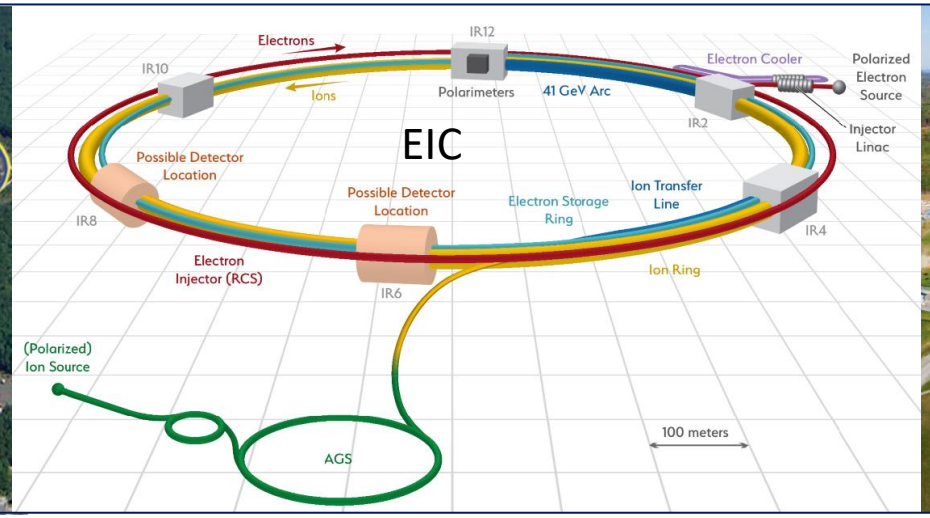
AI/ML for Streaming Readout

Jin Huang

Brookhaven National Lab

Brookhaven National Laboratory

Experiments deploying streaming readout DAQs



Examples in focus of this talk: RHIC (PHENIX, sPHENIX), CEBAF (BDX, CLAS12), EIC (EPIC)

- This talk is an in-complete review of the field,
  see also experiments including at LHC (LHCb, ALICE, AMBER), at FAIR (CBM)
- Streaming Readout Workshop series: [link]
- Talks in this session: Marco Battaglieri, Diana McSpadden

Nuclear collider experiments: unique real-time system challenges leads to streaming DAQ

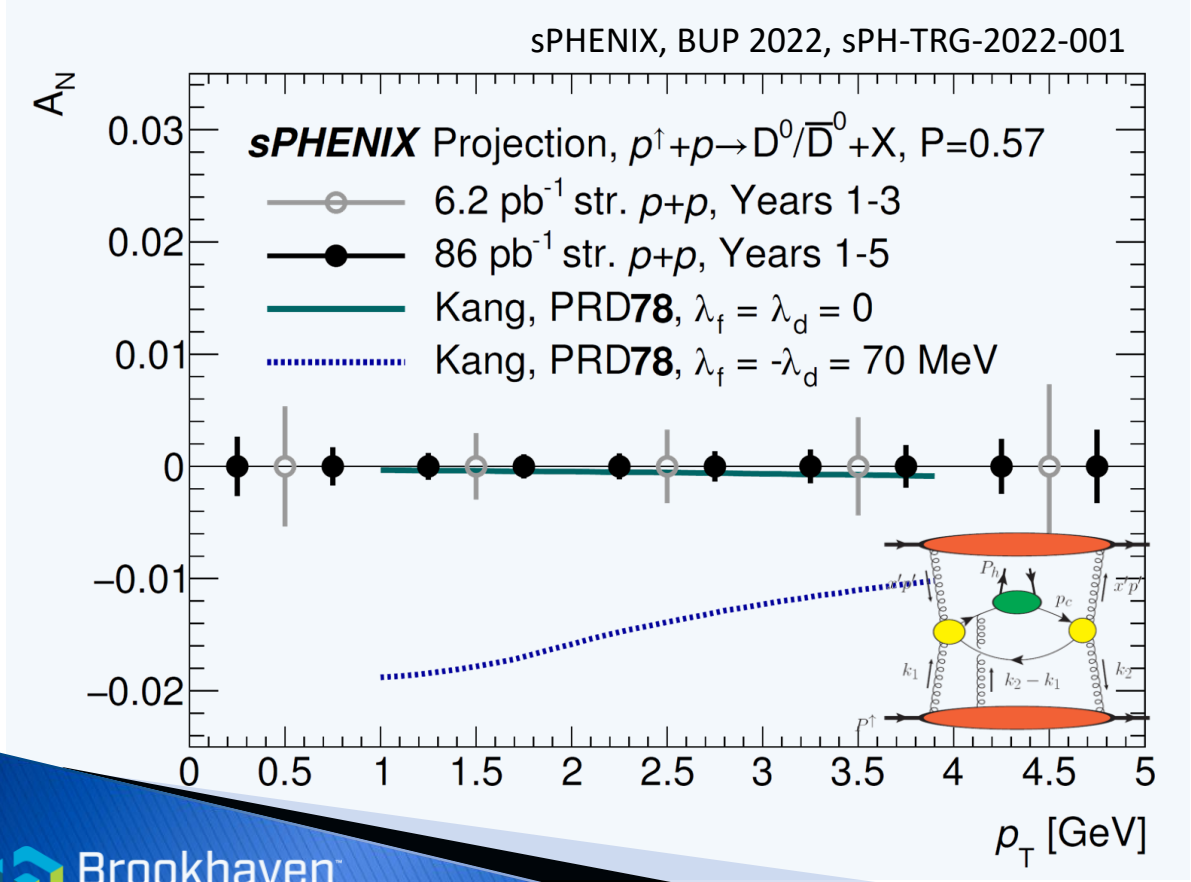
	EIC	RHIC	LHC → HL-LHC
Collision species	$\vec{e}+\vec{p}, \vec{e}+A$	$\vec{p}+\vec{p}/A, A+A$	$p+p/A, A+A$
Top x-N C.M. energy	140 GeV	510 GeV	13 TeV
Bunch spacing	10 ns	100 ns	25 ns
Peak x-N luminosity	10^{34} cm^{-2} s^{-1}	10^{32} cm^{-2} s^{-1}	$10^{34} \rightarrow 10^{35}$ cm^{-2} s^{-1}
x-N cross section	**50 µb**	**40 mb**	**80 mb**
Top collision rate	500 kHz	10 MHz	1-6 GHz
$dN_{ch}/d\eta$ in p+p/e+p	0.1-Few	~3	~6
Charged particle rate	**4M N_{ch}/s**	**60M N_{ch}/s**	**30G+ N_{ch}/s**

- Signal data rate is moderate → possible to streaming recording all collision signal
- But events are precious and have diverse topology → hard to trigger on all process
- Background and systematic control is crucial → avoiding a trigger bias

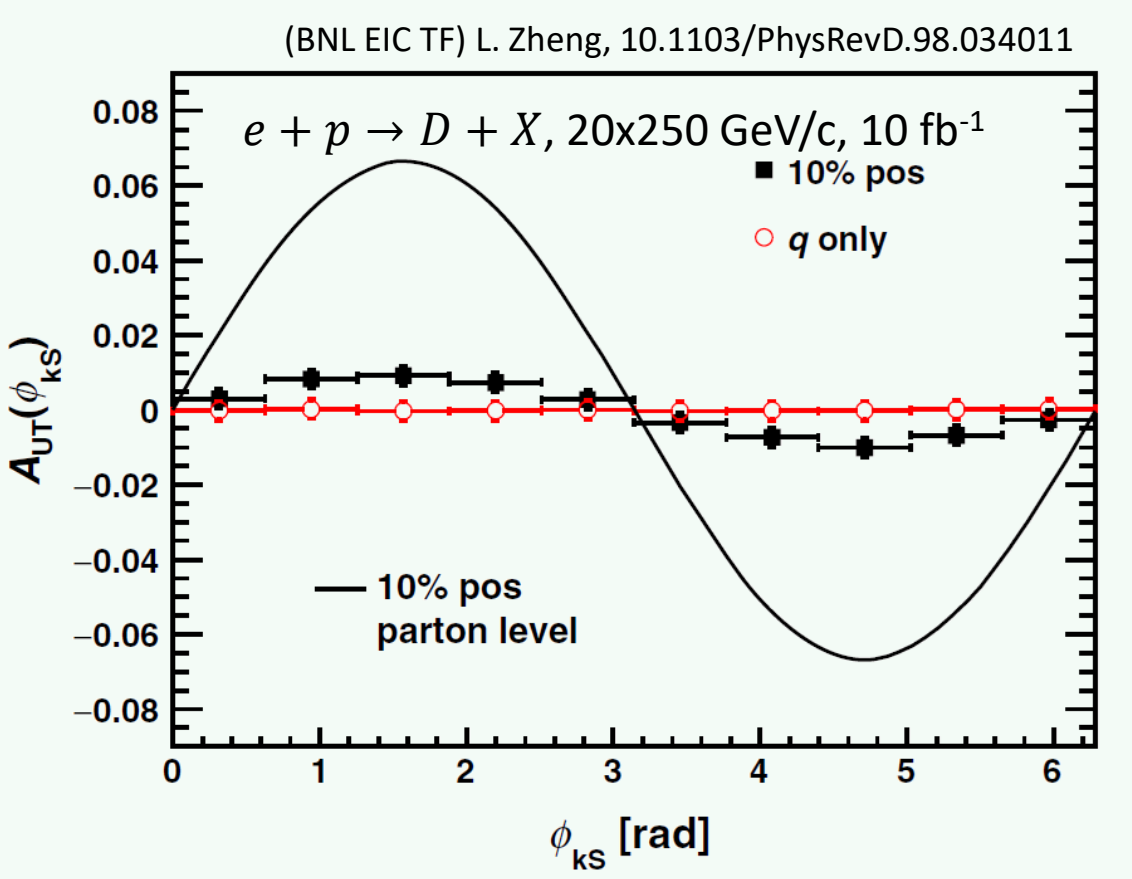
Physics only accessible with a streaming DAQ:
e.g. low p_T HF in hadronic decay → window to gluon dynamics

Universality test on gluon Sievers eff.

