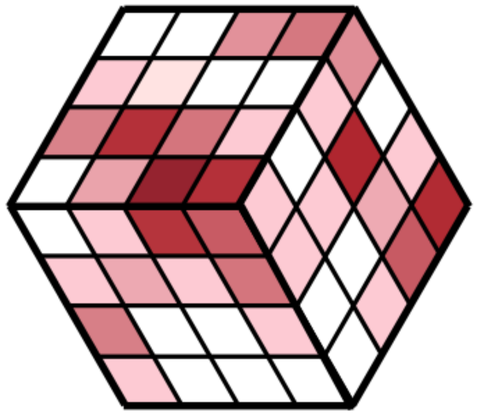
Variable latency	Fixed latency		
Threshold	Best Choice	Super Trigger Cell	Encoder Neural Network
Chooses TC above programmable threshold	Sorts TC by charge Q, sends N with largest Q	Groups TC and forms larger super TCs	Encodes TC with fully reconfigurable weights

Default algorithm

AI on chip

Encoder in ASIC

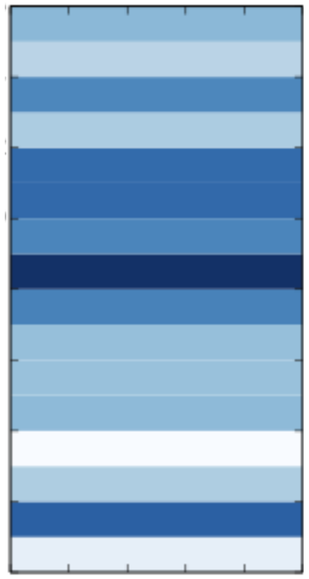
Original input



48x7bit input
336 bits

Encoder on-detector ASIC

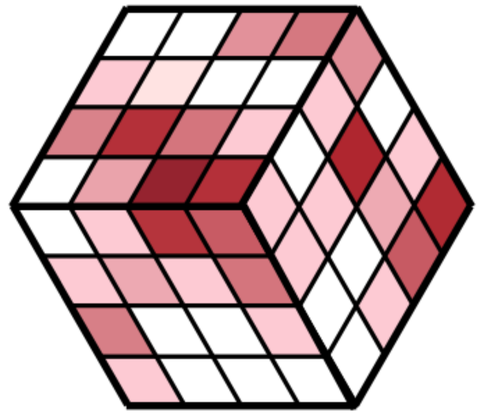
Compressed representation



Transmit 16 x 3bit outputs*
48 bits

Decoder off-detector FPGA

Reconstructed output



Decoded 48-pixel image

*for low occupancy zones

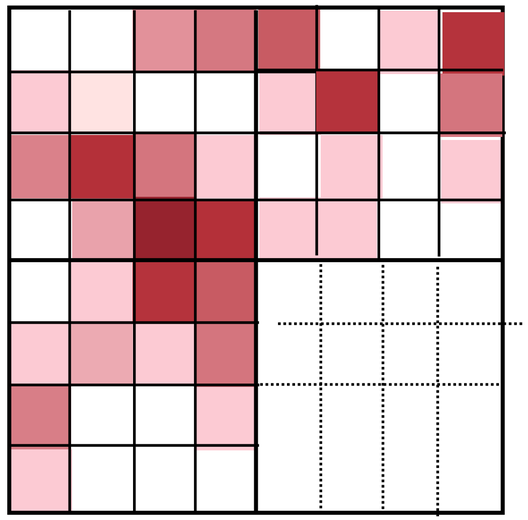
Encoder NN architecture

re-map the geometry

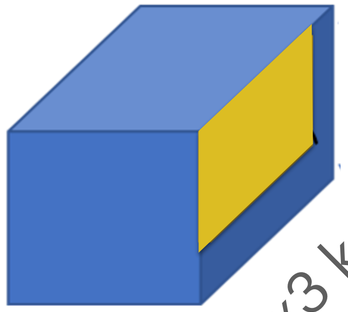
Convolutional layer

Dense layers

Outputs



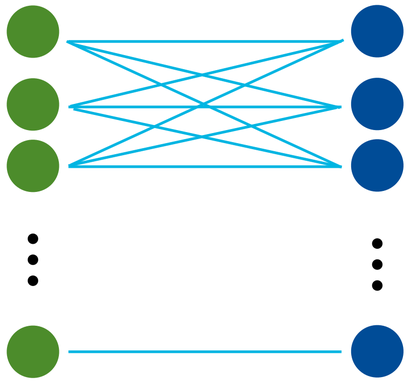
8 x 8



8 filters

3x3 kernel

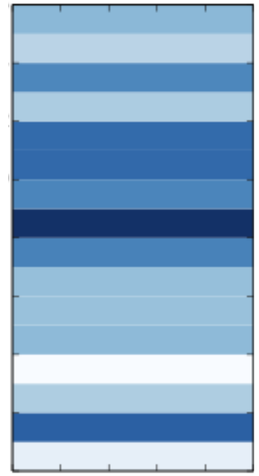
ReLU



Flatten
128 features

Dense
16 outputs

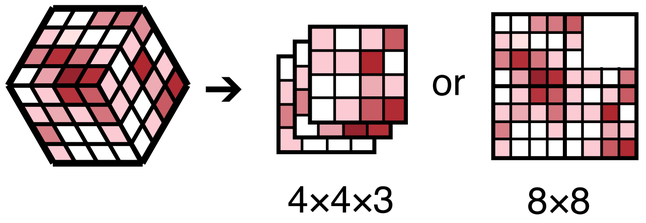
ReLU



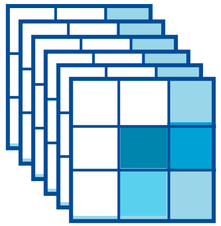
16 outputs
x
3 (up to 9)
bits per
output

