Smart pixel sensors
Towards on-sensor filtering of pixel clusters with deep learning

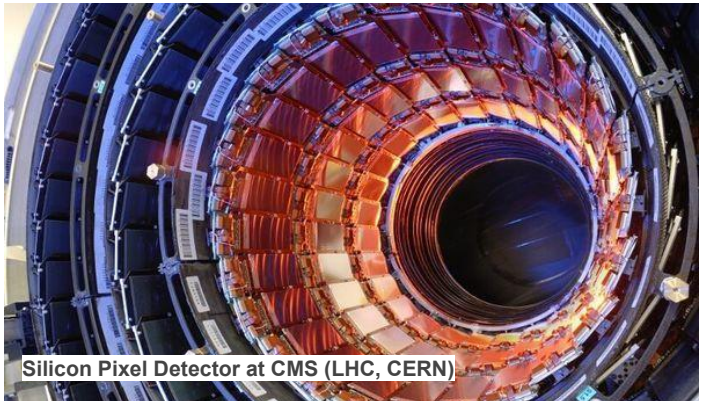
Jieun Yoo, Jennet Dickinson, Morris Swartz, Giuseppe Di Guglielmo, Alice Bean, Douglas Berry, Manuel Blanco Valentin, Karri DiPetrillo, Farah Fahim, Lindsey Gray, James Hirschauer, Shruti R. Kulkarni, Ron Lipton, Petar Maksimovic, Corrinne Mills, Mark S. Neubauer, Benjamin Parpillon, Gauri Pradhan, Chinar Syal, Nhan Tran, Dahai Wen, Aaron Young

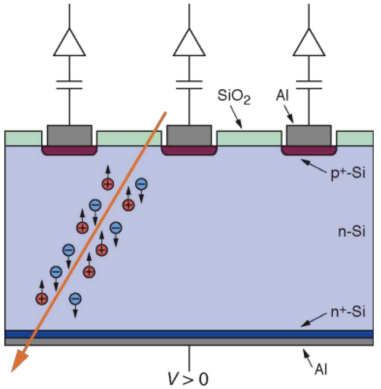
University of Illinois Chicago, Fermi National Accelerator Laboratory, Johns Hopkins University, Northwestern University, University of Kansas, The University of Chicago, Oak Ridge National Laboratory, University of Illinois Urbana-Champaign

FERMILAB-SLIDES-23-350-ETD

Silicon pixel detectors

- Experiments at colliders typically have a silicon pixel detector at the center
  - Concentric rings tiled with sensors
- Silicon sensors are depleted of charge carriers by high voltage
- When a charged particle from a collision passes through, it creates e/h pairs
- Charge is read out and transferred off-detector
  - Charge cluster information is used for physics analysis offline

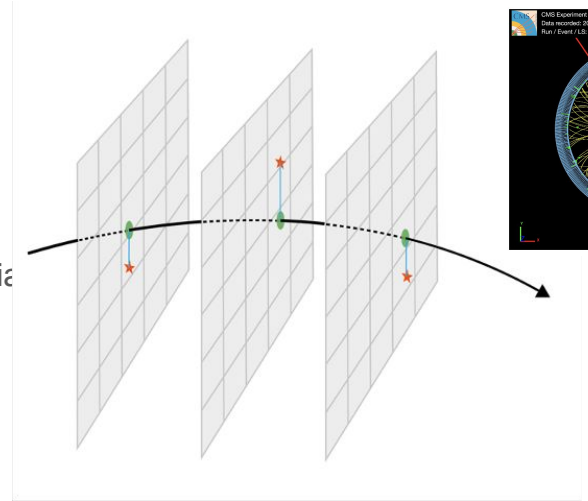

Silicon Pixel Detector at CMS (LHC, CERN)