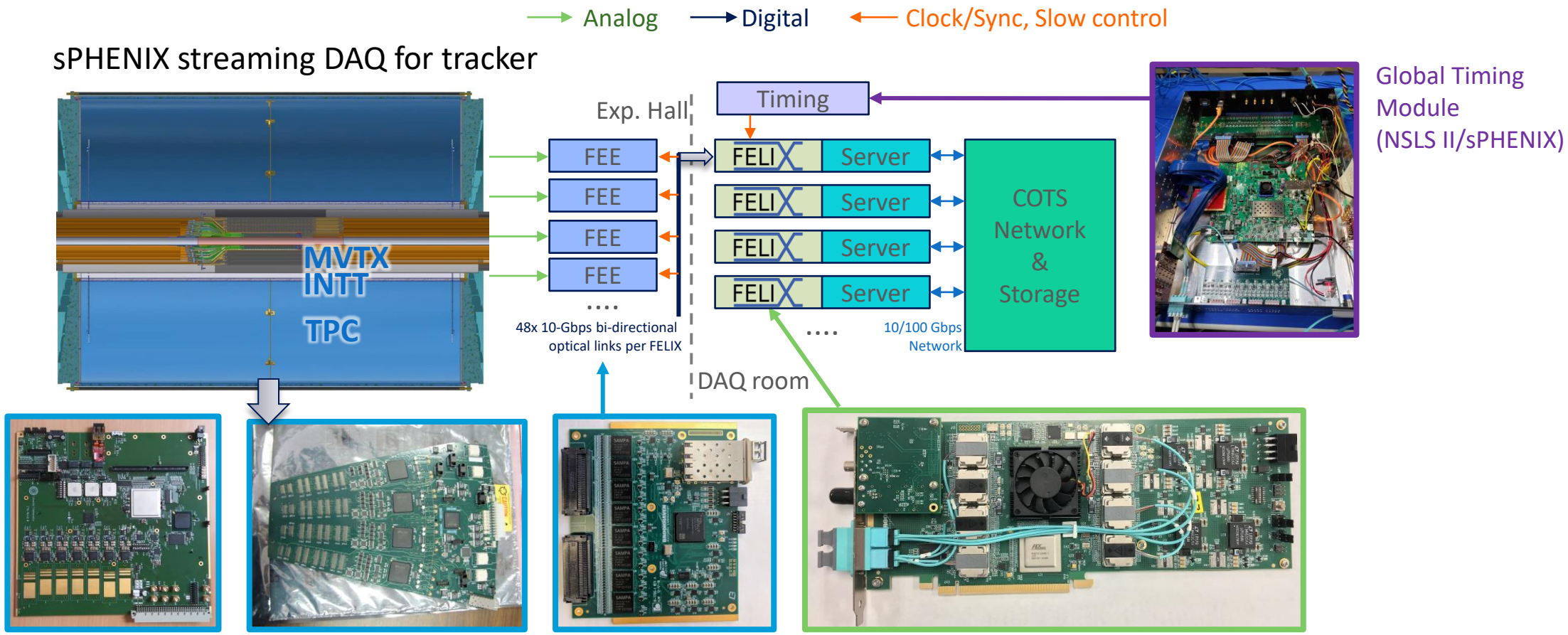
sPHENIX D^0 trans. spin asymmetry, A_N→ Gluon Sievers via tri-*g* cor.



EIC SIDIS D^0 transverse spin asymmetry → Gluon Sievers



Streaming readout electronics: sPHENIX as example



MVTX RU, 200M ch INTT ROC, 400k ch

ALPIDE (ALICE/sPHENIX), FPHX (PHENIX)

TPC FEE, 160k ch

SAMPAv5 (ALICE/sPHENIX)

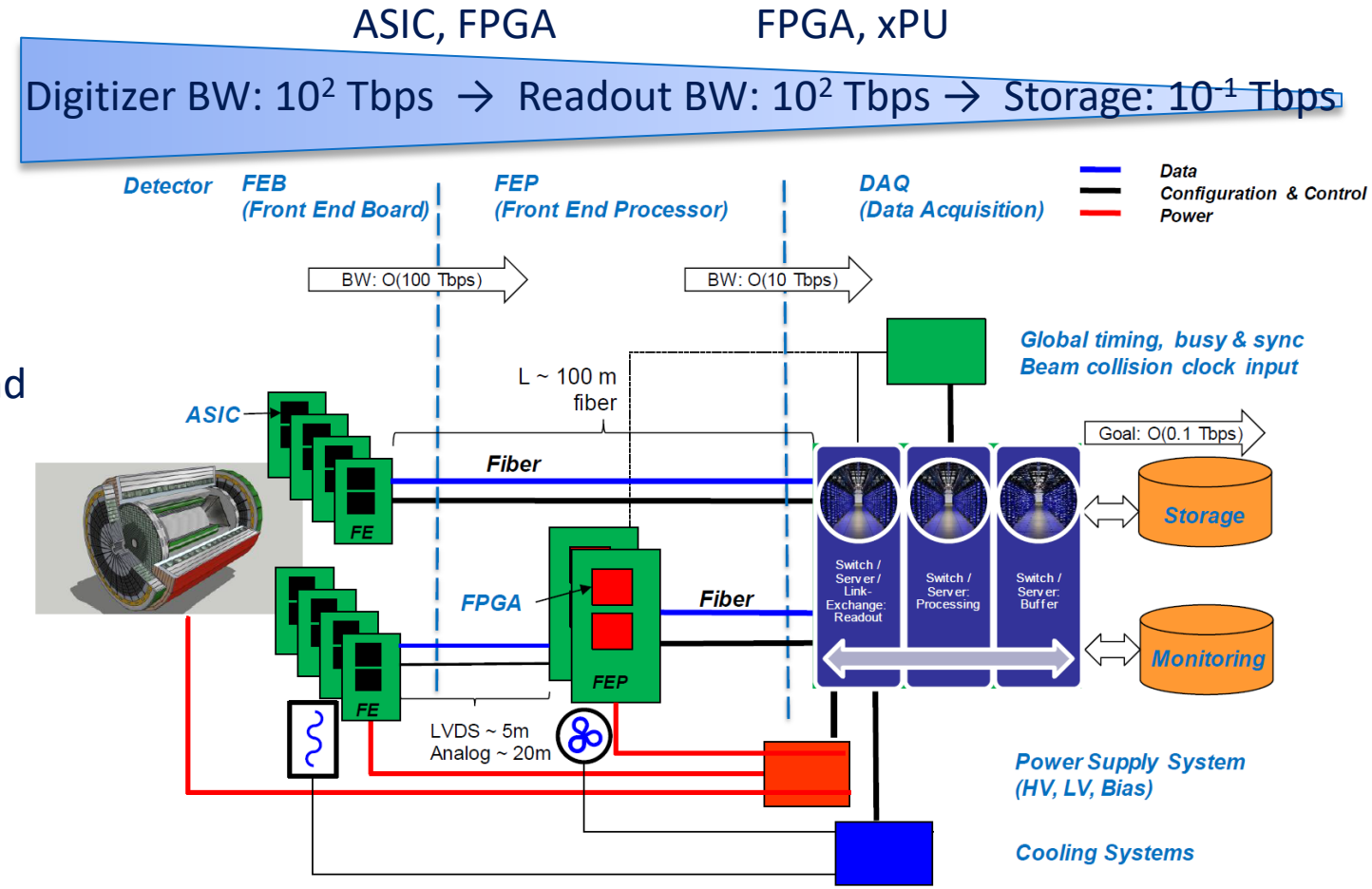
BNL-712 / FELIX v2 x38 (ATLAS/sPHENIX/CBM)

FELIX Ref: 10.1109/tim.2019.2947972

Similar role as PICe40 in LHCb / ALICE

Streaming readout data flow: EIC as example

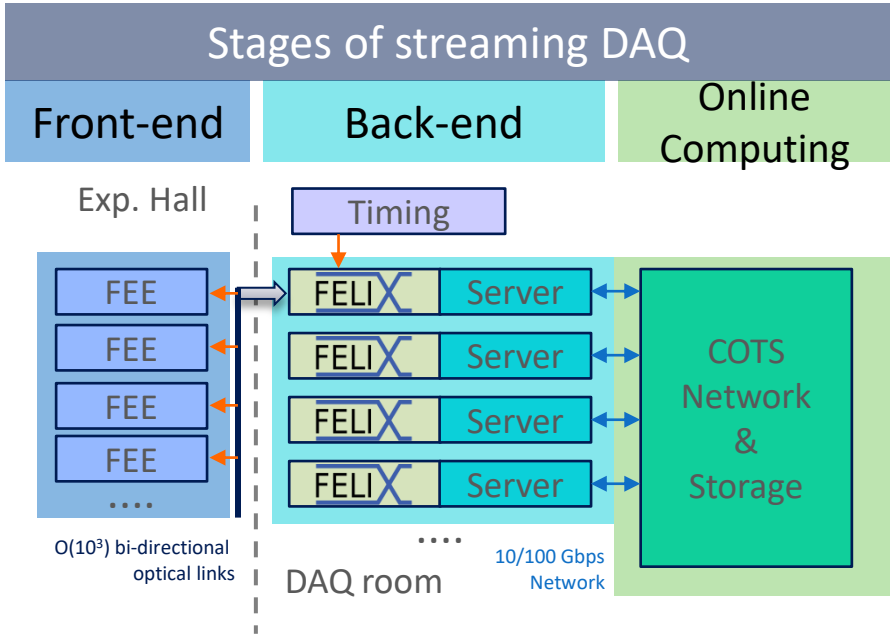
- **EIC streaming DAQ**
  - → Triggerless readout front-end (buffer length : μs)
  - → DAQ interface to commodity computing (FELIX as the candidate in all EIC proposals) Background filter if excessive background rate
  - → Disk/tape storage of streaming time-framed zero-suppressed raw data (buffer length : s)
  - → Online monitoring and calibration (latency : minutes)
  - → Final Collision event tagging in offline production (latency : days+)



[EIC-CDR]

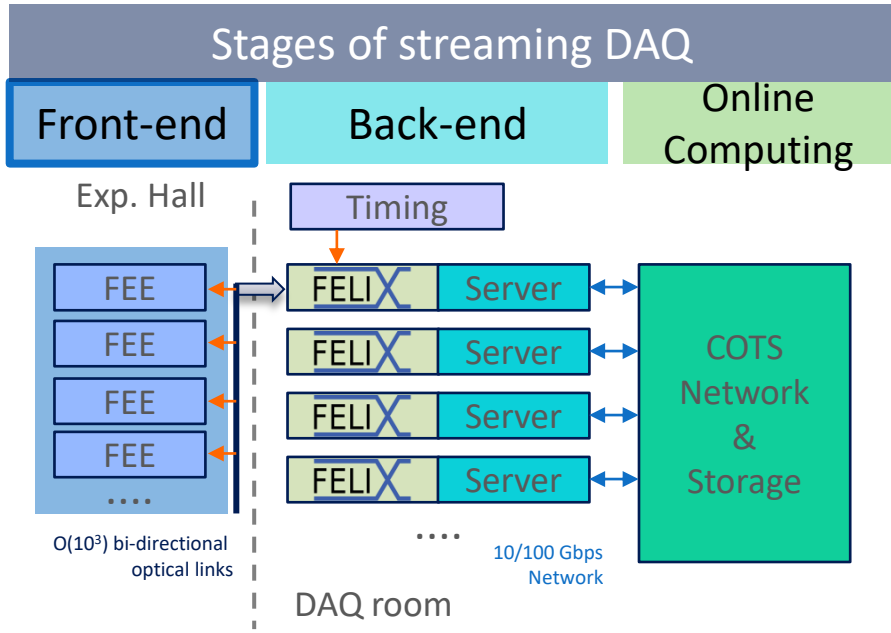
AI in streaming readout DAQ

- Main challenge: data reduction
  - Traditional DAQ: triggering was the main method of data reduction, assisted by high level triggering/reconstruction, compression
  - Streaming DAQ need to reduce data computationally: zero-suppression, feature building, lossy compression
- Opportunities for Real-time AI
  - Emphasize on reliable data reduction, applicable at each stages of streaming DAQ: Front-end electronics, Readout Back-end , Online computing
  - Data quality monitoring, fast calibration/reconstruction/ feedback
    - Also applicable to triggered DAQ.
    - Not focus of this talk, nonetheless important for streaming experiments