Architecture optimization

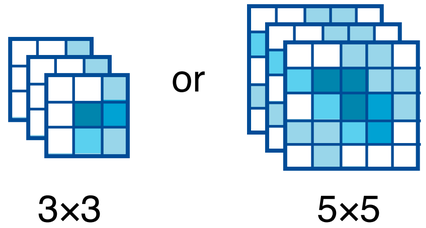
Geometry mapping



of conv2D filters



conv2D kernel size

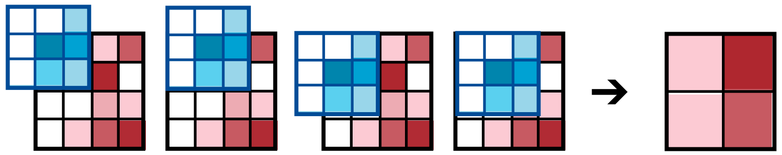


Quantization parameters

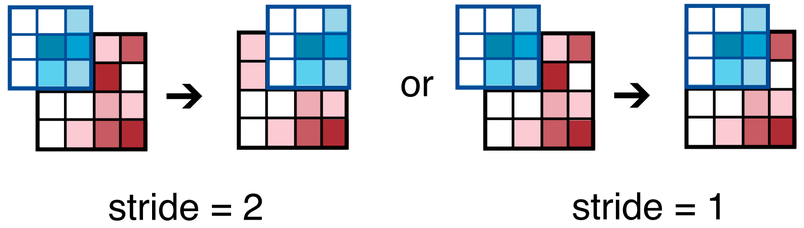
Precision of weights

Number and precision of outputs

Max pooling conv2D outputs

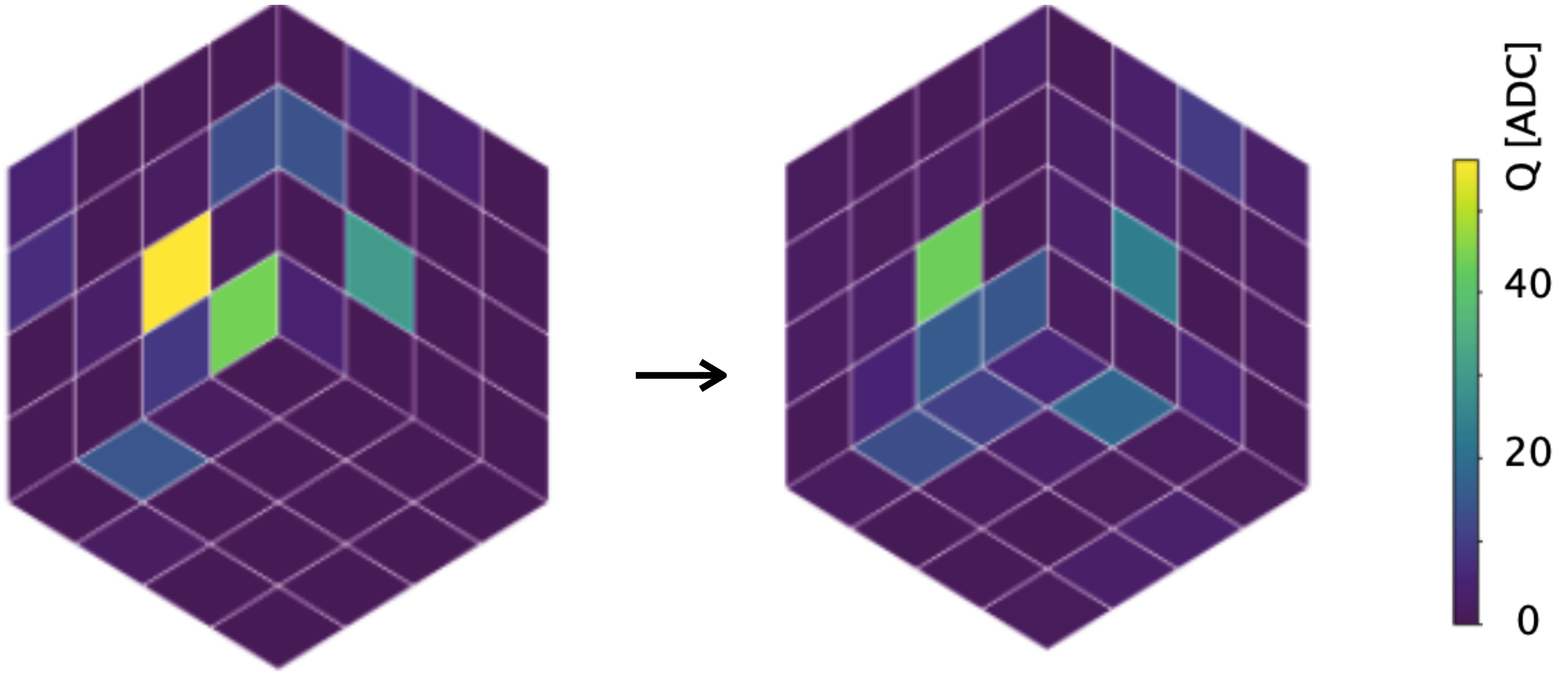


conv2D kernel stride



How to optimize the architecture rapidly?

Use energy mover's distance as metric for physics performance:
the "work" required to rearrange one radiation pattern into another.