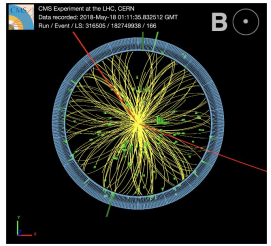


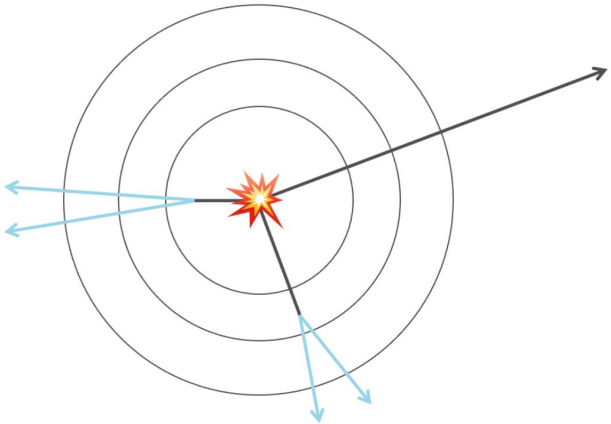
https://cms.cern/detector

Particle tracks and vertices

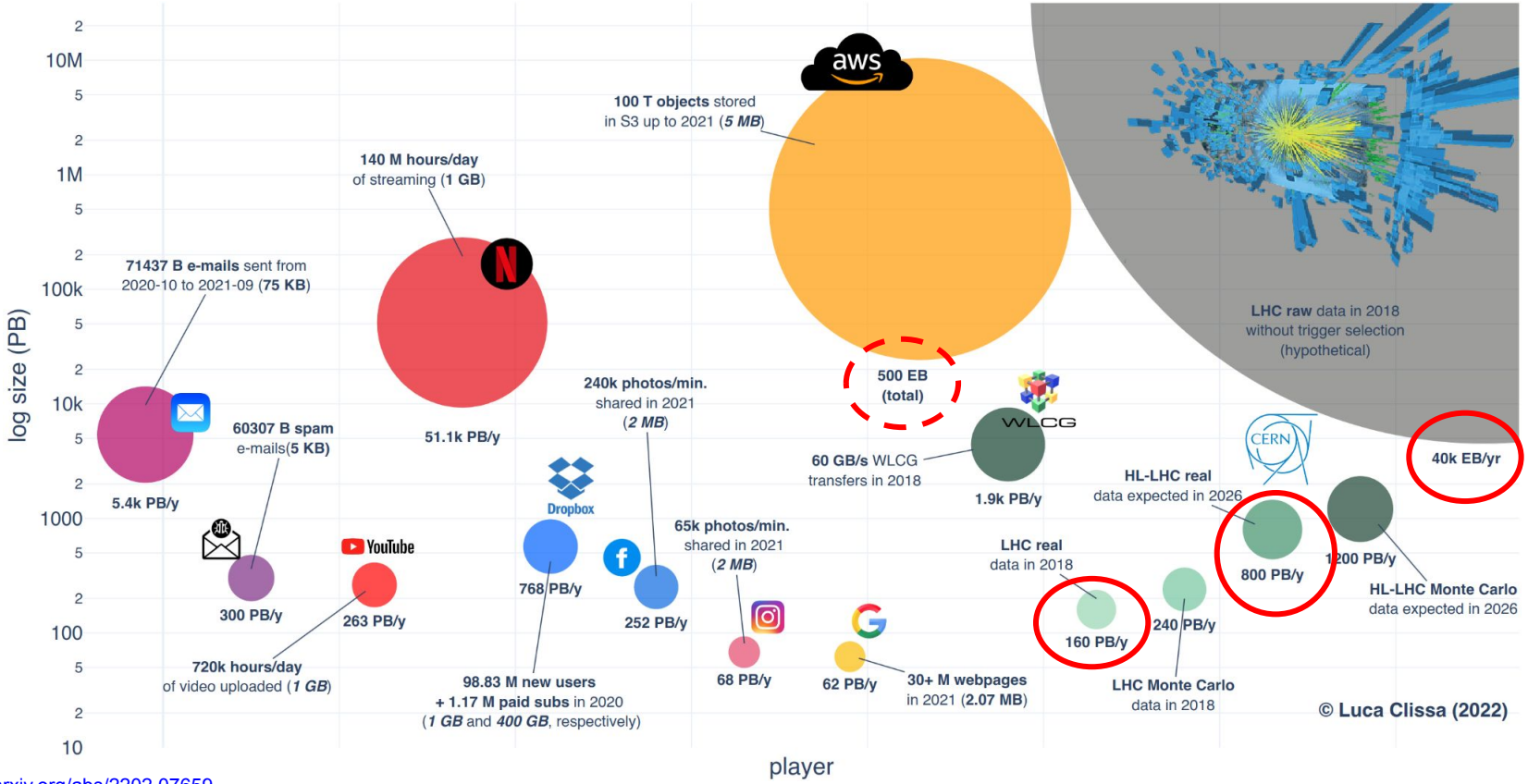
- Connecting the dots between charge collected in different pixel layers creates a particle track
  - Detector should be low-mass so interactions in inactive materia doesn't disrupt this trajectory
- Solenoid magnet immerses the pixel detector in a magnetic-field, causing tracks to curve
  - Very curved → low transverse momentum (low-p_T)
  - Almost straight → high transverse momentum (high-p_T)
- Reconstructing vertices is critical
  - Secondary vertices help identify particles: long, short, medium lifetime?