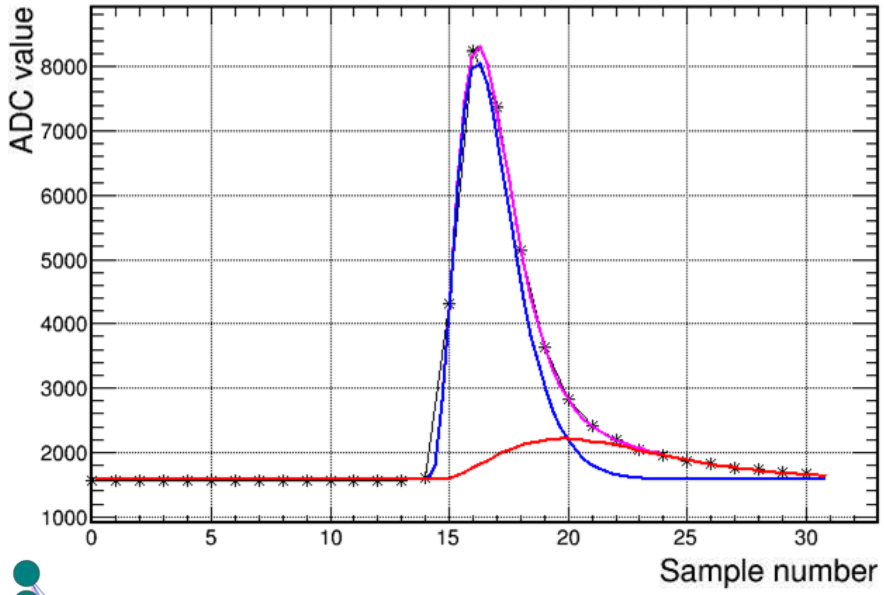
Streaming DAQ stage 1: Front-end electronics

- Perform digitization (ADC, TDC, pixel readout)
  - Common data reduction strategy to immediately apply zero-suppression
- AI opportunities:
  - Improved zero-suppression, e.g. small signal recovery
  - Feature building (example in next slides)
  - Compression (example in later slides)
- Target hardware: ASIC, (smaller) FPGAs
  - Common requirement of low-power consumption, radiation tolerant

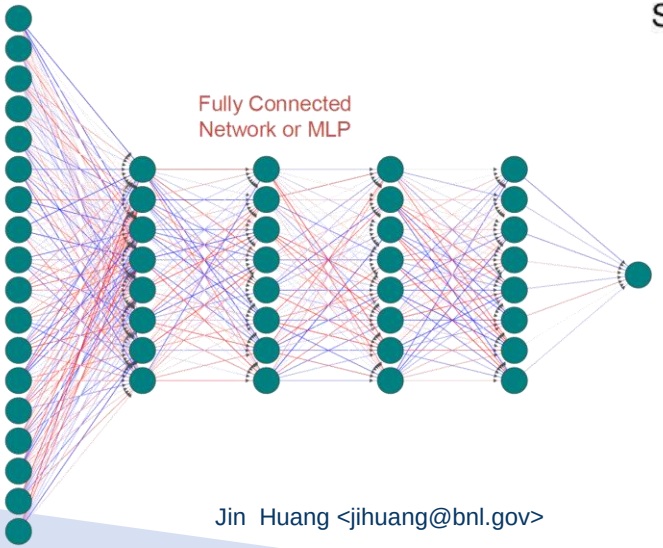


ADC time series → feature of amplitude and time

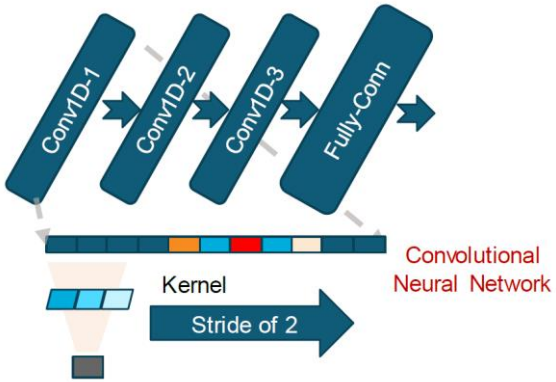
- Wave form digitizer is popular, output data in ADC time series
- In the front-end, NN can be used to extra features such as amplitude and time of arrival
- Fit limited resource in FEE FPGA or ASIC: Emphasizes on quantized-aware training training and pruning



sPHENIX calorimeter Test beam data: [10.1109/TNS.2020.3034643]



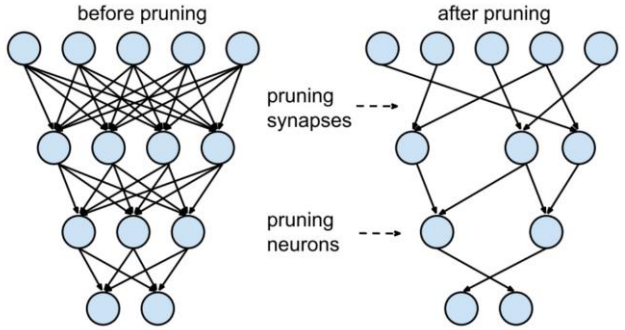
[arXiv:2204.13223]

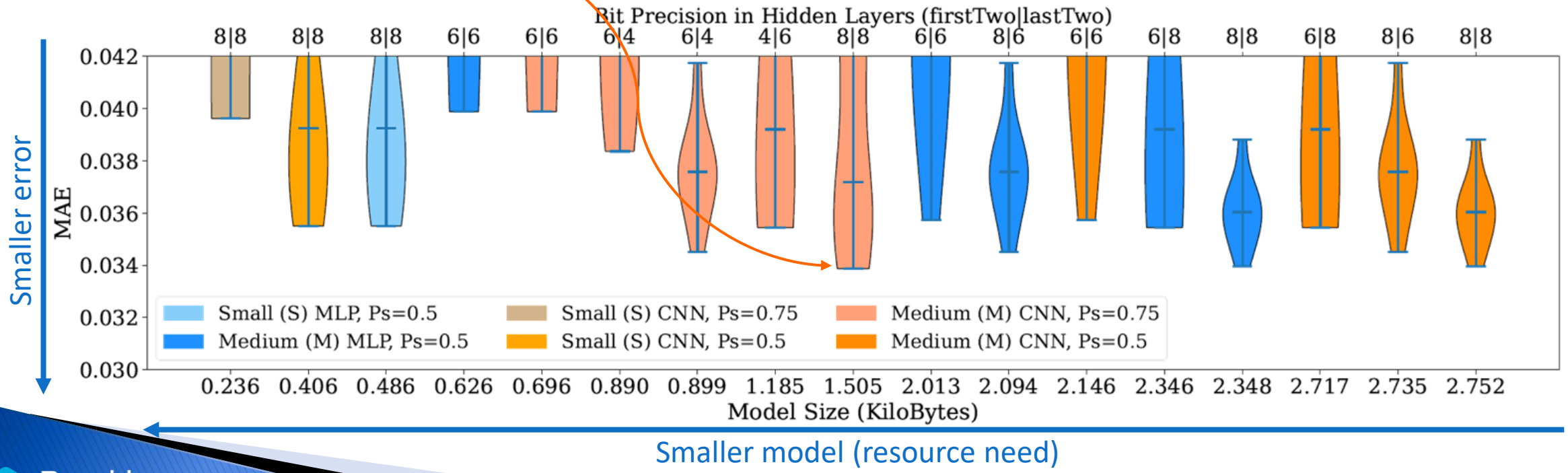


Pruning + Variable Bit Quantization-aware Training

[S. Miryala *et al* 2022 *JINST* **17** C01039]

- Simulated LGAD waveform data
- Highly pruned (sparsity=0.75) CNN with 8bit internal precision strikes good performance (smaller error) and small model size



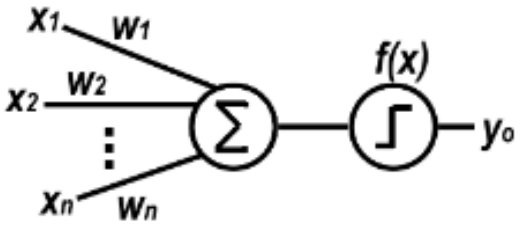


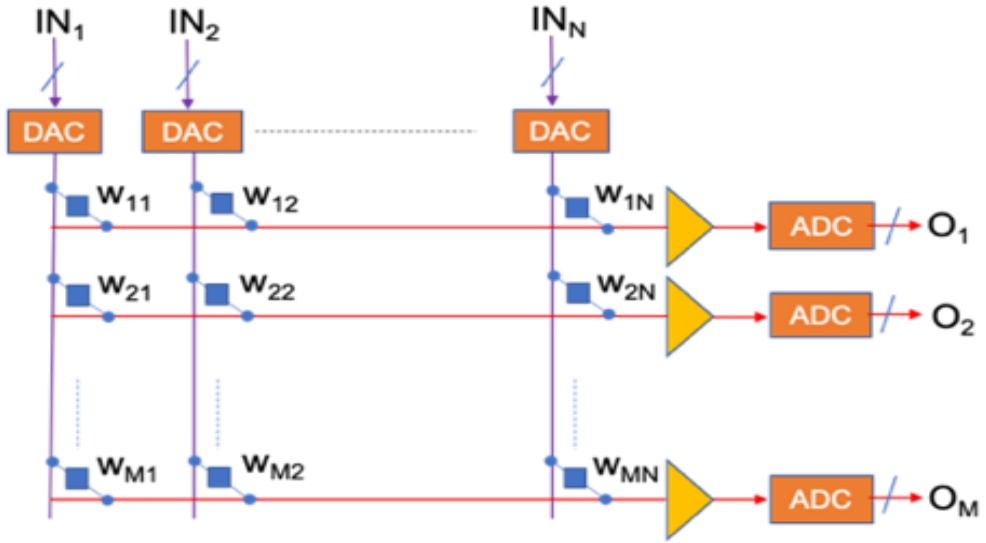
Novel hardware: in-memory computing

[S. Miryala , CPAD21, link]