Input image

Decoded image

First associated with "optimal transport" problem

arXiv:1902.02346
Komiske, Metodiev, Thaler

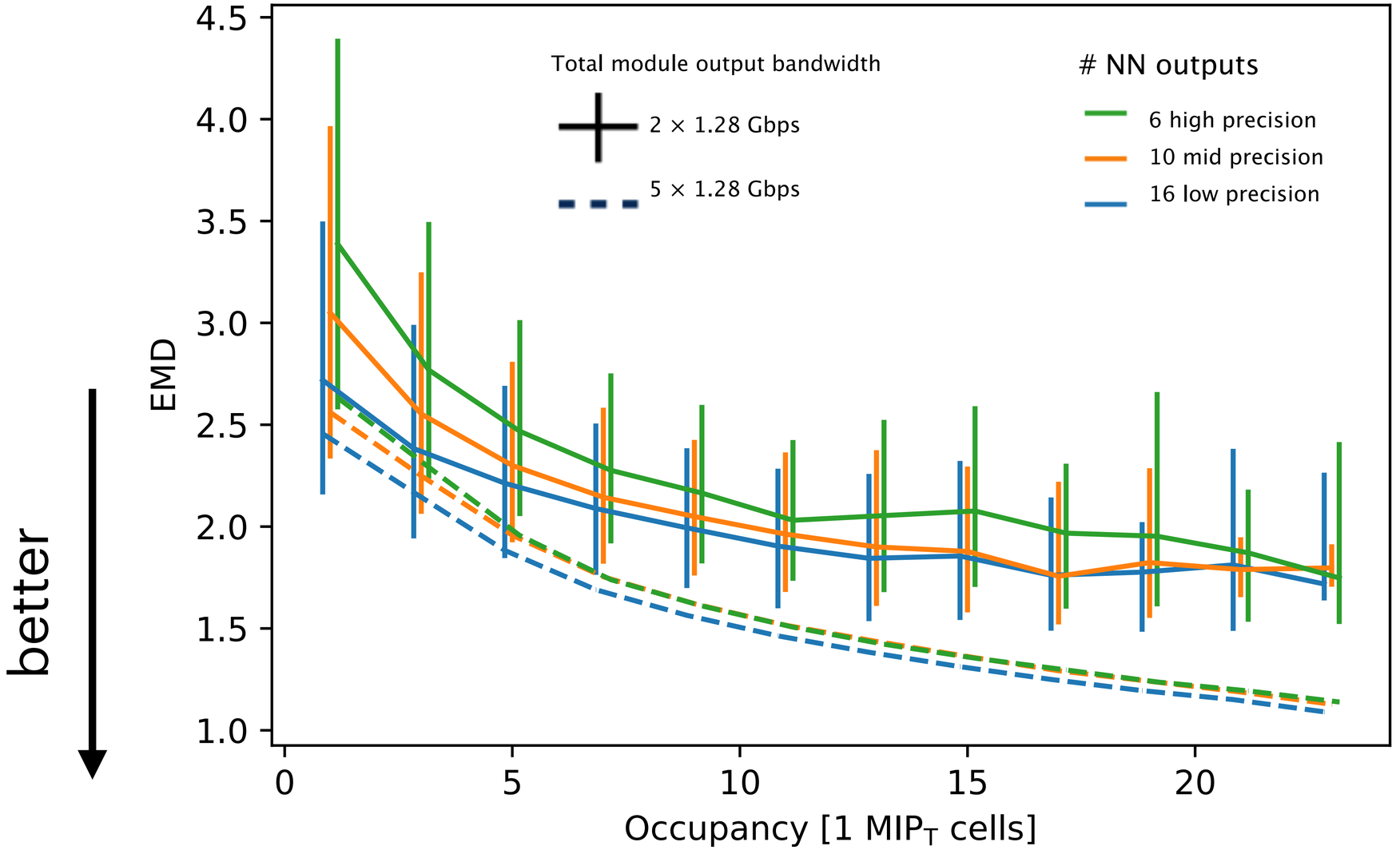
Step	Type	Run Time	Iterations	Size
Model generation	D	1s	50–100	1.1k lines of C++
C Simulation	V	1s		
HLS	D	30 min	3–100	40k lines of verilog
RTL simulation	V	1 min		
Logic synthesis	D	6 hrs		750k gates
Gate-level sim	V	30 min		
Place and route	D	50 hrs	6	780k gates
Post-layout sim	V	1 hrs		
Post-layout parasitic sim	V	2 hrs		
SEE simulation	V	4 hrs		
Layout	D	20 min	1	7.6M transistors
LVS and DRC	V	1 hr		