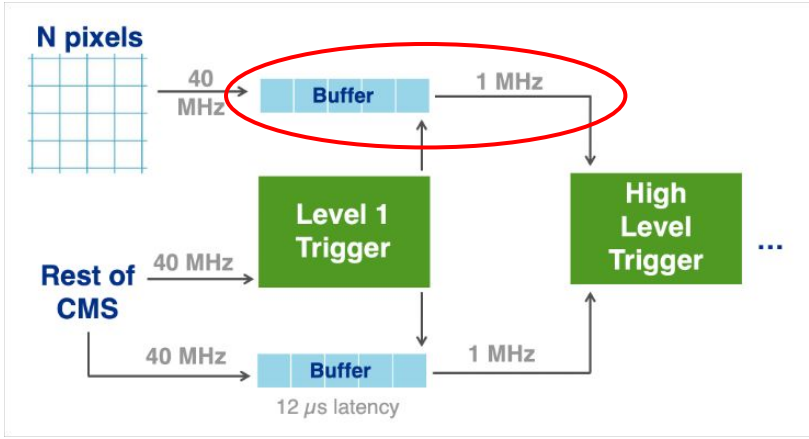
Survey of Big Data sizes in 2021

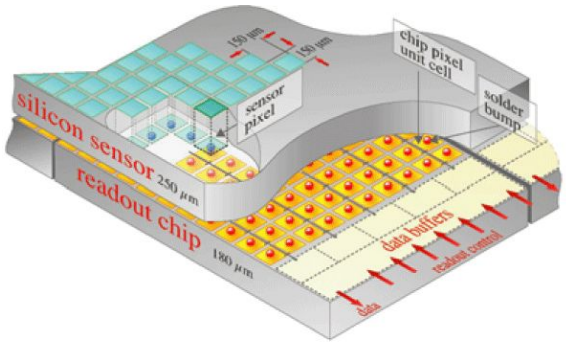


https://arxiv.org/abs/2202.07659

Designing hardware for the LHC is challenging

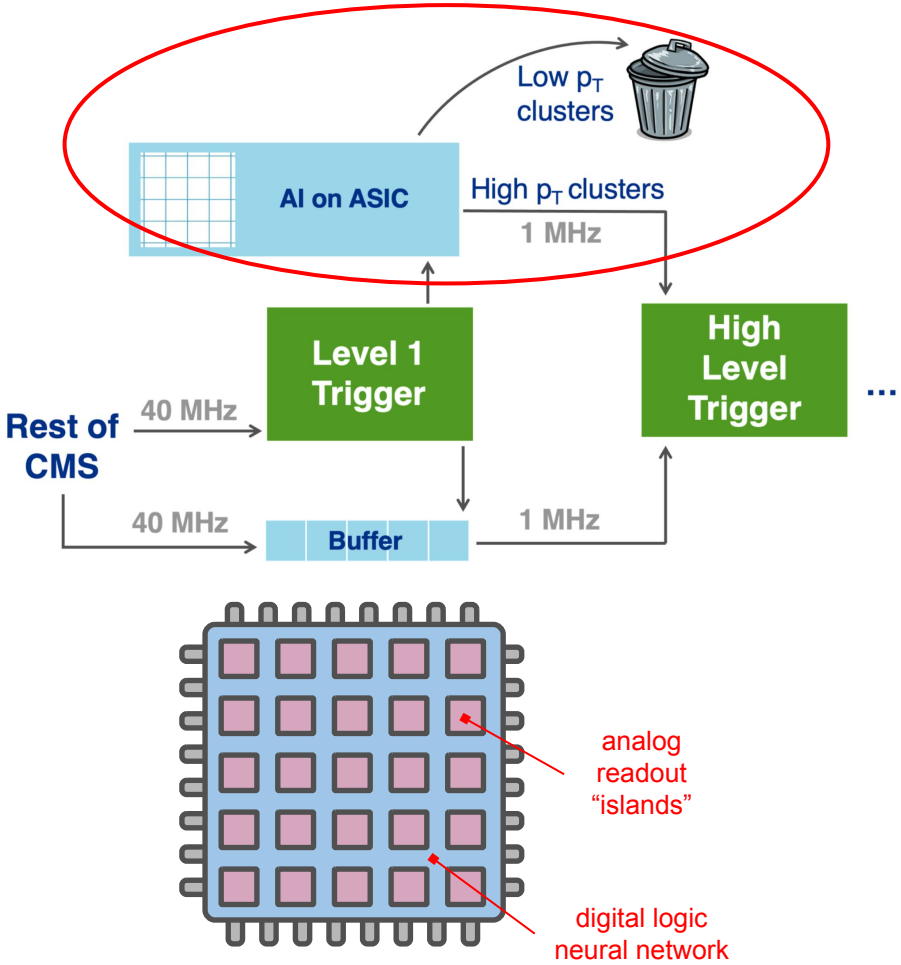
- LHC/CMS produces a lot of data
  - New data every 25 ns (p-p collision)
  - Physicists have to throw most of it away
    - Physically and financially challenging
    - Risk to throw away significant information
- Detector is continuously being sprayed with particles
  - Need radiation tolerant on-detector electronics
- High voltage and low temperature requirements
  - Up to -800 V, -35 C



