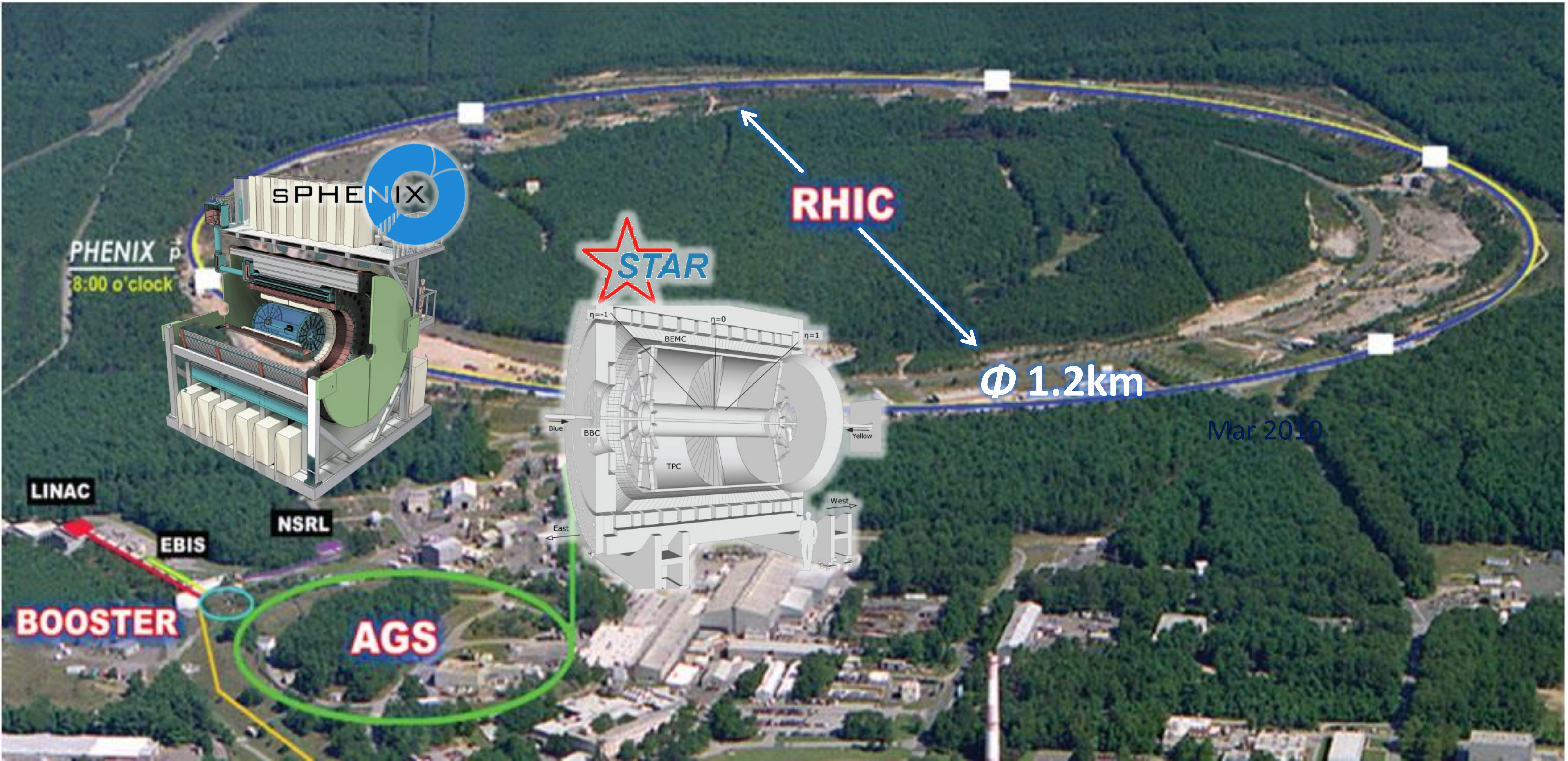
Real-time AI for sPHENIX and EIC

Jin Huang

Brookhaven National Lab

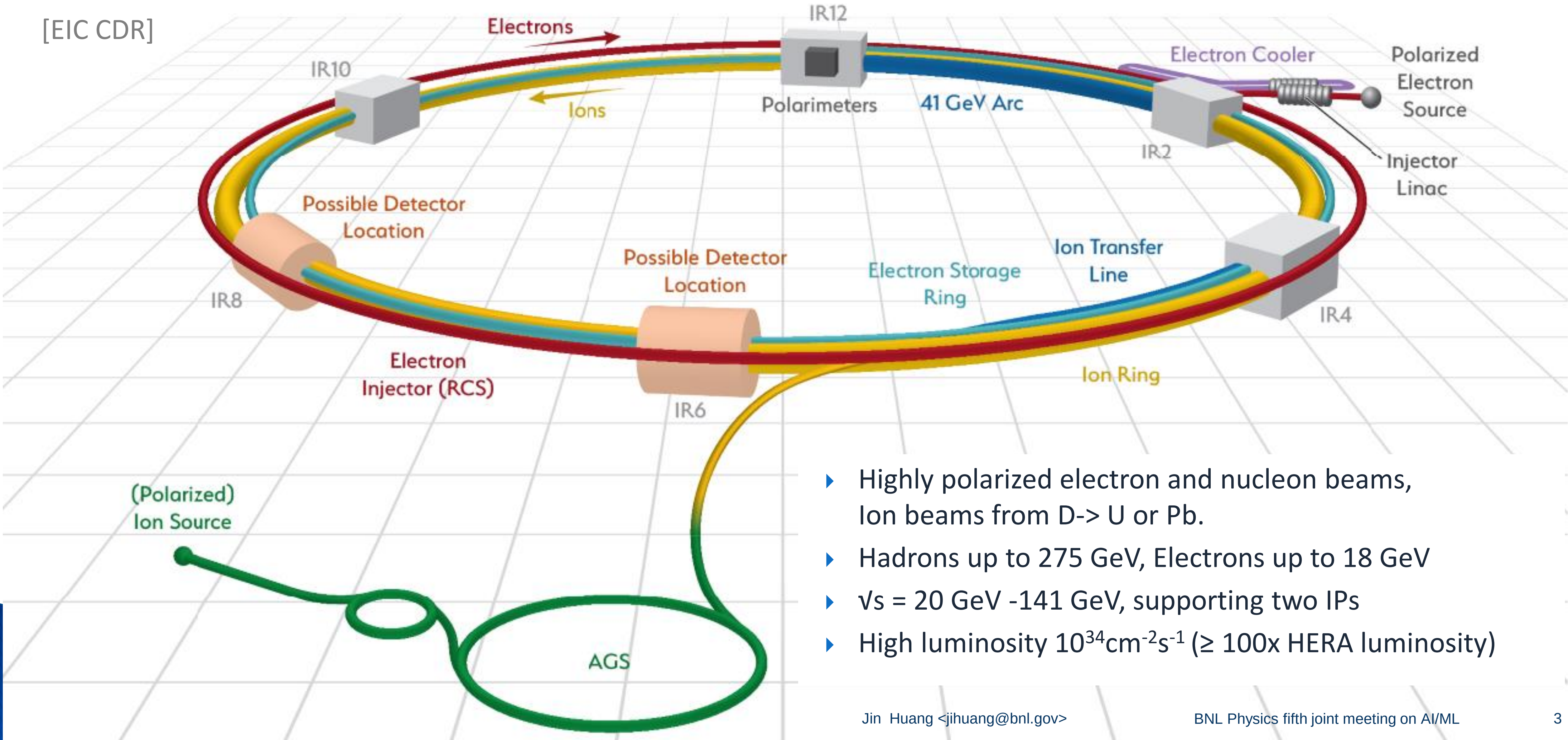
With focus on LDRD 19-028, 21-023, and collaborating work at CSI, IO and PO

Relativistic Heavy Ion Collider in 2023+



RHIC transition to the EIC

[EIC CDR]



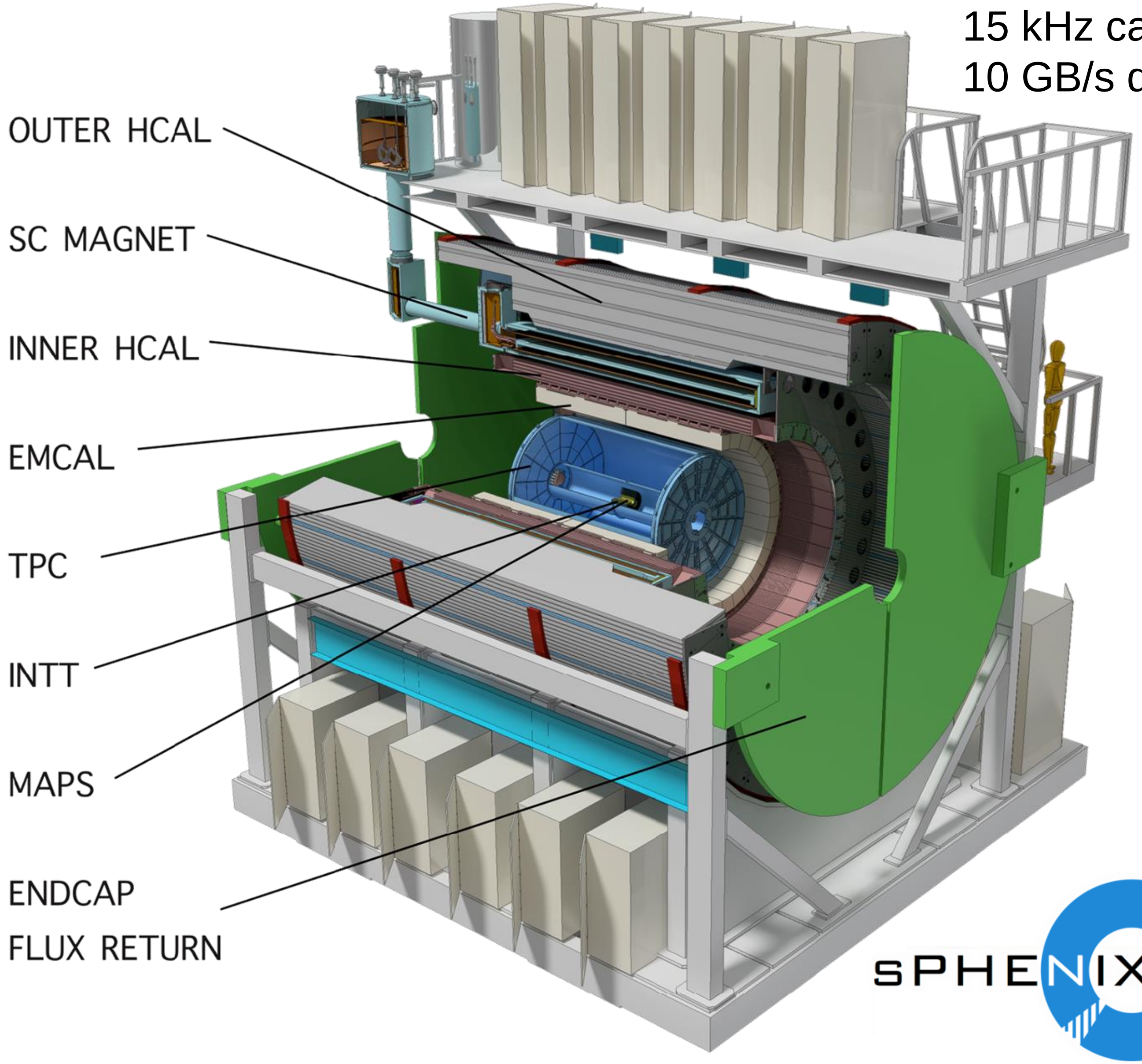
- Highly polarized electron and nucleon beams, Ion beams from D-> U or Pb.
- Hadrons up to 275 GeV, Electrons up to 18 GeV
- √s = 20 GeV -141 GeV, supporting two IPs
- High luminosity $10^{34}cm^{-2}s^{-1}$ (≥ 100x HERA luminosity)