Increasing time and complexity

First steps of ASIC design repeated several times during optimization

Optimizing for:	Metric
Physics performance	EMD
Area / Power consumption	Number of registers and operations

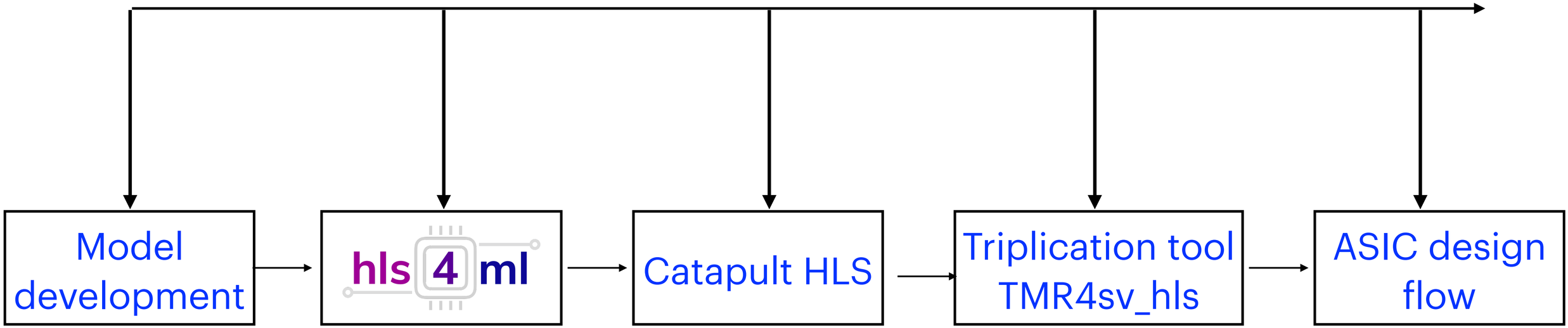
One example of what we learned during the optimization:



more low-precision outputs is better than few high-precision outputs

Physics-driven hardware co-design

Independent verification of Encoder NN



QKeras/TF
Training based on
LHC simulation

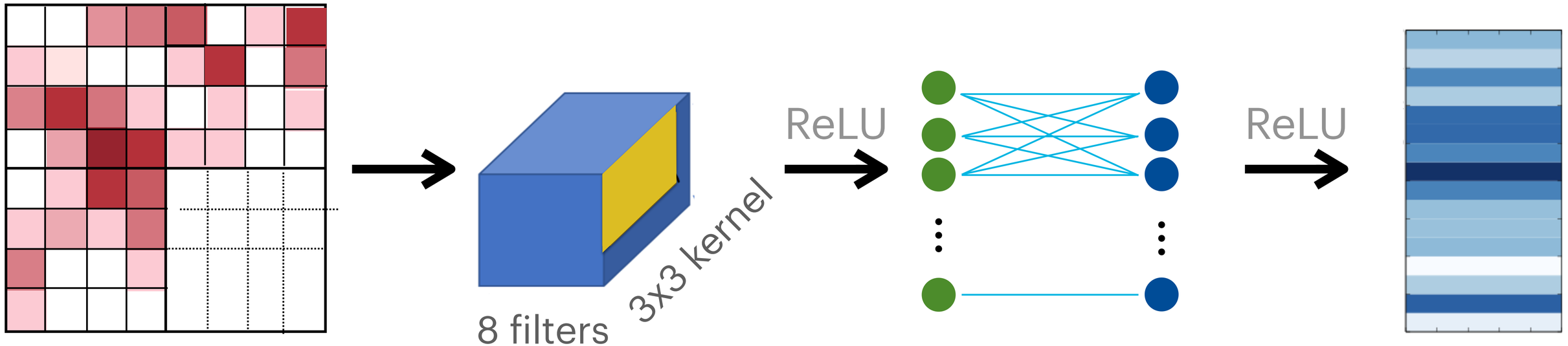
Optimized Encoder NN architecture

re-map the geometry

Convolutional layer

Dense layers

Outputs



8 x 8

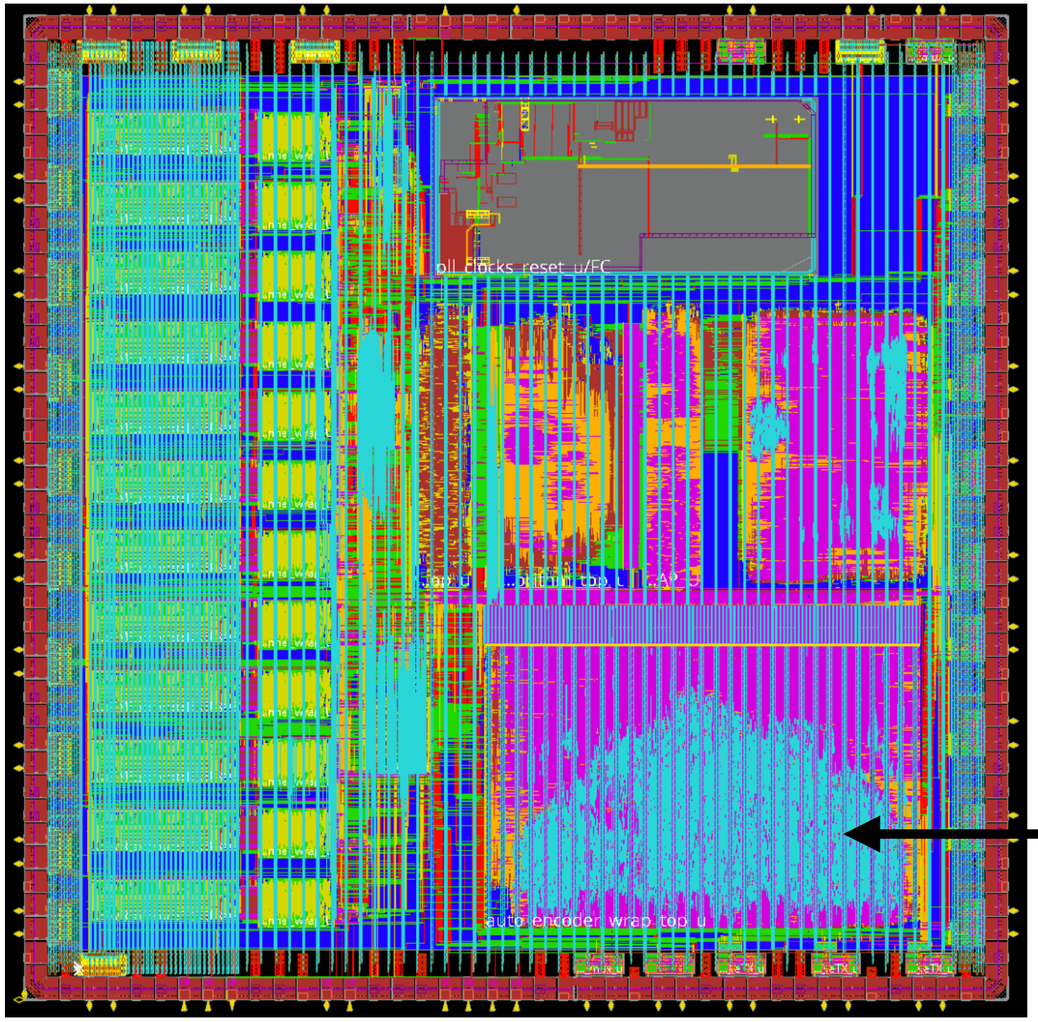