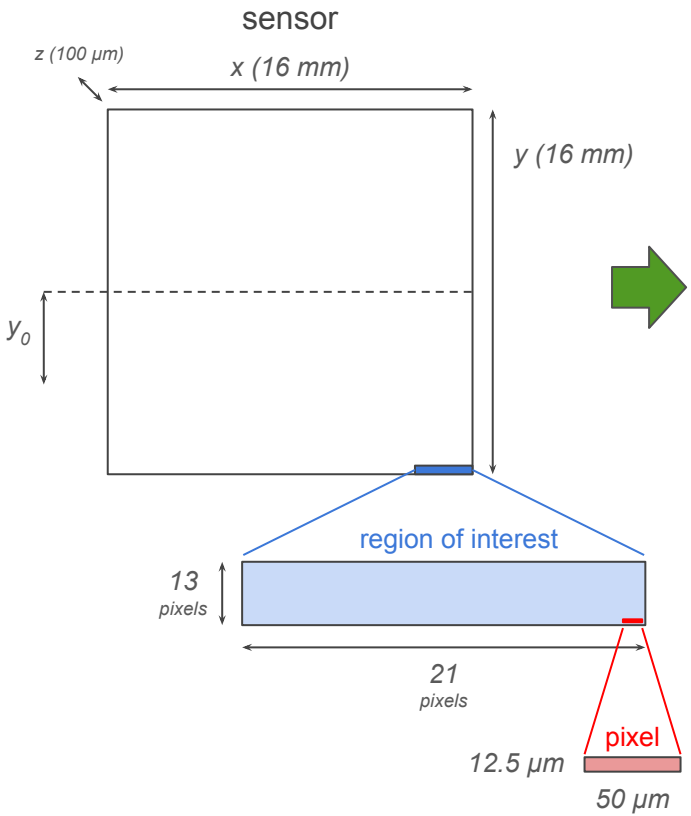


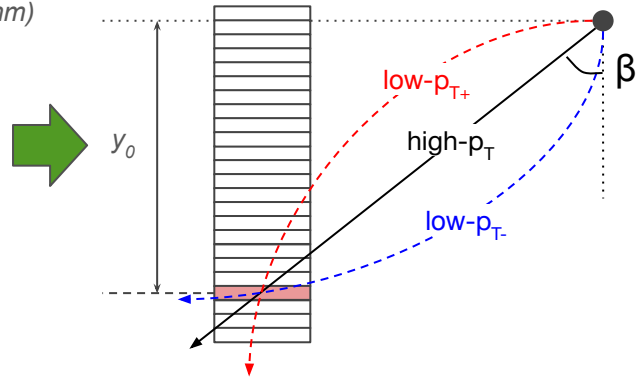
Goal of the Smart Pixel team

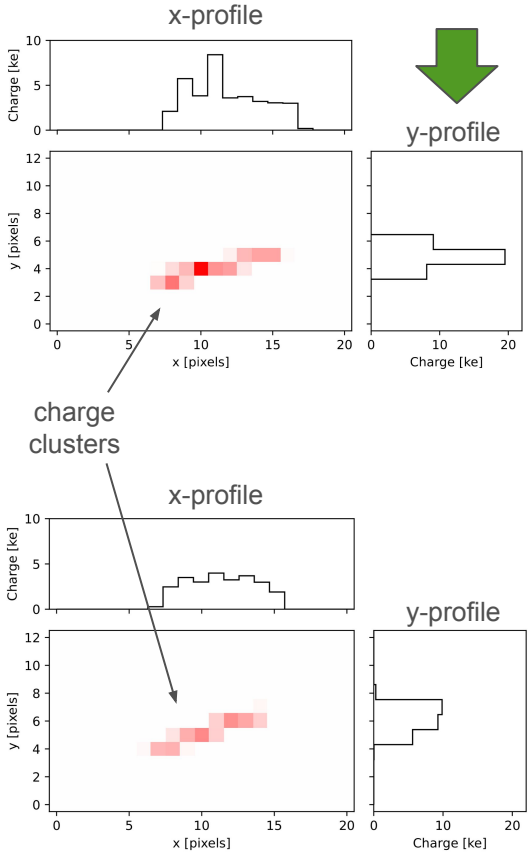
- On-chip data filtering at rate (40 MHz)
- AI algorithms
- Reconfigurable algorithms
- Hybrid pixel detector
  - Silicon sensor
  - Pixelated ROIC
    - Analog front-end + ADC
    - AI in digital logic



