Machine-learning-based regression for edge data reduction of small pixel, high-bandwidth silicon detectors

Mathieu Benoit [1], Aaron Young [1], Shruti Kulkarni [1], Rao Narasinga Miniskar [1], Jeffrey Vetter [1], Alice Bean [2]

[1]**Oak Ridge National Laboratory**

[2] **Kansas University**

Introduction

- Practical problems with high bandwidth pixel data streams
- Classifying, encoding and compressing pixel data
- Neuromorphic Computing
  - SNN
  - Practical examples
  - Work at ORNL
- Perspectives

Practical problems with high bandwidth pixel data streams

- Modern pixel detectors can produce zero-suppressed data stream with very large bandwidth (100's of Gb/cm^2)
- There is a rich source of information on the underlying events in the data stream
  - Multiple pixels belong to an event
  - Each pixel carry information (Energy, timing)
  - Geometrical and charge distribution tells us more about an interaction
- The event generating the data stream can be complex

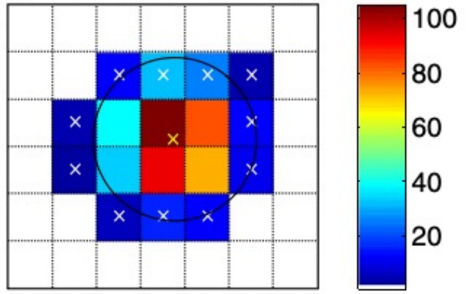
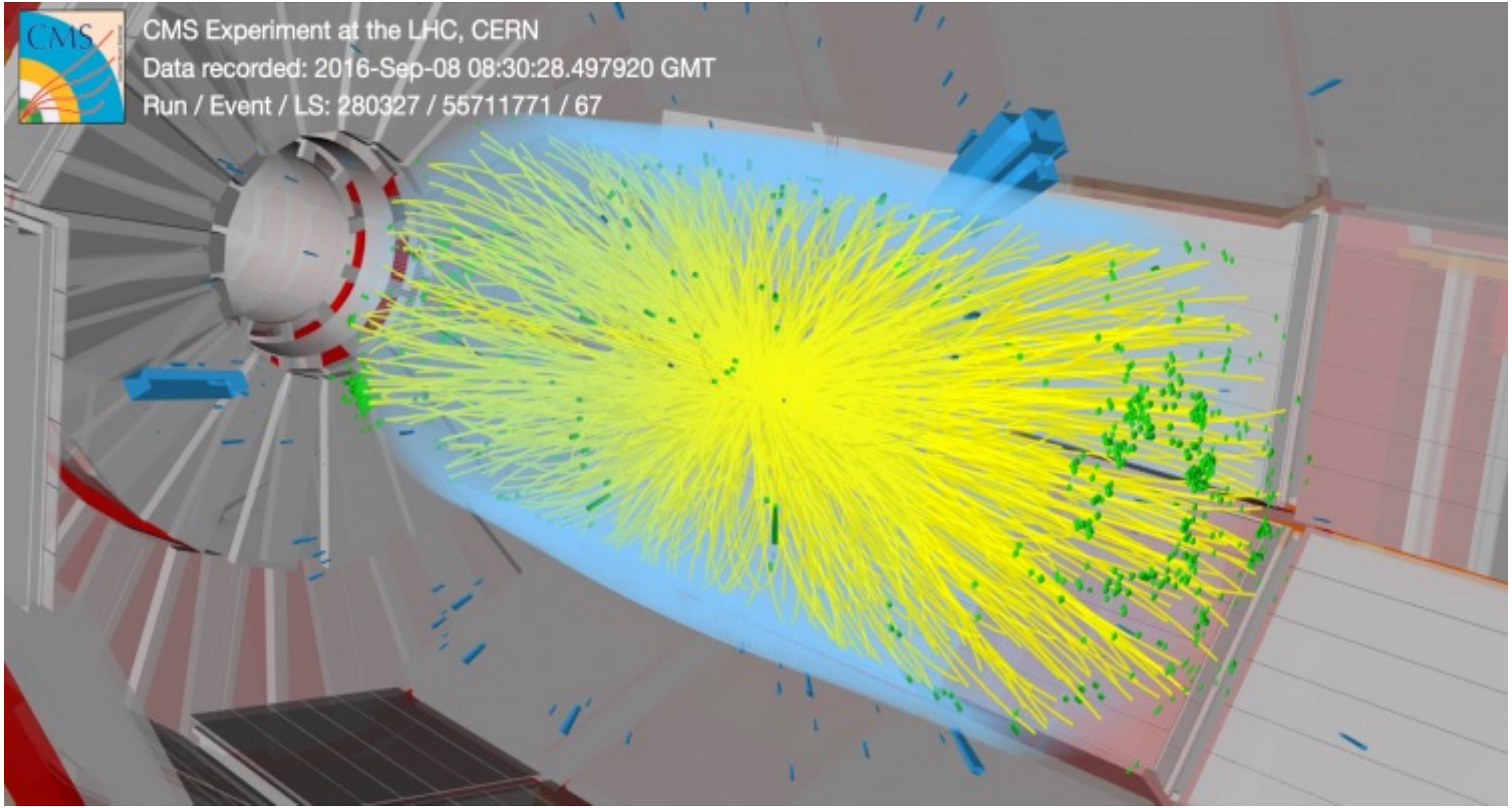
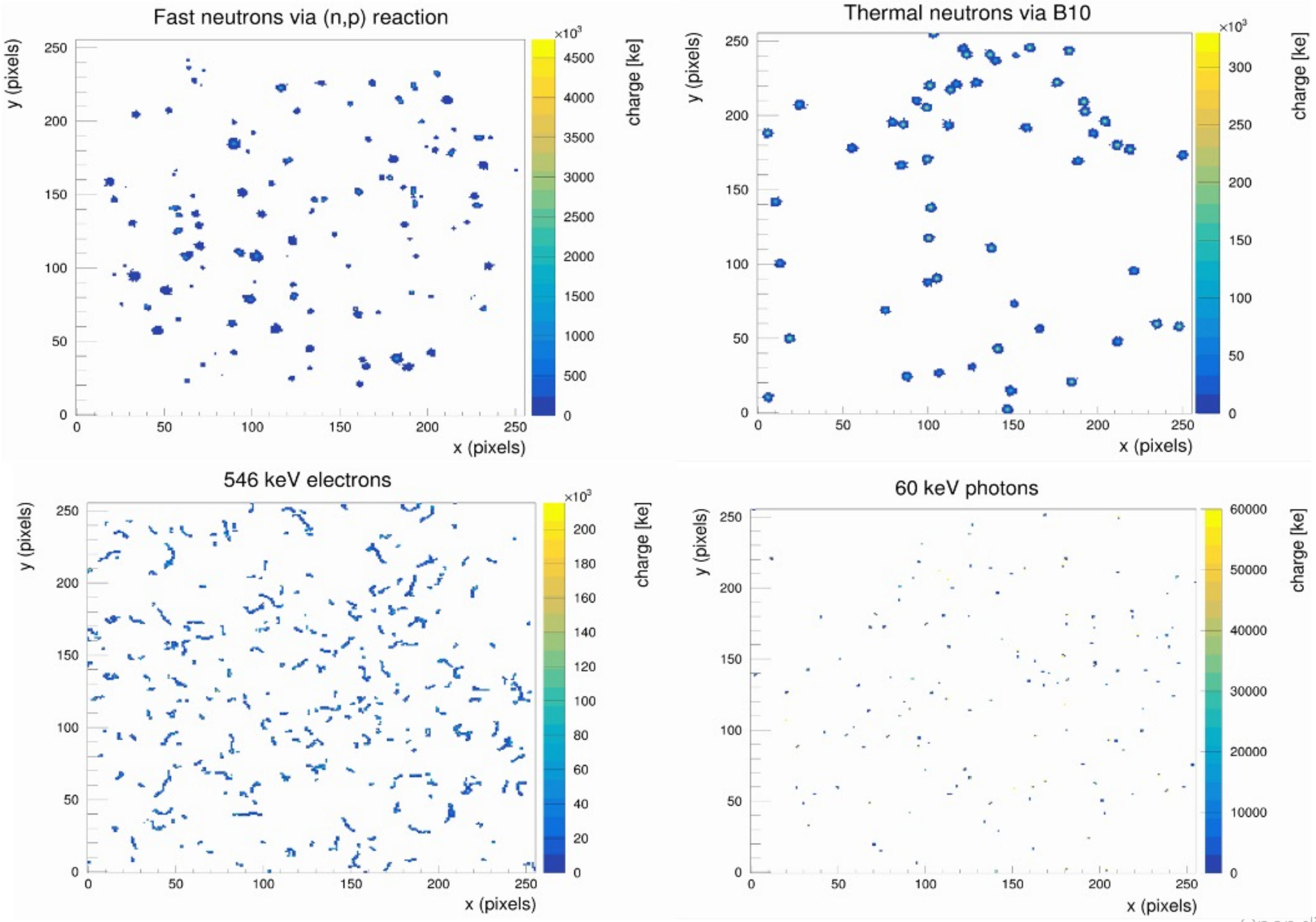


