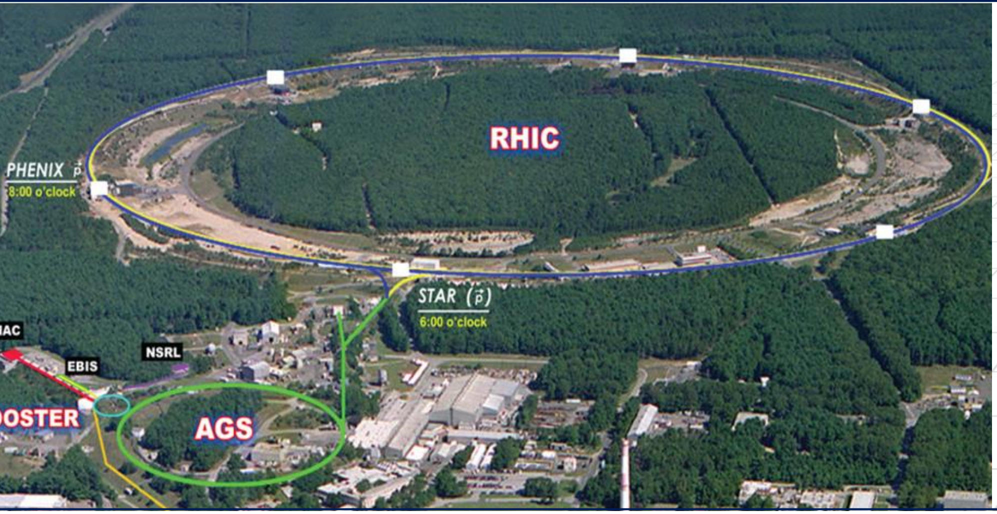
FastML: Application and Opportunities in Nuclear Physics (NP)

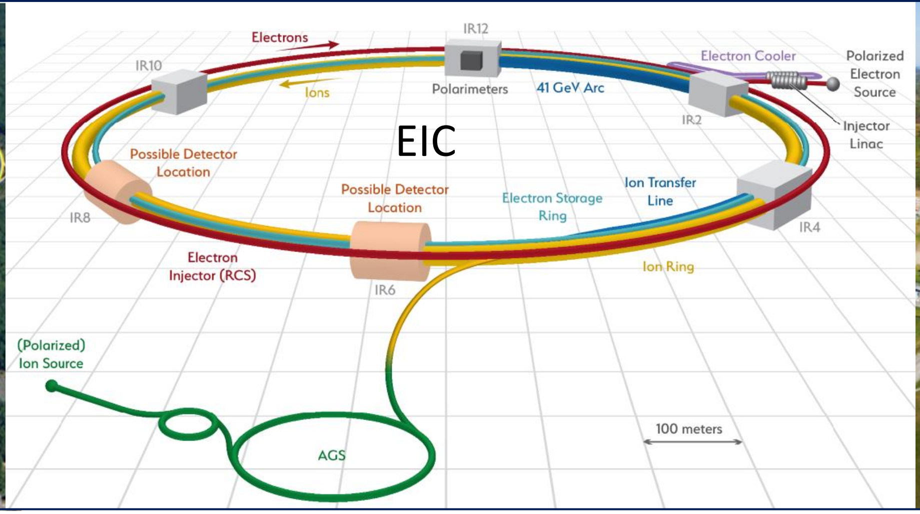
Outline: ● Opportunities in NP Exp. Faculties ● Sample applications of Realtime AI ● Summary

Jin Huang

Brookhaven National Lab

Nuclear Physics Facilities in focus for this talk





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

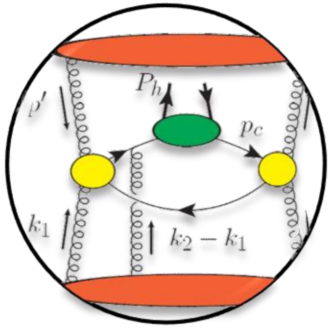
- This talk is an in-complete review of the field, see also experiments including at LHC (LHCb, ALICE, AMBER), at FAIR (CBM)
- FastML application with strong connection to the evolution to Streaming DAQ for next generation NP experiment
- See also Streaming Readout Workshop series [link], AI4EIC series [link]

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} → 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**

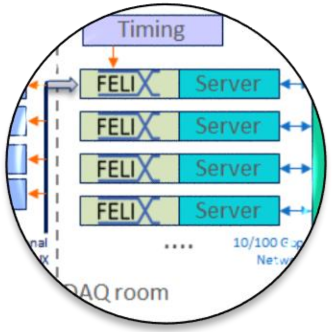
- Signal data rate is moderate → possible to streaming recording all collision signal
- But events are precious and have diverse topology → hard to trigger on all process
- Background and systematic control is crucial → avoiding a trigger bias; reliable data reduction

Streaming DAQ and real-time AI:
A new and paradigm shift for experiments in next NP LRP



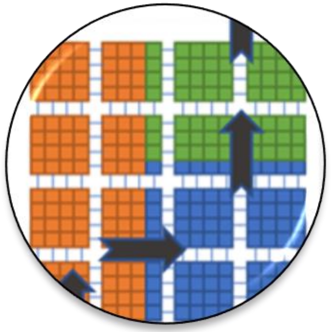
NP Physics

- Diverse topology
- Stringent sys. Ctrl
- Max data preservation



Streaming DAQ

- New physic capability accessible only via streaming DAQ
- Example: adopted for sPHENIX and EIC
- Require data reduction computationally

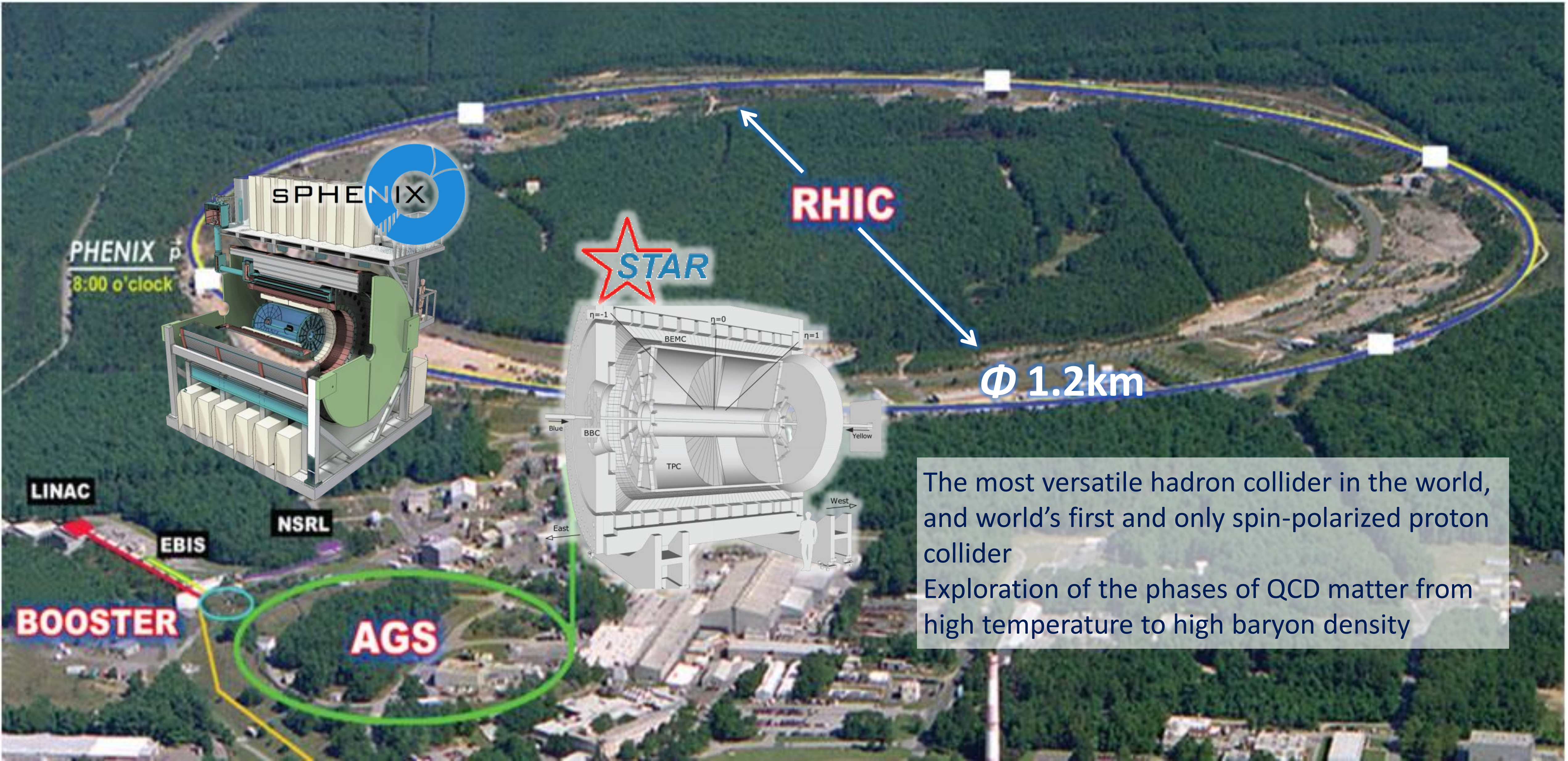


Opportunities for FastML

- Specialized AI algorithm for reliable and high-performance data reduction
- Novel hardware emerging for high-throughput AI computing

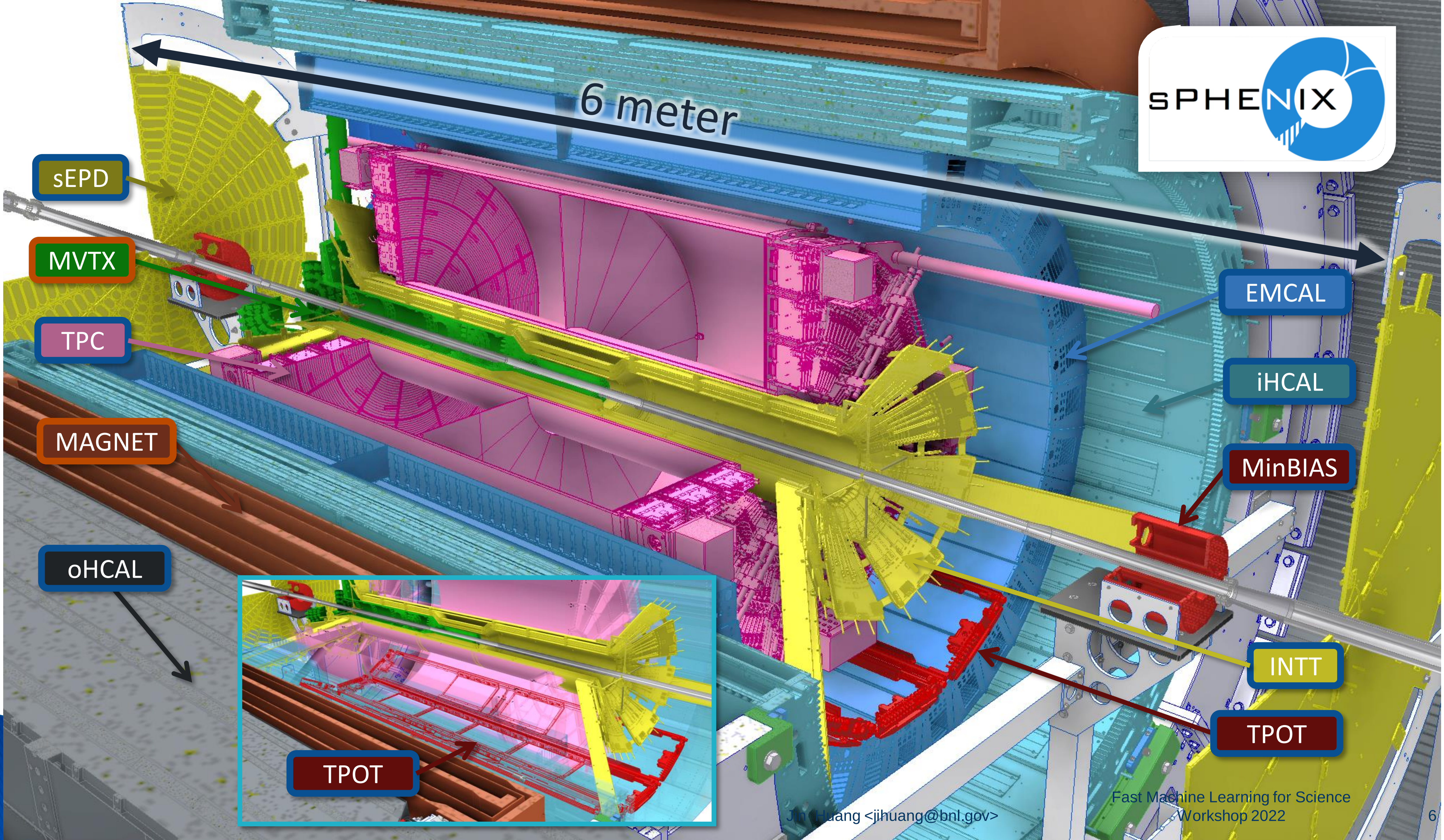
Physics need → Streaming DAQ → Opportunity for real-time AI → Enhanced physics program

Relativistic Heavy Ion Collider (RHIC) in 2023+



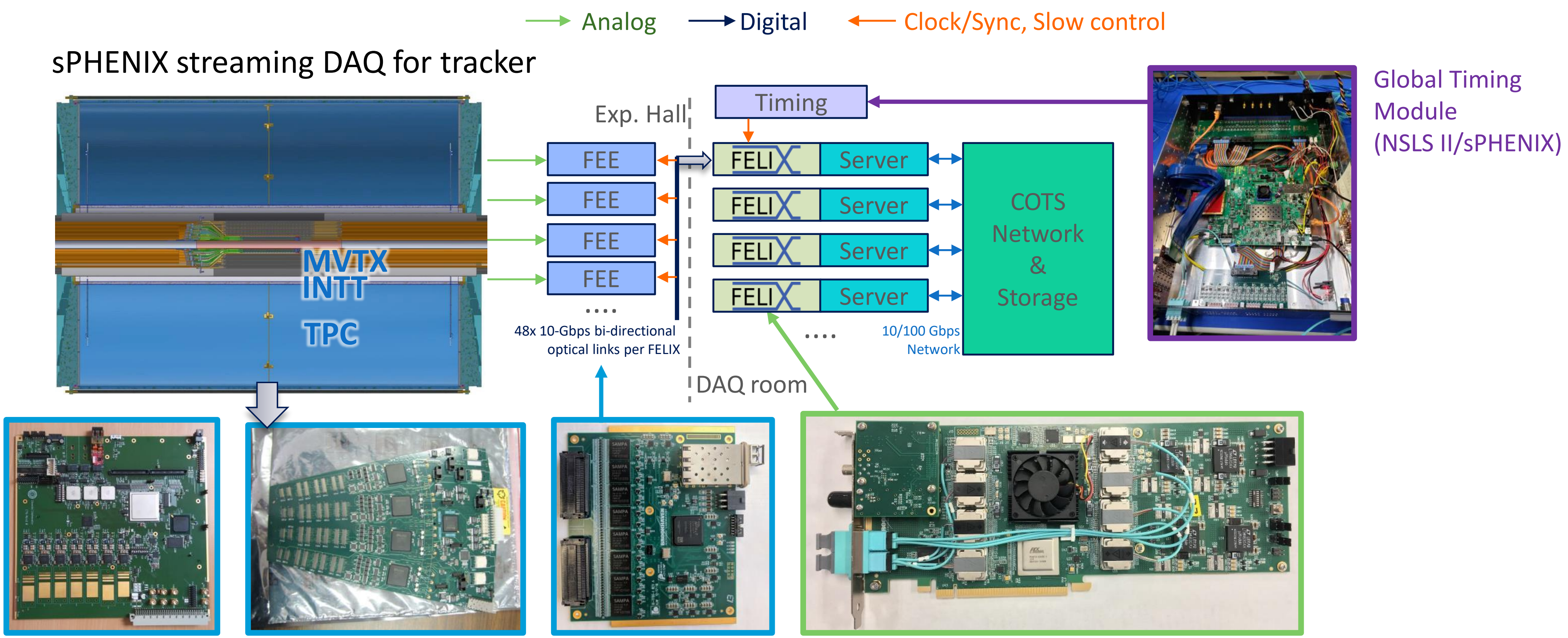
The most versatile hadron collider in the world, and world's first and only spin-polarized proton collider

Exploration of the phases of QCD matter from high temperature to high baryon density


sPHENIX
6 meter
sEPD
MVTX
TPC
MAGNET
oHCAL
TPOT
EMCAL
iHCAL
MinBIAS
INTT
TPOT


sPHENIX installation on going in RHIC IR8
Data taking start in spring 2023!