EIC: unique collider
→ unique real-time system challenges

	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**

- EIC luminosity is high, but collision cross section is small ($\propto \alpha_{EM}^2$) → low collision rate
- But events are precious and have diverse topology → hard to trigger on all process
- Background and systematic control is crucial → avoiding a trigger bias

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

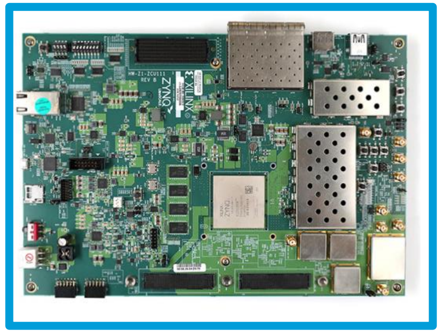


sPHENIX Detector

Related streaming readout electronics

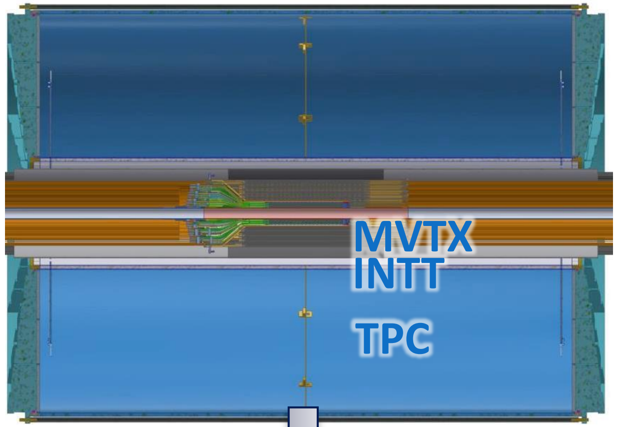
Associated test projects

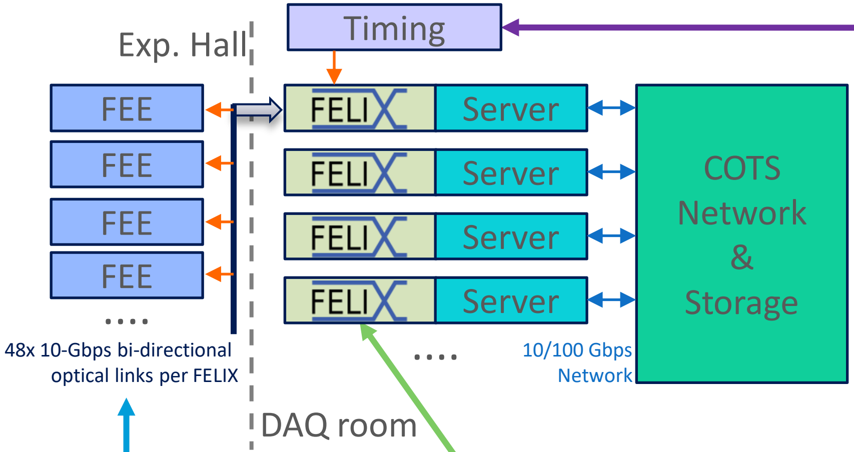
Precision timing digitizer
DRS4GlO (SBIR/LDRD)



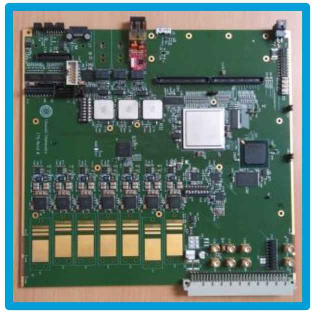
High density multiplexer+ ADC
RFSoC Digitizer (LDRD)

sPHENIX streaming DAQ for tracker



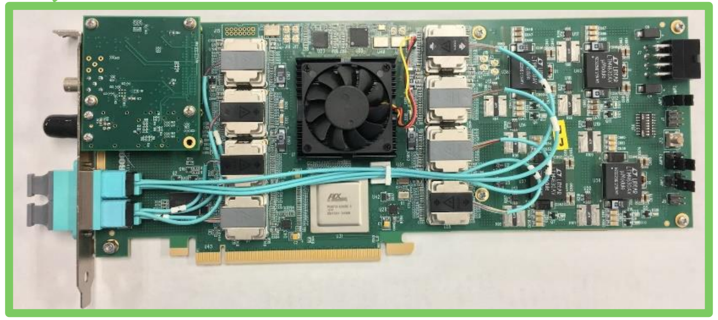


Global Timing Module (NSLS II/sPHENIX)
Receiving from RHIC RF low glitter clock source



MVTX RU, 200M ch — INTT ROC, 400k ch
ALPIDE (ALICE/sPHENIX), FPHX (PHENIX)

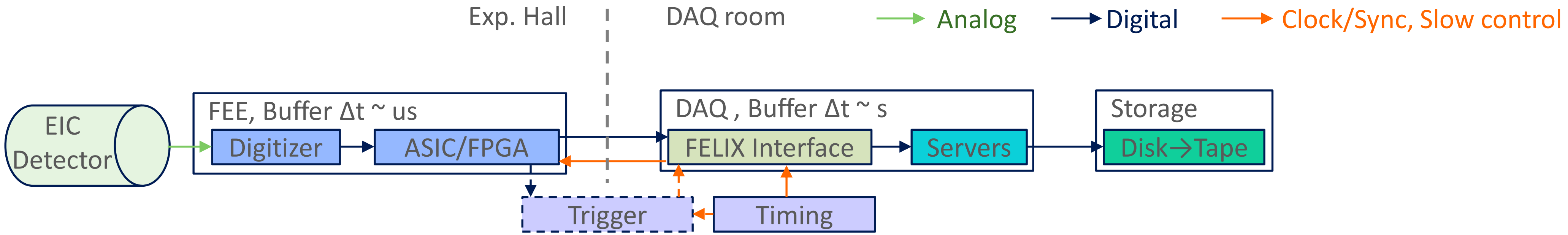
TPC FEE, 160k ch
SAMPAv5 (ALICE/sPHENIX)

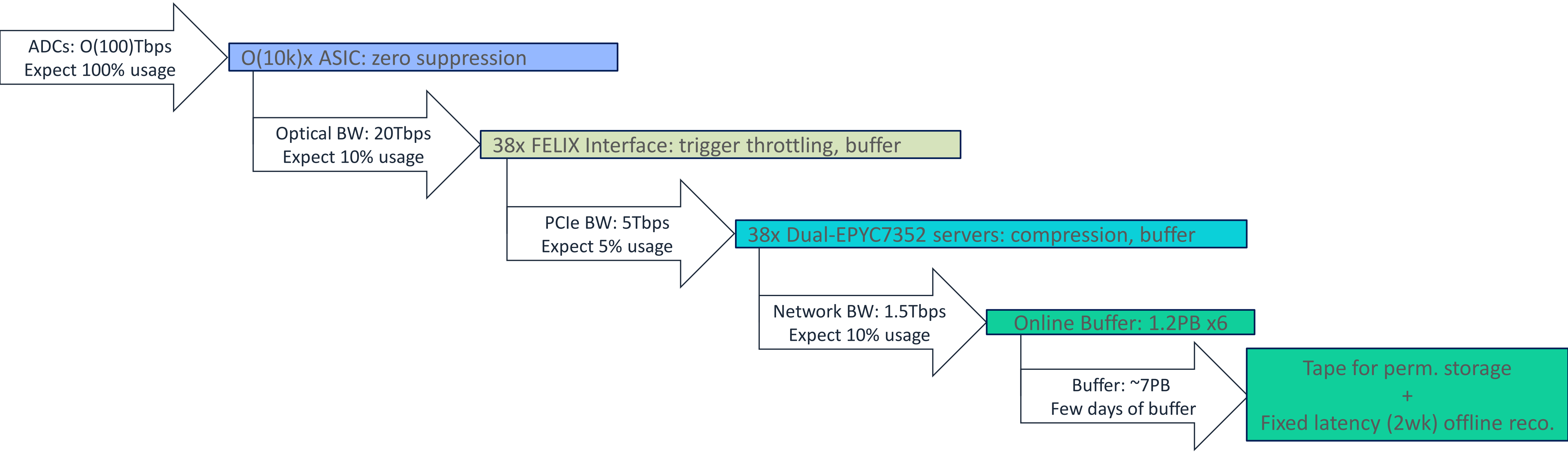


BNL-712 / FELIX v2 x38 (ATLAS/sPHENIX)

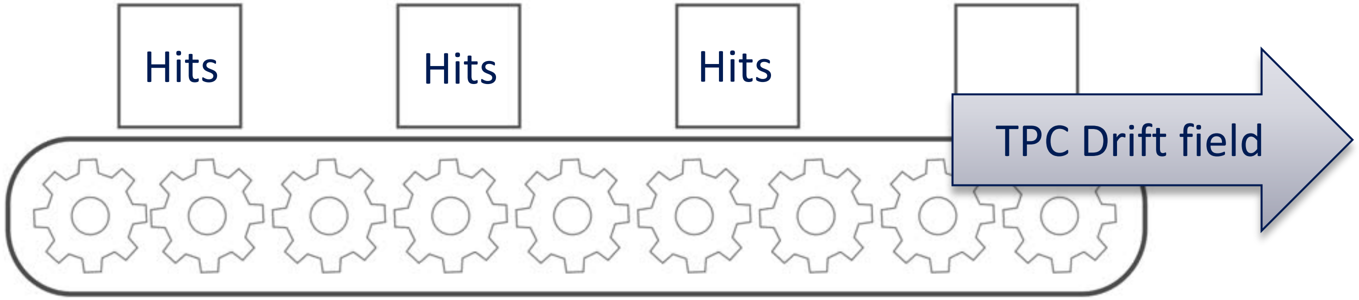
FELIX Ref: [10.1109/tim.2019.2947972](https://doi.org/10.1109/tim.2019.2947972)
Developed at BNL by Omega group!