FIG. 3. Discrete Gaussian track imaged by Timepix. The color represents the energy deposited in each pixel. The border pixels are labeled by white crosses. This track has parameters (Tables I and II): *Area* = 17, *Maximum distance from the circle* = 0.51, *Roundness* = 0.78, *Cluster volume* = 0.59 MeV, *Registered amplitude* = 103 keV.

Practical problems with high bandwidth pixel data streams



Classifying, encoding and compressing pixel data



Classifying, encoding/decoding and compressing pixel data

On Detector

- **Pixel raw data is collected, serialized and transmitted to DAQ**

DAQ

- **Data is aggregated, event assembled and written to disk**

Offline

- **Pixel are aggregated in clusters**
- **position is extracted**
- **unrelated data and noise is discarded**

Large amount of of information are carried to disk, offline reconstruction perform data decoding and reduction in multiple manner :

- Data **calibration**
- Clustering of pixels and calculation of a **(X,Y,T,t) coordinate** of the event using energy and timing information
- **Removal of noise related data** and physical data related to **unrelated events**

Classifying, encoding and compressing pixel data

On Detector

- **Pixel raw data is collected, serialized and transmitted to DAQ**

DAQ

- **Data is aggregated, event assembled and written to disk**

Offline

- **Pixel are aggregated in clusters**
- **position is extracted**
- **unrelated data and noise is discarded**

As detector and ASIC technology improved, the data stream that can be produced increases faster that our capabilities to handle it, for various reasons :

- Small experiments with new opportunities
- Restriction in budget, supply chain
- Cost/benefit analysis

Data storage, computing and transmission comes to a cost that can be better used in the project

- Potential cost savings for small medium and large scale experiments

Classifying, encoding/decoding and compressing pixel data , **at the edge**

On Detector

- **Pixel raw data is collected, serialized and transmitted to DAQ**

DAQ + Edge Data reduction

- **Data is aggregated, event assembled, classified and simple quantities reconstructed**

Offline

- **Computing receives ready to used qualified logical data**
- **Event are reconstructed and higher order objects are recorded to disk**

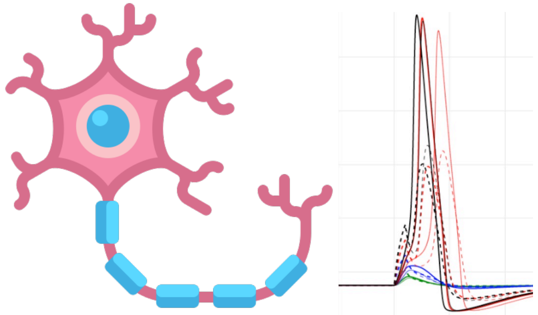
Change of paradigm ! Yes, we must throw away data, we already do !