Streaming readout electronics for sPHENIX tracker



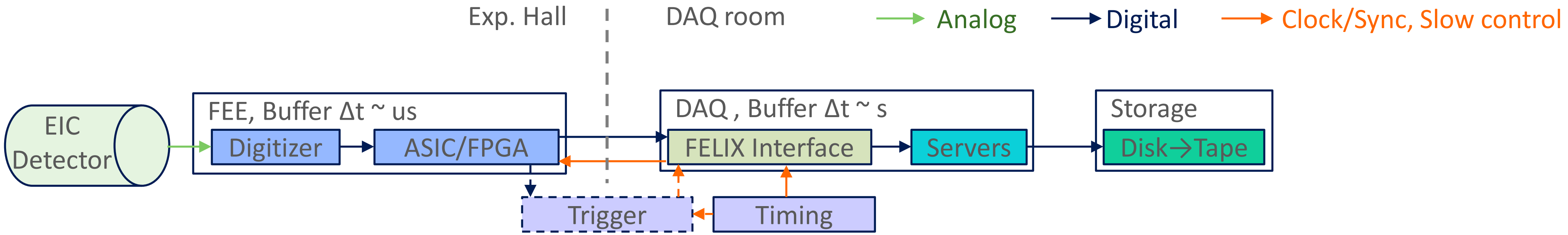
MVTX RU, 200M ch INTT ROC, 400k ch TPC FEE, 160k ch BNL-712 / FELIX v2 x38 (ATLAS/sPHENIX/CBM)

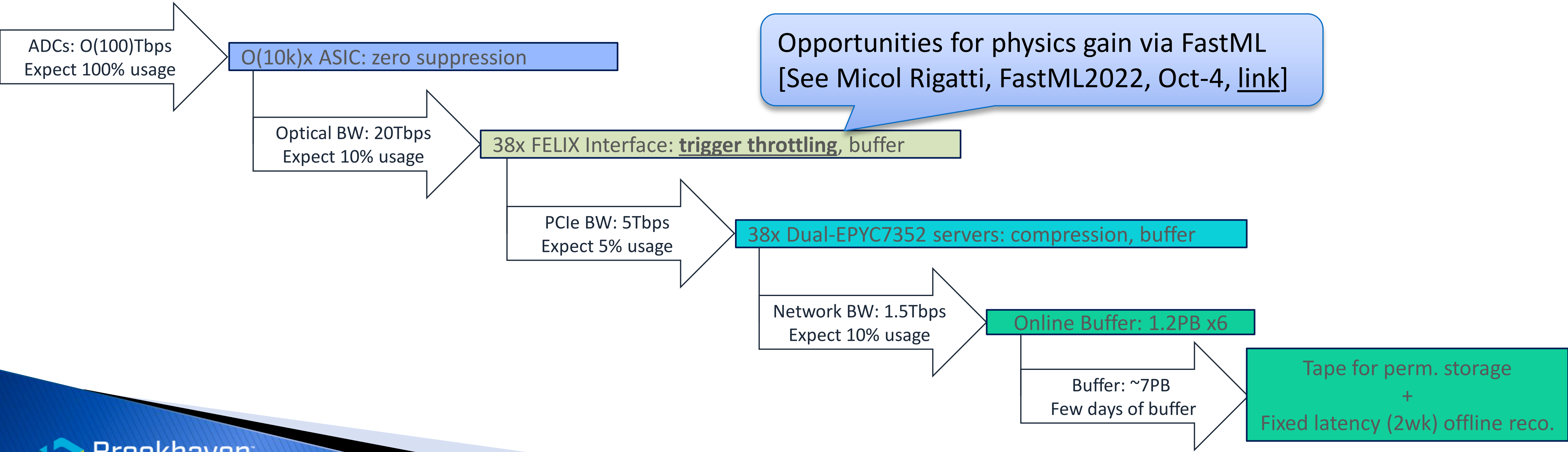
ALPIDE (ALICE/sPHENIX), FPHX (PHENIX) SAMPAv5 (ALICE/sPHENIX)

FELIX Ref: 10.1109/tim.2019.2947972

Similar role as PICe40 in LHCb / ALICE

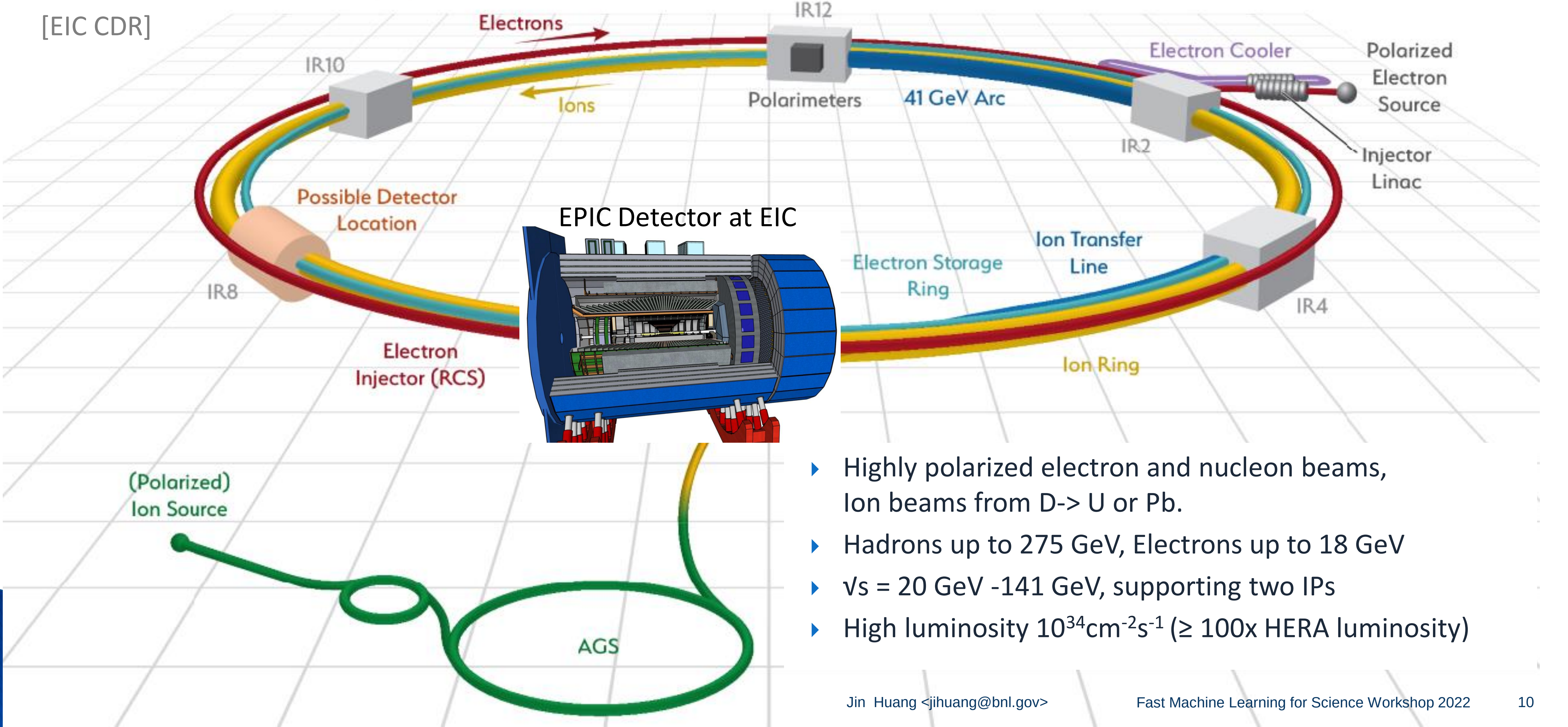
sPHENIX Streaming data flow





RHIC transition to the Electron Ion Collider (EIC)
CD-1 Approval in 2021, Science Phase in 2030+

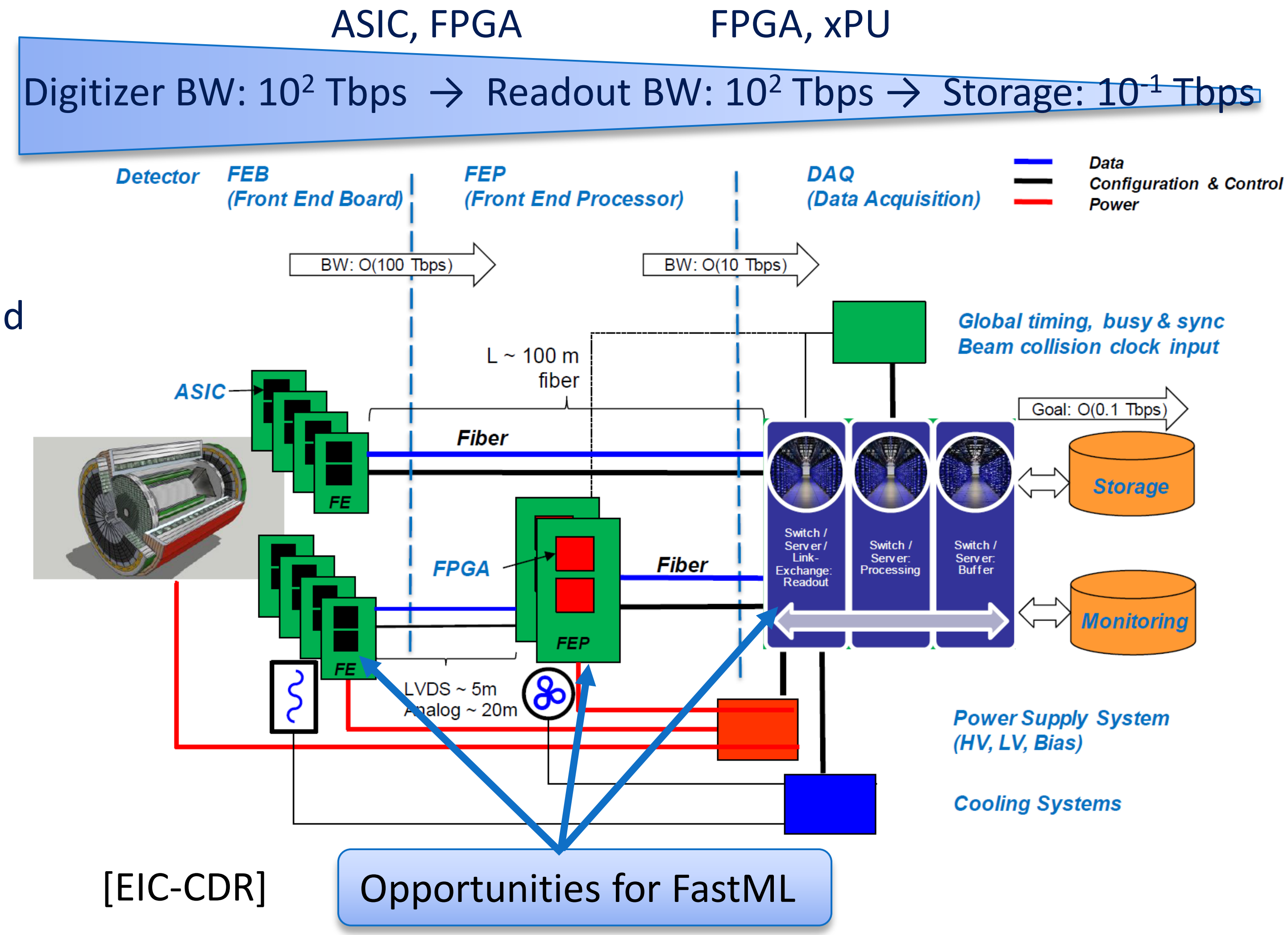
[EIC CDR]



- Highly polarized electron and nucleon beams, Ion beams from D-> U or Pb.
- Hadrons up to 275 GeV, Electrons up to 18 GeV
- $\sqrt{s}$ = 20 GeV -141 GeV, supporting two IPs
- High luminosity $10^{34}cm^{-2}s^{-1}$ (≥ 100x HERA luminosity)

Streaming readout data flow: EIC

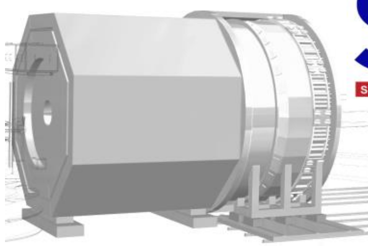
- **EIC streaming DAQ**
  - → Triggerless readout front-end (buffer length : μs)
  - → DAQ interface to commodity computing (FELIX-type interface as the candidate) Background filter if excessive background rate
  - → Disk/tape storage of streaming time-framed zero-suppressed raw data (buffer length :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



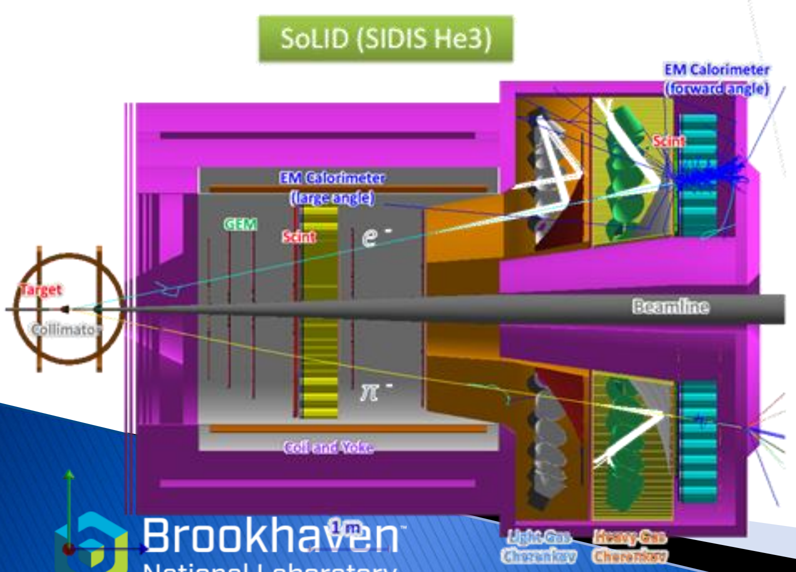
[EIC-CDR]

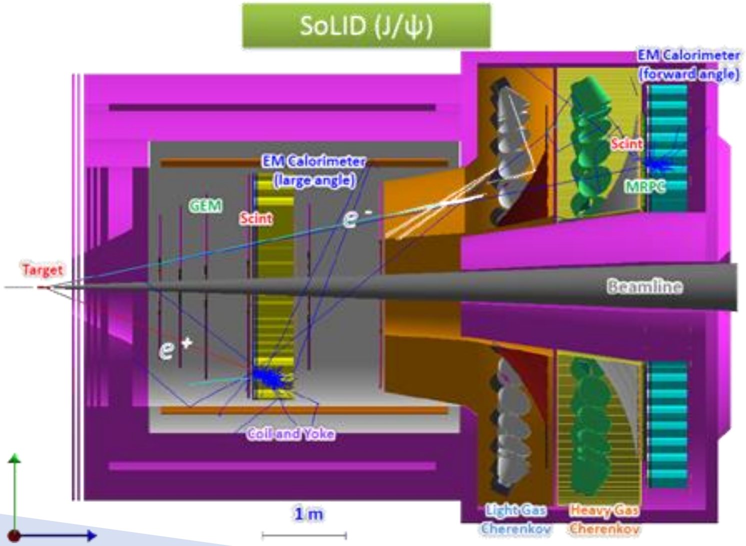
SoLID @ Jlab (Solenoidal Large Intencity Device)

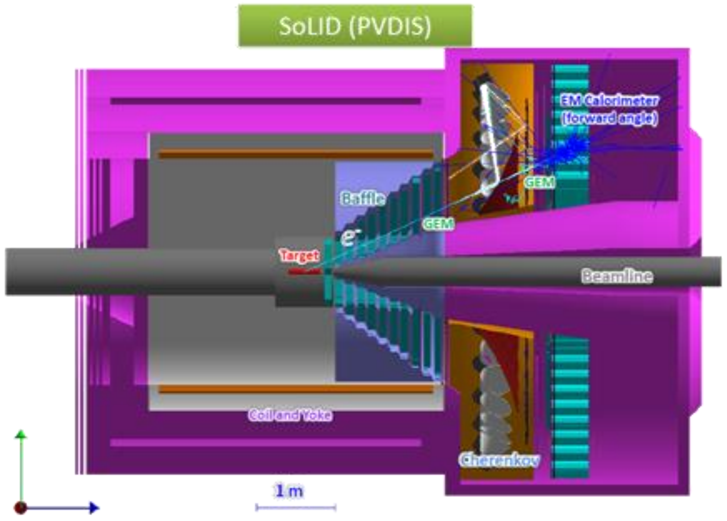
[Zein-Eddine Meziani, 2022 QCD Town hall, link]



- SoLID is proposed to fully utilize the intensity frontier at JLab
  - 10^{37}-10^{39} /cm^2/s + large acceptance fixed target experiment with electron beam
  - DOE Science Review in 2021
- Opportunities for FastML:
  - Clustering and particle ID on waveform digitizer/FPGA readout pipeline
  - Tracking based trigger / data filtering

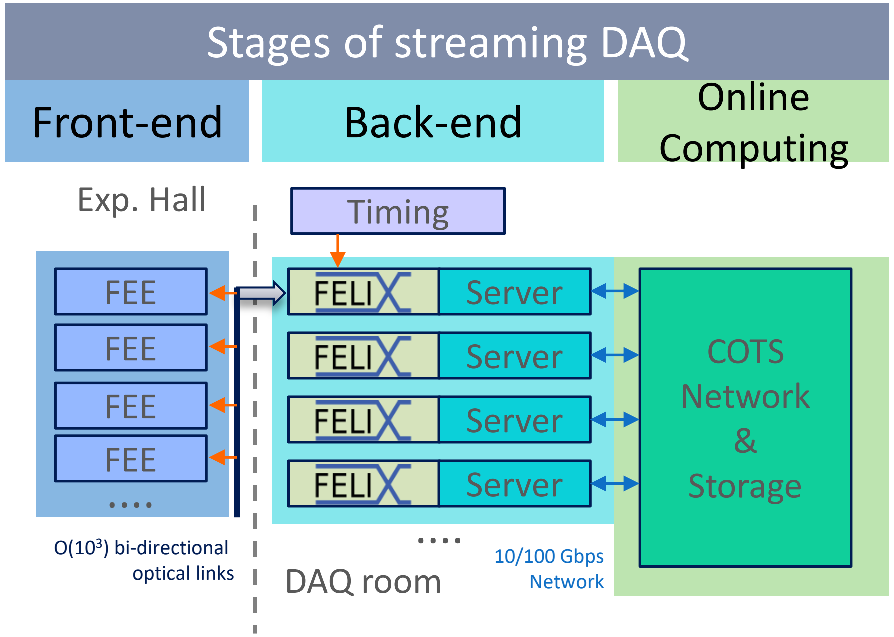






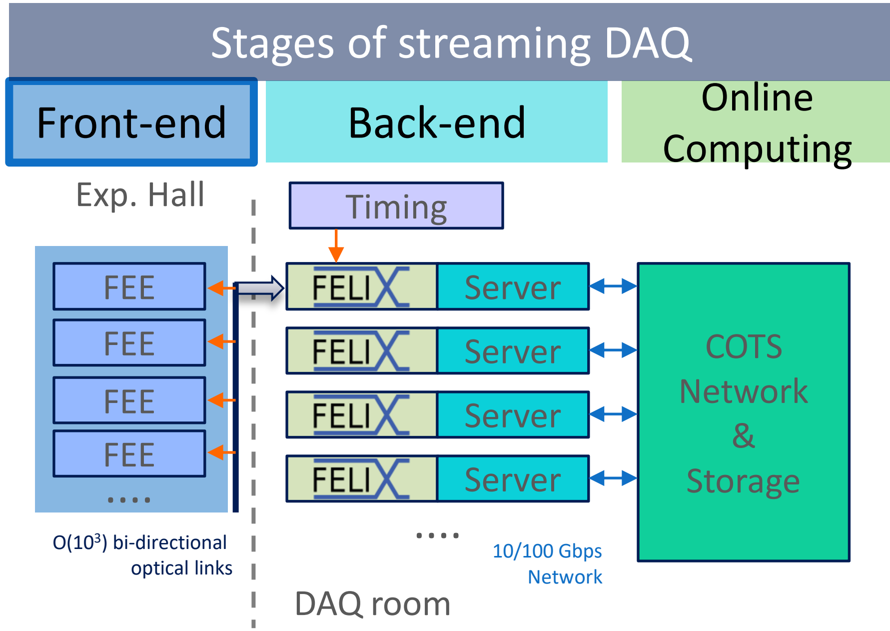
AI in streaming readout DAQ

- 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
    - Could use "traditional" computing
    - Not focus of this talk, nonetheless important for NP experiments