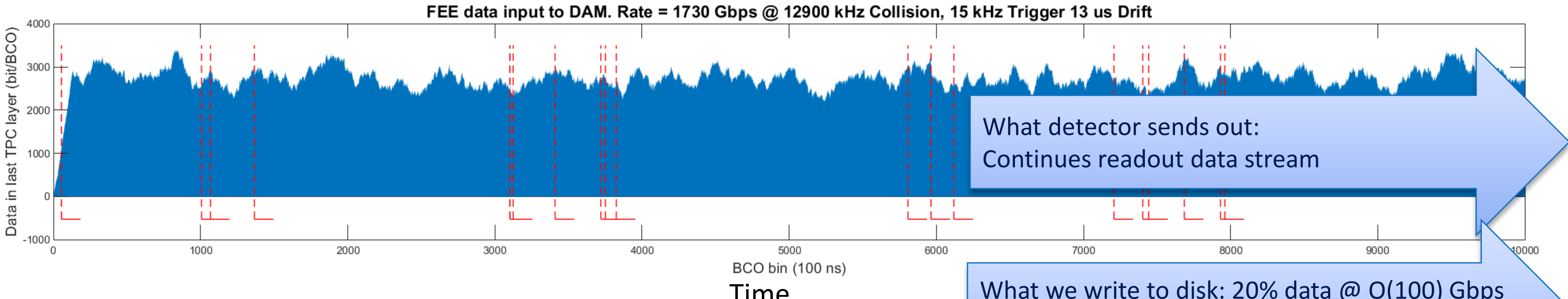
sPHENIX Streaming data flow

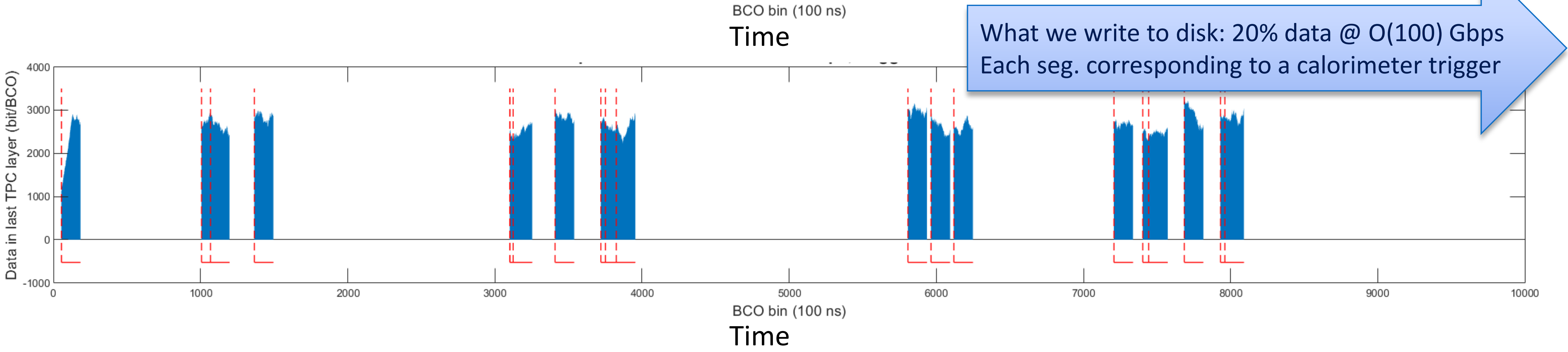




TPC data stream in sPHENIX triggered DAQ

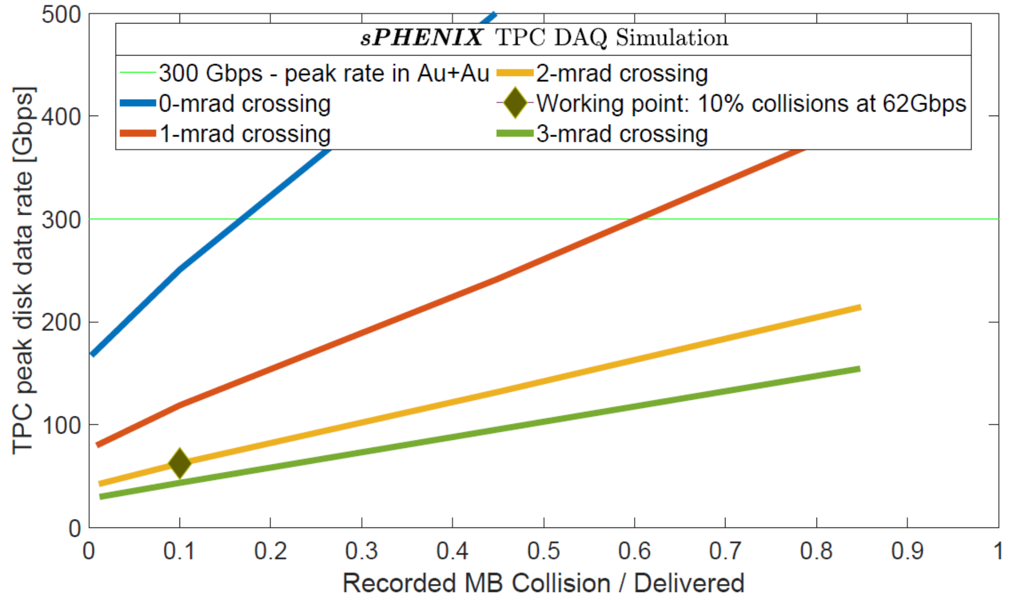






Streaming readout status at sPHENIX

- 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!

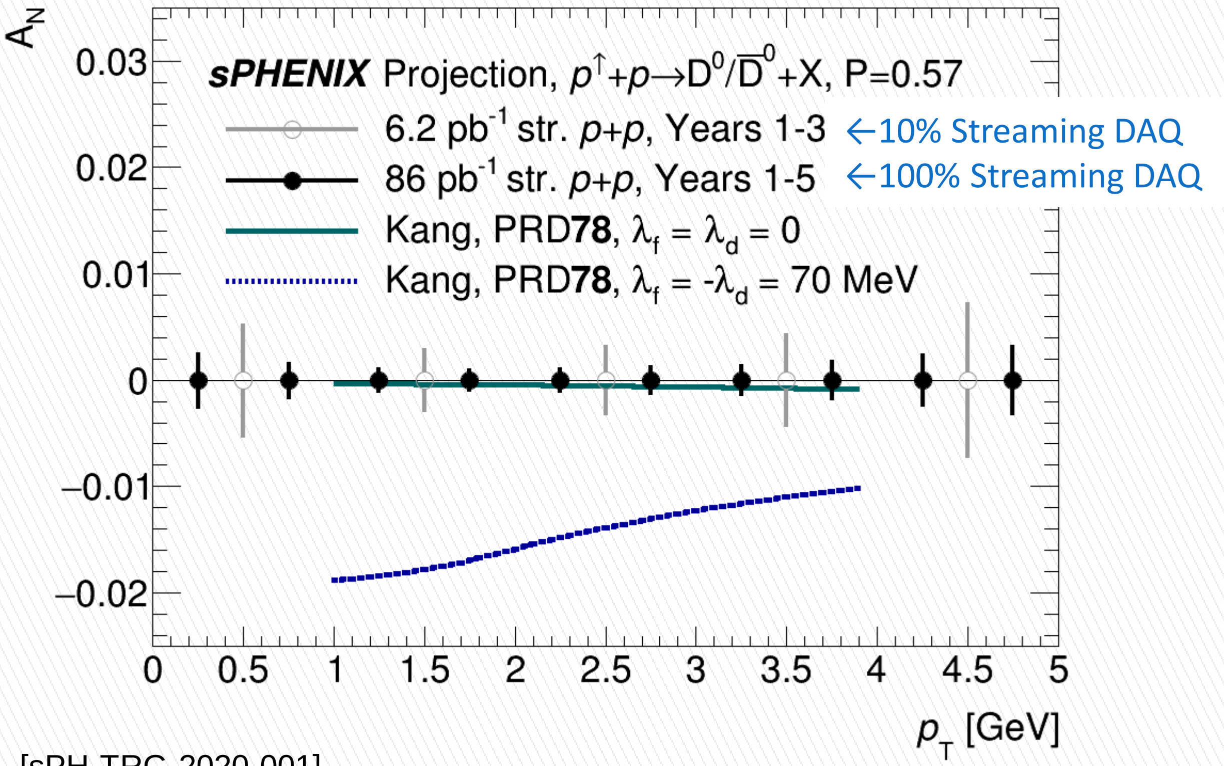


RHIC PAC 2020 report

> 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

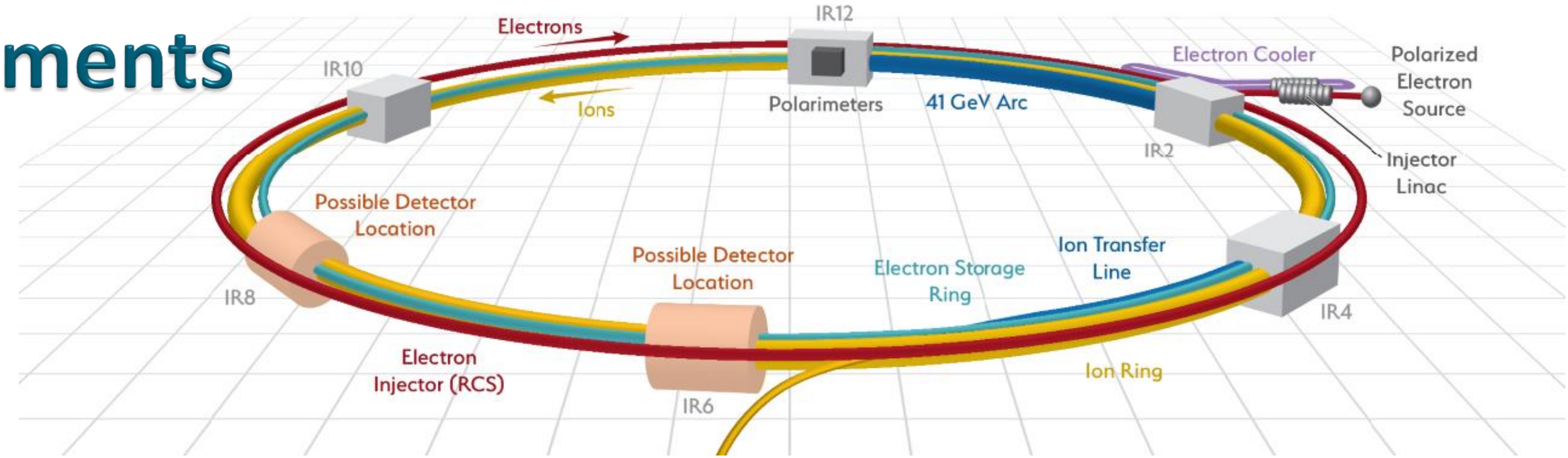
sPHENIX D^0 trans. spin asymmetry, A_N → Gluon Sievers via tri-g cor.