- Traditional AI-target hardware in FEE including digital processing in ASIC and FPGAs
- New opportunity emerges to perform in-memory computing that is low latency and energy efficient
- Example is Memristor-based crossbar arrays that perform Multiply & Accumulate (MAC) in one cycle

MAC in a neuron

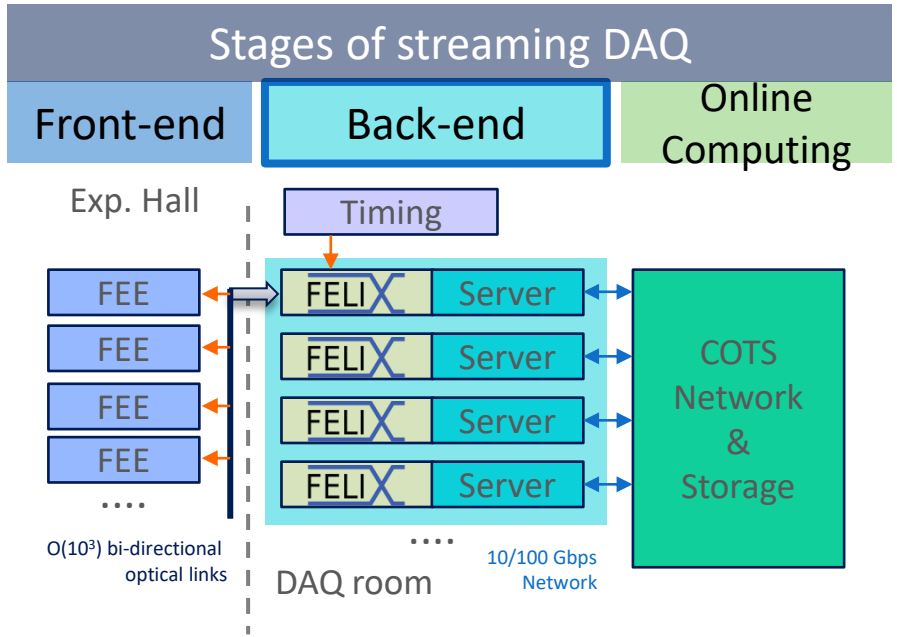




Memristor crossbar array, a Non-Von Neumann architecture for in-memory computing of neural networks

Streaming DAQ stage 2: Readout back-end

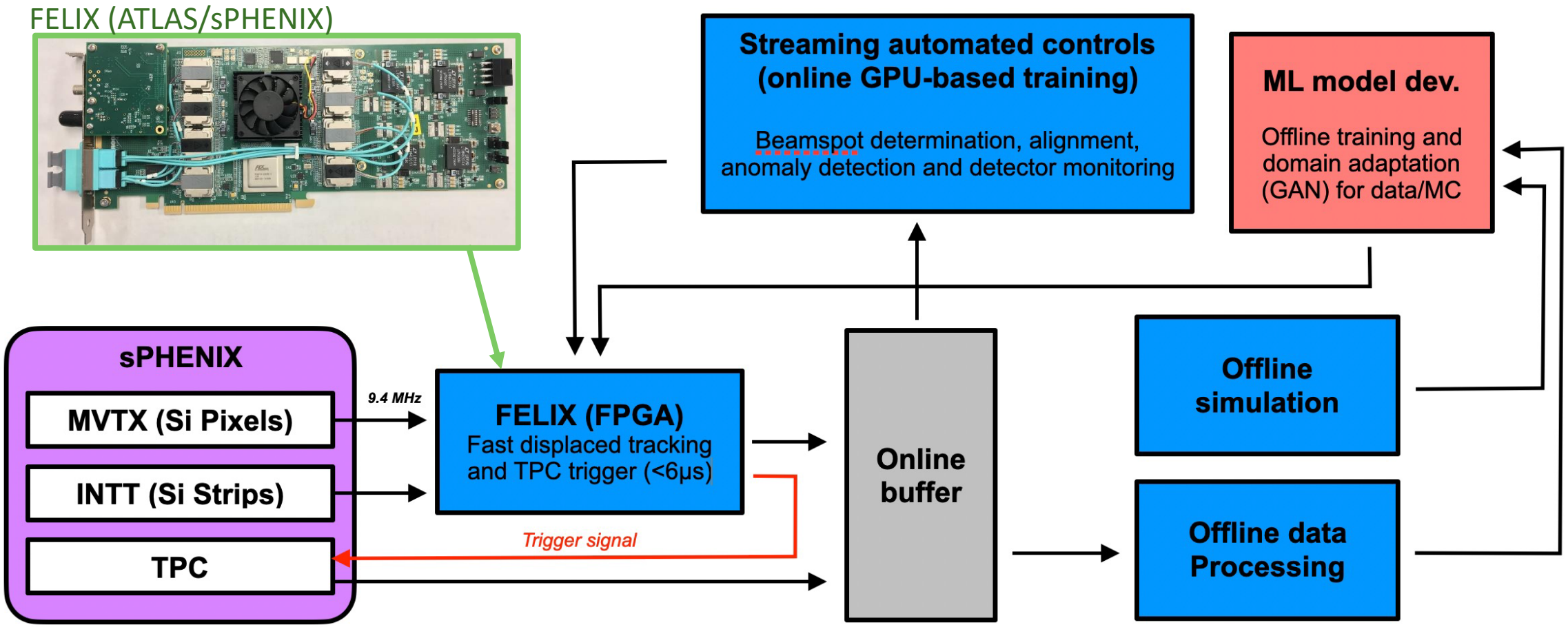
- Perform data aggregation and flow control
  - Common strategy include optical data receiver in large FPGA, routing data to server memory
- AI opportunities:
  - Higher level feature building
  - Selection of interesting time slices, background/noise rejection
  - Two example projects in next slides
- Target hardware: large-scale FPGAs



FPGA based data filter for sPHENIX and EIC

[Y. Corrales Morales, RHIC AUM 22, link]

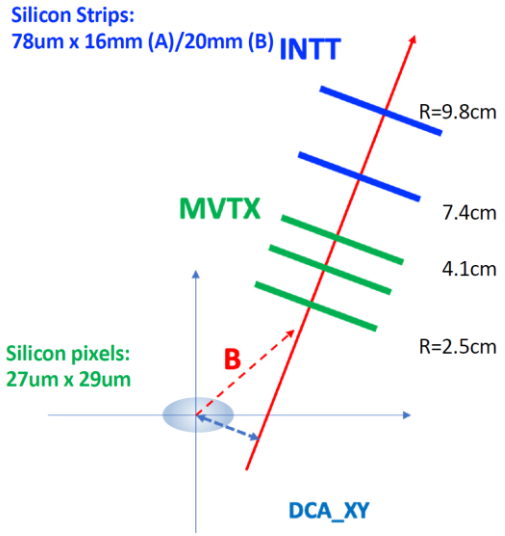
DOE Funded project on streaming readout data reconstruction on FGPA, initiated by LANL, MIT, FNAL and NJIT



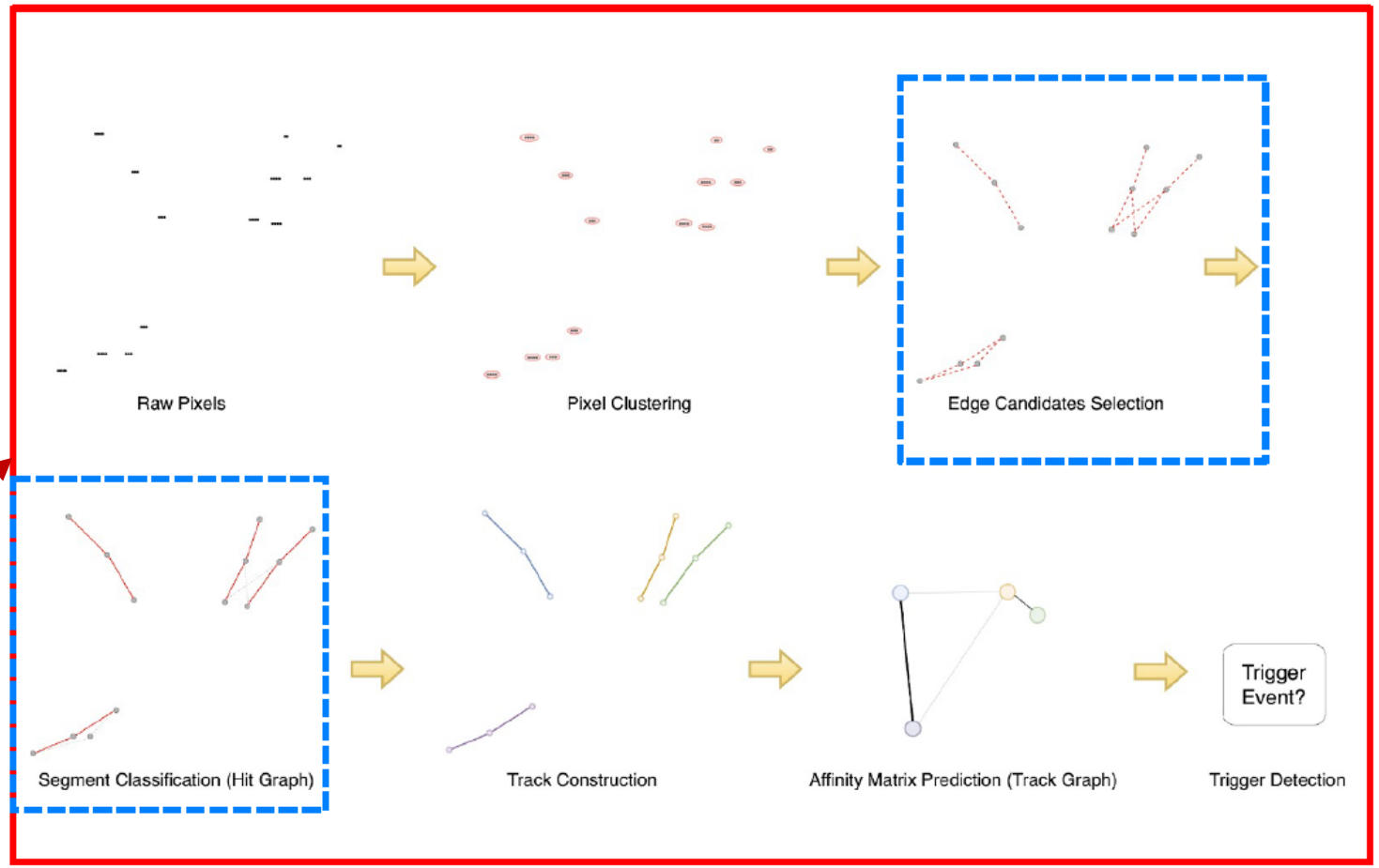
FPGA based data filter for sPHENIX and EIC