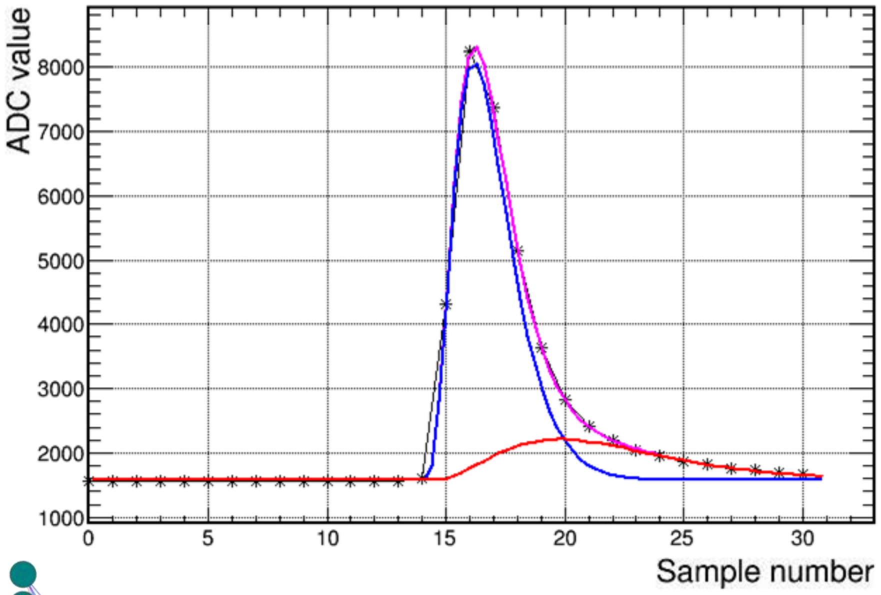
Streaming DAQ stage 1: Front-end electronics

- Perform digitization (ADC, TDC, pixel readout)
  - Common data reduction strategy to immediately apply zero-suppression
- FastML 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



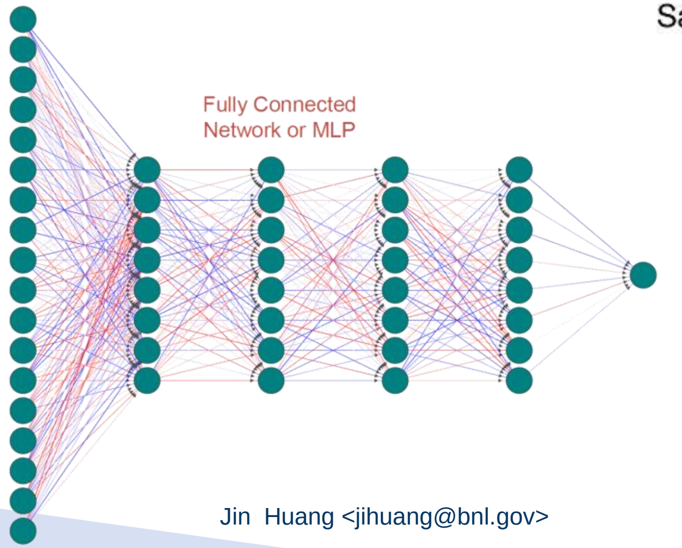
ADC time series → feature of amplitude and time

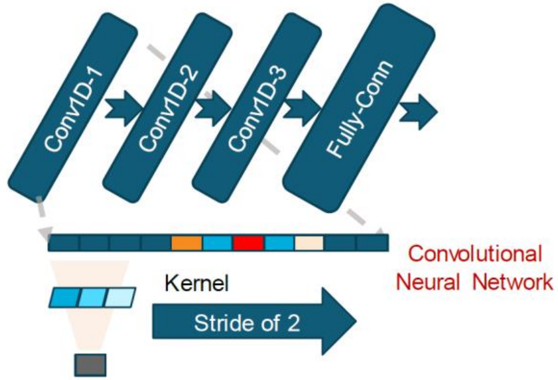
- Wave form digitizer is popular, output data in ADC time series
- In the front-end, NN can be used to extra features such as amplitude and time of arrival
- Fit limited resource in FEE FPGA or ASIC: Emphasizes on quantized-aware training training and pruning



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

[arXiv:2204.13223]

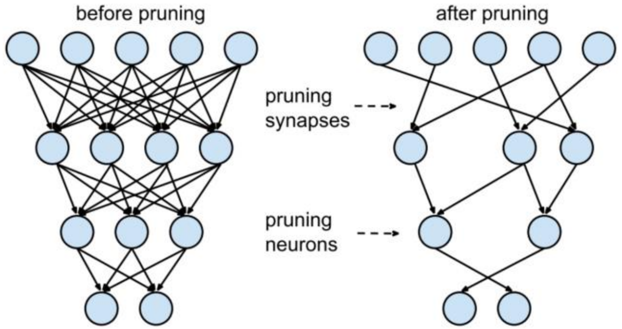


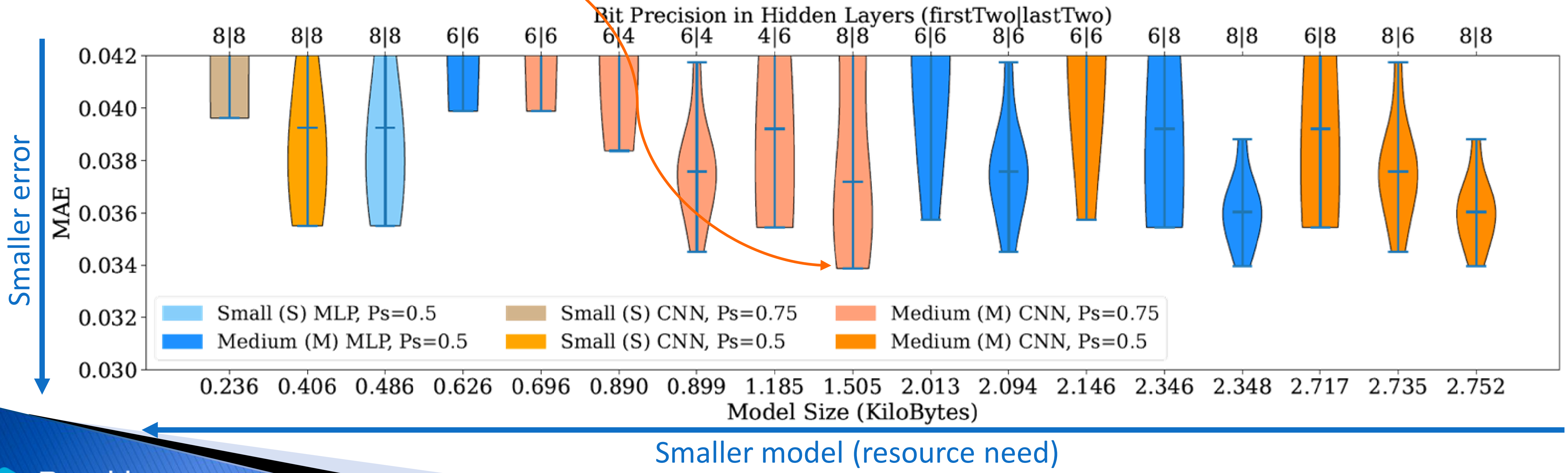


Pruning + Variable Bit Quantization-aware Training

[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



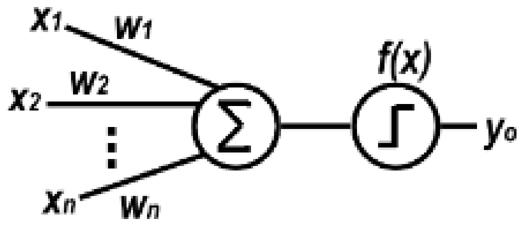


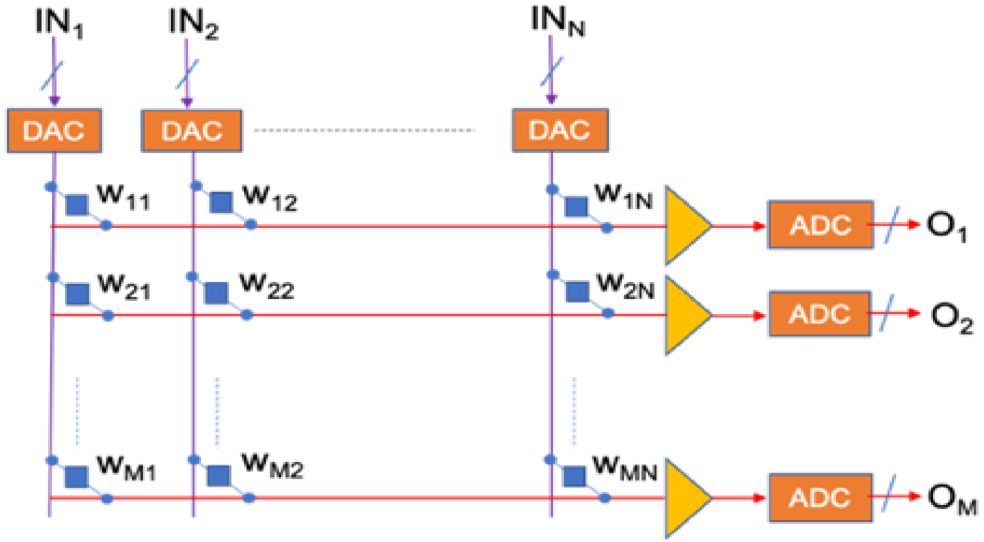
Novel hardware: in-memory computing

[S. Miryala , CPAD21, link]

- One viable 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

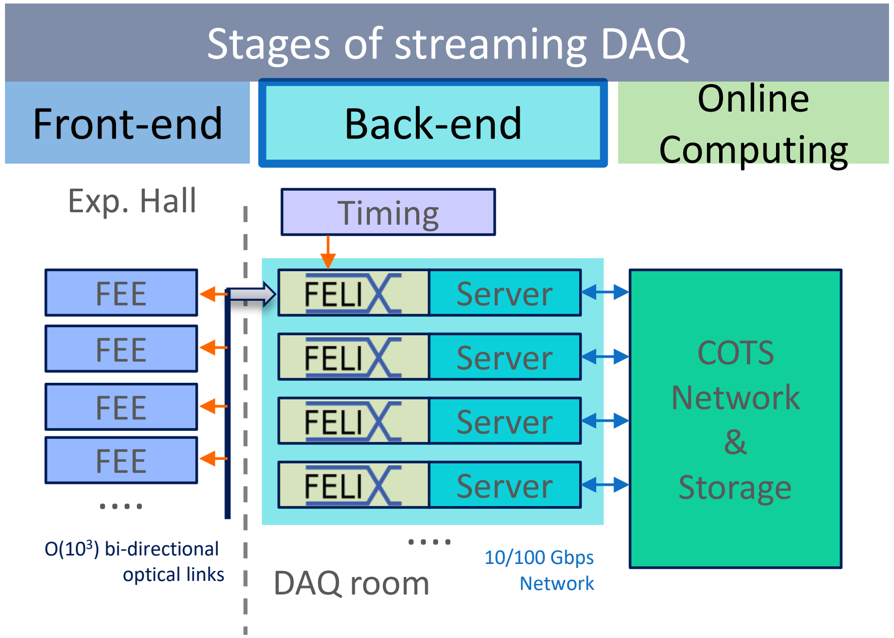




Memristor crossbar array, a Non-Von Neumann architecture for in-memory computing of neural networks

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
- FastML opportunities:
  - Higher level feature building
  - Selection of interesting time slices, background/noise rejection
  - Two example projects in next slides
- Target hardware: large-scale FPGAs

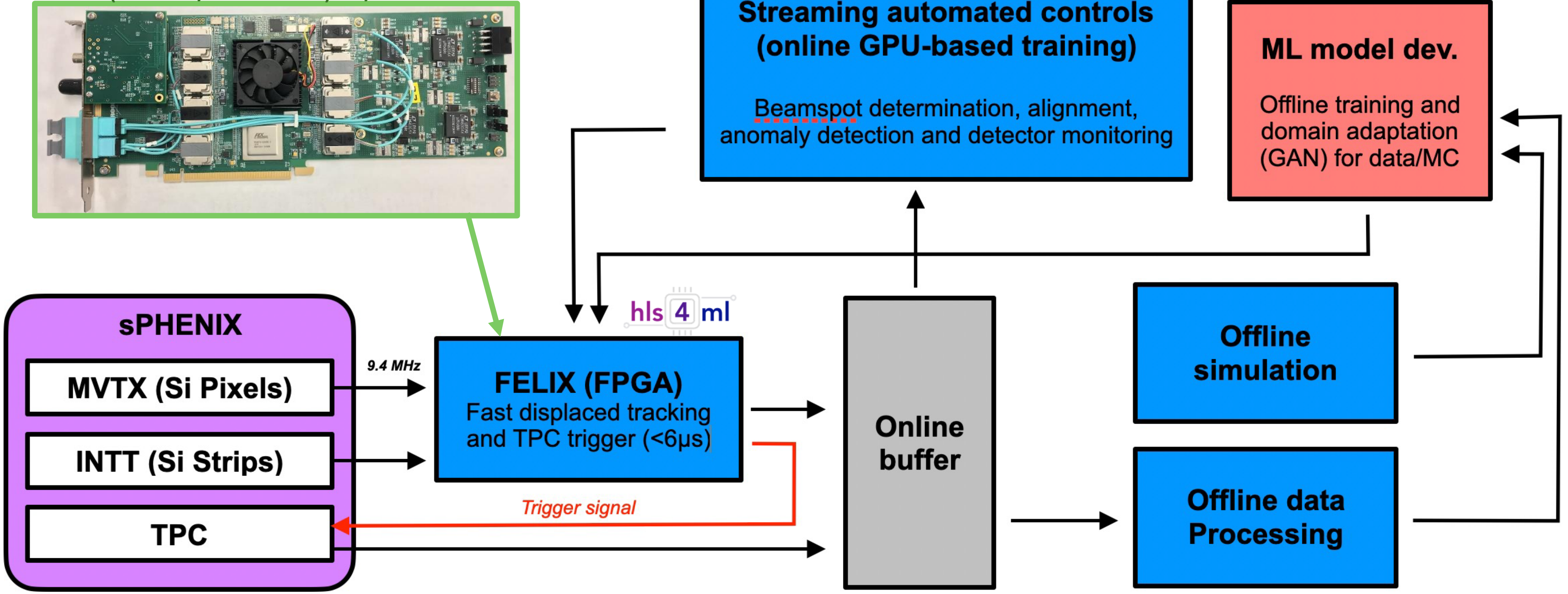


FPGA based trigger/data filter for sPHENIX and EIC

[See Micol Rigatti, FastML2022, Oct-4, link]

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

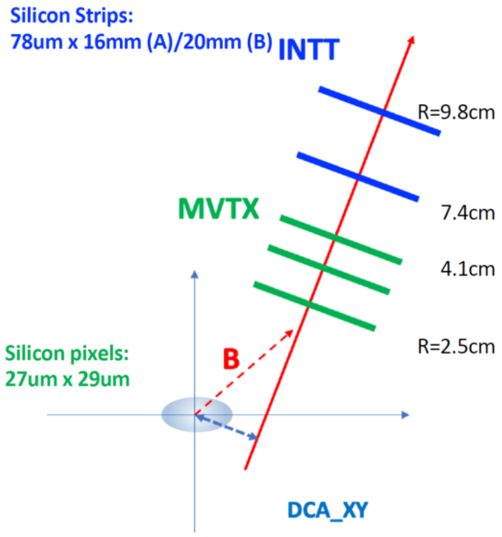
FELIX (ATLAS/sPHENIX) w/ Xilinx Kintex UltraScale



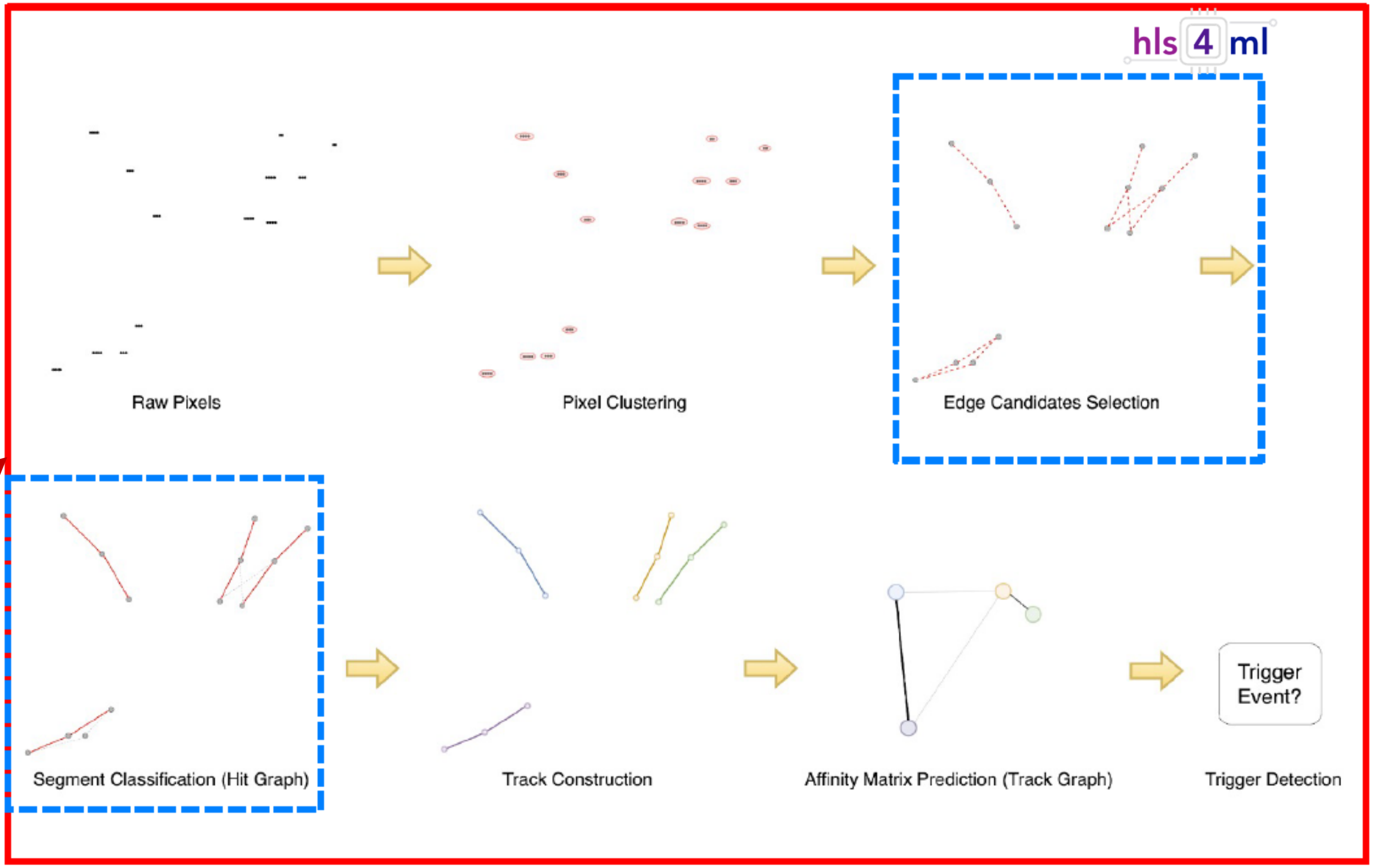
FPGA based trigger/data filter for sPHENIX and EIC