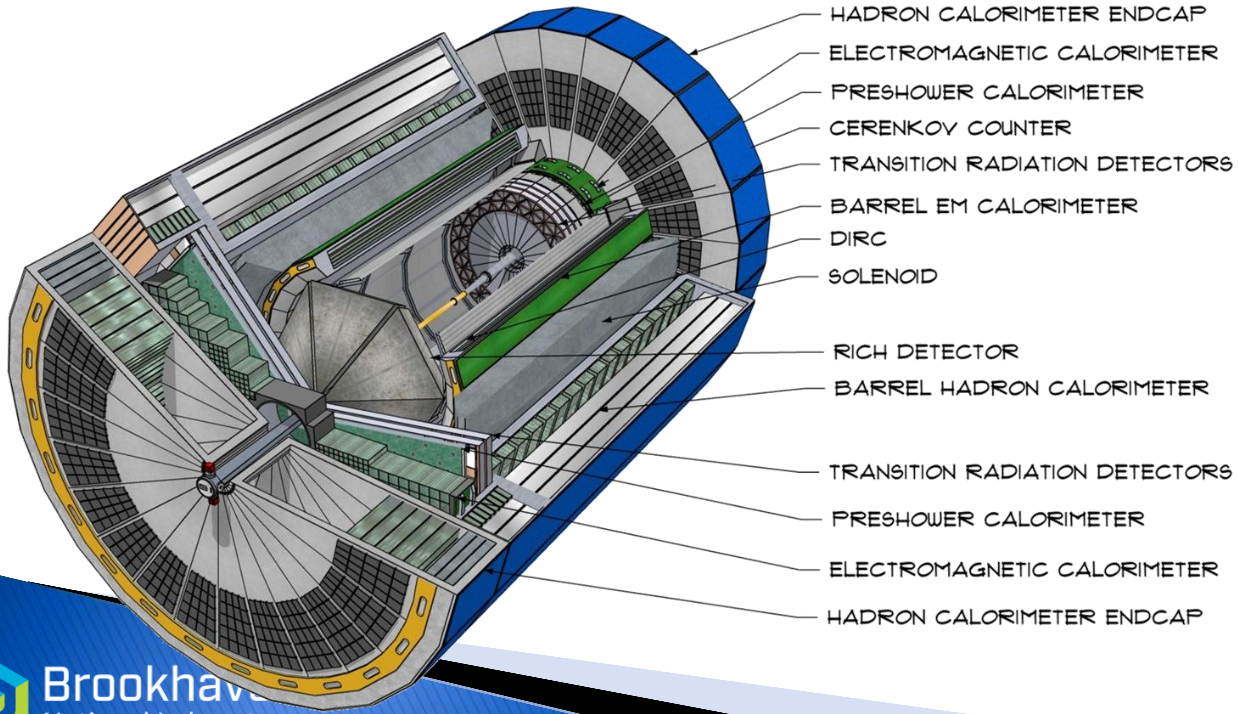


[sPH-TRG-2020-001]
Model: 10.1103/PhysRevD.78.114013

- 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

EIC experiments





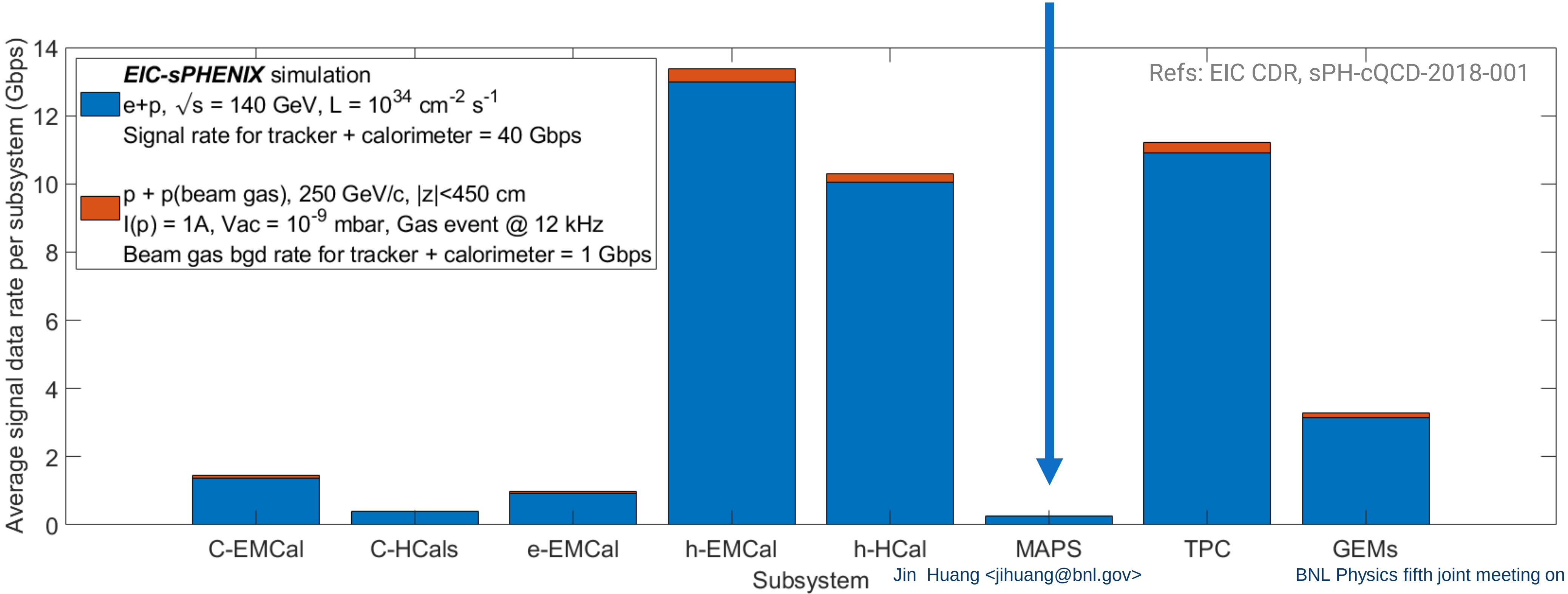
Generic EIC Detector model [EIC YR]

See also: proposal detectors

- ATHENA: athena-eic.org
- CORE: eic.jlab.org/core
- ECCE: ecce-eic.org

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)

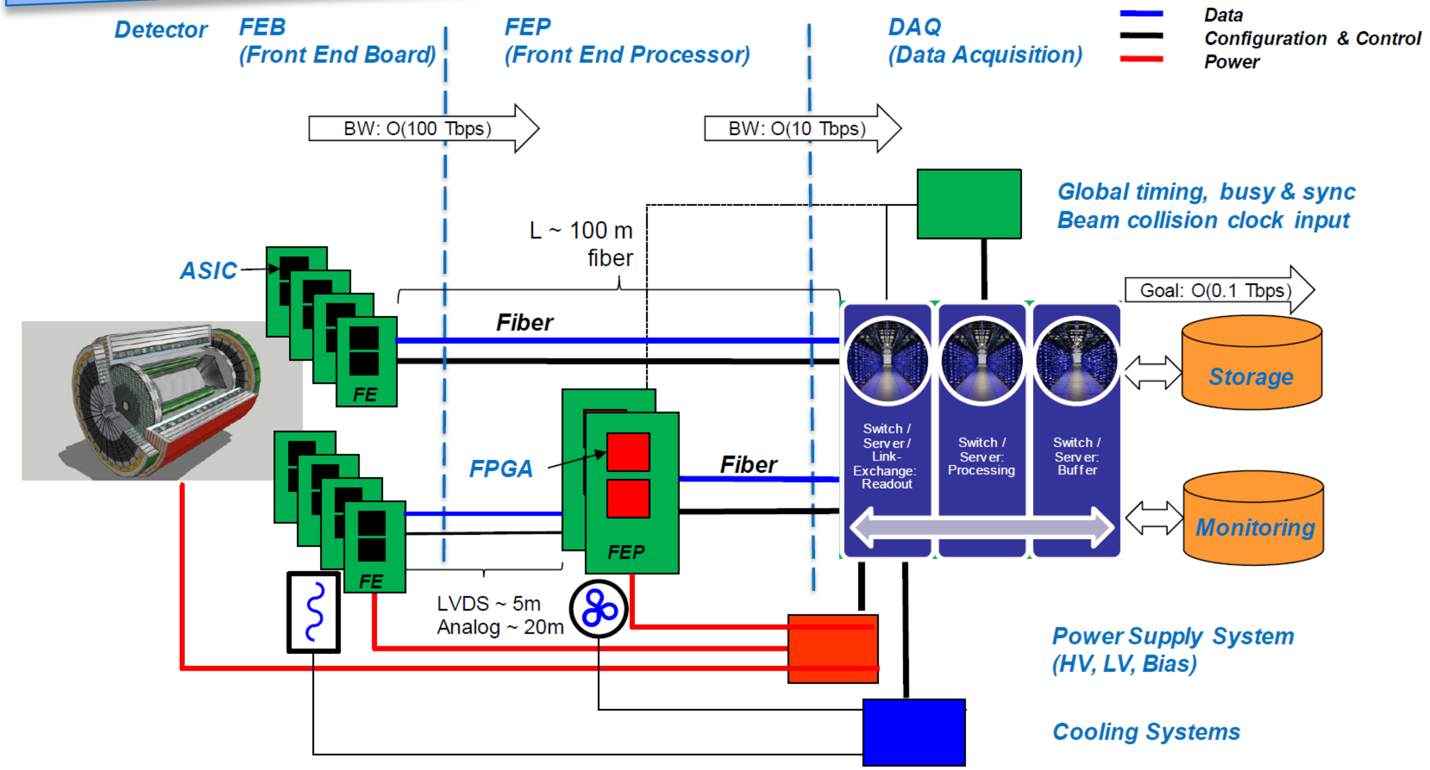


Strategy for an EIC real-time system

ASIC, FPGA — FPGA, TPU/IPU, GPU, CPU

Digitizer BW: 10^2 Tbps → Readout BW: 10^2 Tbps → Storage: 10^{-1} Tbps

- **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+)
- An essential job of EIC real-time computing: reliable streaming data reduction to fit permanent storage