With sufficiently smart and efficient algorithms , we can process data as it stream and reduce the data stream efficiently close to data acquisition

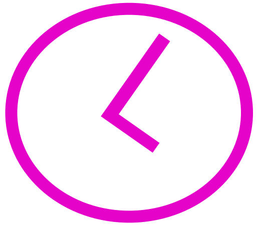
- Data is aggregated and transmitted to DAQ and stored in High-Speed Memory, accessible by a high-speed bus (For example PCI express and DDR4 memory)
- Commodity FPGA/ Custom cards consume the data stream and reduce the data
- High performance network consumes the reduced data and transmit to offline

Neuromorphic Computing

20W !!



Neurons process and store information as needed



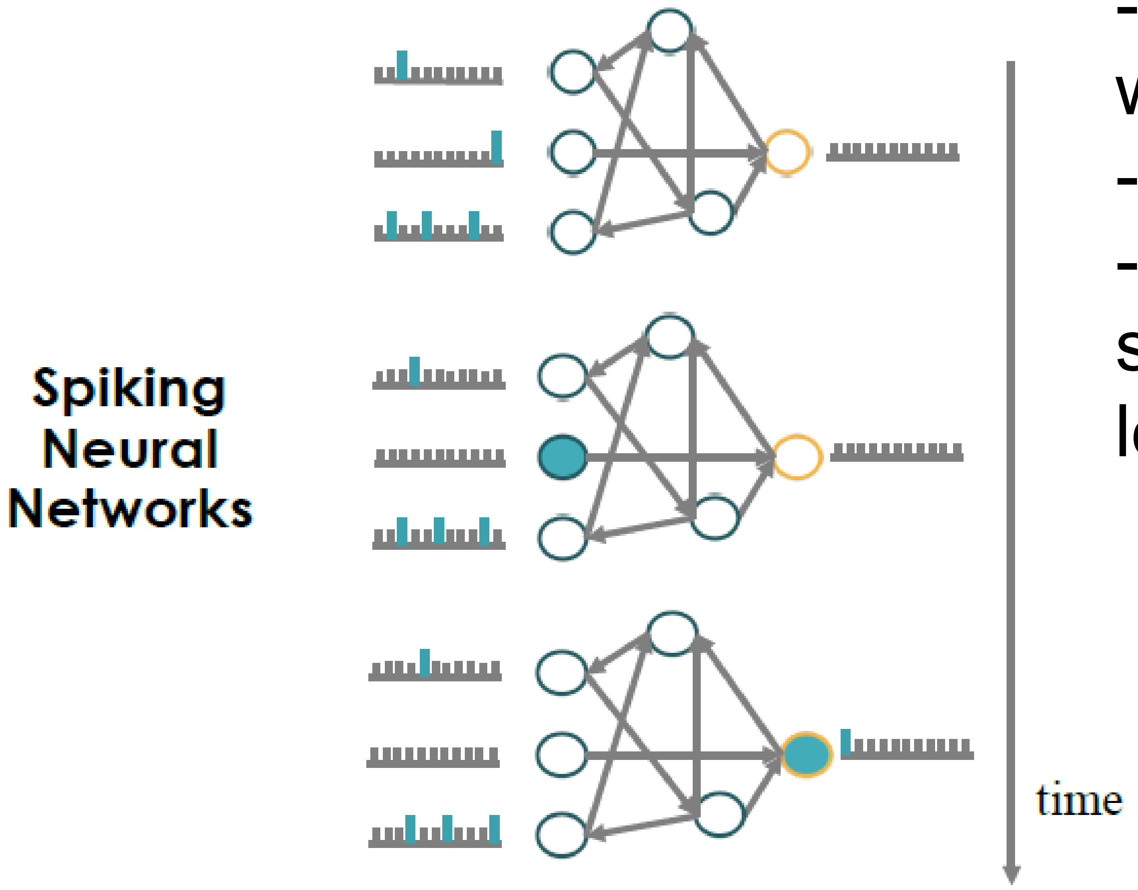
Processing Unit

Memory

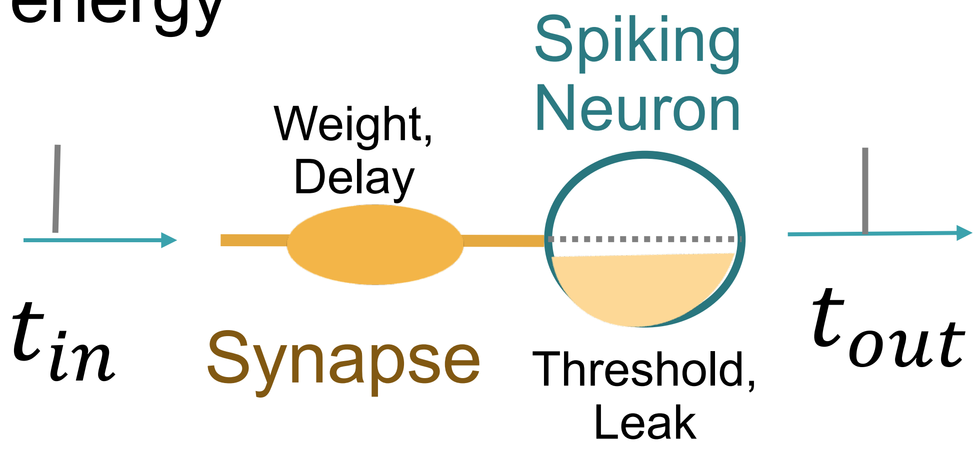
-Processing and Memory Separated
-Clocked data uses lots of power

Lots of watts!!

Spiking Neural Networks (SNN)



- Data spikes excite “neurons” which communicate via synapses
- Use temporal information
- Event driven cameras have been shown to process data faster with less energy