[Y. Corrales Morales, RHIC AUM 22, link]

Produce real-time selection of HF events: hit input → clustering → seeding→ trak reco → displaced vertex tagger



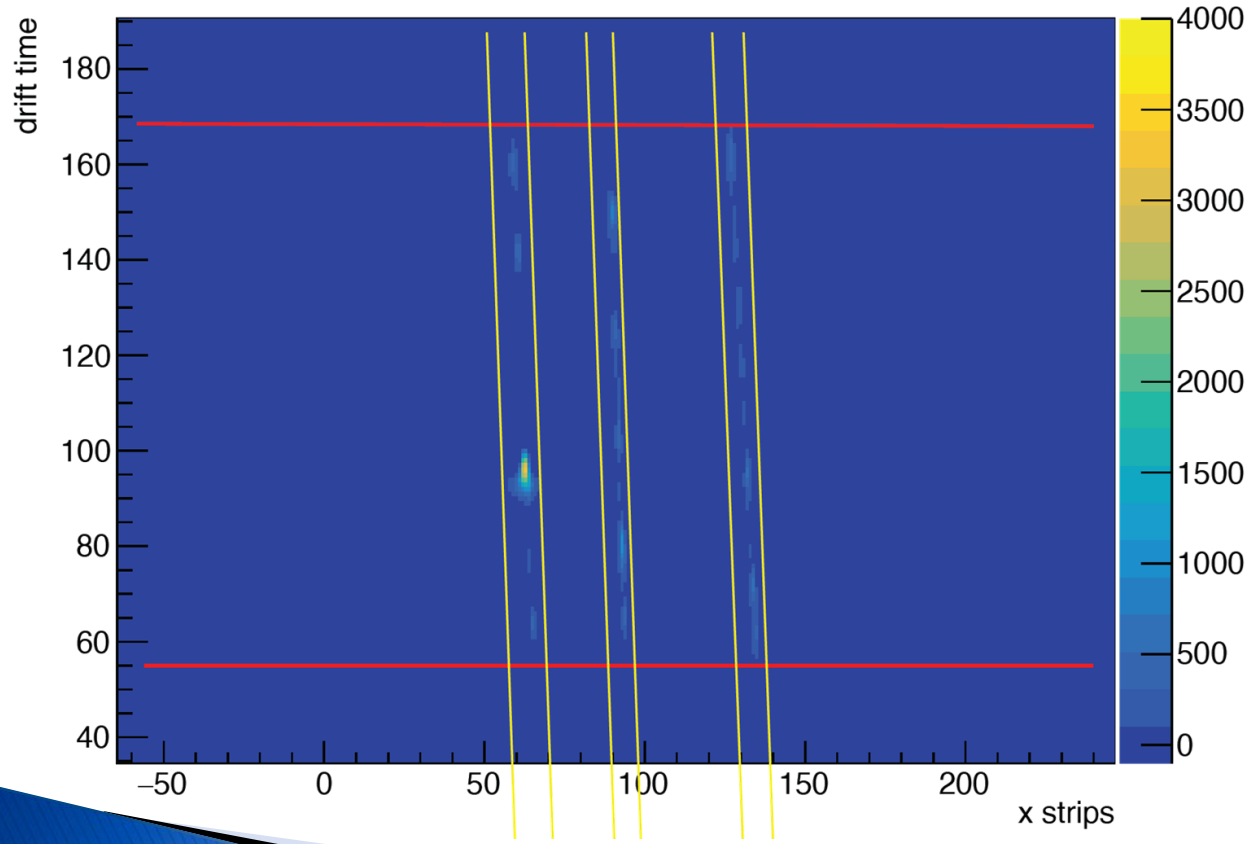
FELIX (ATLAS/sPHENIX)



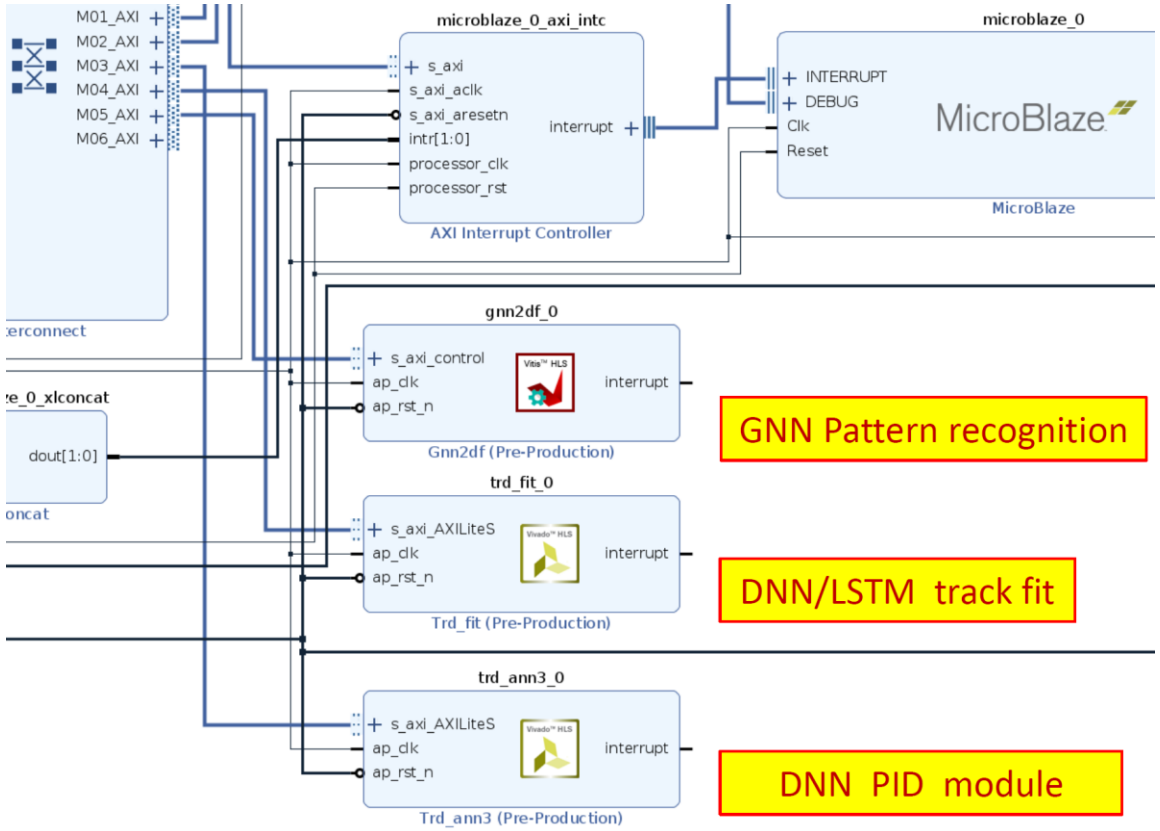
Another example: GEM TRD tracking/PID

[S. Furletov, IEEE RT22, link]

GEM TRD tracks

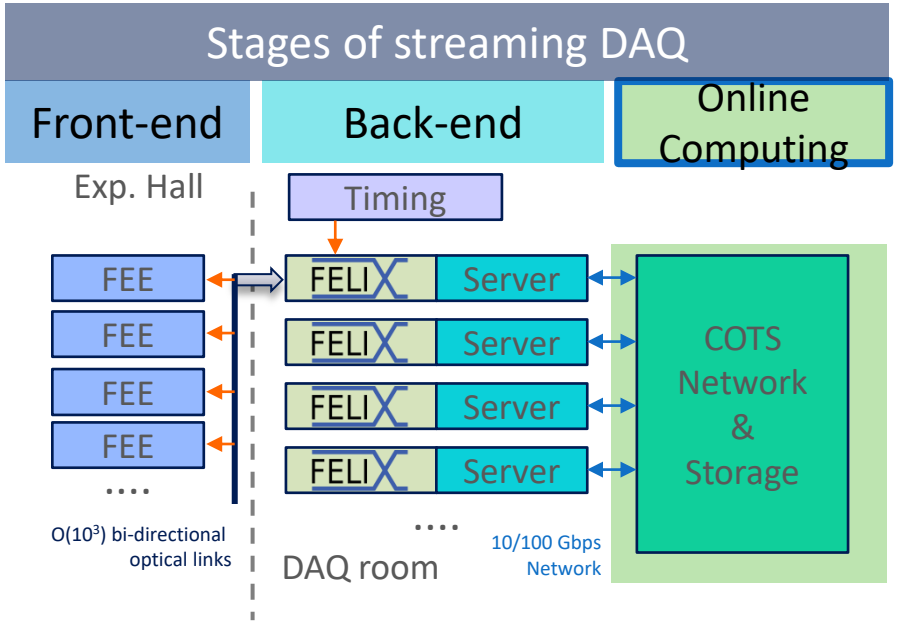


GNN Pattern reco, track fit and PID on FPGA test bench



Streaming DAQ stage 3: Online computing

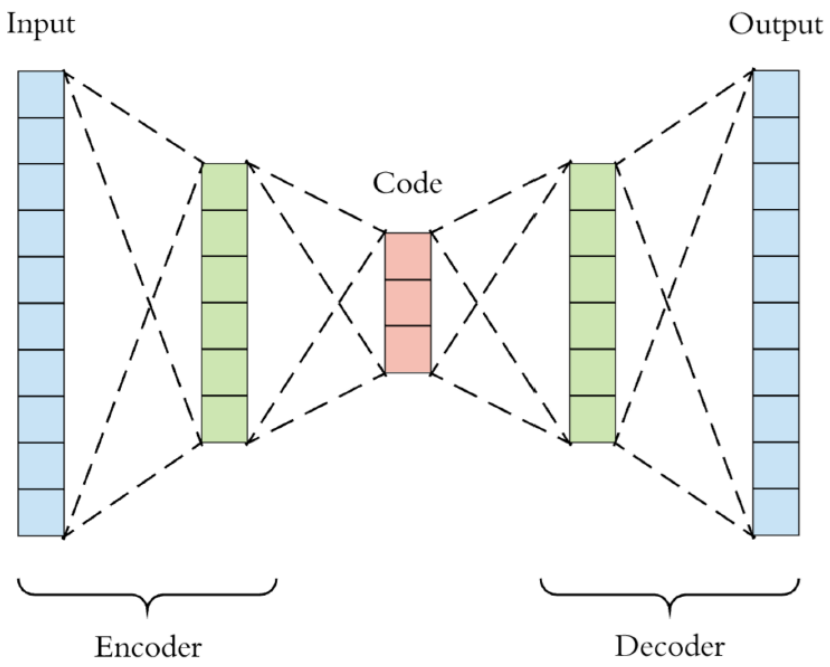
- Online computing is an integral part of streaming DAQ
  - Blending the boundary of online/offline computing
- AI opportunities:
  - Lossy compression
  - Noise and background filtering
  - Higher level reconstruction
- Target hardware:
  - Traditional computing: CPU, GPU
  - Novel AI Accelerators (next slides)