Ref: EIC-CDR

Physics driven need for Real-time AI applications for sPHENIX and EIC

- Both RHIC and EIC has much lower collision signal data rate comparing to LHC
  - But background is important and can be dominating
  - We DO NOT want to drop any event for systematic uncertainty control and broader physics interests
- Opportunities for Real-time AI, e.g.
  - Lossy compression of data, noise filtering (LDRD 19-028)
  - Feature extraction: Energy time extraction from ADC time-series (LDRD 21-023, NPPS)
  - Feature extraction: tracking, vertexing, HF signal selection (sPHENIX ML Open Data [GitHub]→SBIRs→sPHENIX demo. See seminar Feb 1st D.T. Yu [link])

Lossy compression and noise filtering

- Team members, supported under LDRD 19-028
  - Yi Huang (CSI)
  - Thomas Marshall (UCLA)
  - Yihui Ren (CSI)
  - Shinjae Yoo (CSI)
  - Jin Huang (PO)

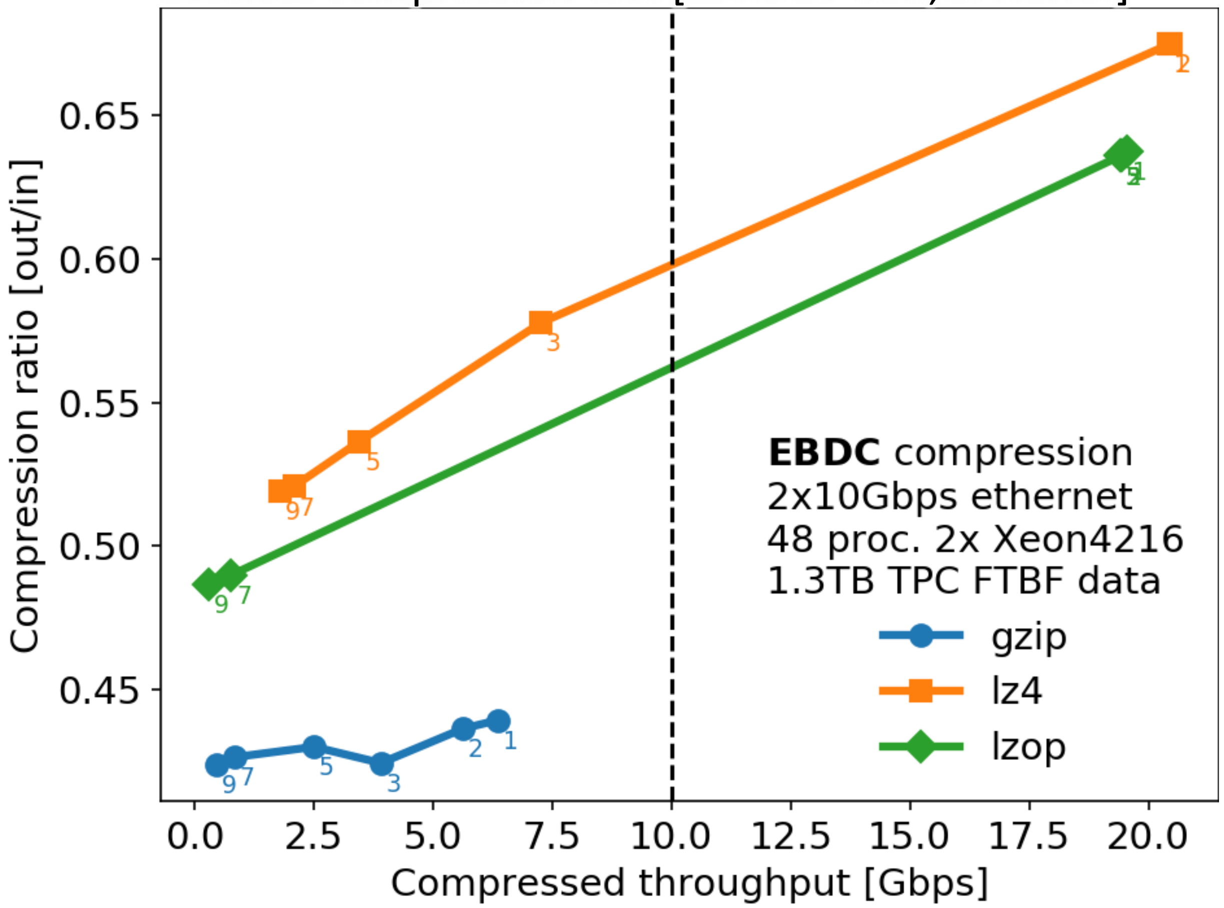
Reference:

- arXiv: 2111.05423, in print
- Yi Huang, IEEE ICMLA2021, AI4EIC, Streaming Readout VIII

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

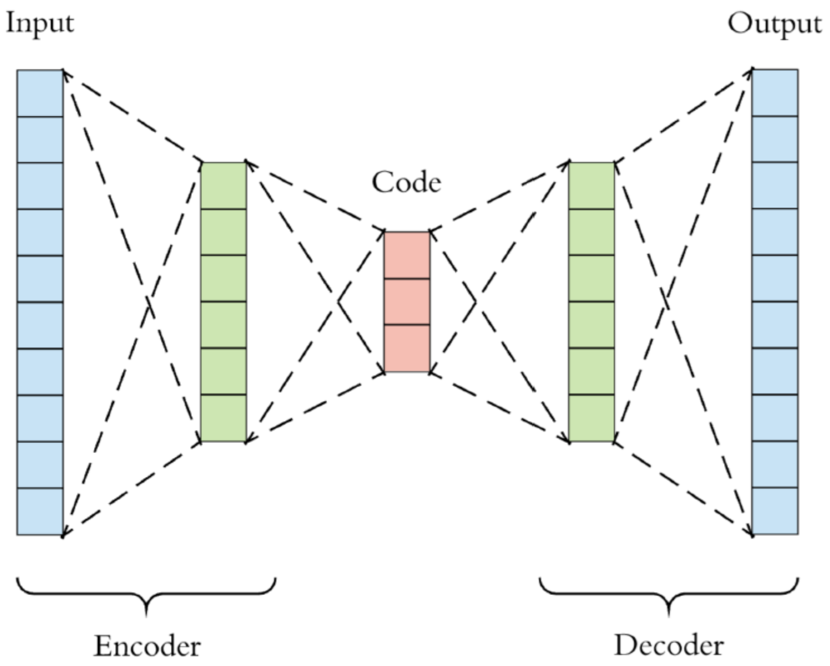
Lossless compression test [LDRD19-028, sPHENIX]



Lossy compression of data, noise filtering

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

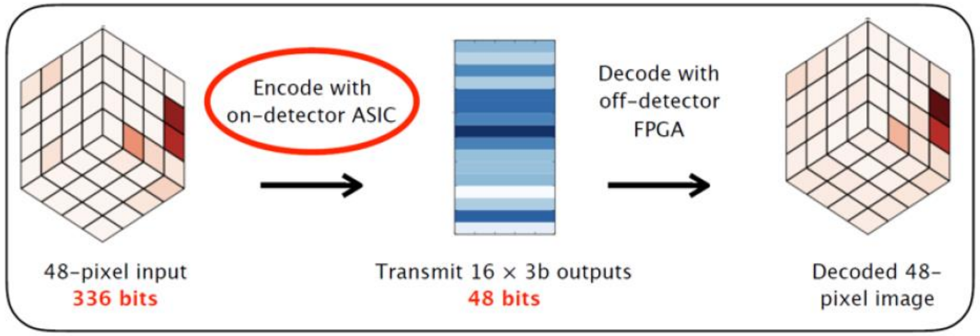
Simple auto-encode neural network



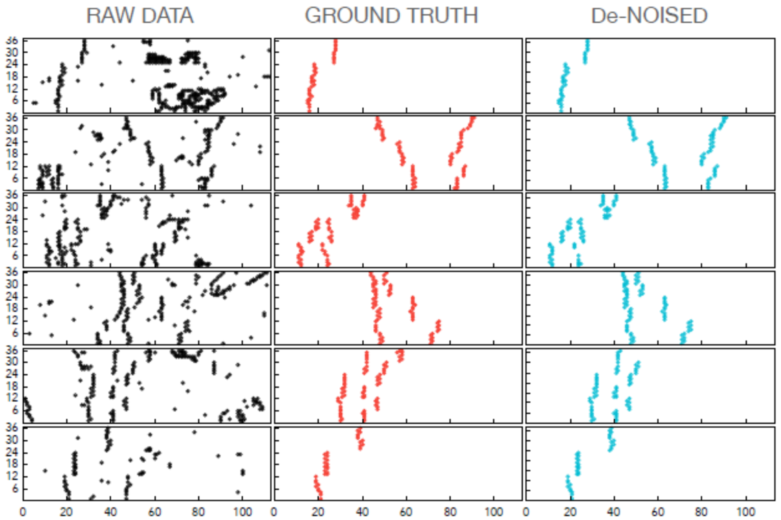
Lossy compression of data, noise filtering

- 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
- We are not alone in this research: see also in CMS HGCal ASIC, CLAS12 tracker offline reco.