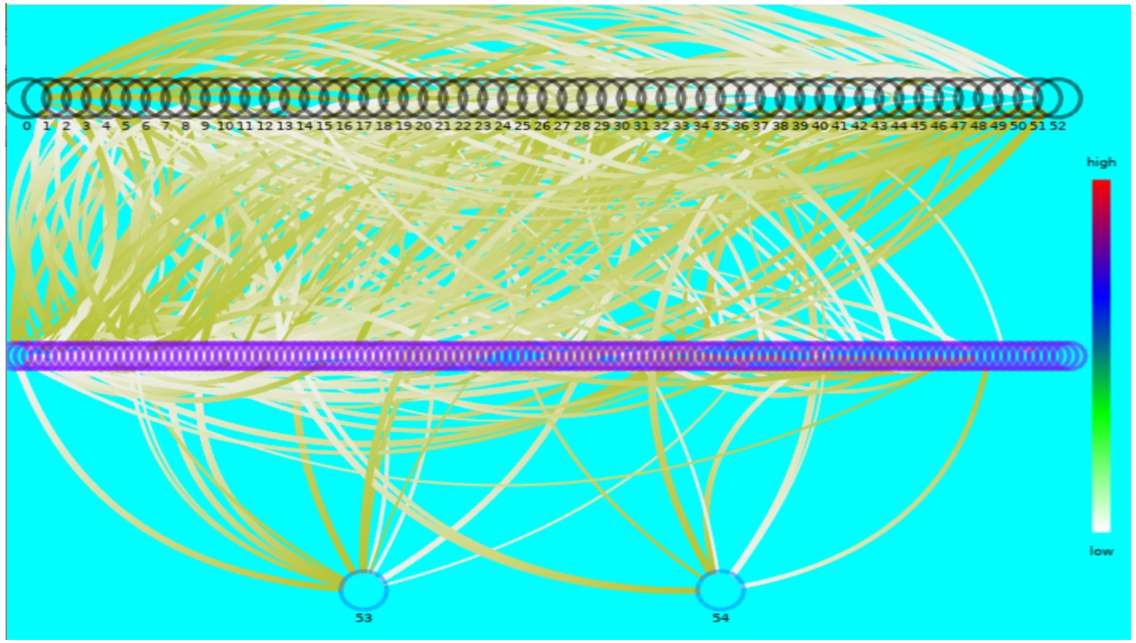


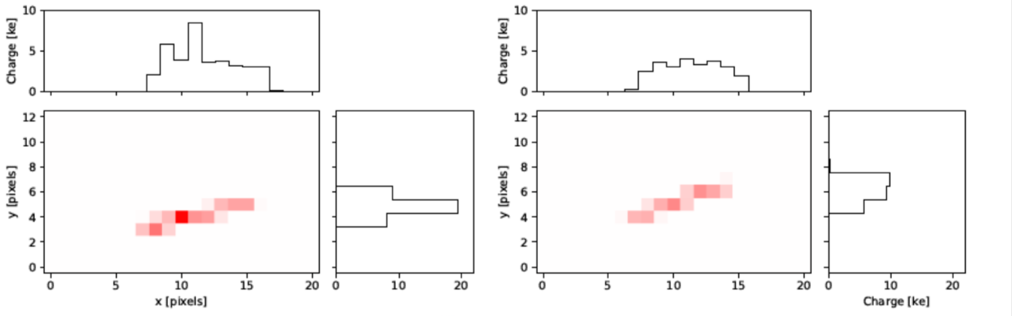
Leaky Integrate and Fire Neuron

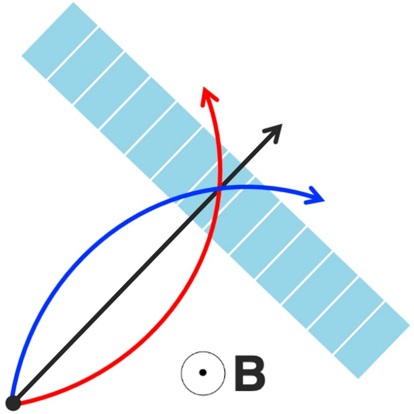
Results of SNN trained classifier for Smart Pixels

- Use Network with highest signal efficiency for p_T> 2.0GeV clusters
  - Trained Network size: 84 neurons, 493 synapses

Models	DNN	DNN (quantized)	SNN (this work)
Signal Efficiency	94.8 %	91.7 %	**91.89%**
Data Reduction	24.02 %	25.71 %	**25.47%**
Neurons	128	128	**84**
Parameters	2049	2561	**930**

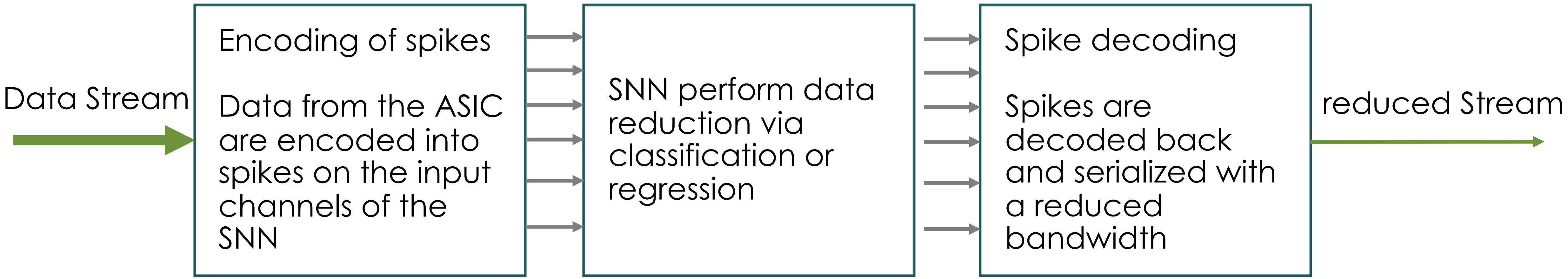






Ongoing work at ORNL/KU

- ORNL team is working on the implementation of Neural Networks for pixel detector data reduction using State-Of-The-Art Spiking Neural Network models and training tools
  - SNN design and simulation tools are used to generate networks of interest using Simulation Data and Monte-Carlo Truth
  - Generated networks can be implemented in RTL using HLS tools
  - RTL Models can be implemented in FPGA for data processing
  - Ultimately, they can be integrated in ASIC for ultimate performance, latency and power consumption