[See Micol Rigatti, FastML2022, Oct-4, link]

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

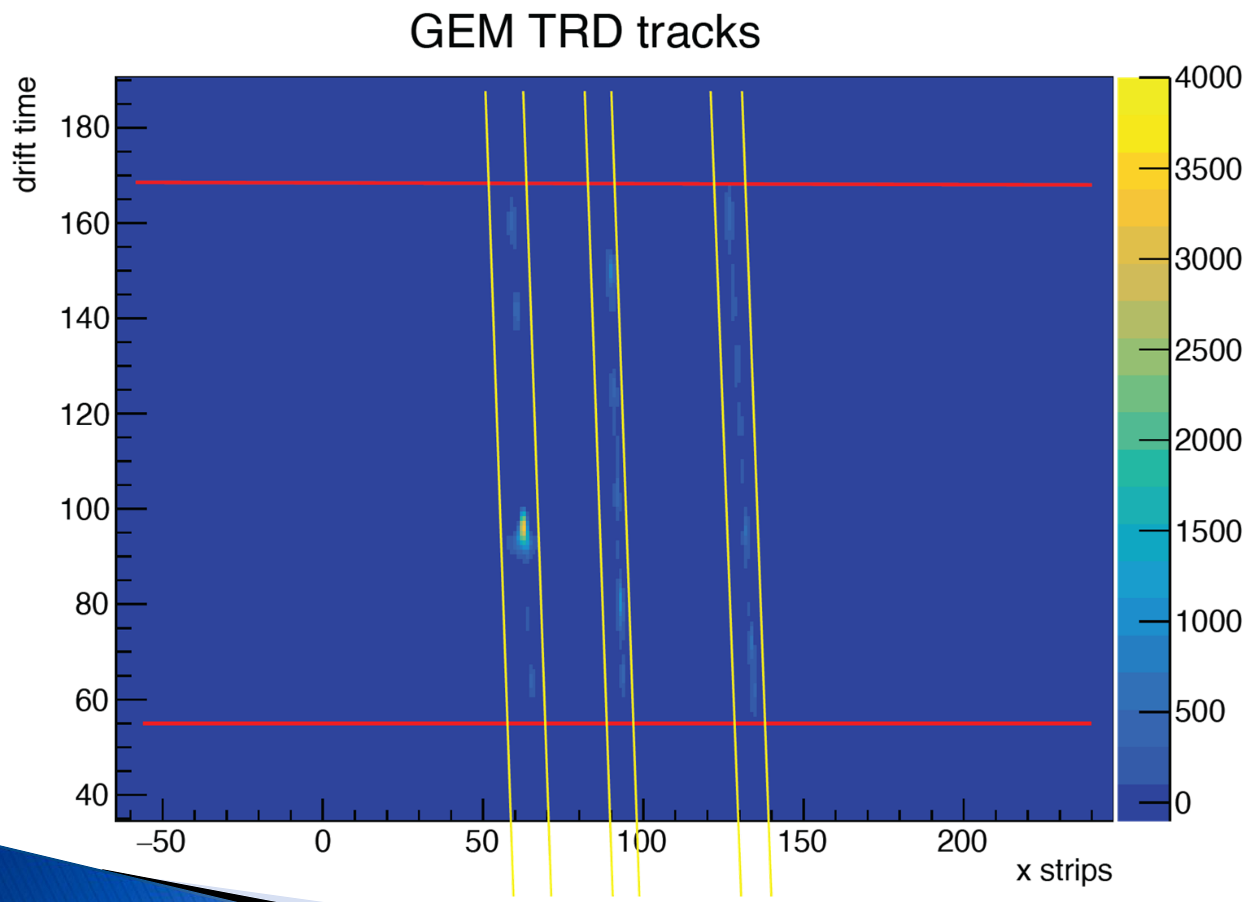


FELIX (ATLAS/sPHENIX)

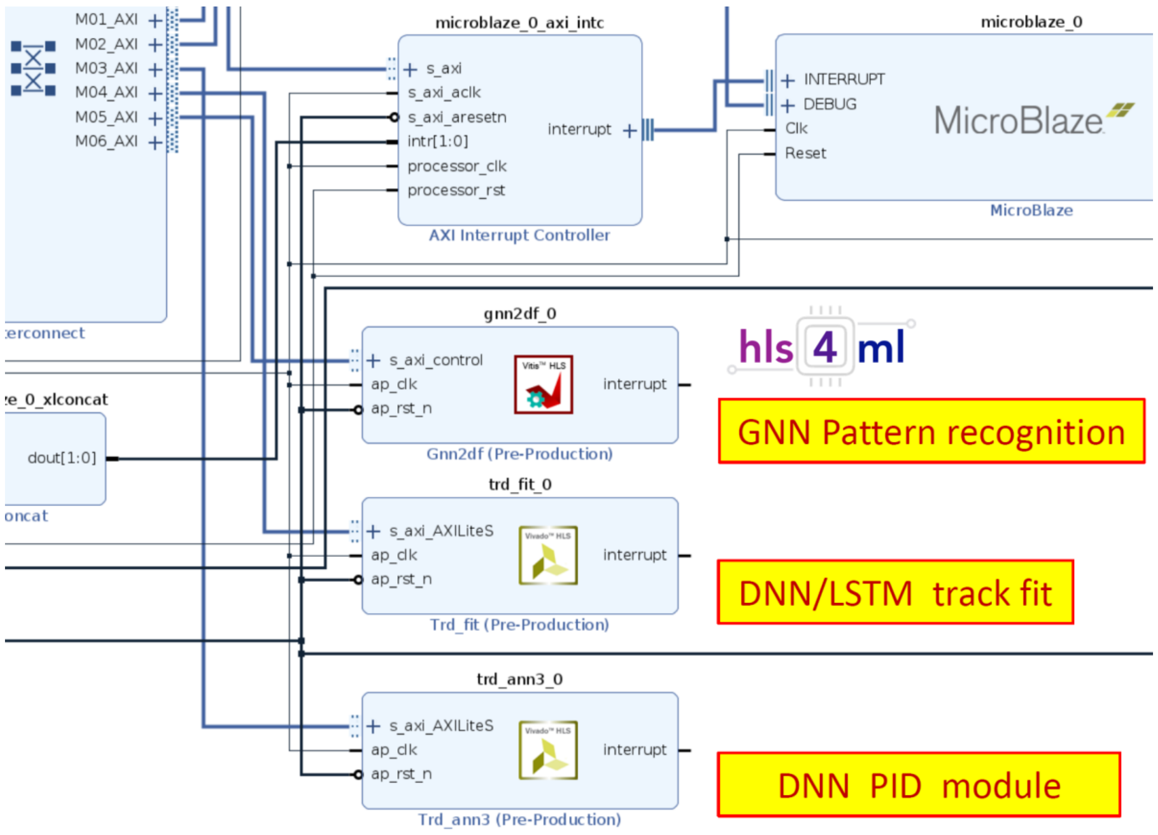


Example 2: GEM TRD tracking/PID

[S. Furletov, IEEE RT22, link]

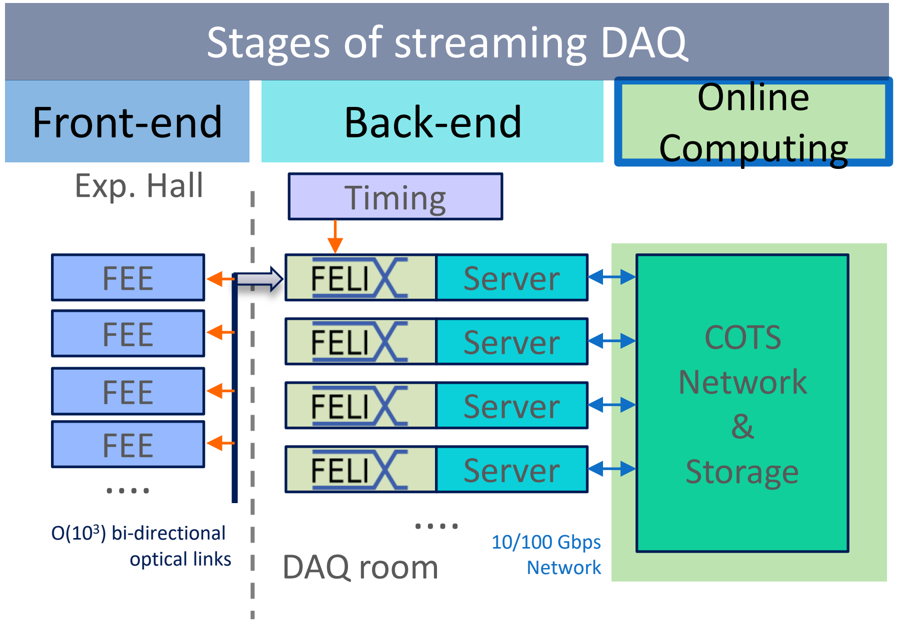


GNN Pattern reco, track fit and PID on FPGA test bench



Streaming DAQ stage 3: Online computing

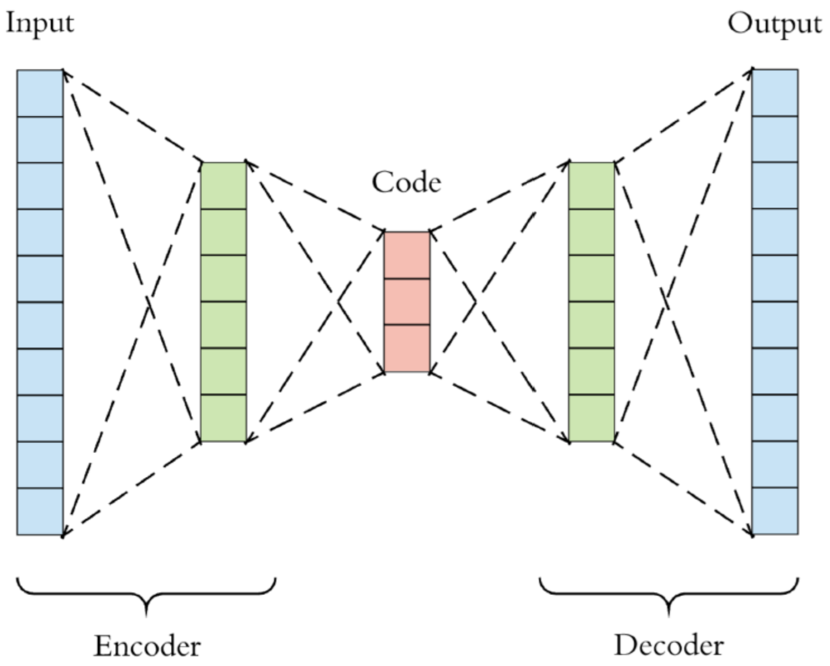
- Online computing is an integral part of streaming DAQ
  - Blending the boundary of online/offline computing
- FastML opportunities:
  - Lossy compression
  - Noise and background filtering
  - Higher level reconstruction
- Target hardware:
  - Traditional computing: CPU, GPU
  - Novel AI Accelerators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

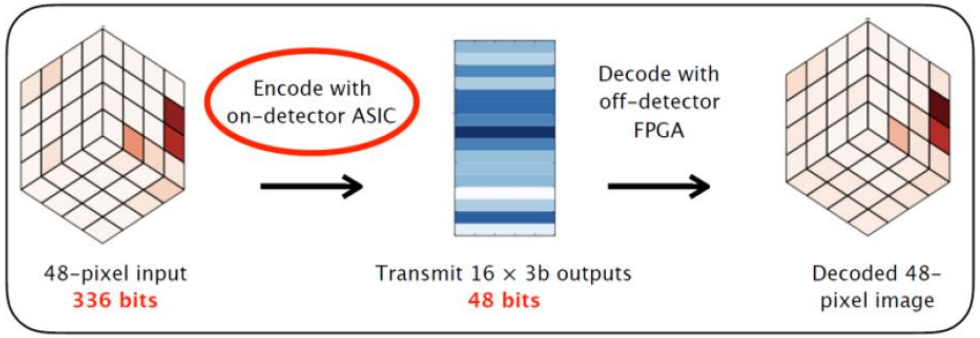
Simple auto-encode neural network



Lossy compression of data, noise filtering

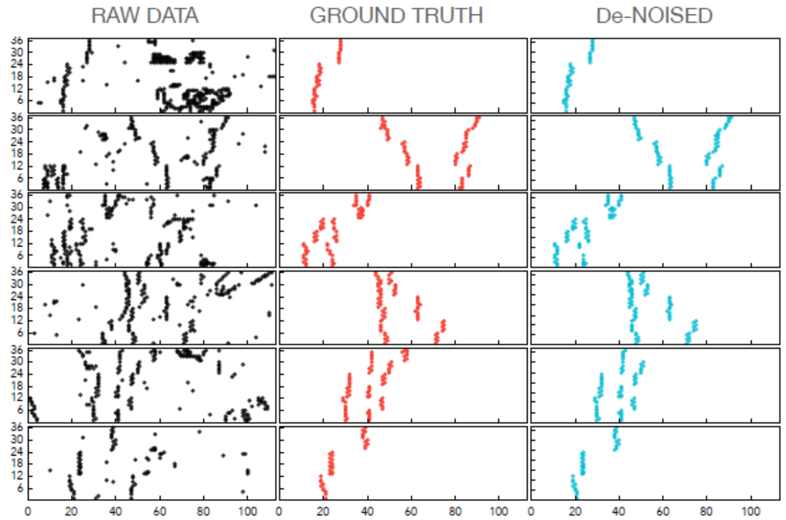
- Auto-encoder (AE) is a natural choice for unsupervised learning for lossy data compression: streaming data reduction
- Same network architecture can be adopted with supervised learning to filter out noise: further data reduction, speed up reconstruction
- See also in CMS HGCal ASIC, CLAS12 tracker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



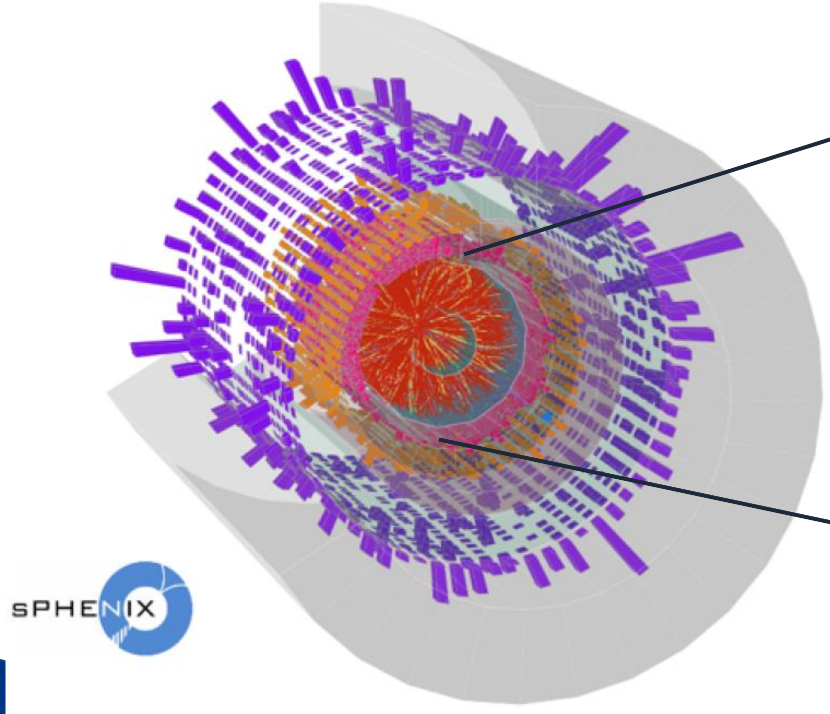
Data of time projection tracker at sPHENIX