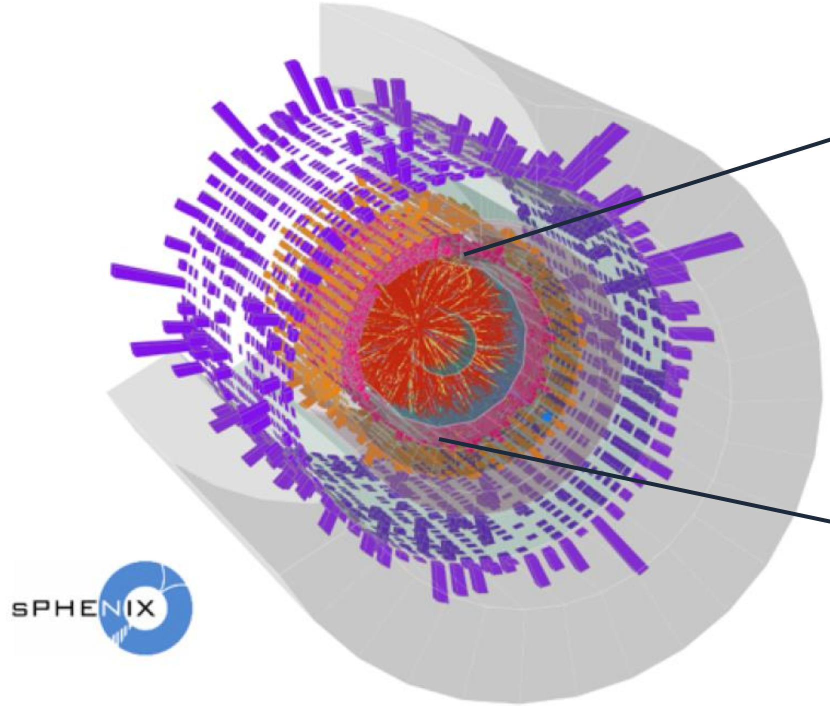
CMS HGCal compression ASIC, [10.1109/TNS.2021.3087100]



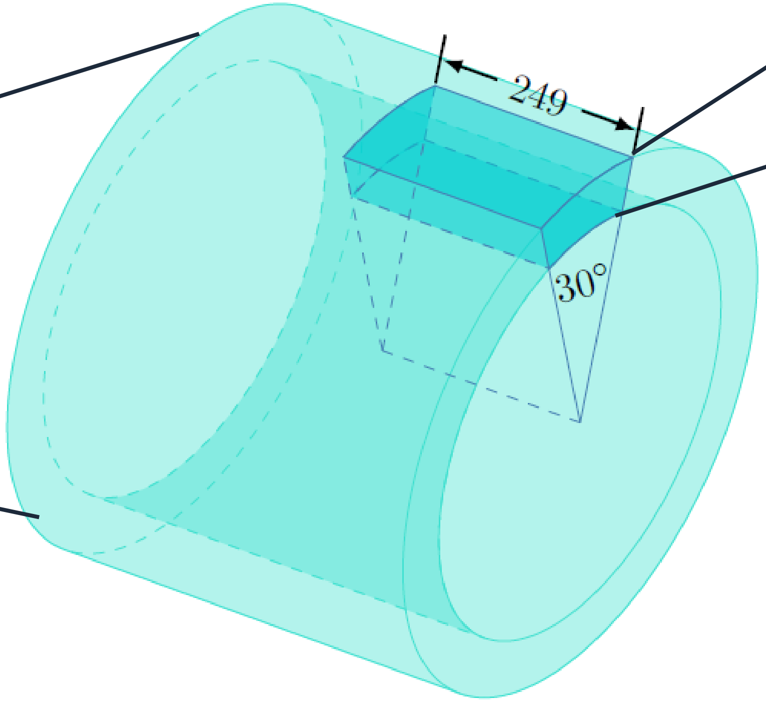
CLAS12 Drift Chamber offline AE de-noise [link]

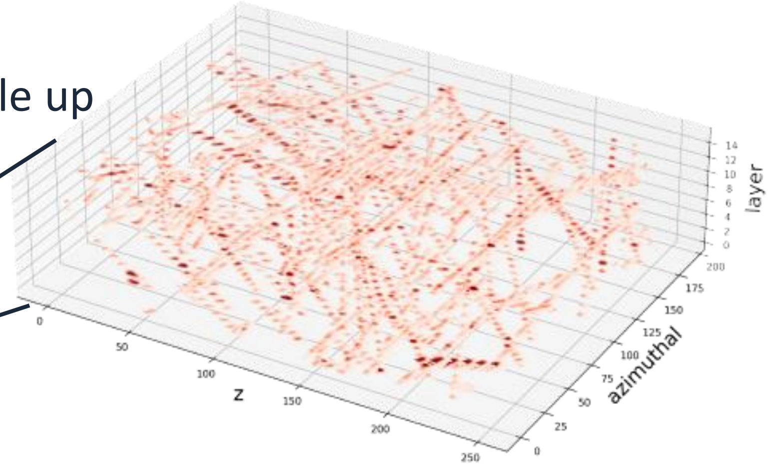


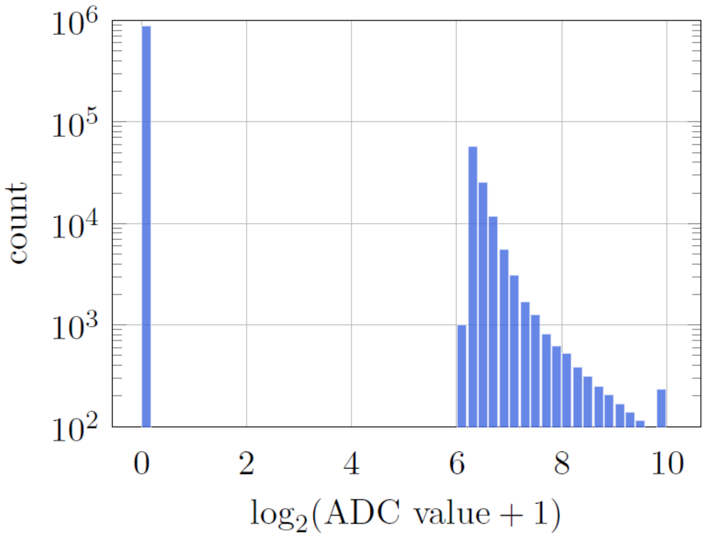
Data of time projection tracker at sPHENIX



Busiest event in sPHENIX TPC
10% central Au + Au collision with 170kHz pile up
Data frame for 1/12 azimuth sector



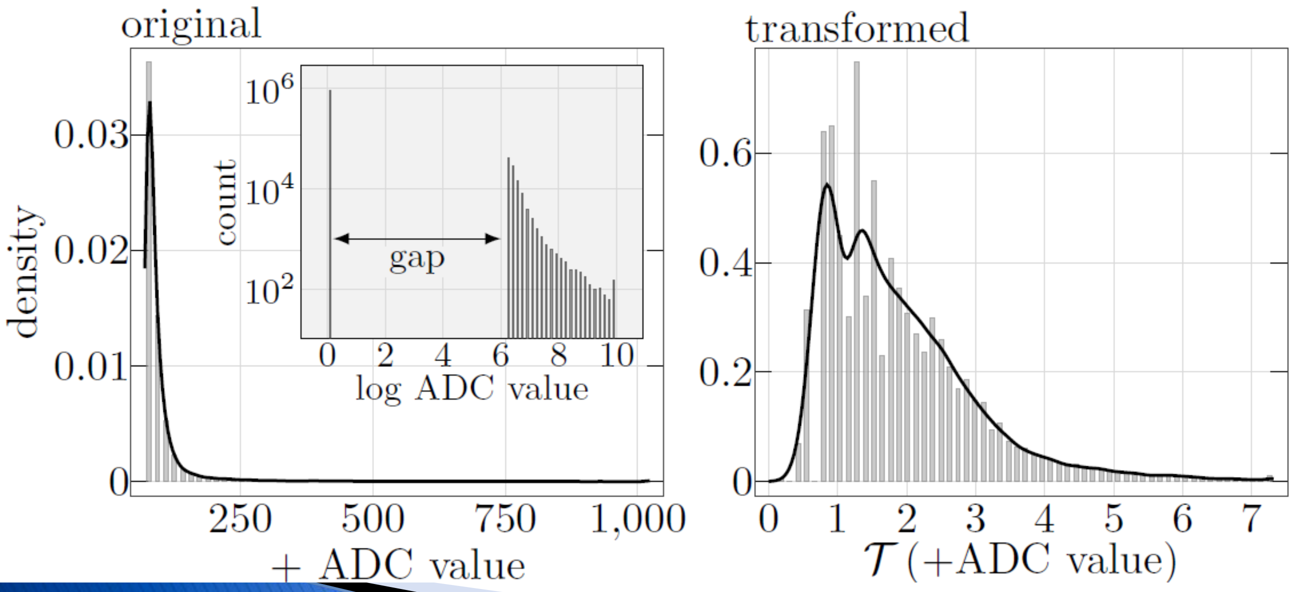


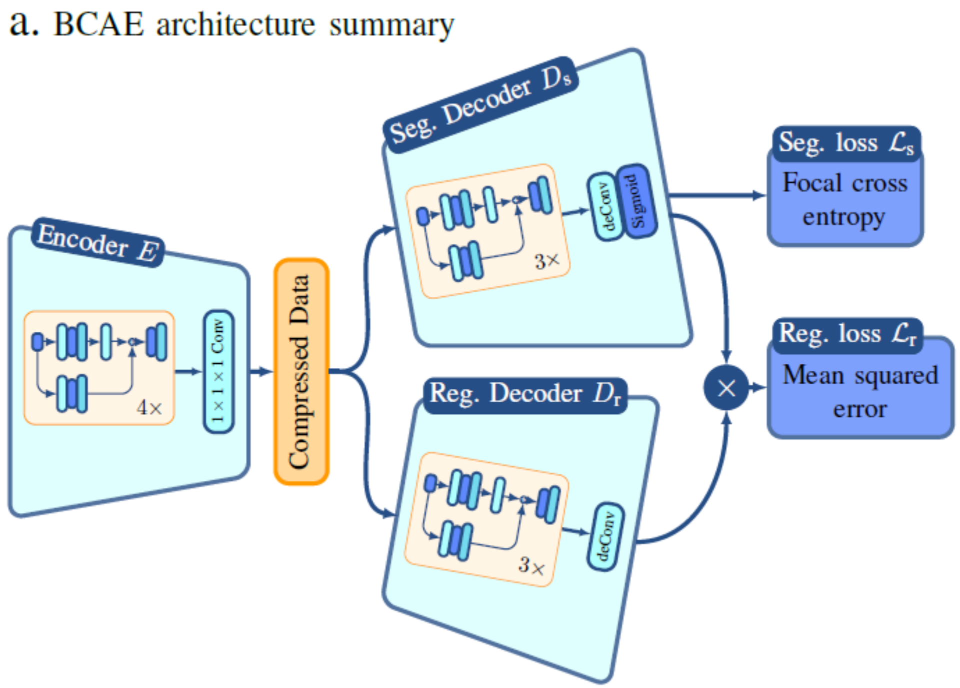


Bicephalous Convolutional Auto-Encoder (BCAE) and input transform [arXiv:2111.05423]