Sensor geometry, charge clusters, and profiles

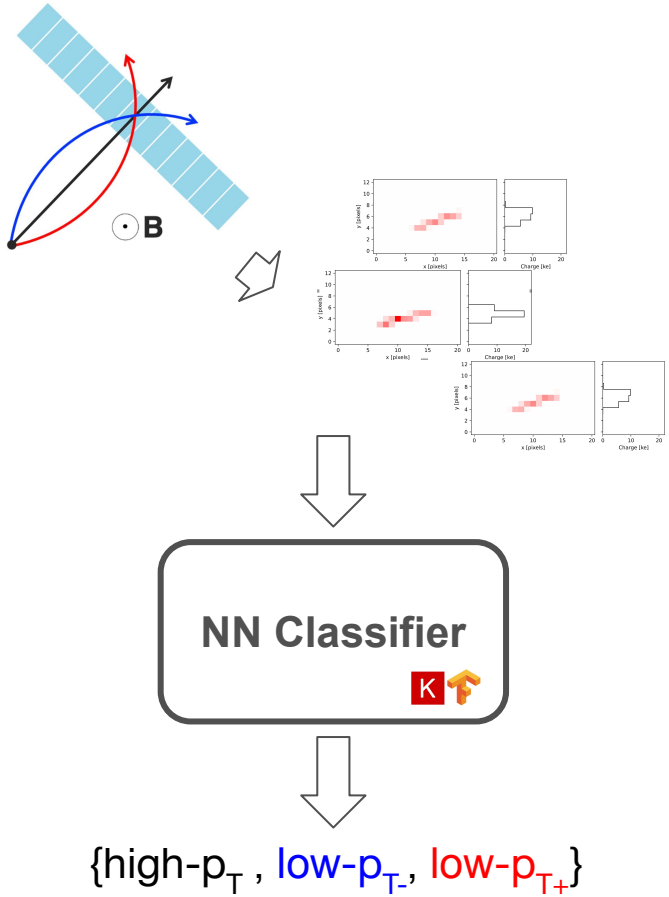






Neural network classifier (filter)

- Inputs are cluster images projected onto y-axis and the associated y_0
- Three output categories
  - high-momentum (> 200 MeV)
  - low-momentum, negatively charged
  - low-momentum, positively charged
- Simulated dataset of 800,000 clusters
- Classical training and testing set split 80%-20%
- Tensorflow/Keras, 200 epochs for training, 20 epochs of early stopping, 1024 batch size, Adam optimizer



Neural-network design space exploration

- Three models of increasing complexity
  - Model 1
    - Two input features: **cluster y-size** and position y_0
    - Single dense layer, 128 neurons, 771 parameters
  - Model 2
    - 14 input features: **cluster y-profile** (13) and position y_0
    - Single dense layer, 128 neurons, 2,307 parameters
  - Model 3
    - 105 input features: **cluster y-profile** at **8 time instants** (13 x 8) and position y_0
    - CNN (time-lapse picture), 83,331 parameters

Model	Sig. efficiency	Bkg. rejection
Model 1	84.8 %	26.6 %
Model 2	93.3 %	25.1 %
Model 3	97.6 %	21.7 %

→ Hardware implementation

Bandwidth saving
54.4% - 75.4%

HLS-based hardware design flow

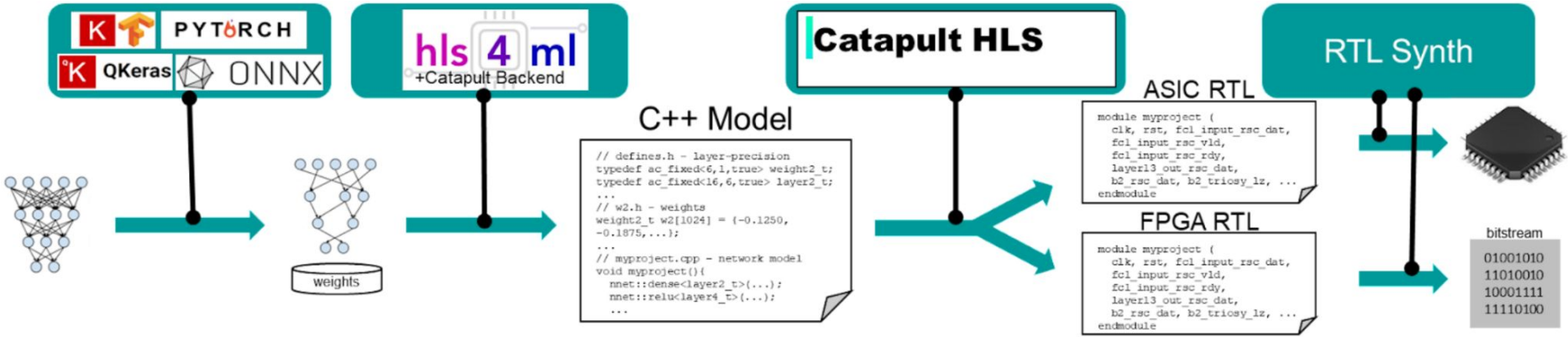
Model Development
- Optimization
- Quantization
- Training

Model Conversion
- Parallelism
- I/O Style/BRAM

C++ Model Gen
- AC Datatypes
- AC Math/Linebuffer

High-Level Synthesis
- Micro-Architecture
- Memory Opt
- Pipelining
- PPA
- ASIC or FPGA Target

RTL Synthesis
- Timing Closure
- Gate Netlist/ Bitstream



Pre-HLs Validation
- Static Checks
- Code Coverage (100-1000x faster than RTL coverage)