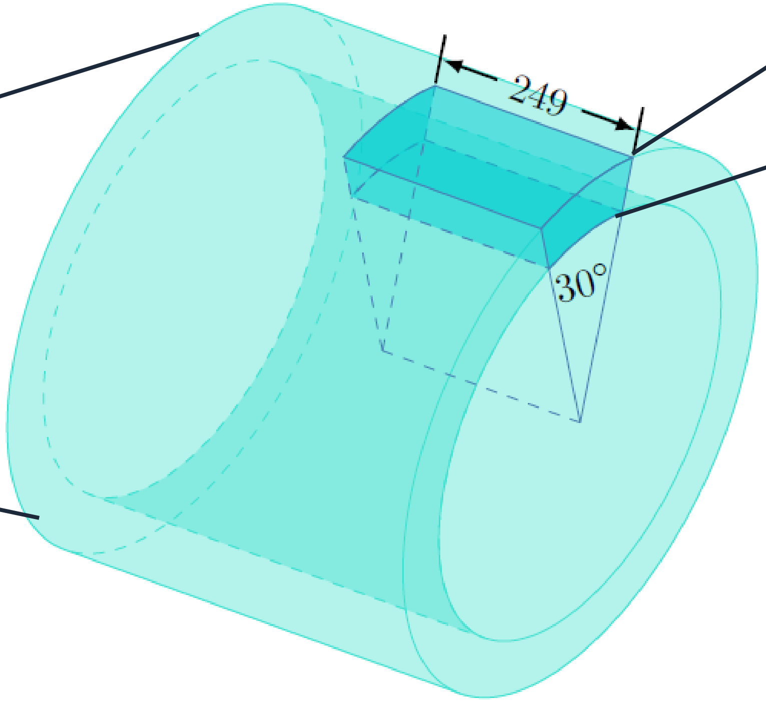
Busiest event in sPHENIX TPC

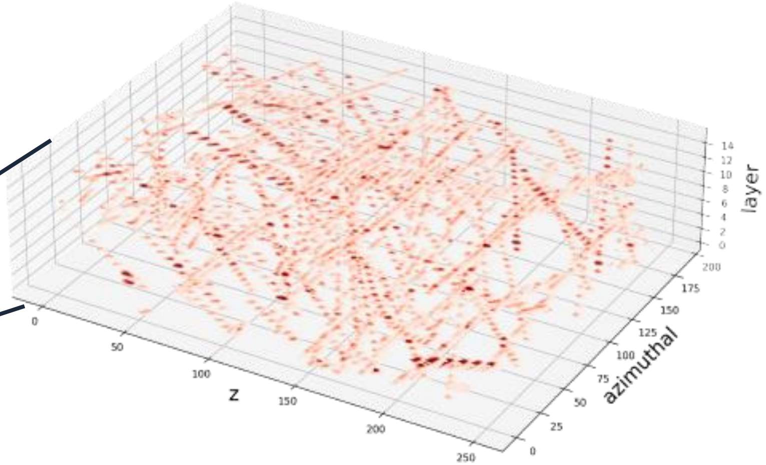
3D X-Y-Time time frame at 50Tbps prior to zero-suppression

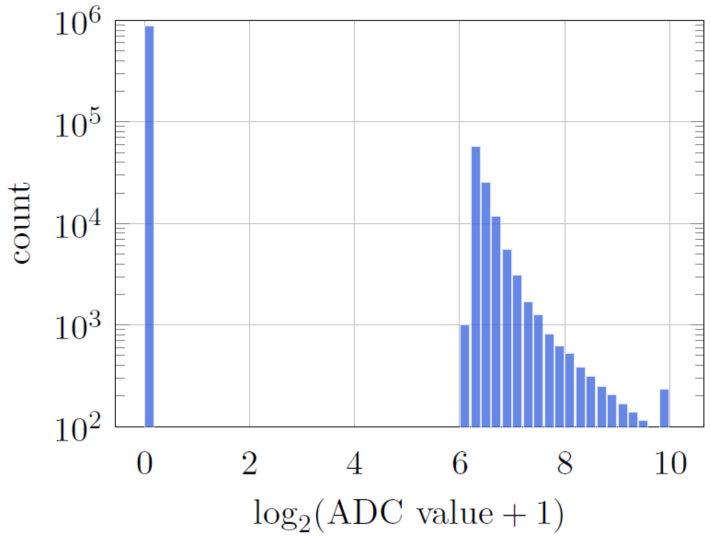
10% central Au + Au collision with 170kHz pile up

Data frame for 1/12 azimuth sector shown here







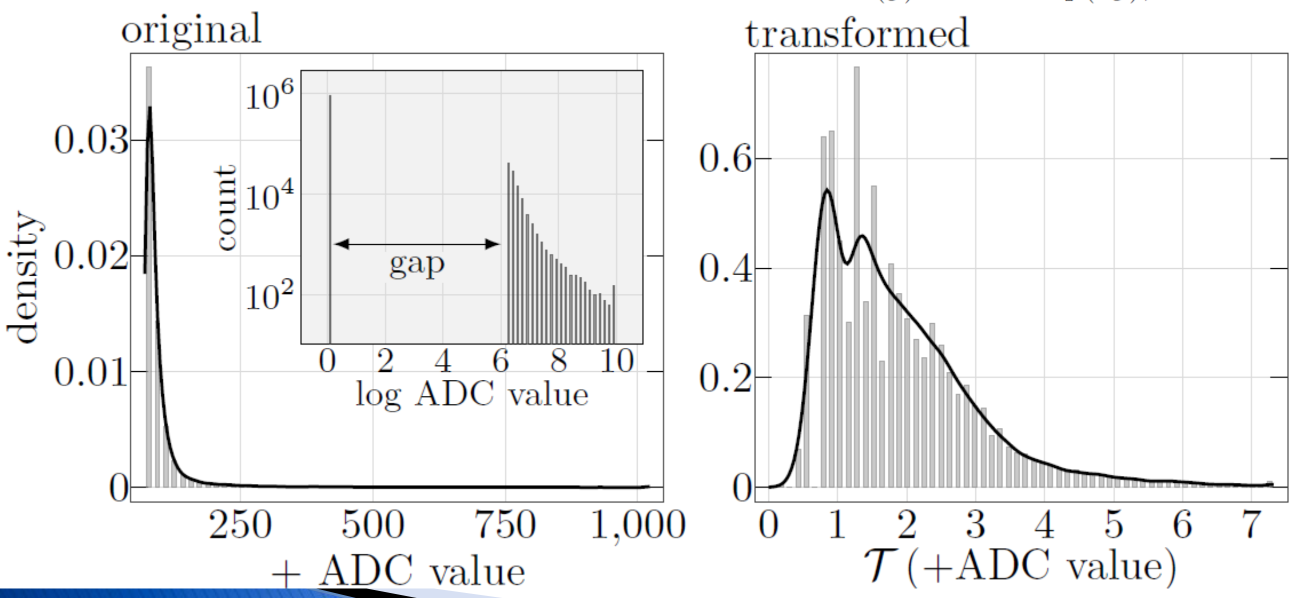


Bicephalous Convolutional Auto-Encoder (BCAE) and input transform [arXiv:2111.05423]

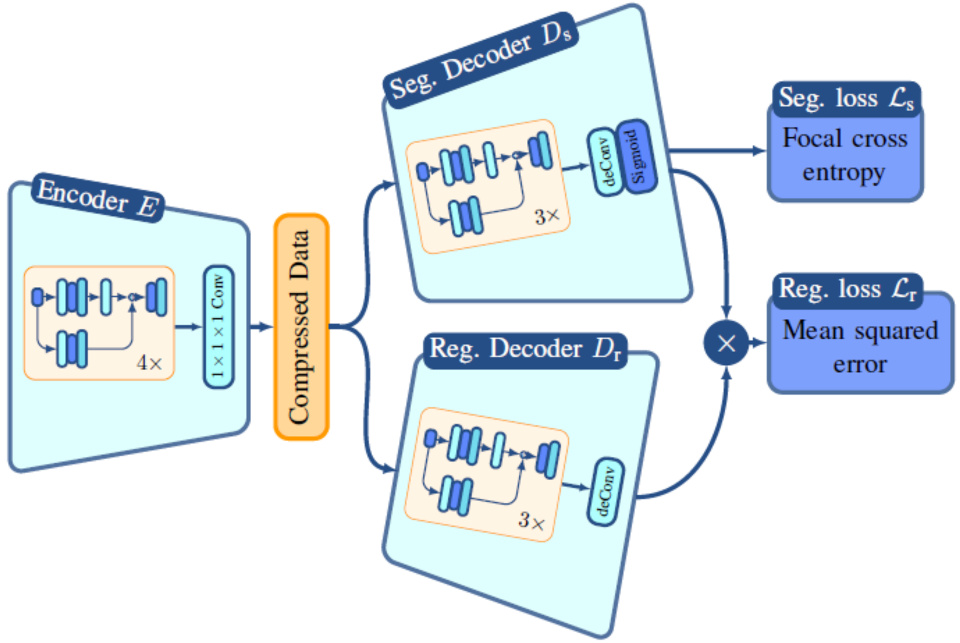
- Input transform: fill in the zero-suppression gap and make ADC distribution much less steep
- Bicephalous decoder: +classification decoder to note the zero-suppressed ADC voxels and +noise voxels in TPC

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

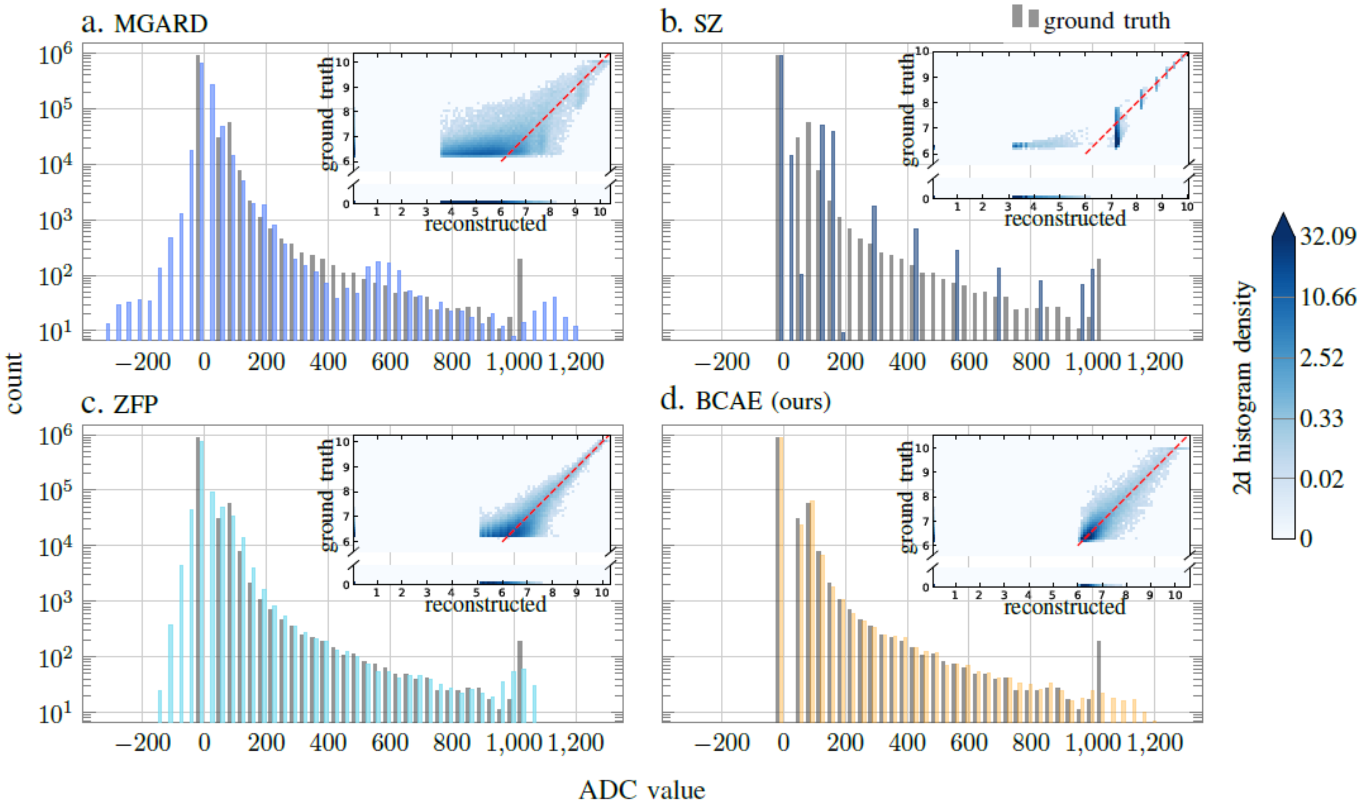
Inverse transform: $\mathcal{T}^{-1}(y) = 64 + \exp(6y), \quad x \in \mathbb{R}$

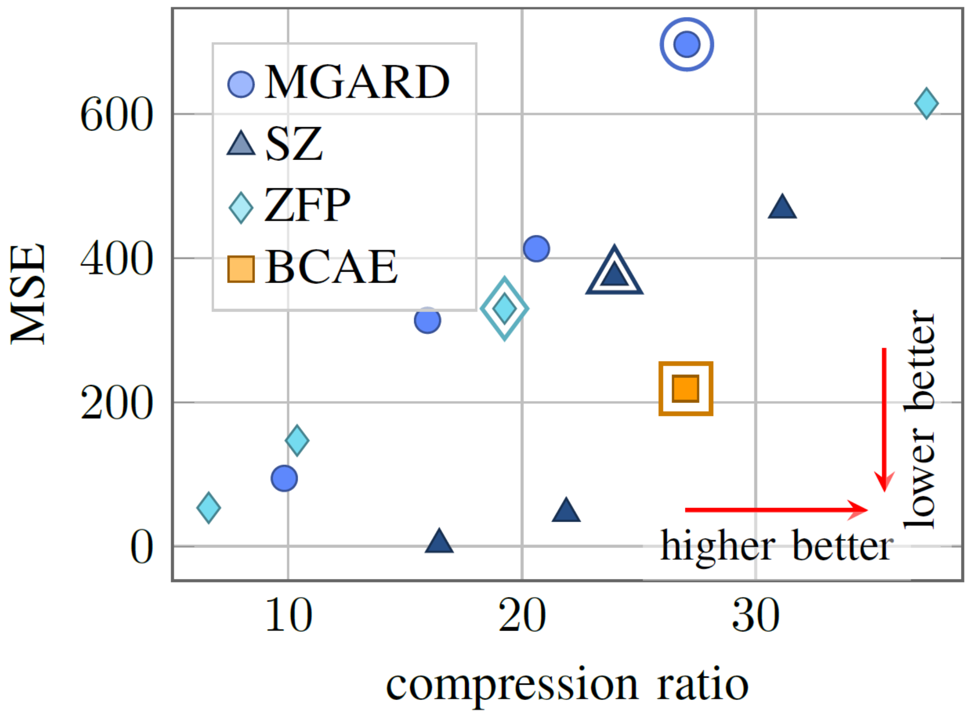


a. BCAE architecture summary



Comparison with existing algorithm [arXiv:2111.05423]

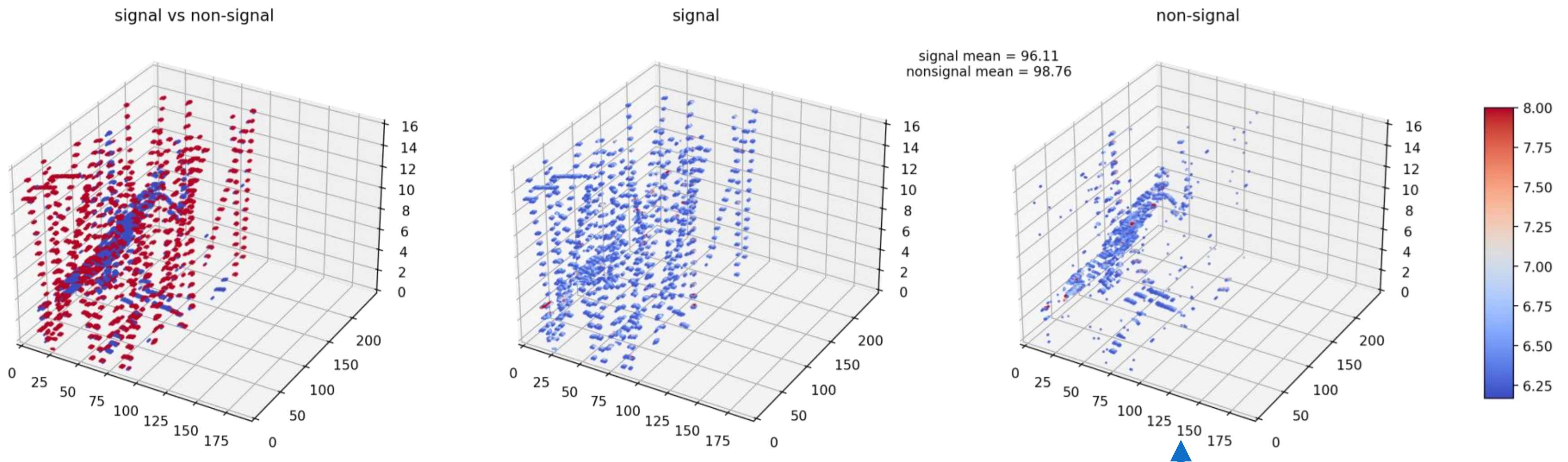


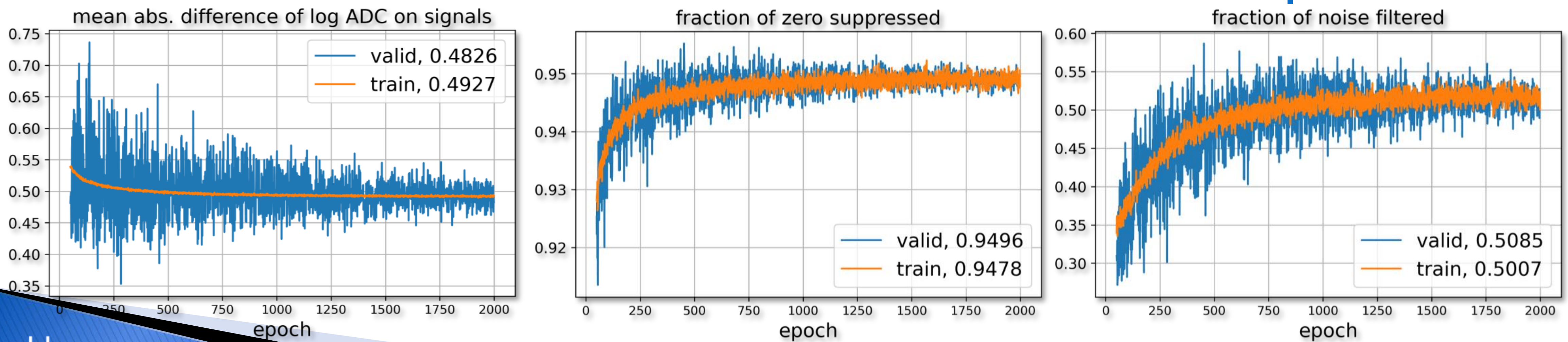


BCAE Compressor with noise filtering

[Y. Huang, IEEE RT22, link]