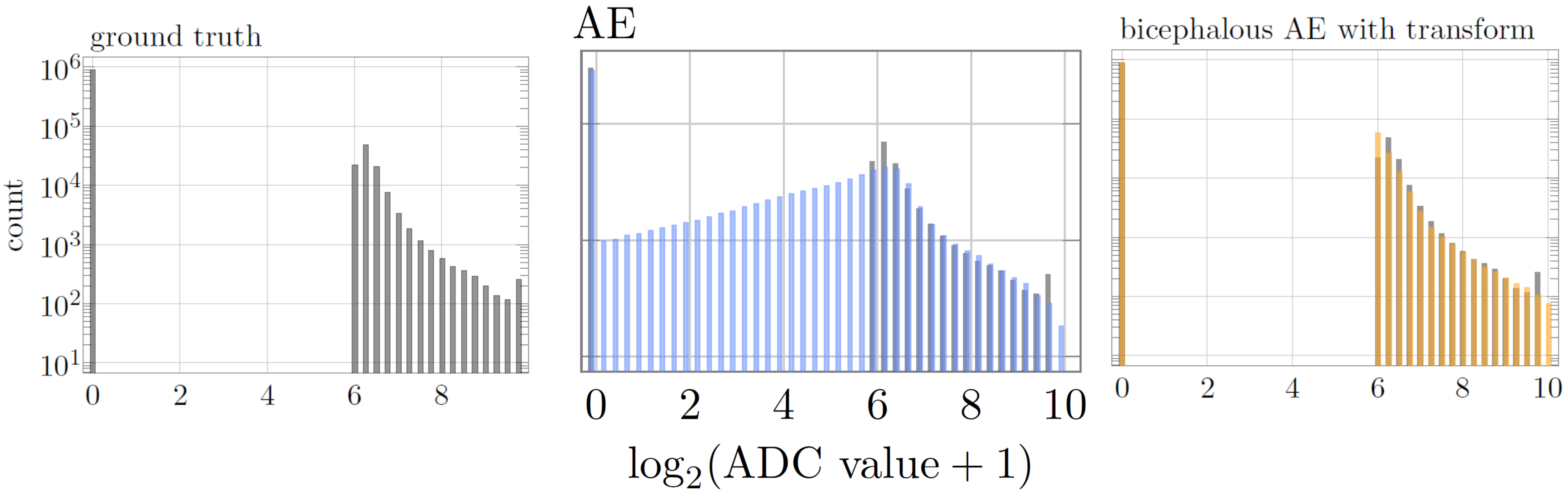
- Input transform: fill in the zero-suppression gap and make ADC distribution much less steep
- Bicephalous decoder: +classification decoder to note the zero-suppressed ADC bins (unsupervised training) and +noise bins (supervised training)

$$\text{Input transform: } \mathcal{T}(x) = \log(x-64)/6, \qquad x > 64$$
$$\text{Inverse transform: } \mathcal{T}^{-1}(y) = 64 + \exp(6y), \qquad x \in \mathbb{R}$$

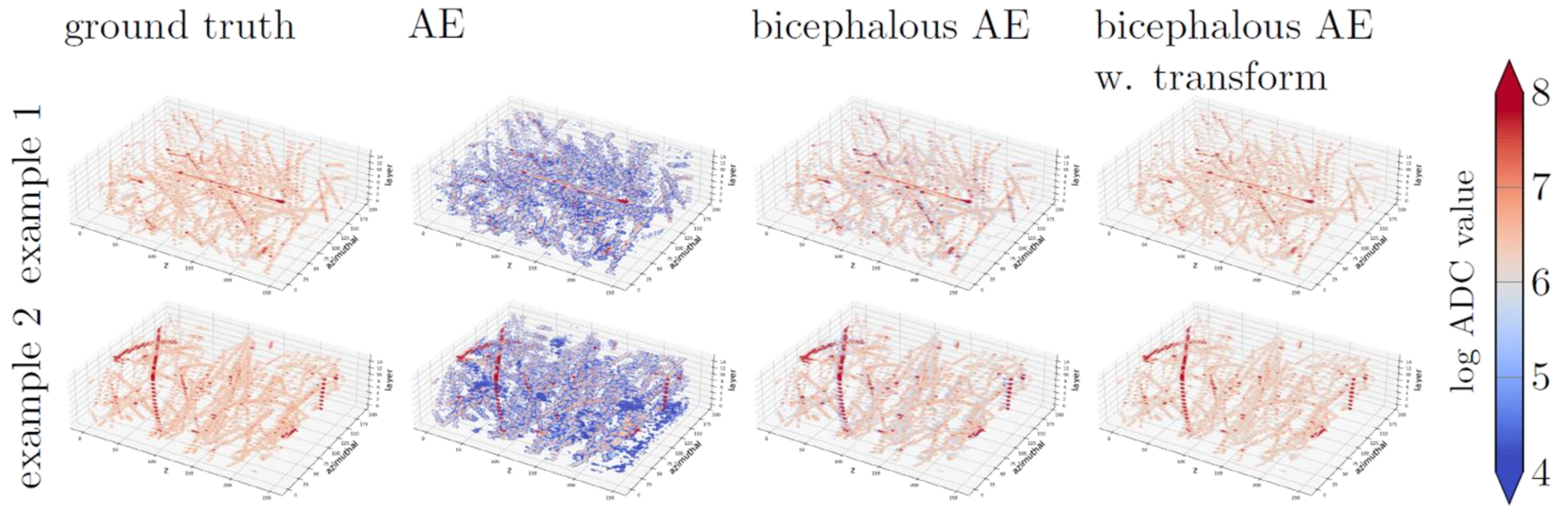




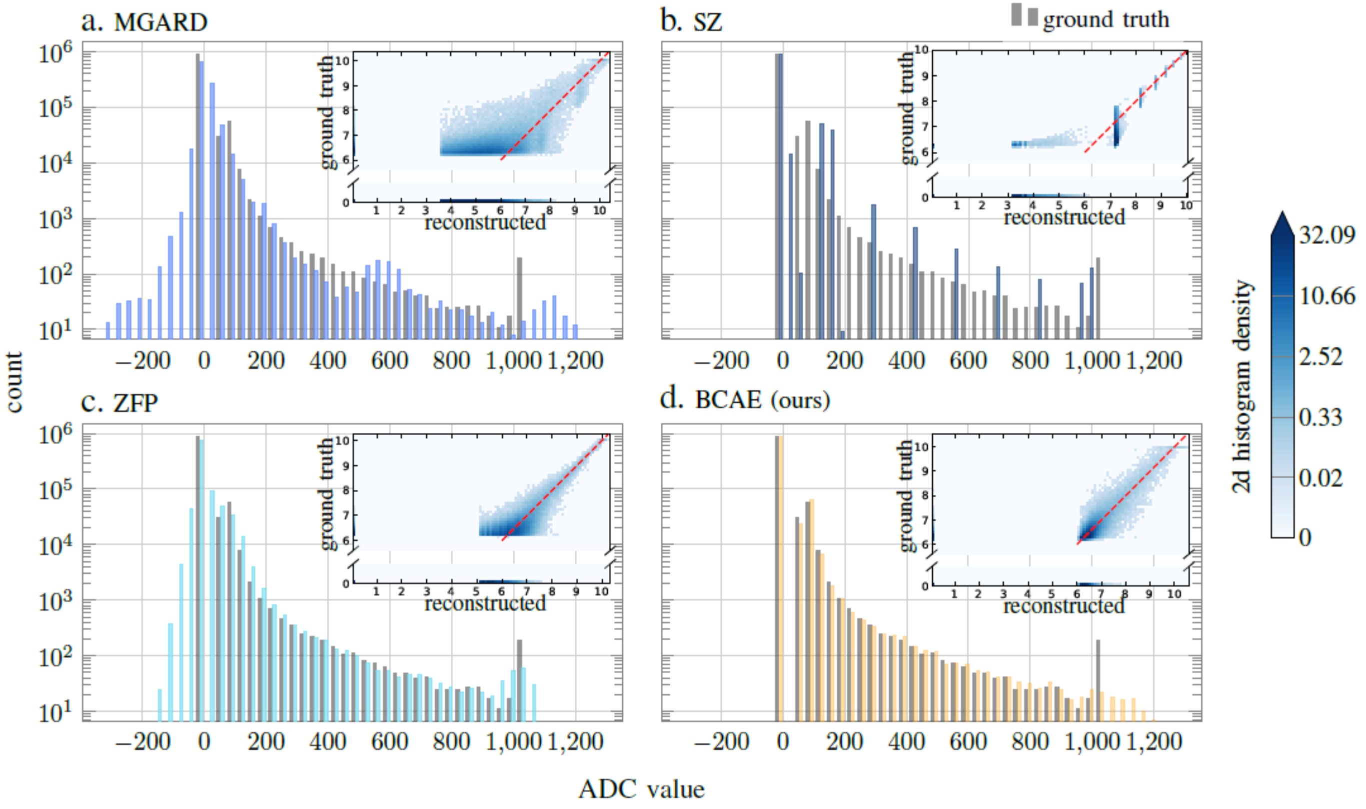
Results from Bicephalous AE with transform [arXiv:2111.05423]