Lossy compression of data, noise filtering

- Auto-encoder (AE) is a natural choice for unsupervised learning for lossy data compression: streaming data reduction

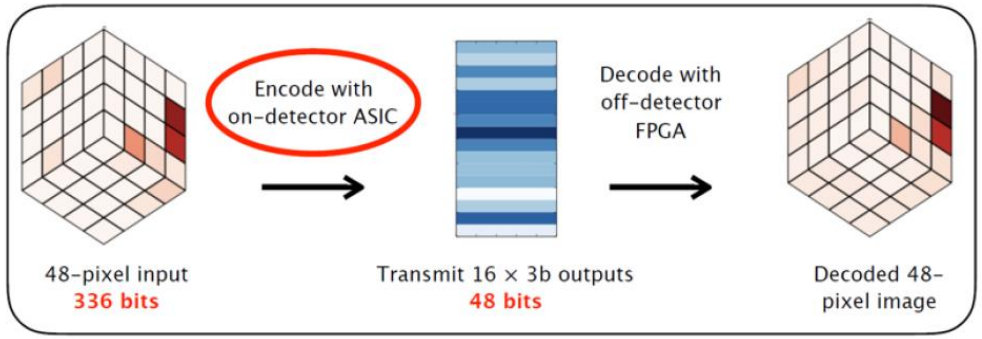
Simple auto-encode neural network



Lossy compression of data, noise filtering

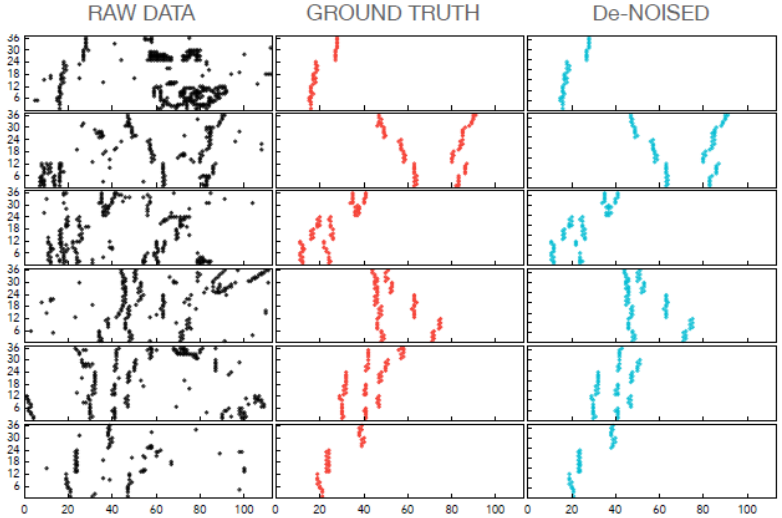
- Auto-encoder (AE) is a natural choice for unsupervised learning for lossy data compression: streaming data reduction
- Same network architecture can be adopted with supervised learning to filter out noise: further data reduction, speed up reconstruction
- See also in CMS HGCal ASIC, CLAS12 tracker offline reco.

CMS HGCal compression ASIC, [10.1109/TNS.2021.3087100]



CLAS12 Drift Chamber offline AE de-noise [link]

See also: talk by Diana McSpadden



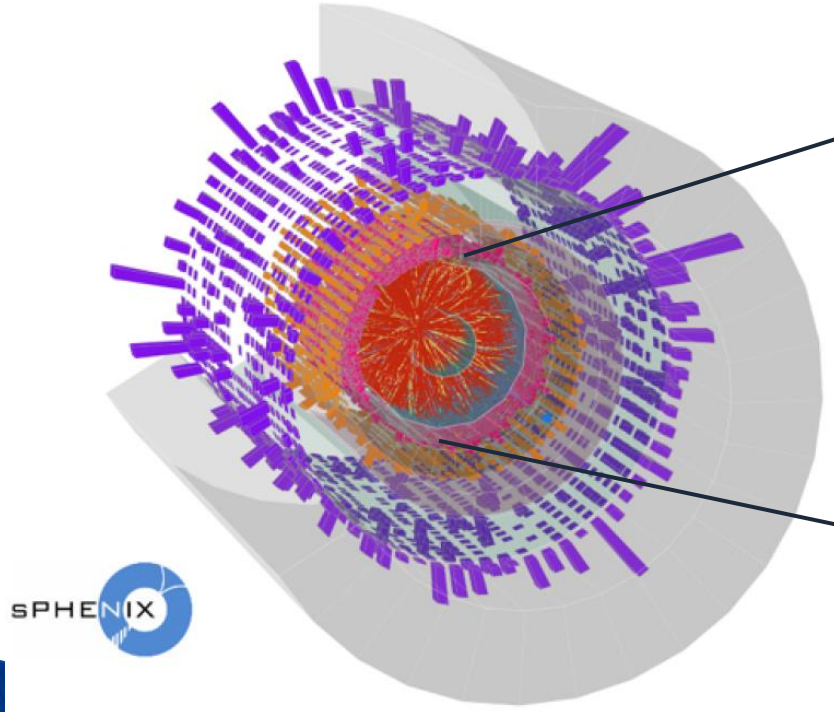
Data of time projection tracker at sPHENIX

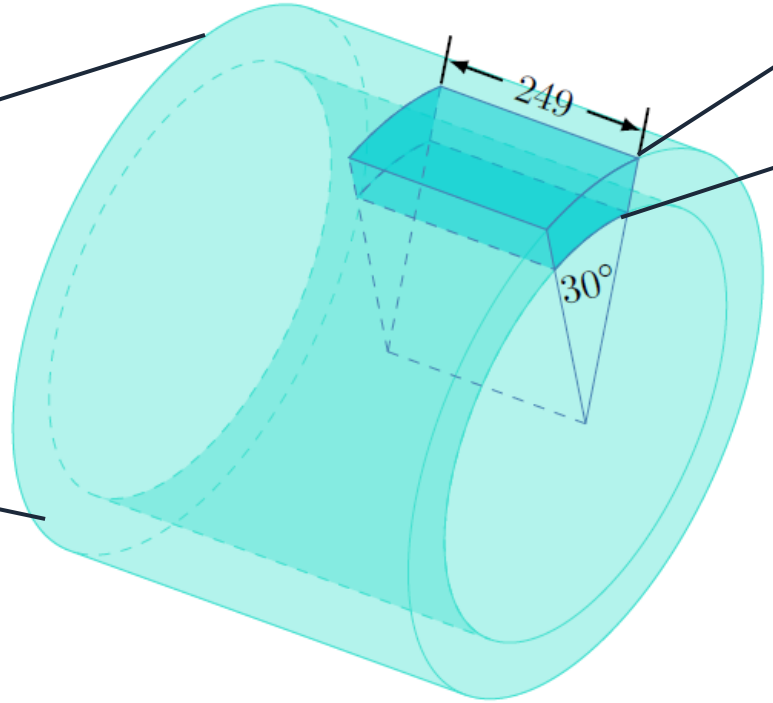
Busiest event in sPHENIX TPC

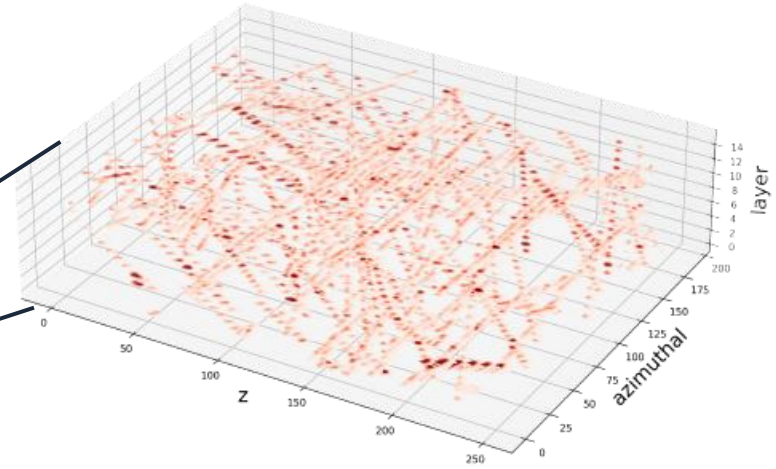
3D X-Y-Time time frame at 50Tbps prior to zero-suppression

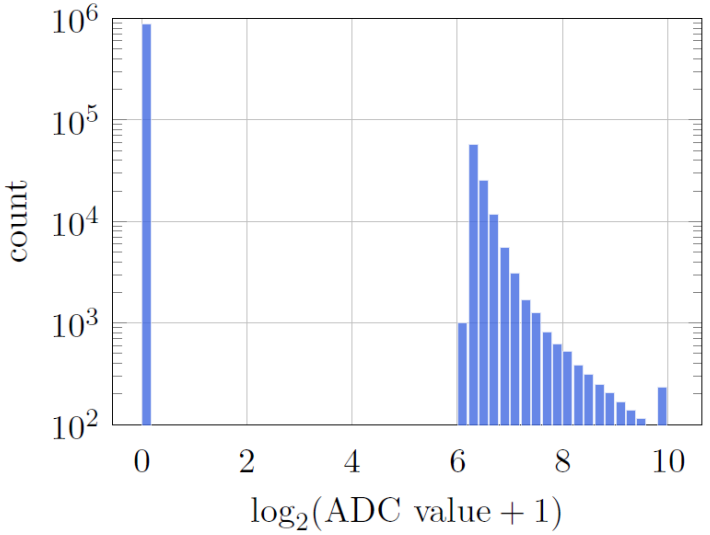
10% central Au + Au collision with 170kHz pile up