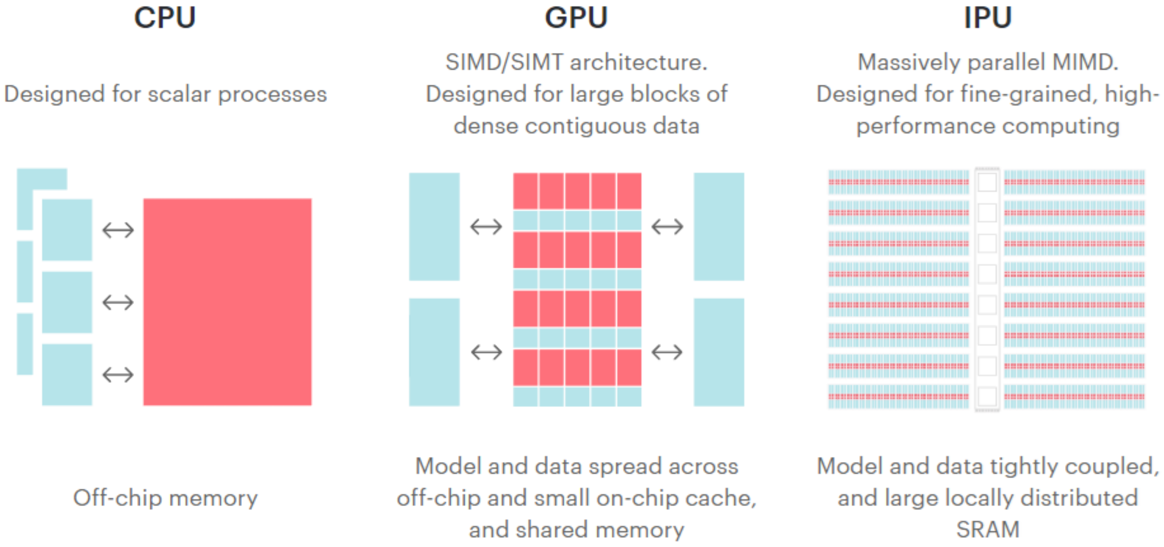
sPHENIX simulation
3 MHz $p + p$ TPC
streaming data
BCAE with compression
ratio 204:1 and 95% signal
retention (recall)



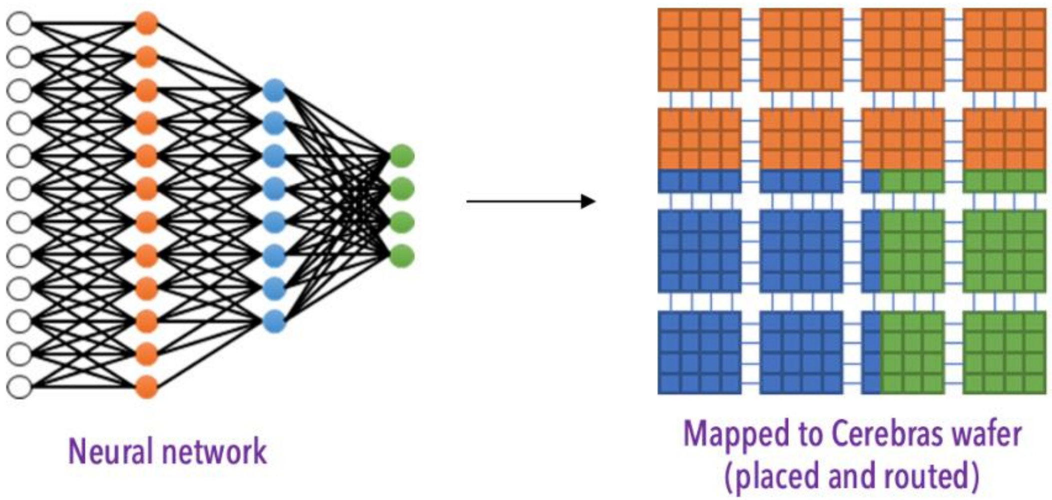


Novel AI Accelerators for streaming DAQ

- A new family of AI chips is emerging with non-von Neumann Architectures
  - Designed for NN computing, similarities to ML on FPGA
  - 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



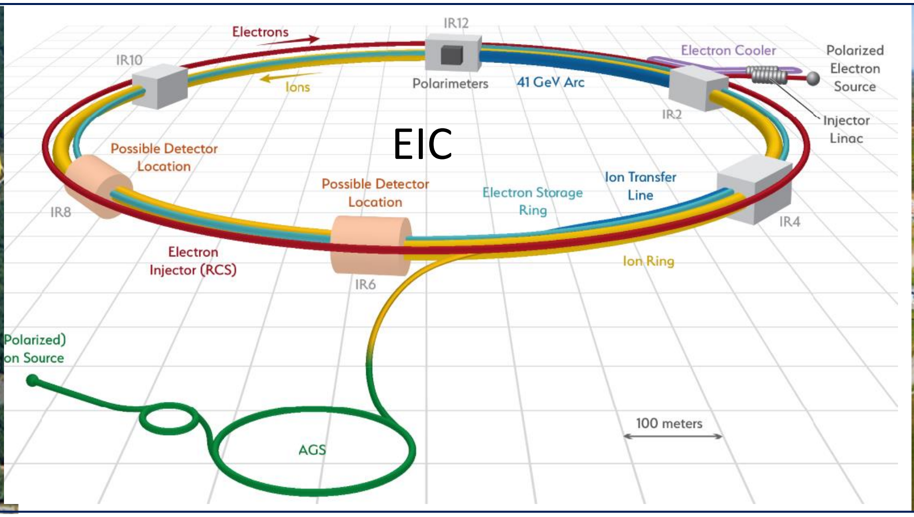
[GraphCore Web, link]



[Cerebras Compiler Docs, link]

Summary

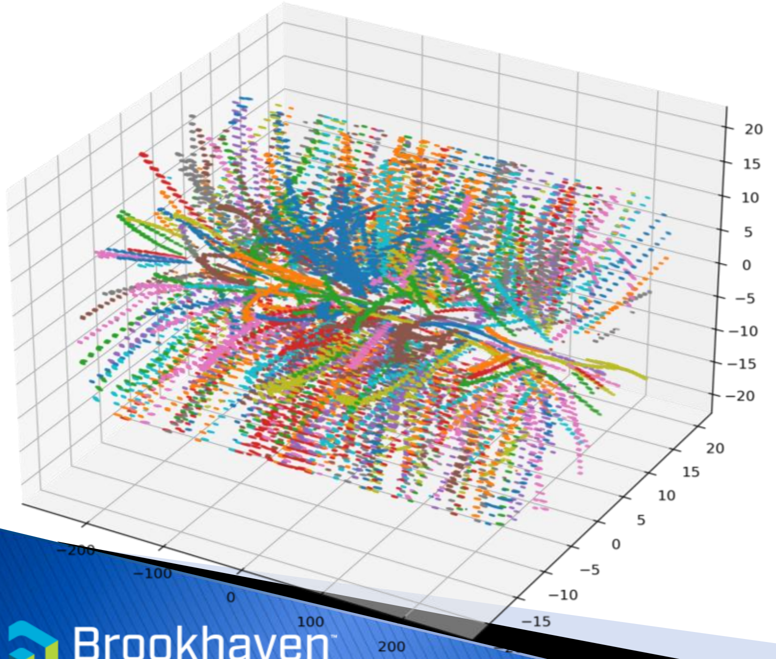
- Streaming readout is a paradigm shift adopted by many modern Nuclear Physics (NP) experiments, driven by diverse event topologies and stringent bias control
- Requiring large factors of data reduction computationally and at high throughput
- Driving the need of AI-based algorithms and platforms
  →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

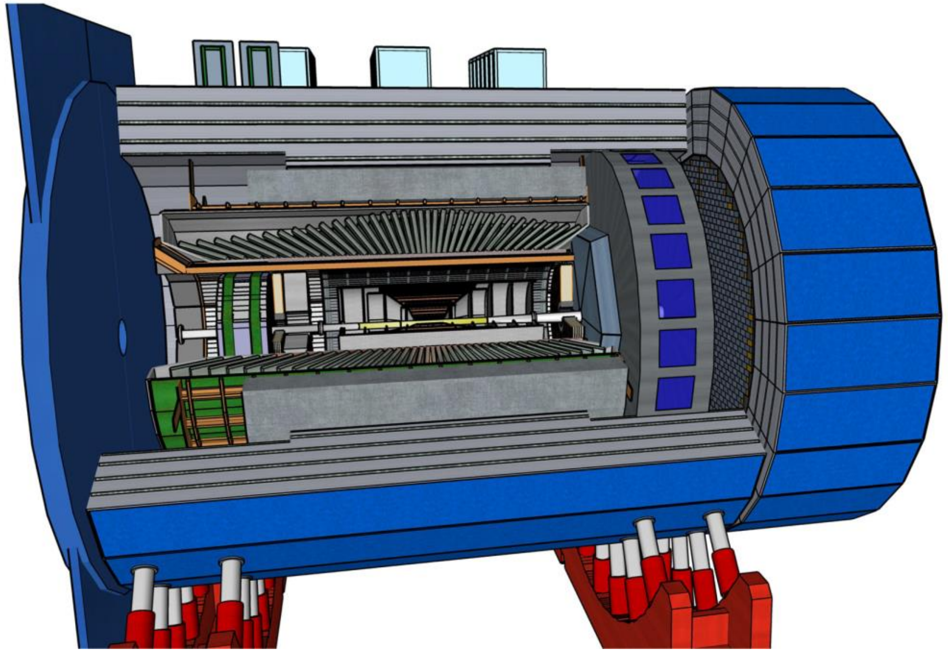


Join us! A Postdoc Advertisement

- BNL plan to open a postdoc position in coming months on real-time AI-based data reduction for sPHENIX and EIC
- Interested candidate please contact jhuang@bnl.gov

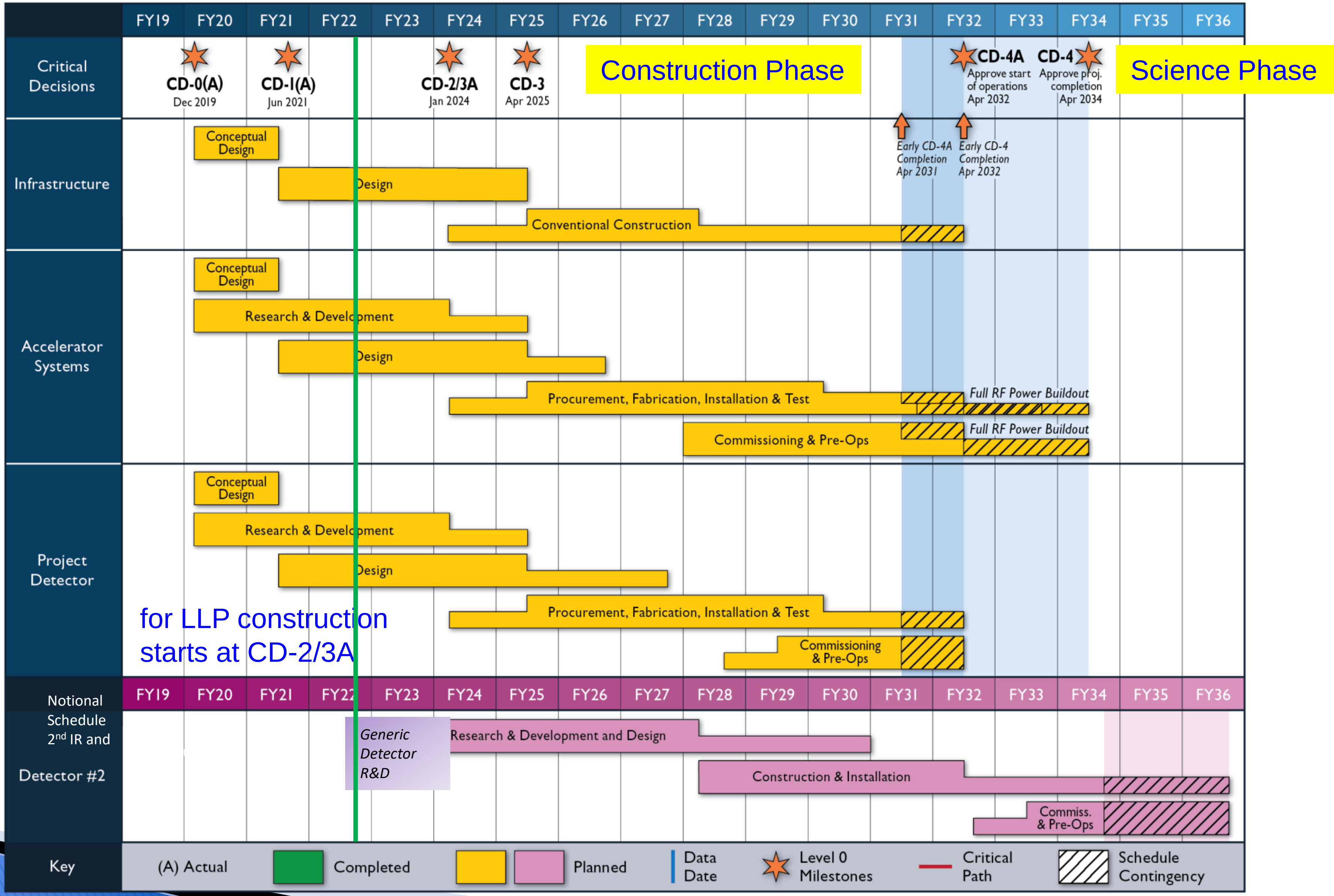
sPHENIX TPC data frame | sPHENIX detector, first data in 2023 | EPIC detector for EIC in 2030+





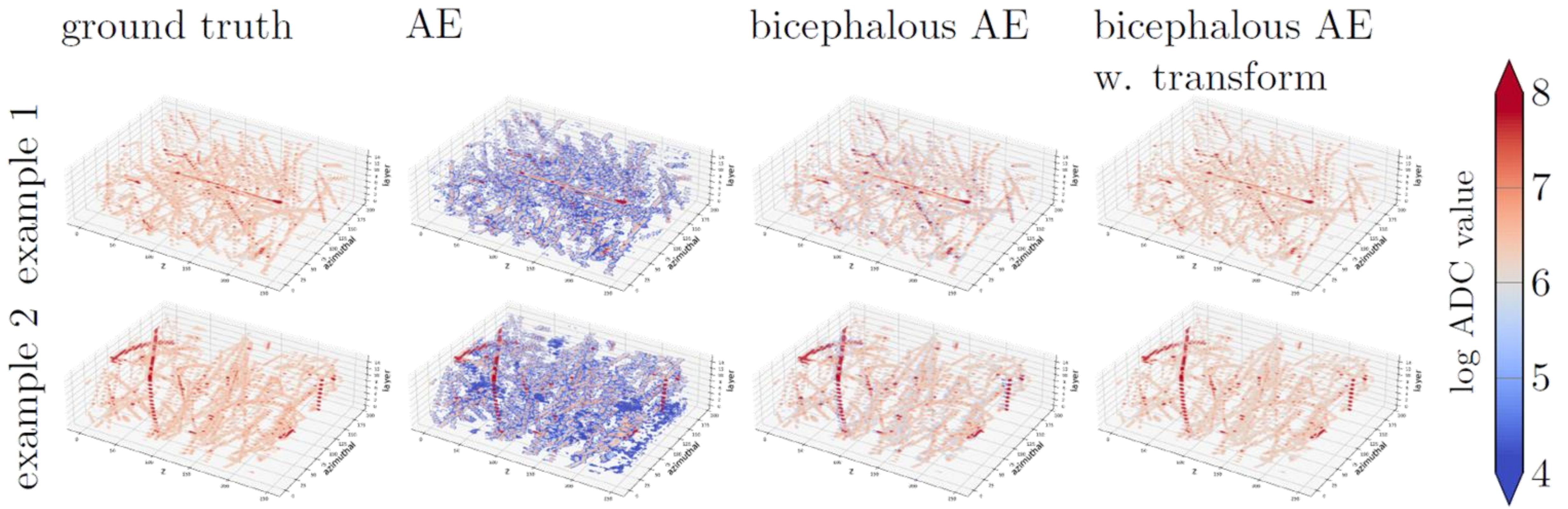
Extra information

High Level EIC Reference Schedule

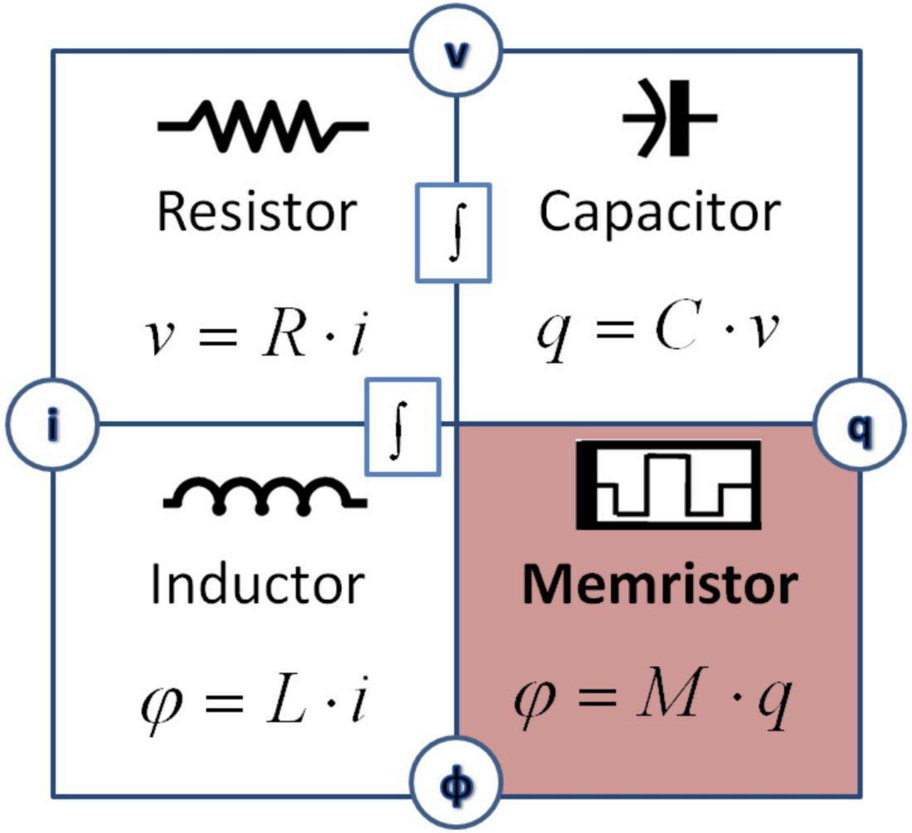


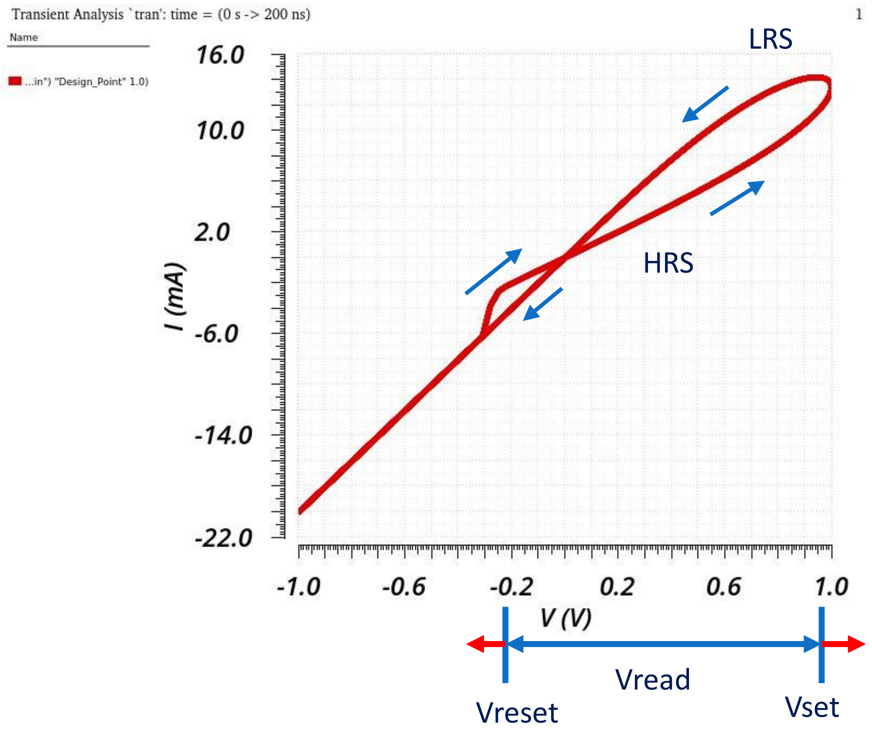
Brookhaven National Laboratory

Results from Bicephalous AE with transform [arXiv:2111.05423]