8 filters

Flatten
128 features

Dense
16 outputs

16 outputs
x
3 (up to 9)
bits per
output

ECON ASIC place and route

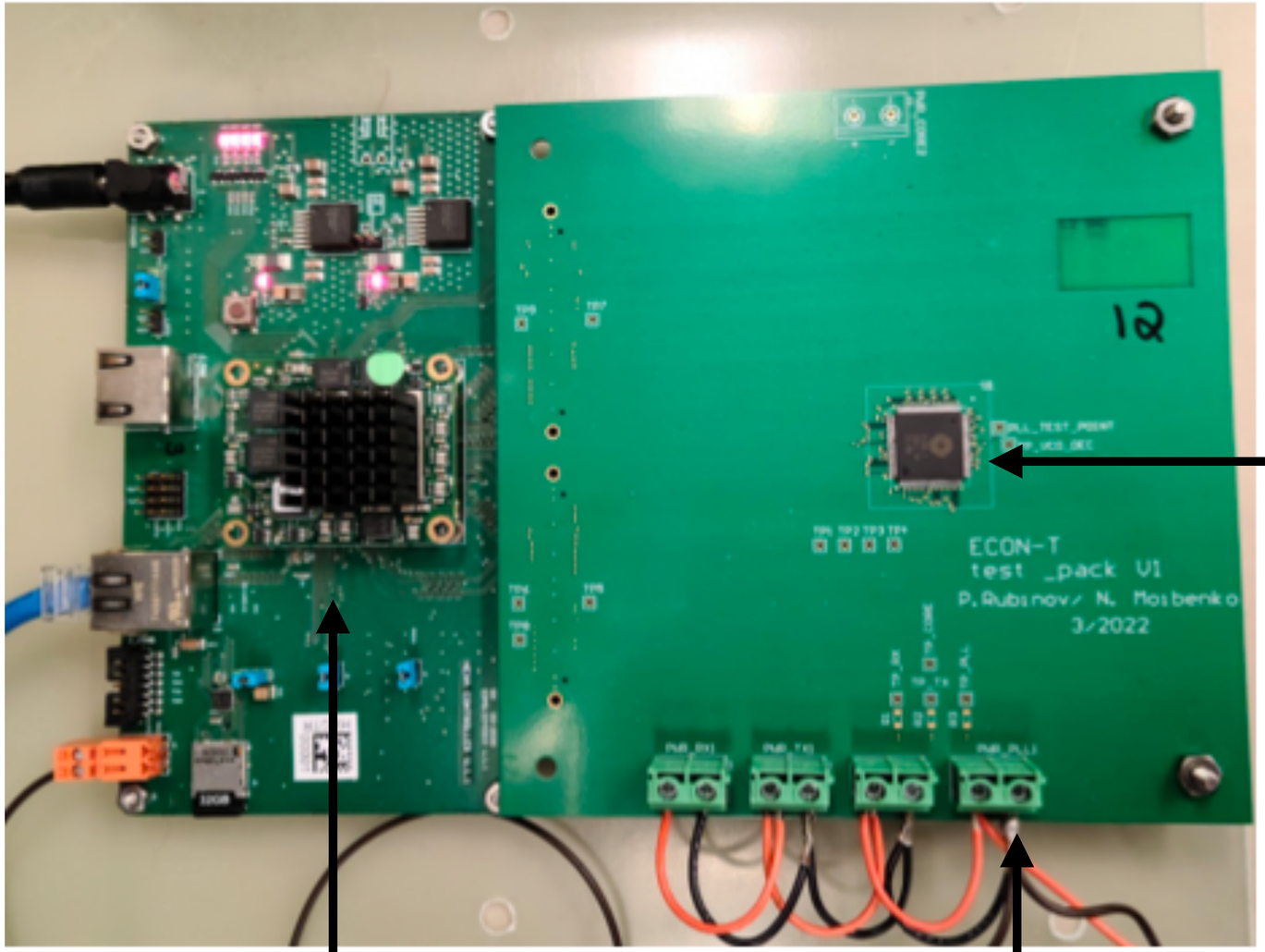


Metric	Simulation	Target
Power	48 mW	<100 mW
Energy / inference	1.2 nJ	N/A
Area	2.88 mm²	<4 mm²
Gates	780k	N/A
Latency	50 ns	<100 ns

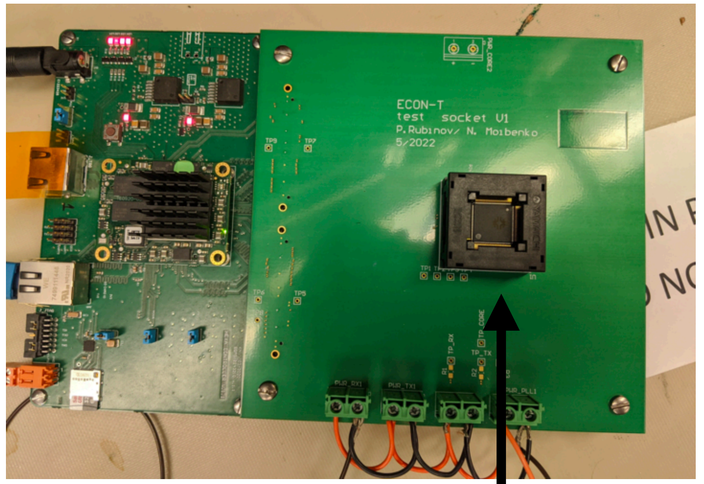
Encoder NN block (distributed i2c)