Data frame for 1/12 azimuth sector shown here







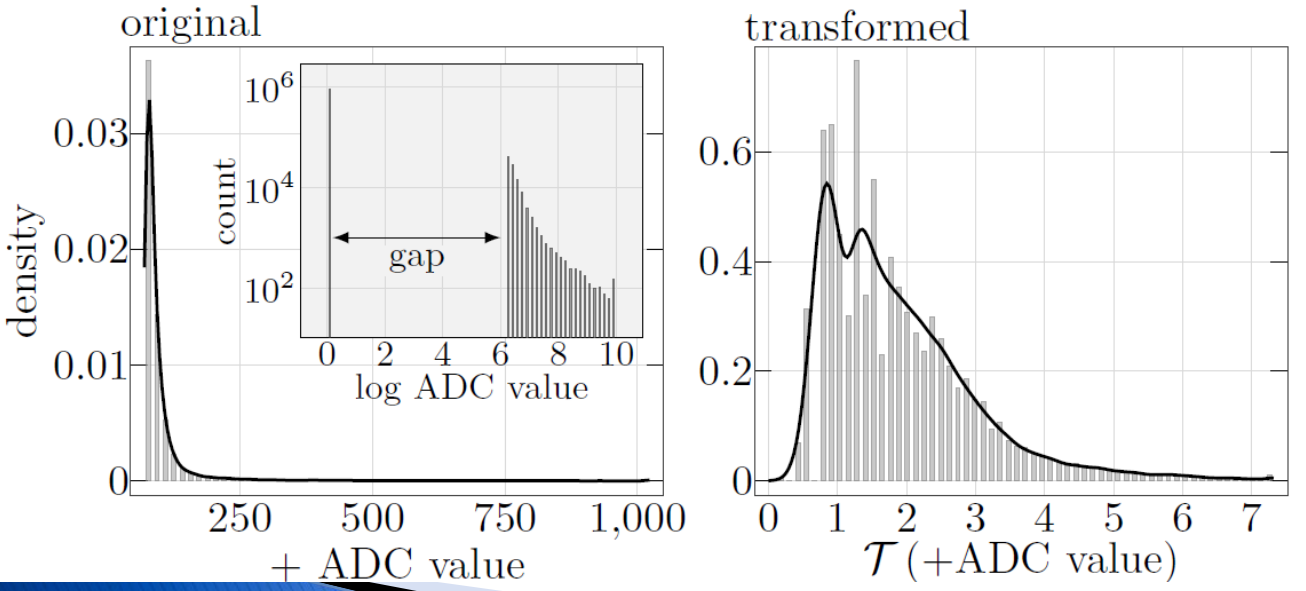


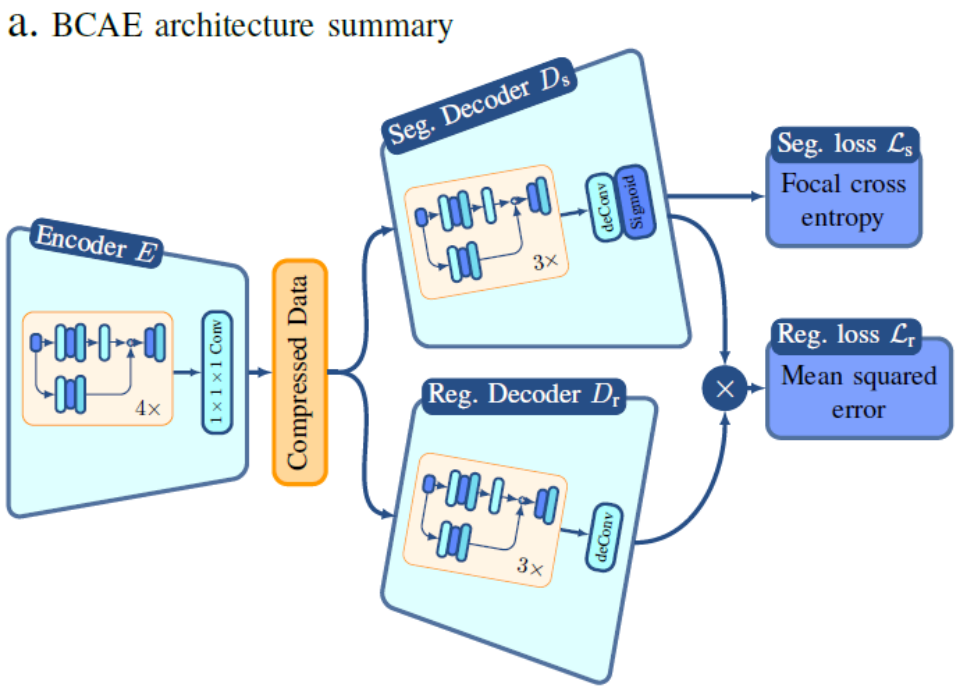
Bicephalous Convolutional Auto-Encoder (BCAE) and input transform [arXiv:2111.05423]

- Input transform: fill in the zero-suppression gap and make ADC distribution much less steep
- Bicephalous decoder: +classification decoder to note the zero-suppressed ADC voxels and +noise voxels in TPC

Input transform: $\mathcal{T}(x) = \log(x - 64)/6, \quad x > 64$

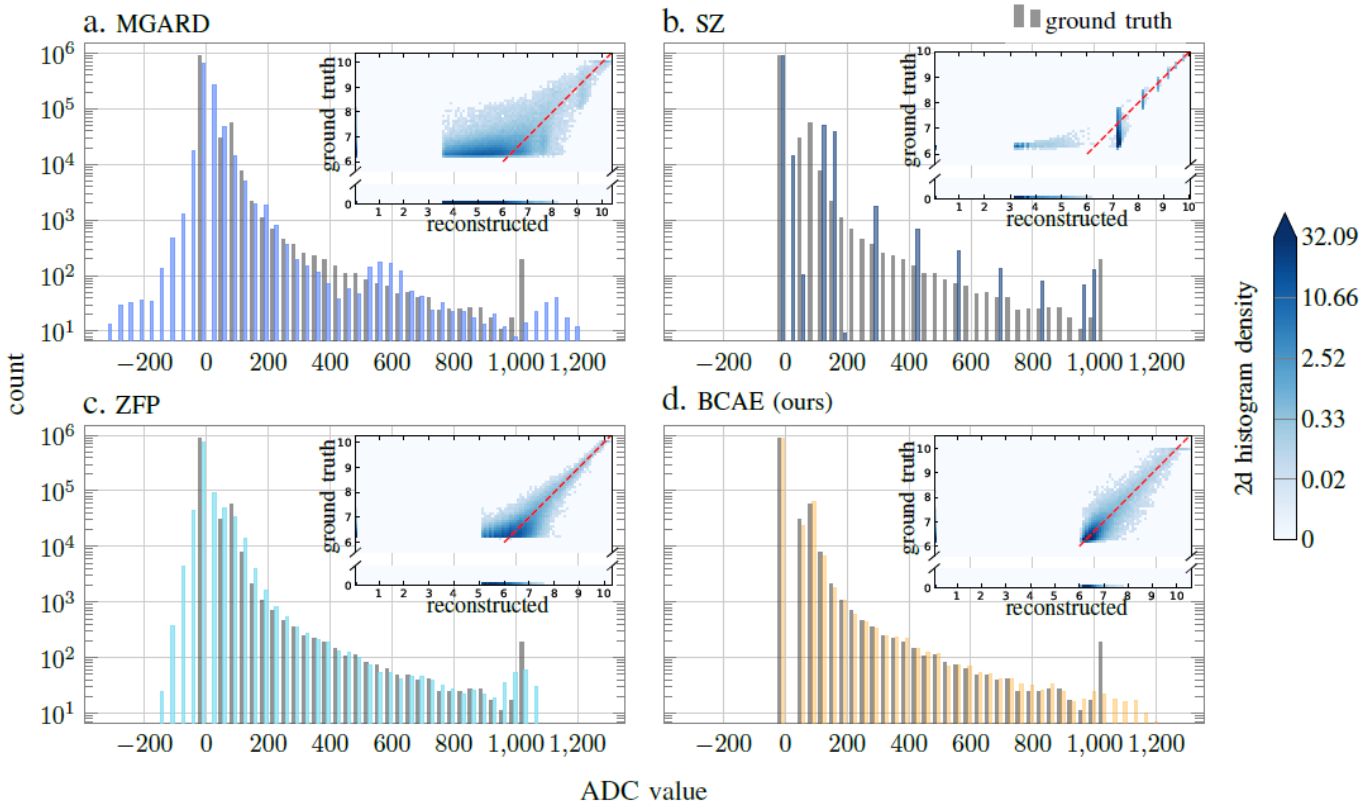
Inverse transform: $\mathcal{T}^{-1}(y) = 64 + \exp(6y), \quad x \in \mathbb{R}$

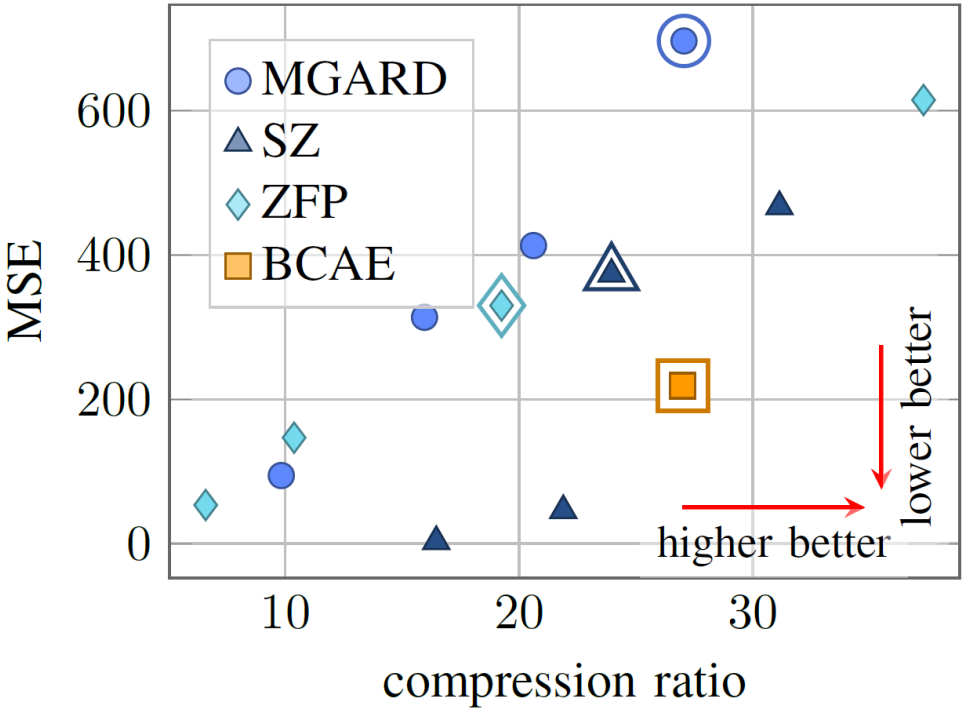




Comparison with existing algorithm

[arXiv:2111.05423]