Memristor





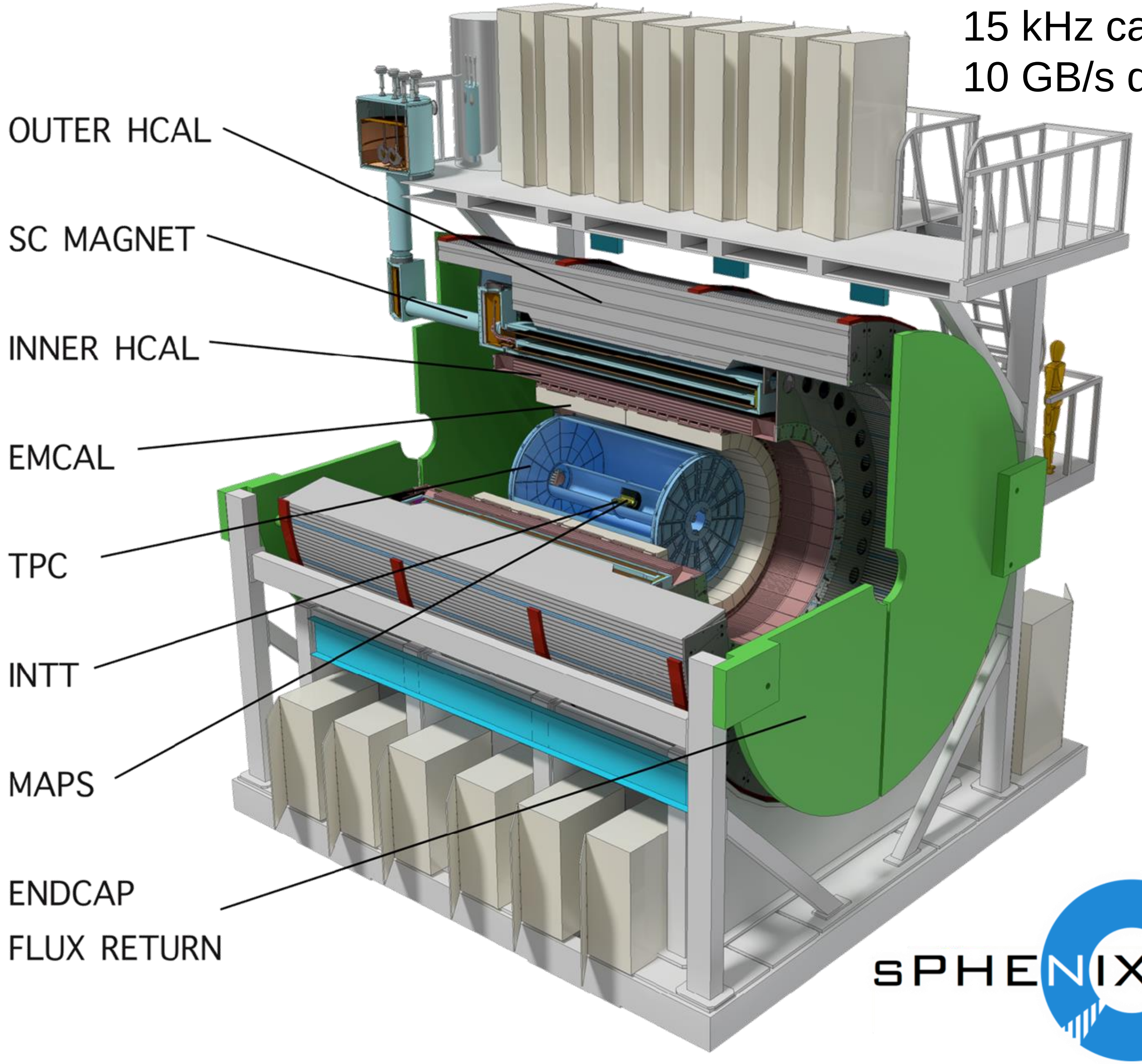
IEEE TRANSACTIONS ON CIRCUIT THEORY, VOL. CT-18, NO. 5, SEPTEMBER 1971 507

Memristor—The Missing Circuit Element

LEON O. CHUA, SENIOR MEMBER, IEEE

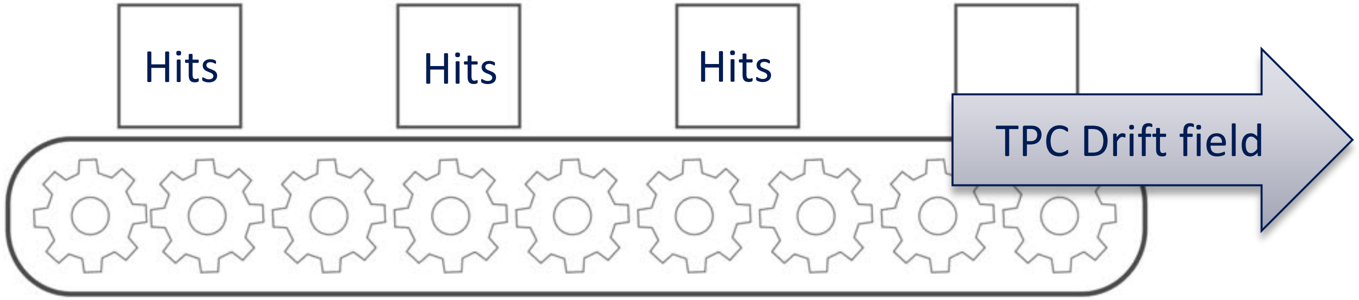
- Resistor with varying resistance
- Low Resistive State (LRS)
- High Resistive State (HRS)

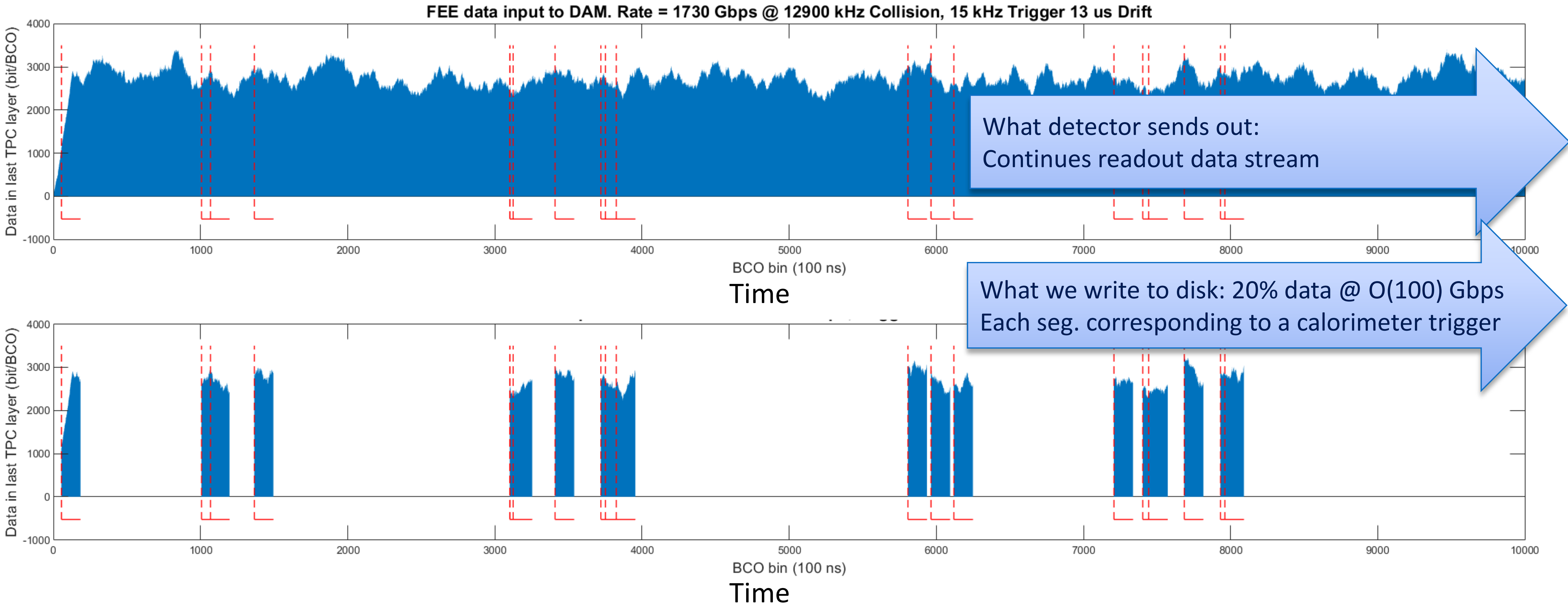
15 kHz calo trigger + 10% streaming DAQ
10 GB/s data logging



sPHENIX Detector

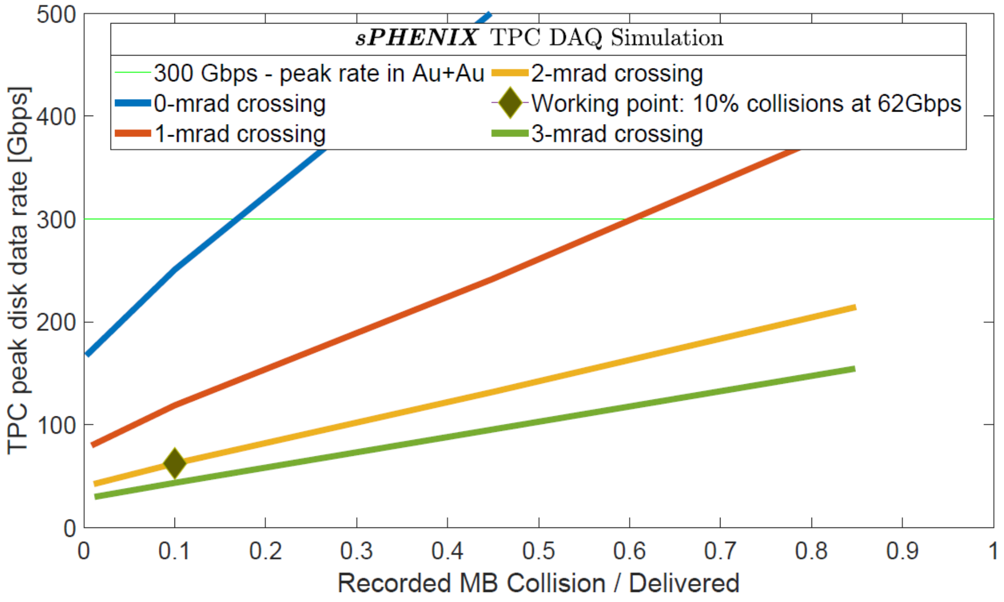
TPC data stream in sPHENIX triggered DAQ





Streaming readout status at sPHENIX

- All three sPHENIX tracking detector uses streaming readout
- Developed plan to take 10% streaming data for heavy flavor physics program commended by RHIC PAC.
- Data taking start in 2023!

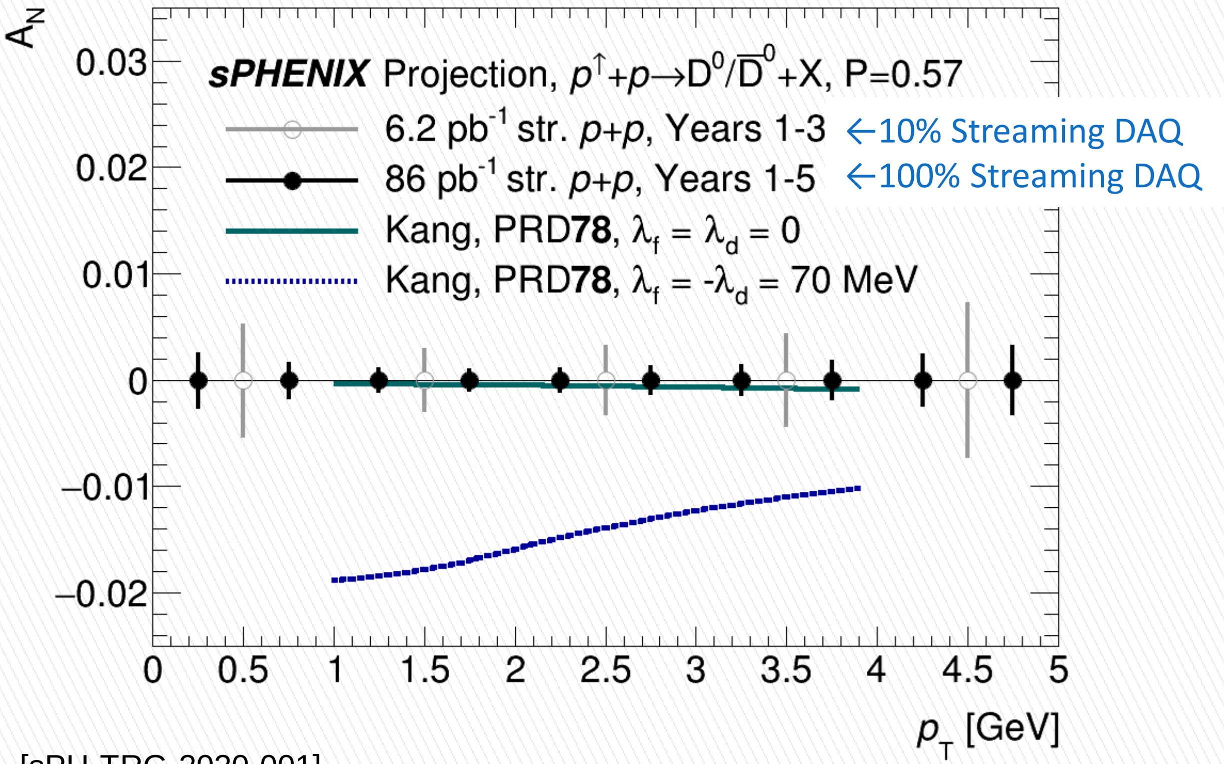


RHIC PAC 2020 report

> We commend sPHENIX for developing the continuous streaming readout option for the detector, which increases the amount of data that can be collected in Run-24 by orders of magnitude. In particular in the sector of open heavy flavor, this technique will give access to a set of qualitatively novel measurements that would otherwise not be accessible. Given the tight timeline for completing the RHIC physics program before construction of the EIC begins, this is a tremendous and highly welcome achievement.

Expanding the streaming data would given much better physics output

sPHENIX D^0 trans. spin asymmetry, $A_N \rightarrow$ Gluon Sievers via tri-g cor.