CDesignChecker

CCoverage

SCVerify Questa

CFormal Apps

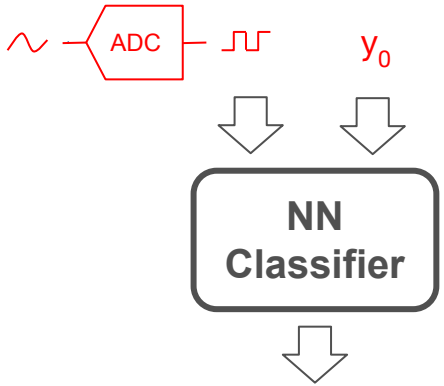
Post-HLS Validation
- C vs RTL Simulation
- RTL Coverage
- UVM Support

SIEMENS
Fermilab

Neural-network hardware co-design (model inputs)

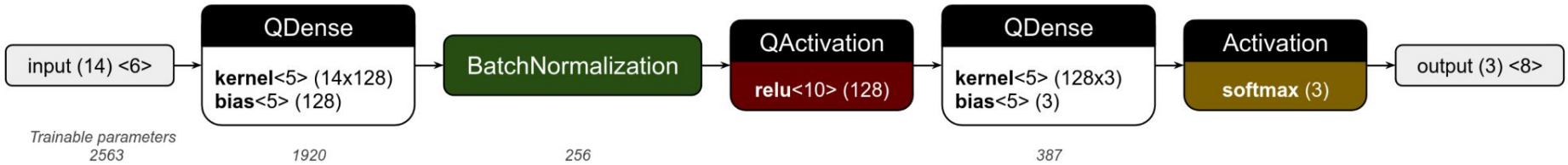
- Input quantization
  - The charge collected by the analog readout goes through an ADC
    - 2-bit quantization for the y-profile
    - Ranges of collected charges
  - y_0 coordinate on 6 bits
    - 64 bins of 250 μm width

ADC output	Charge interval $[e^-]$
00	< 400
01	$400 - 1600$
10	$1600 - 2400$
11	> 2400



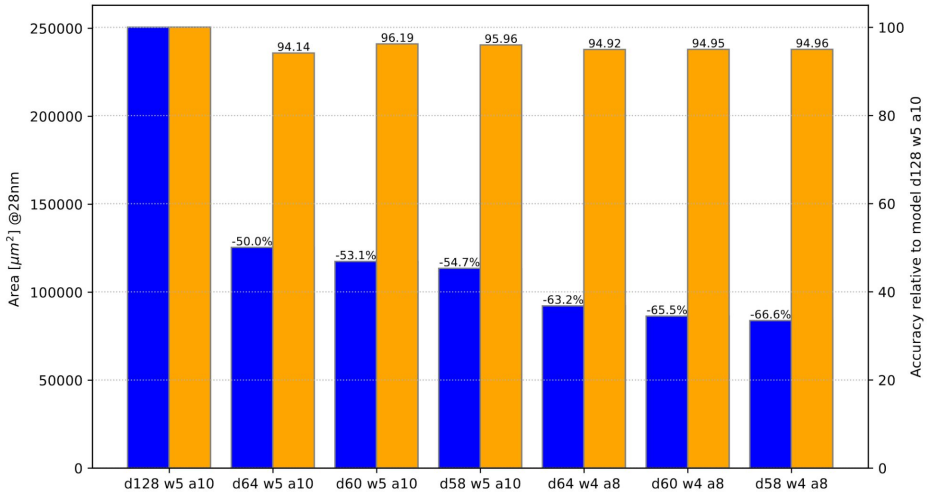
Neural-network hardware co-design (model weights)

- Model quantization
  - Quantization-aware training with QKeras
  - Model “inspired” on the previous design space exploration
    - Extra dense layer
    - Batch normalization (~Dropout) prevent overfitting in training
  - 5 bits for weights, 10 bits for activation provides best results



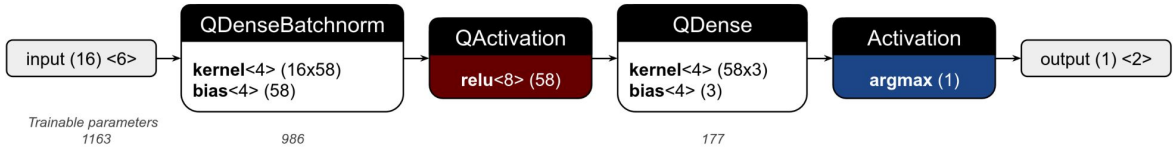
Neural-network hardware co-design (optimizations)