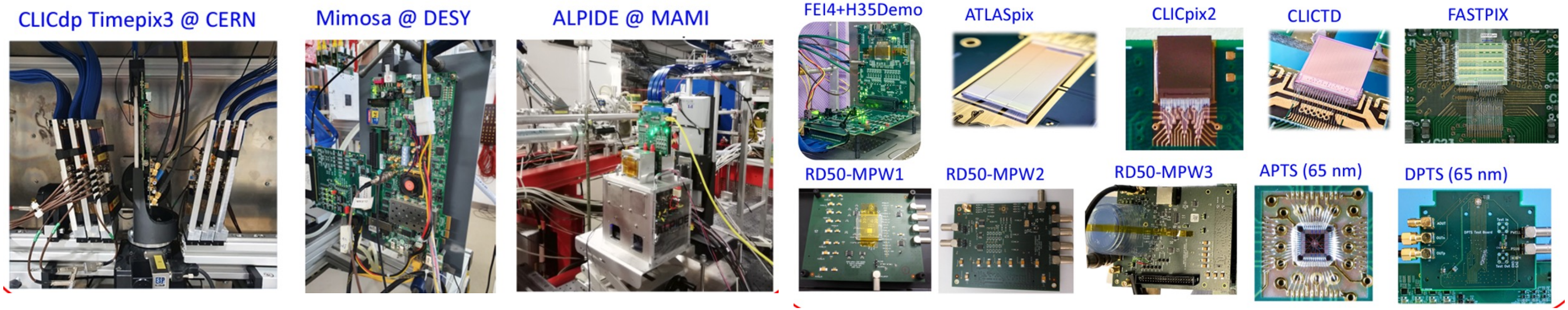
The CARIBOu 2.0 Framework

The **CaRIBOu 1.x** framework was originally developed in 2014, in collaboration between BNL (HW and Felix integration), UNIGe, and CERN (FW and SW design) as a versatile platform for DAQ development of our prototypes

- Over **50 systems in use across the world**
- **Large base of users** across multiple experiments , multiple ASIC and sensors
- Focus on sharing code, experience in system design to **reduce time to first test** by avoiding hardware, firmware and software work duplication and providing **robust design tested by fire** by the large base of users.

Collaboration between **ORNL, BNL OMEGA group, University of Carleton, Canada** in the development of the next generation system hardware **, CARIBOu 2.0**

- Strong requirements from pixel R&D for large bandwidth, analog signal processing, timing and more system integration, scalability to large array, AI/ML algorithm integration
- Software and firmware design in collaboration with CERN EP, DESY, RD50/AIDANova, DRD3
- ORNL wants to deliver a scalable Timepix4 readout using CARIBOu 2.0



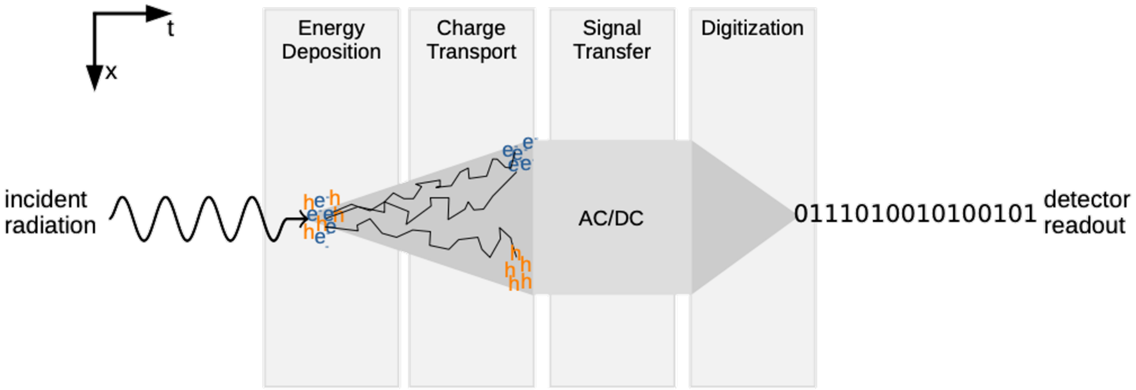
https://indico.cern.ch/event/1214410/contributions/5325456/attachments/2616558/4522699/Caribou_DRD3_Workshop-23Mar2023.pdf