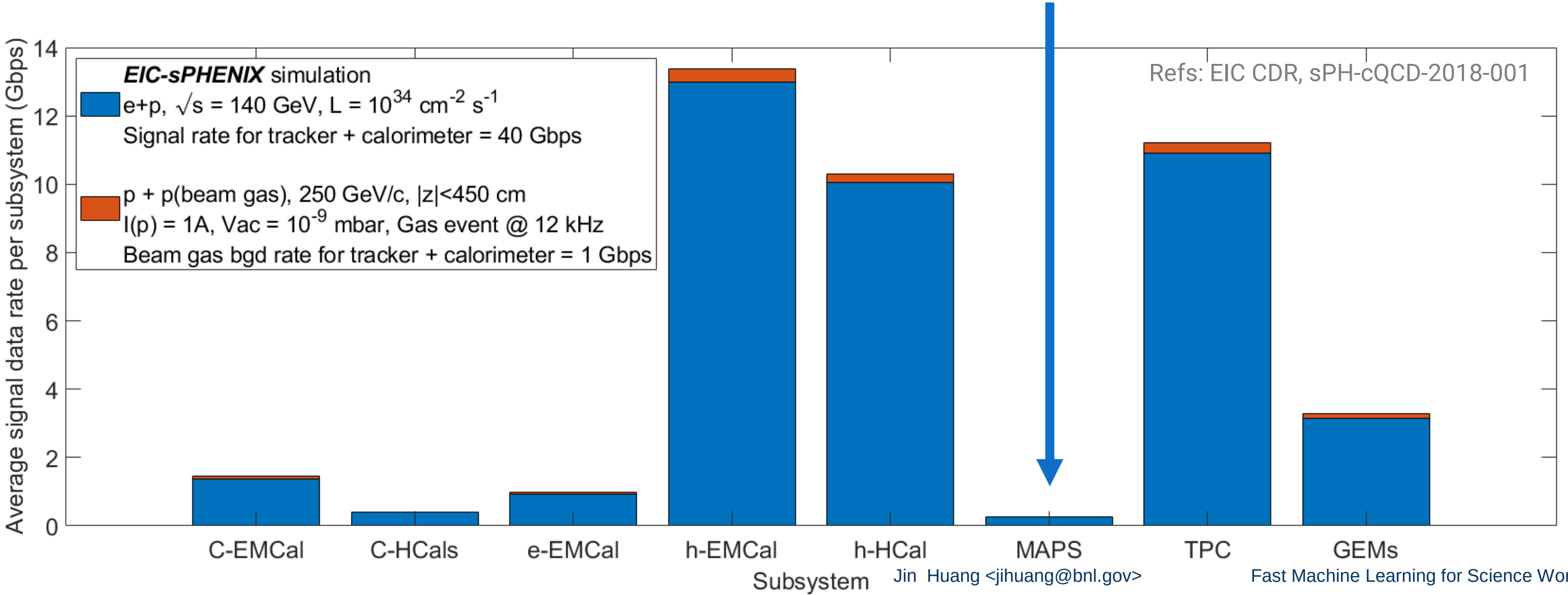


[sPH-TRG-2020-001]
Model: 10.1103/PhysRevD.78.114013

- sPHENIX default to record 10% streaming data in tracker
- By increasing to 100% streaming data, we can significantly improve reach of D0 access to tri-gluon correlation
- However, 100% recording is significantly bump to data rate, >250Gbps (sPHENIX expect to log at ~100Gbps)
- Requires some real-time data reduction, opportunity for AI application
  - Lossy compression, focus of later this talk
  - Signal selection: seminar D.T. Yu Feb 1st

Signal data rate -> DAQ strategy

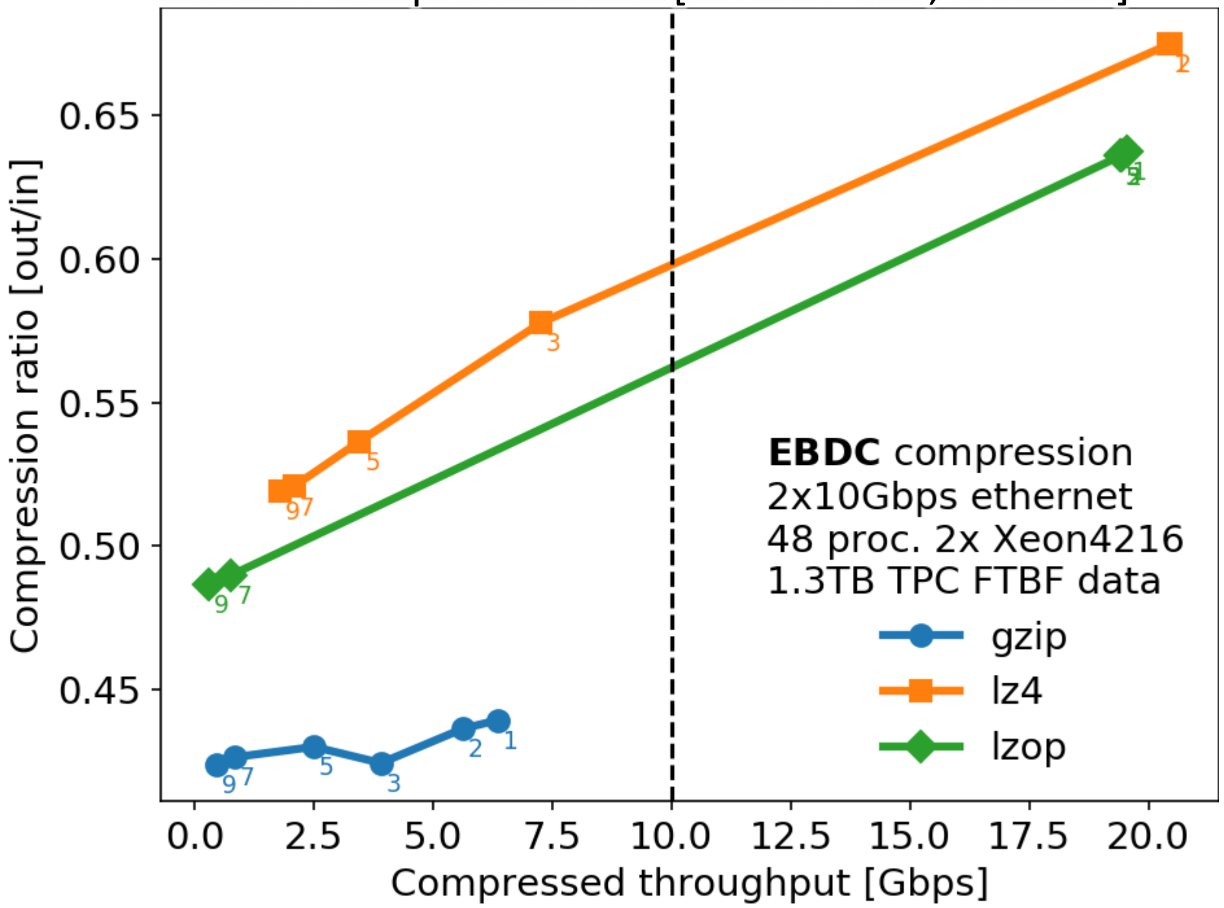
- What we want to record: total collision signal ~ 100 Gbps @ 10^{34} cm^{-2} s^{-1}
  - Assumption: sPHENIX data format, 100% noise, Less than sPHENIX peak disk rate. 10^{-4} comparing to LHC collision
- Therefore, we could choose to stream out all EIC collisions data
  - In addition, DAQ may need to filter out excessive beam background and electronics noise, if they become dominant.
- Very different from LHC, where it is necessary to filter out uninteresting p+p collisions (CMS/ATLAS/LHCb) or highly compress collision data (ALICE)



Online computing for streaming data - compression

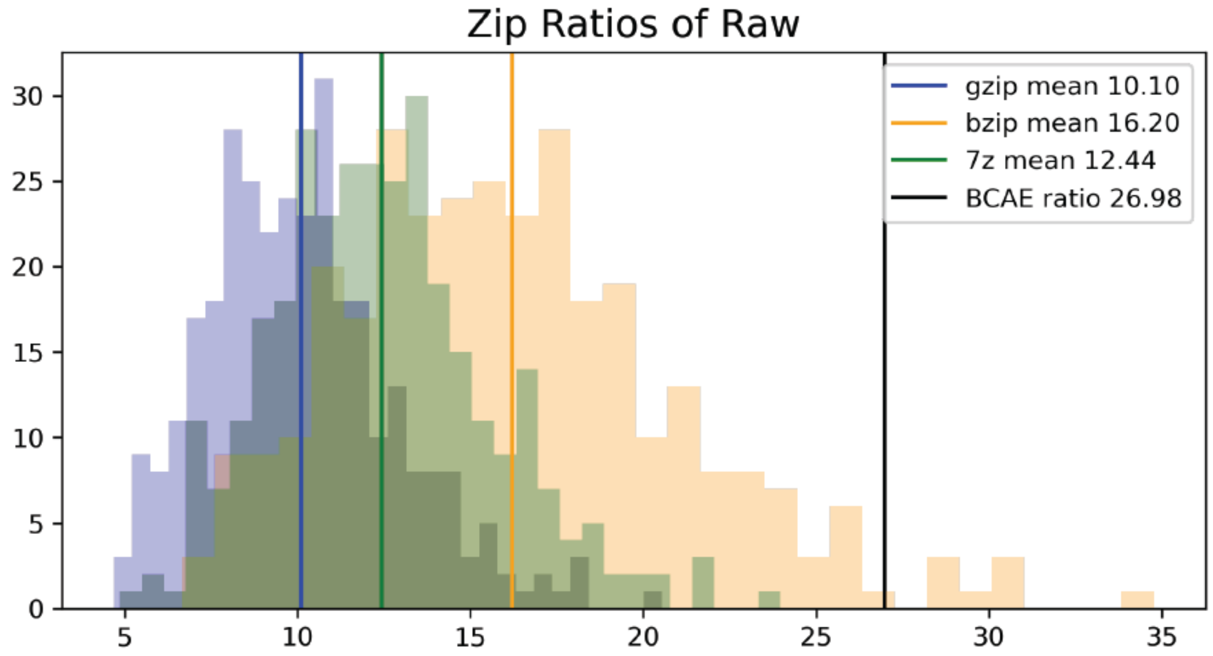
- Lossless compression
  - Compress by ~1/2
  - Well established fast compression algorithm
- Lossy compression
  - Opportunity for unsupervised machine learning based on data
  - This work: Bicephalous Convolutional Neural Encoder for compressing zero-suppressed data and noise filtering

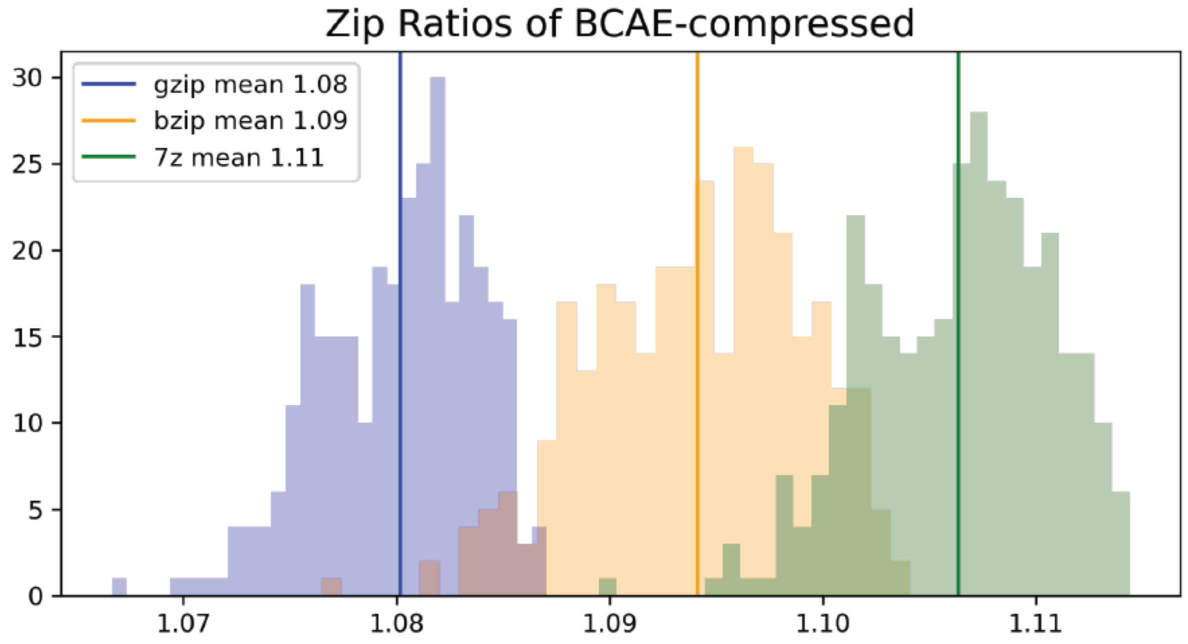
Lossless compression test [LDRD19-028, sPHENIX]



Compressibility check: thanks to suggestion from Brett!

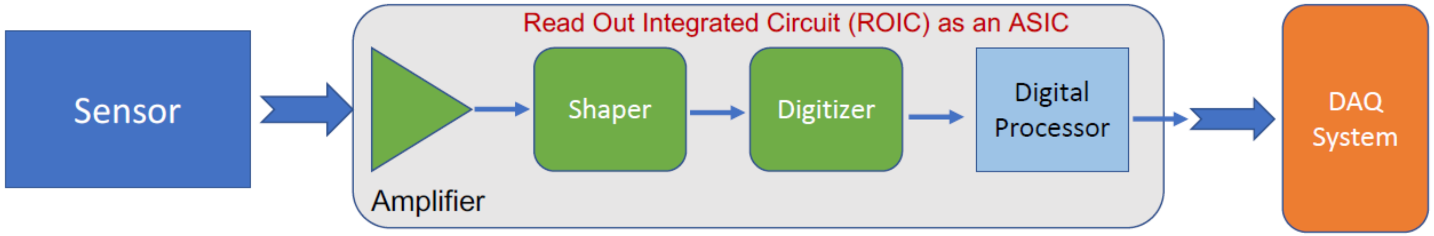
- The lossy-compressed code is hardly compressible further losslessly

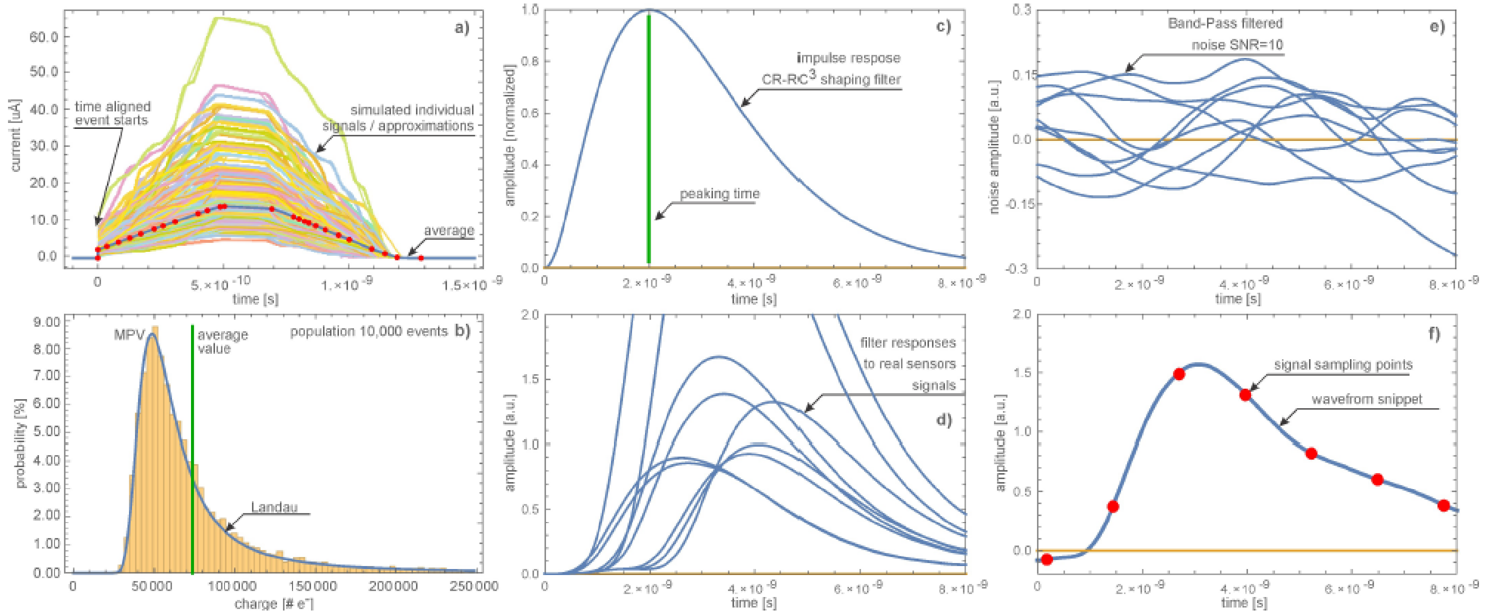




LGAD signal sample [LDRD 21-023, JINST in press]

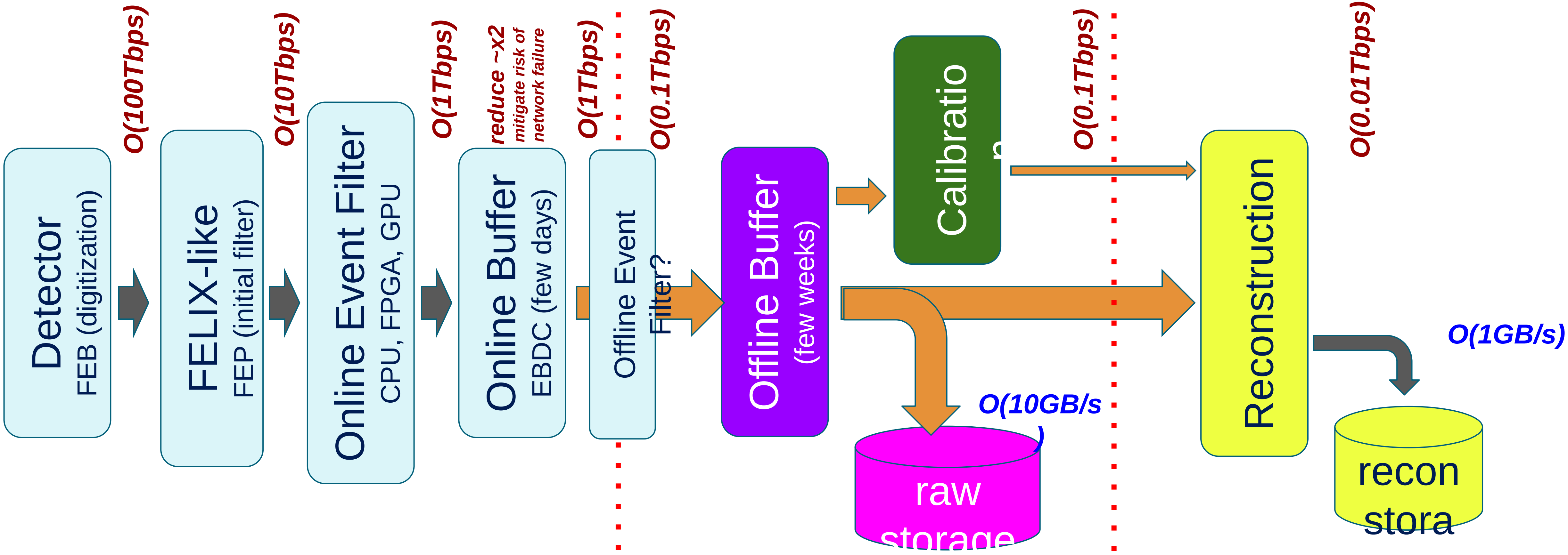
Current focus:
Deep dive into NN regression for LGAD tracker-TOF data





Blured boundary with offline computing

Countesy: David Lawrence
ECCE computing model [link]



Experimental Hall and Counting House (Project Funds)

Data Center(s): SDCC [,JLab, …] (Operations Funds)

HTC Compute Facilities SDCC ,JLab, … (Operations)