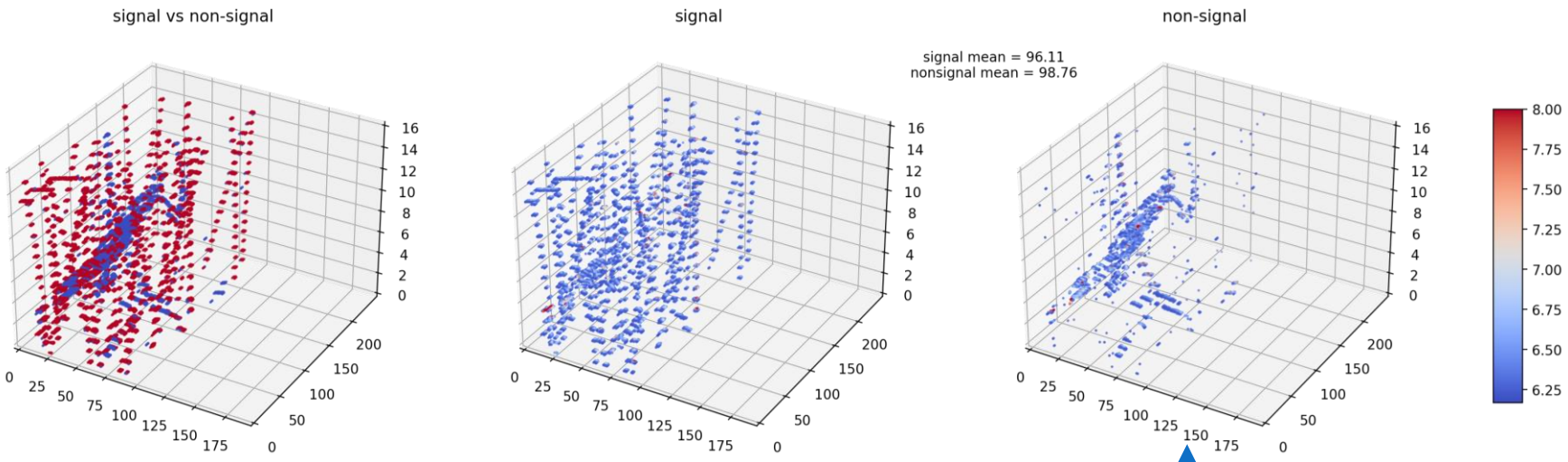


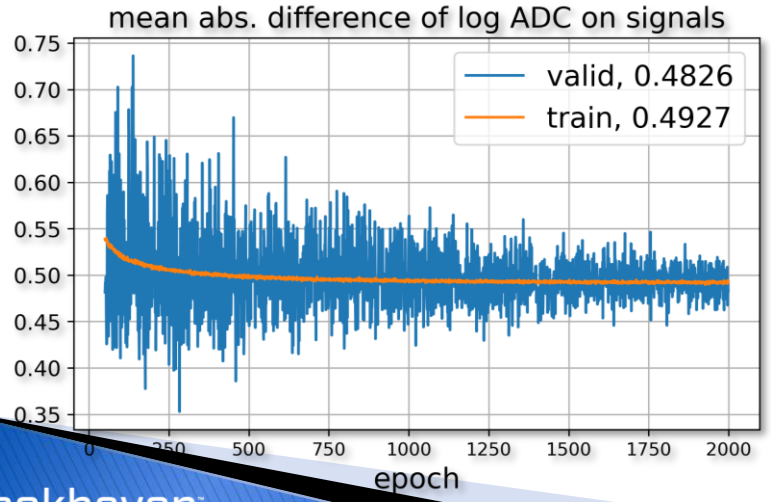


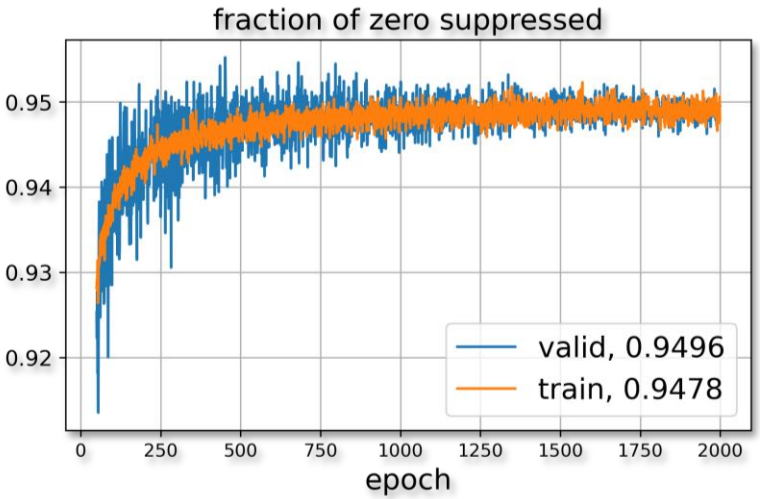
BCAE Compressor with noise filtering

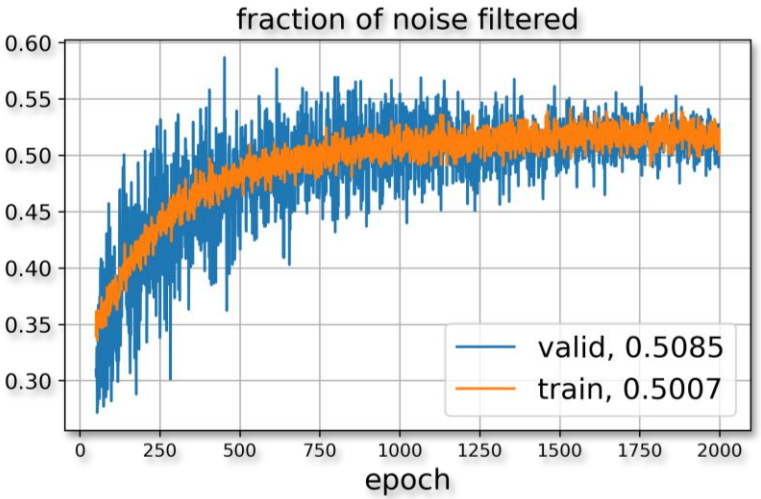
[Y. Huang, IEEE RT22, link]

sPHENIX simulation
3 MHz $p + p$ TPC streaming data
BCAE with compression ratio 204:1 and 95% signal retention (recall)



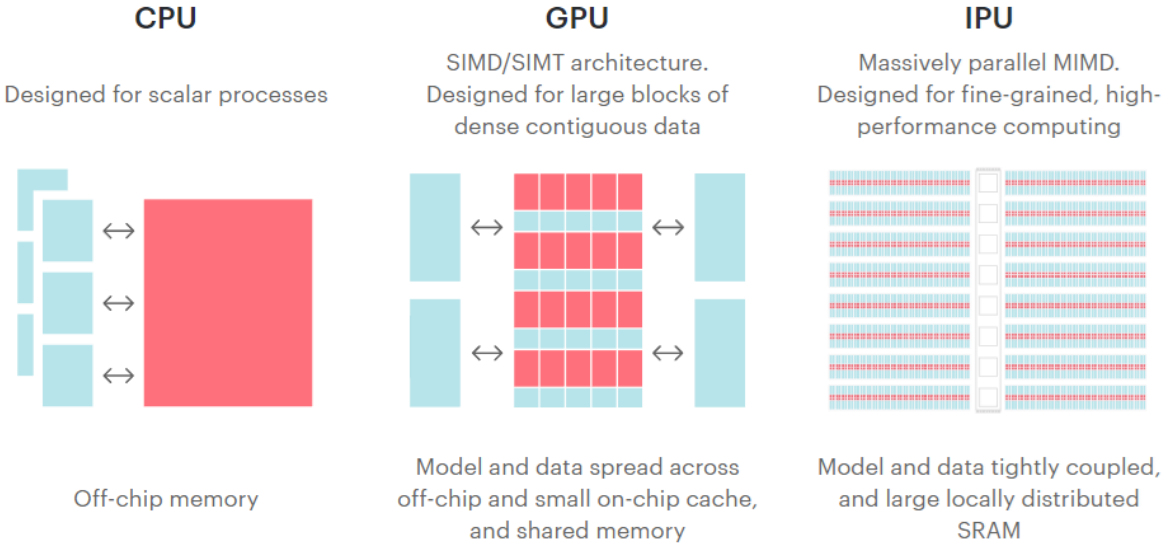




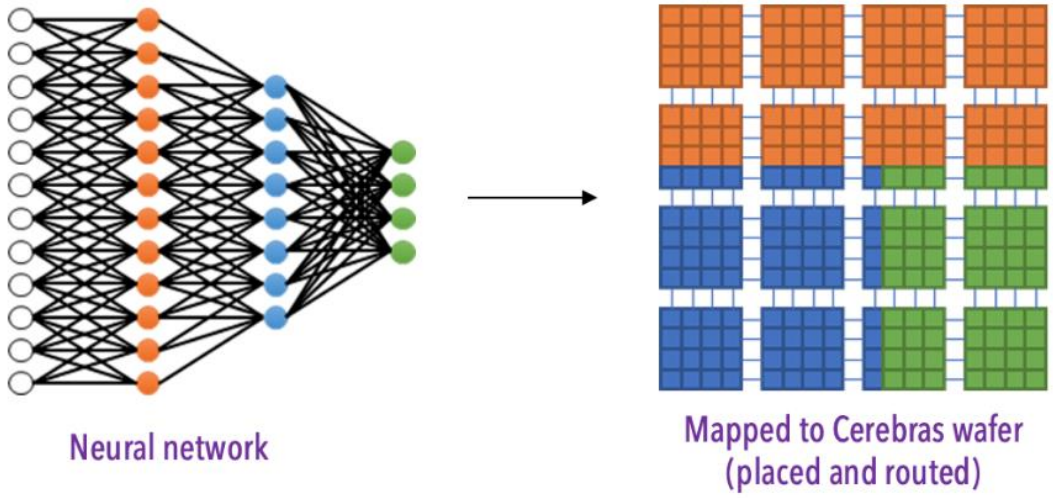


Novel AI Accelerators for streaming DAQ

- A new family of AI chips is emerging with non-von Neumann Architectures
  - Designed for NN computing
  - Massive on-chip activation/weight storage on sRAM
  - Good integration with popular AI tools
  - Energy efficient and high throughput
- Significant throughput gain with testing of BCAE on Graphcore IPUs, a Dataflow Architectures processor for AI application



[GraphCore Web, link]



[Cerebras Compiler Docs, link]

Summary

- Streaming readout is a paradigm shift adopted by many modern experiments, driven by NP physics of diverse event topologies and stringent bias control
- Requiring large factors of data reduction computationally and at high throughput
- Driving the need of AI-based algorithms and platforms:
  - Feature extraction, compression, signal selection/background noise removal, reconstruction
  - Utilizing ASIC, FPGA, and emerging novel AI accelerators