ASIC Testing

Testing setup



ASIC powered @ 1.2 V

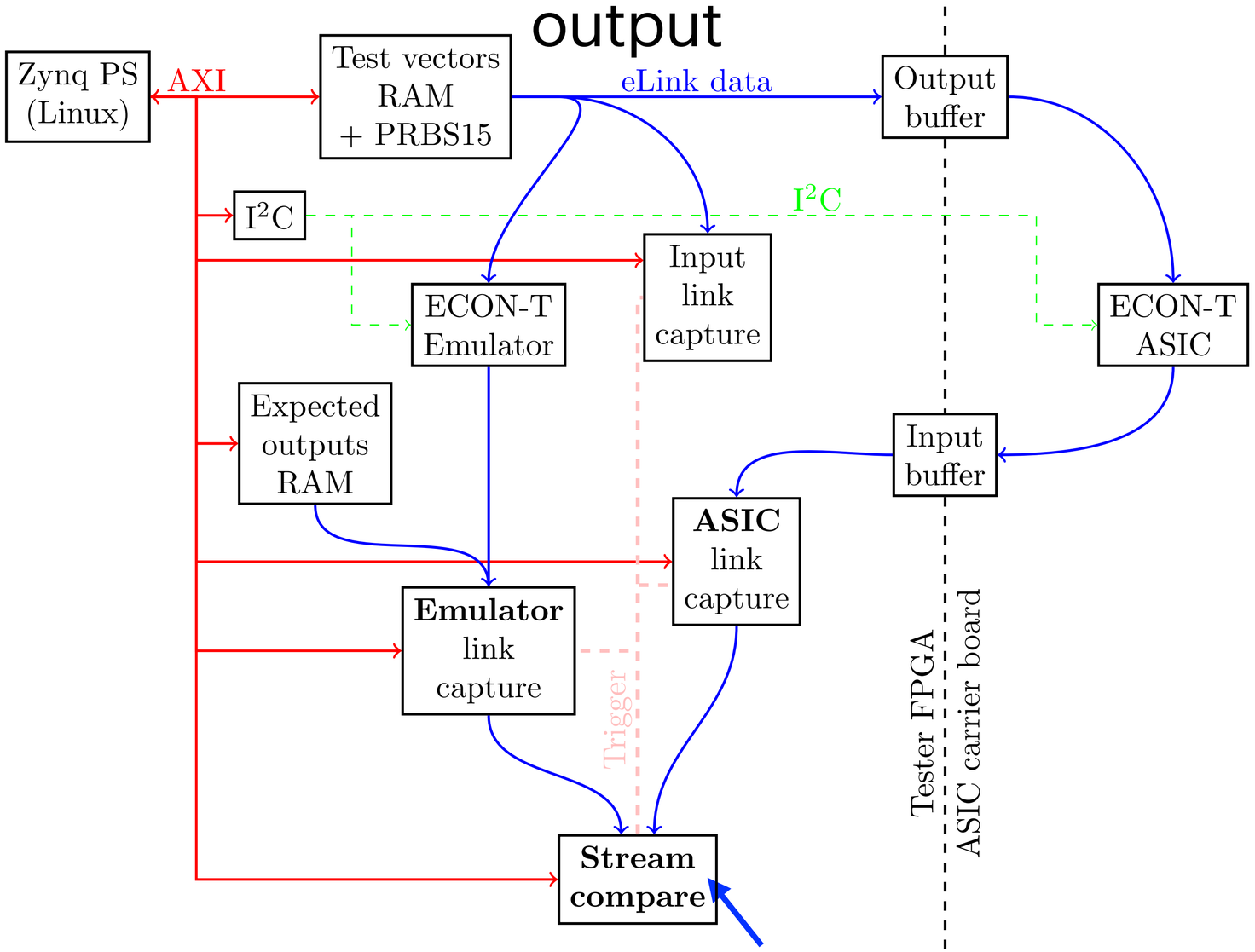


Socket testing

FPGA

Individual power domains

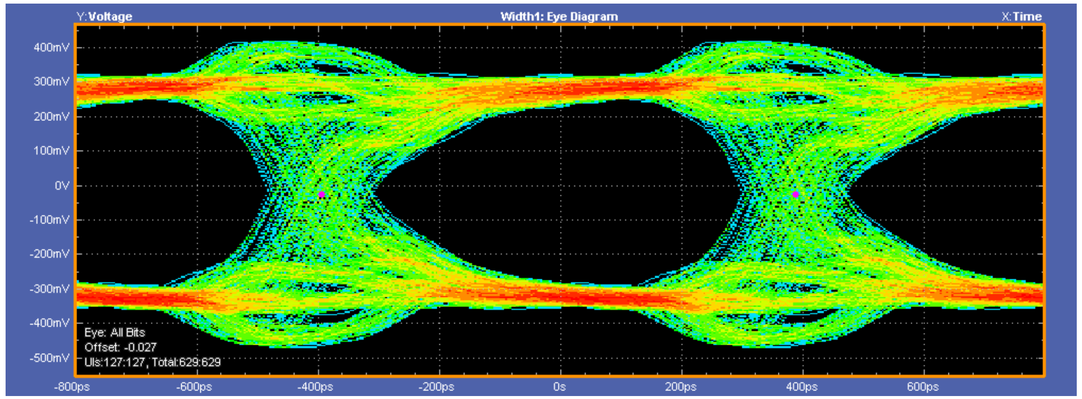
FPGA provides Fast Command (FC) clock, i²c, input and simulated output



Simulation/Emulation is key: live comparison of output data stream, captures data mis-matches

Functionality has been fully verified

- Tested under different configurations (number of eTx, algorithms, ...) and input test vectors (random data, LHC simulation, ...).
- 1.28 Gbps outputs agree perfectly with simulation/emulation in test bench.
- Power-up-state-machine, PLL, eTx, Formatter, buffer: everything works.
- Total power consumption with Encoder NN below 450mW (cf. 500 mW target).