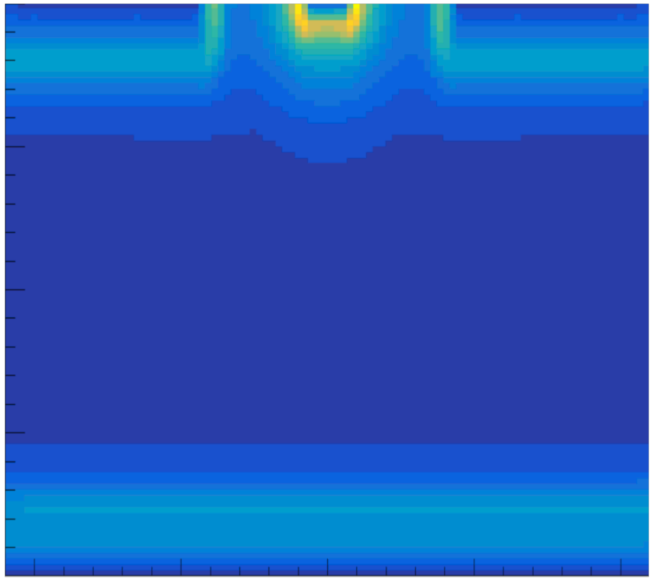
The Allpix^2 Framework

- A **Modular, Generic** Simulation Framework for pixel Detectors

- The framework aims at facilitating the different steps of the simulation of semiconductor detectors
  - Energy deposition in the detector material (GEANT4 etc.)
  - Charge Transport in the semiconductor (TCAD, various substrate and geometries)
  - Transfer, Digitization and Analysis
  - **Excellent training tool for AI/ML**

- The developpers aim at implementing the best practices in semiconductor detector simulation in a generic way to provide to the community verified and standardardized methods and a developement environment for further improvement to simulation methods
  - Framework infrastructure
  - Documentation, examples and code demonstration





Website https://cern.ch/allpix-squared
Repository https://gitlab.cern.ch/allpix-squared/allpix-squared

Possible collaborative opportunities

- Similar issues are common to pixel detectors in various experiments
  - Clustering and position extrapolation of hits
    - ETA correction, merge-hits disentanglement, Delta-electron removal
  - Reduction of background un-associated to interesting physics
    - Sorting of interaction by particle, nature of interaction
    - Sorting hits by geometrical origin (for example removing beam halo!)
- Similar challenges
  - How to obtain good efficiency, deal with imperfect detectors
  - How to deal with pile-up , various pixel shapes
  - Etc...

Conclusion and perspective

- New experiments will produce larger than even data volume that we need to handle in an intelligent way
  - Machine Learning and AI provide new opportunities to design smart data acquisition system that integrate data reduction schemes
  - Similarities between experiments calls for a collective effort
- We propose the formation of a Working group on smart edge data reduction techniques in the framework of this RDC
  - Share common tools and standard for simulation and training
    - Allpix^2, SNN training tools
  - Federate expertise and tackle identified challenges, providing solution for the communities
    - IP for FPGA integration -> Neuromorphic accelerator ASICs etc.

References

- Shruti R. Kulkarni et al., “On-sensor Data Filtering using Neuromorphic Computing for High Energy Physics Experiments”, ICONS 23 Proceedings, arXiv:2307.11242
- Jieun Yoo et al., “Smart pixel sensors: towards on-sensor filtering of pixel clusters with deep learning,” arXiv: 2310.02474v (3 Oct 2023)