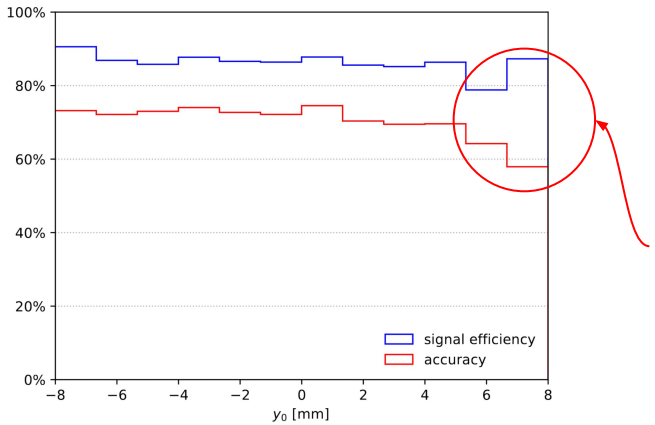
- Goal: reduce area
  - Batch-normalization folding
  - Neurons reduction
  - Weights & activation bit-width reduction
  - Softmax → Argmax



Neural-network hardware co-design (optimizations #2)

- Drop y_0 as a feature and use a region-specific implementation
  - 13 different models provide altogether better performance
  - We can leverage weight reprogrammability

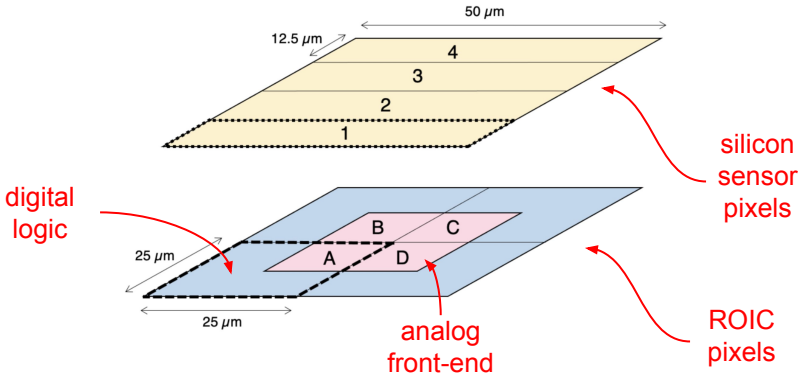


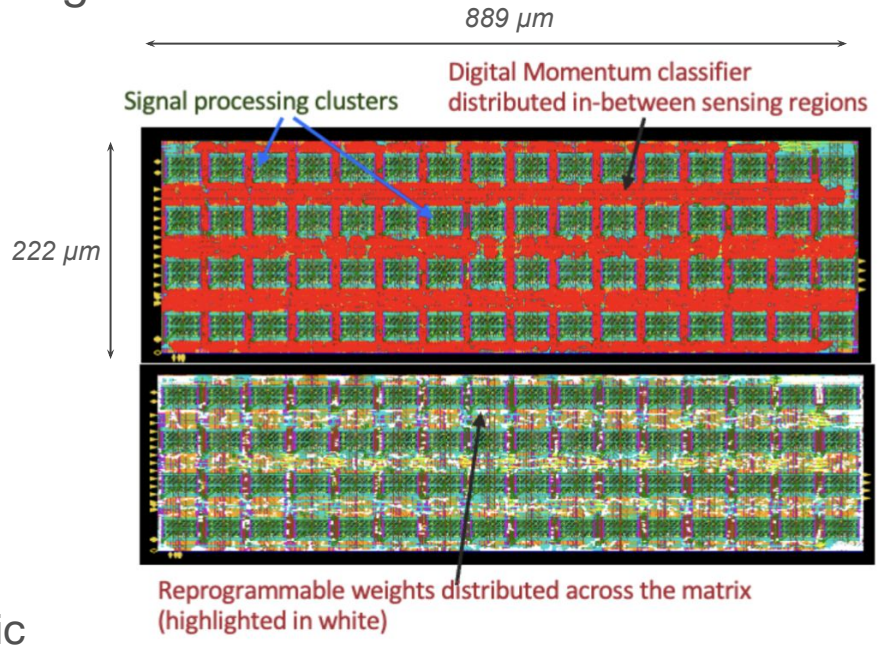


Expected loss in model performance because of Lorentz drift

ROIC integration

- Integration of the ML algorithm as digital logic with the analog front-end into the in the pixelated area (ROIC)