1.28 Gpbs eTx eye diagram

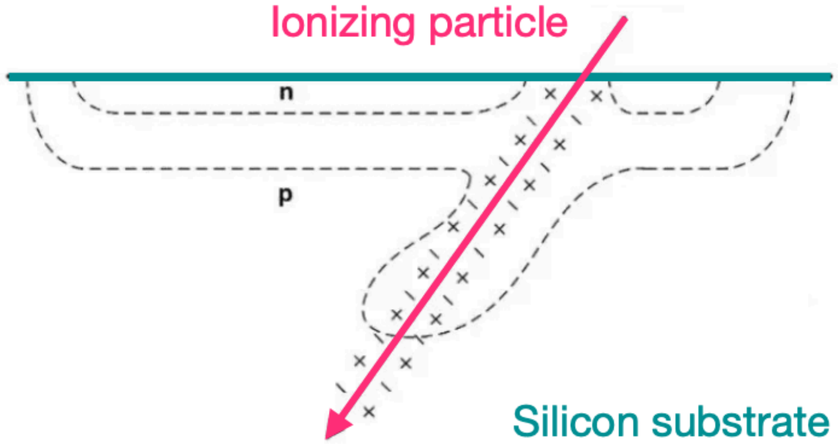
Radiation Tolerance Testing

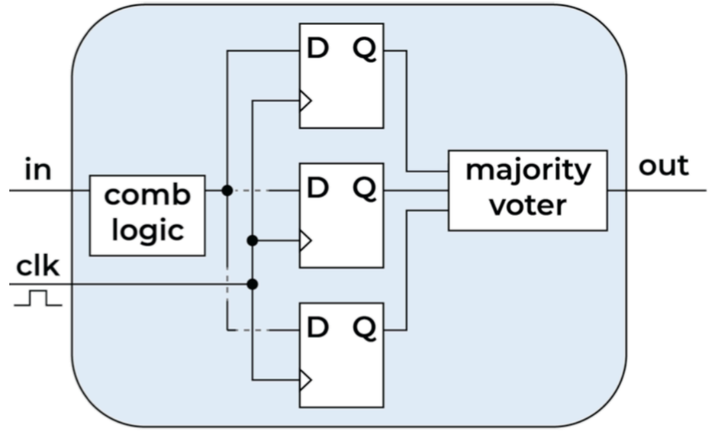
Tests for Single Event Effects (SEE) and Total Ionizing Dose (TID)

Reminder of Single Event Effects (SEE)

*Credit to Elena Vernazza

- A single particle can induce localized and non-cumulative radiation effect: bit flips, clock/ logic transients or permanent damage.
- To protect against bit flips and transients: use triplication (TMR) and hamming based error correction codes (ECC).



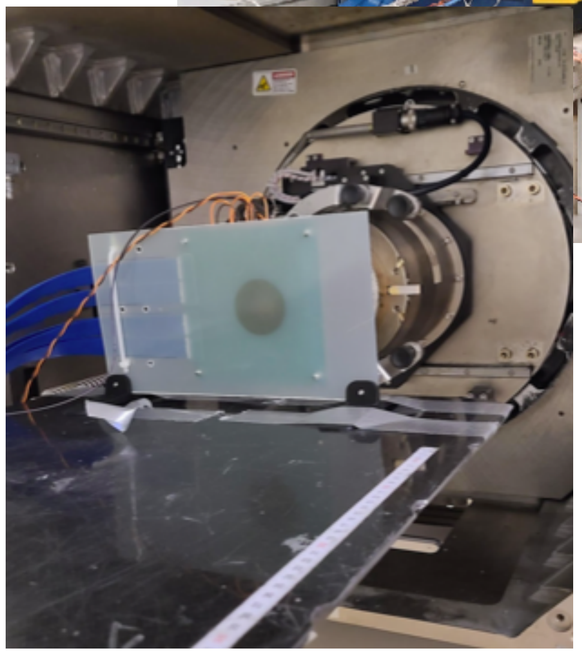


ECON-T P1 SEE protection

	Total	TMR	AutoCorrect	ECC
Data		On flip flops	No	No
i2c registers	675 bytes	On flip flops	Yes	Yes
Encoder i2c weights and biases	1608 bytes	TMR on flip flops, logic and clock	Yes	No

SEE Radiation testing

- ECON-T P1 ASIC tested at two different irradiation facilities:
  - At FNAL with 400 MeV proton beam. Flux for hadrons w. E > 20 MeV: 2E+15 cm^{-2}/s
  - At medical facility with 217 MeV proton beam. Flux for hadrons: 5E+9 cm^{-2}/s
  - Flux HL-LHC: 3E+6 cm^{-2}/s
- Validate overall chip performance (input alignment, PLL, serializer, data logic) by monitoring 1.28GPbs outputs.
- Also, check i2c registers to confirm stability.



Radiation testing results

Facility	Fluence (p/cm²)	Preliminary Observations in i2c (without extracting bit cross section)
ITA	9.6E+12	Bit flips on Encoder RW registers: 4 Increase in ECC error counters but no bit flips: 10
Medical facility	5.4E+12	Bit flips on Encoder RW registers: 10

HL-LHC Fluence: 1E+14

- Overall excellent performance (including Encoder).
- Low / acceptable cross section for bit errors on not-fully triplicated data path.
- Low cross section for serializer errors (serializer design is already improved in v2).

Conclusions

- ECON-T ASIC design is complete and includes an Encoder neural network for on-detector data compression.
  - Low power, low latency, radiation tolerant and fully re-configurable weights.
- Preliminary ECON-T-P1 testing successful:
  - Overall chip and Encoder NN perform very well: algorithm and other block functionality has been fully verified.
    - Minor RTL bugs will be fixed in production.
  - Larger scale testing of 300 parts in progress.