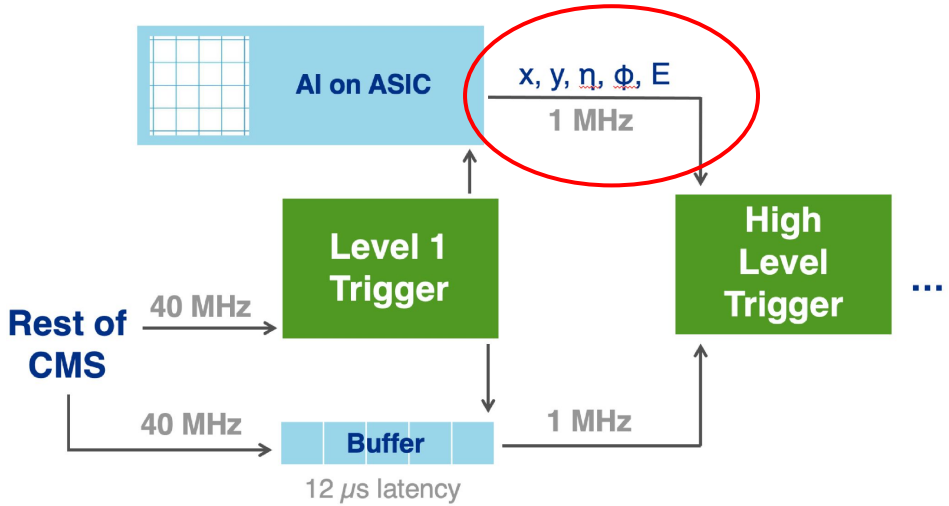


- Low-power 28nm CMOS
  - Early power estimates for the digital logic

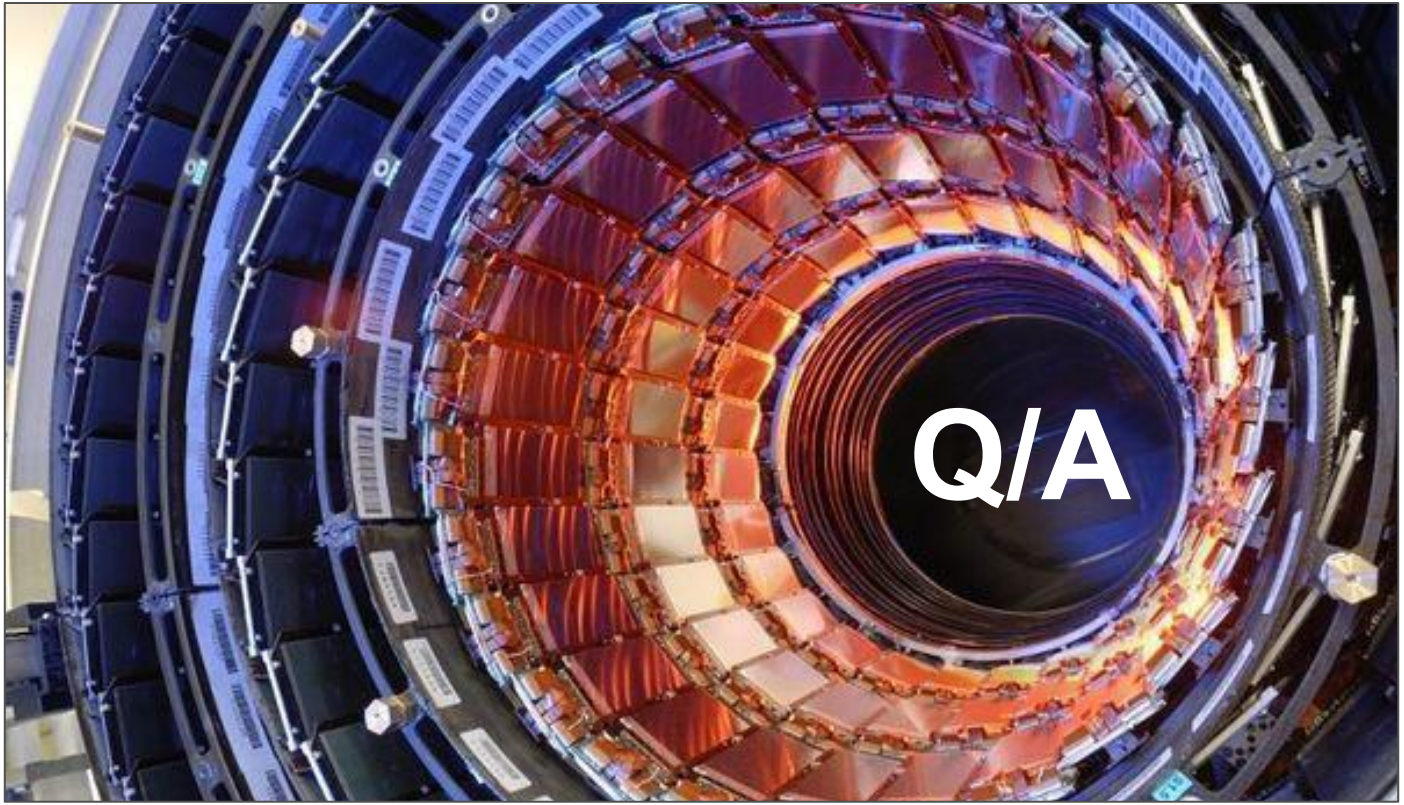
$< 300\ \mu W$

Where we are now

- Tapeout of the current classifier algorithm (**filtering**)
- Moving to **compression** (or **featurization**)
  - Train algorithms to extract properties of the incident particles
  - Read this information out instead of raw (filtered) data



Thank you



Fermilab