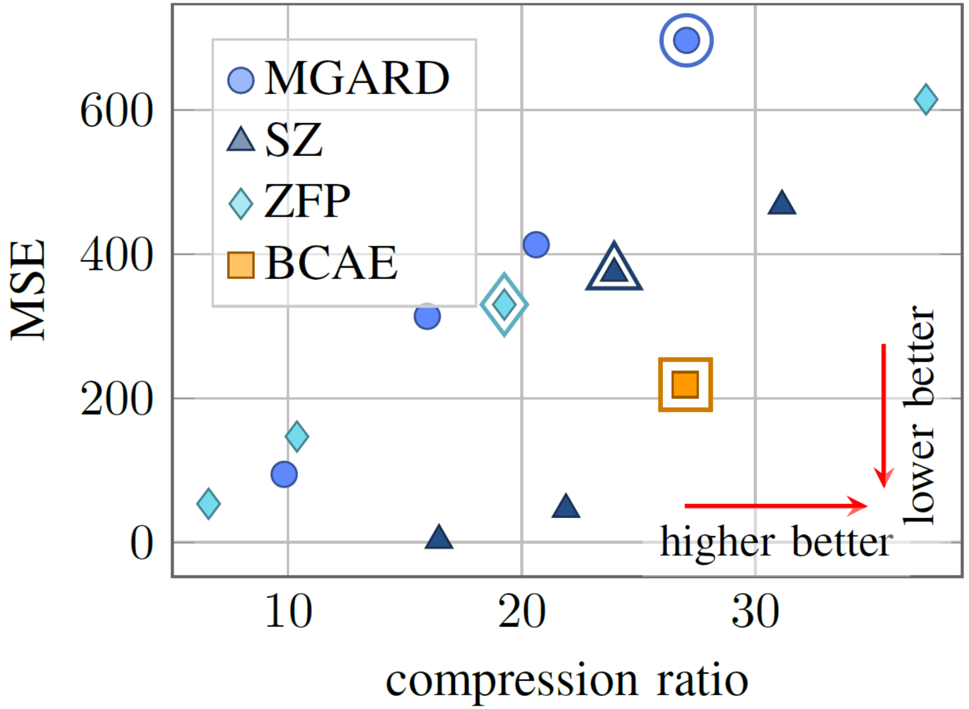


Results from Bicephalous AE with transform [arXiv:2111.05423]



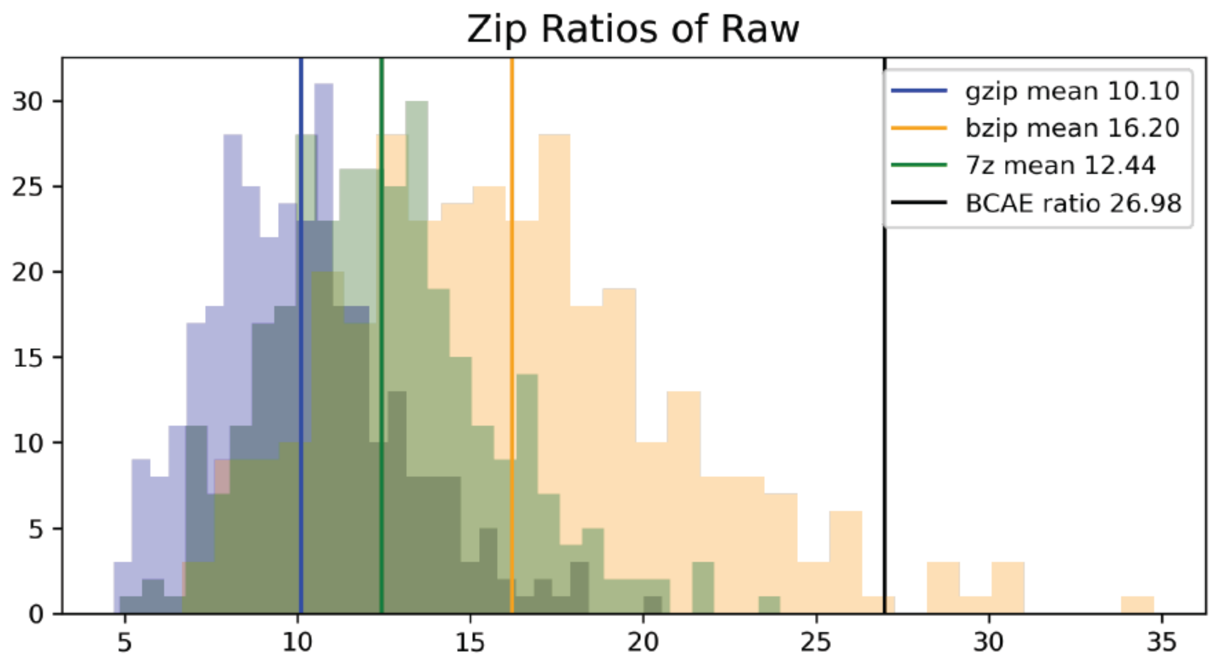
Comparison with existing algorithm [arXiv:2111.05423]

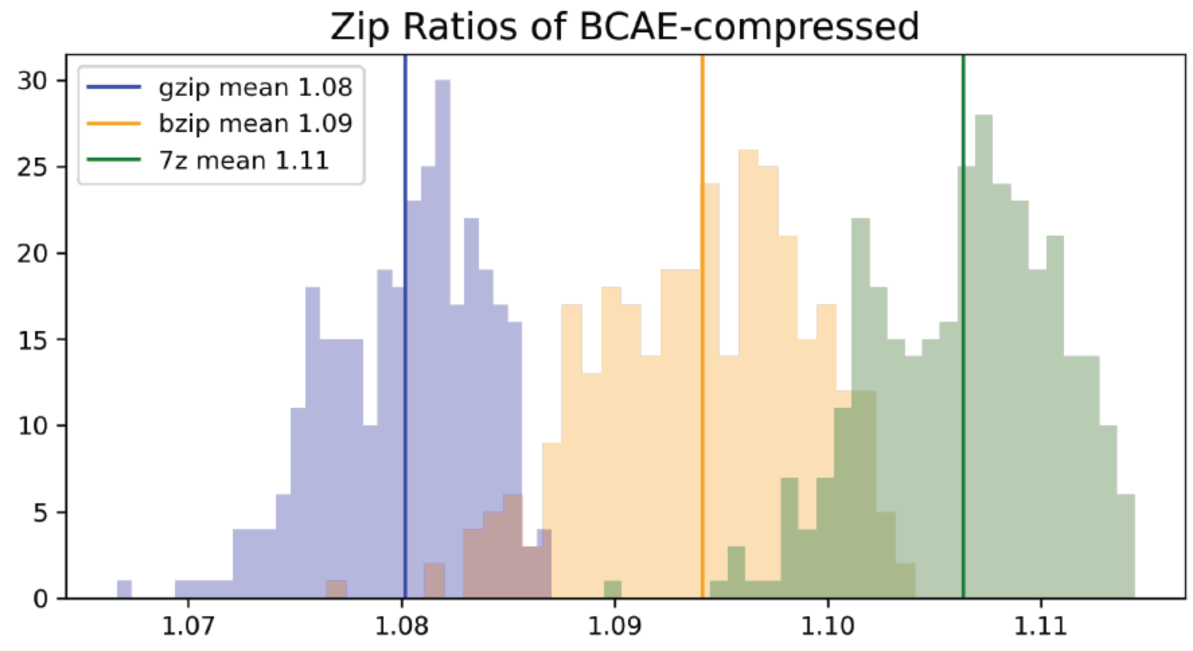




Compressibility check: thanks to suggestion from Brett!

- The lossy-compressed code is hardly compressible further losslessly

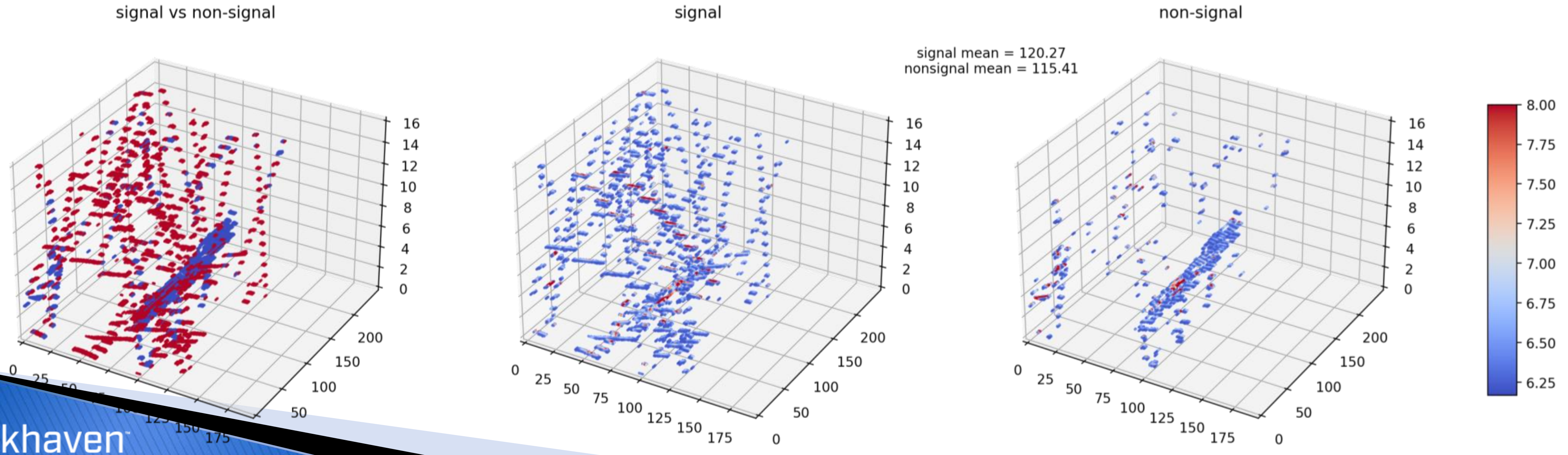




On-going research

- Supervised learning for noise filtering [on-going work by Thomas Marshall (UCLA), Yi Huang(CSI)]
- Throughput demonstration and optimizations on GPU servers (A6000, DGX-2 and DGX-A100)
- Downstream integration and performance/bias evaluation with physics observables, e.g. D0->piK
- Test resilience with TPC distortion, event background, and 2023 real data
- Exploring AI-optimized hardware: e.g. Intelligence Processing Unit (IPUs)

sPHENIX MDC2 data, TPC R3-single sector, pp->D(piK) +X √s=200 GeV with 3MHz pile up



Feature extraction: ADC time series →amplitude, time-of-arrival

- Team members, supported under LDRD 21-023
  - Sandeep Miryala (IO)
  - Sandeep Mittal (CSI)
  - Gabriella Carini (IO)
  - Grzegorz Deptuch (FNAL)
  - Sioan Zohar (IO)
  - Jack Fried (IO)
  - Shinjae Yoo (CSI)
  - Jin Huang (PO)

Reference:

- Sandeep Miryala, CPAD21, 22nd iWoRiD
- JINST, in press

Also work by

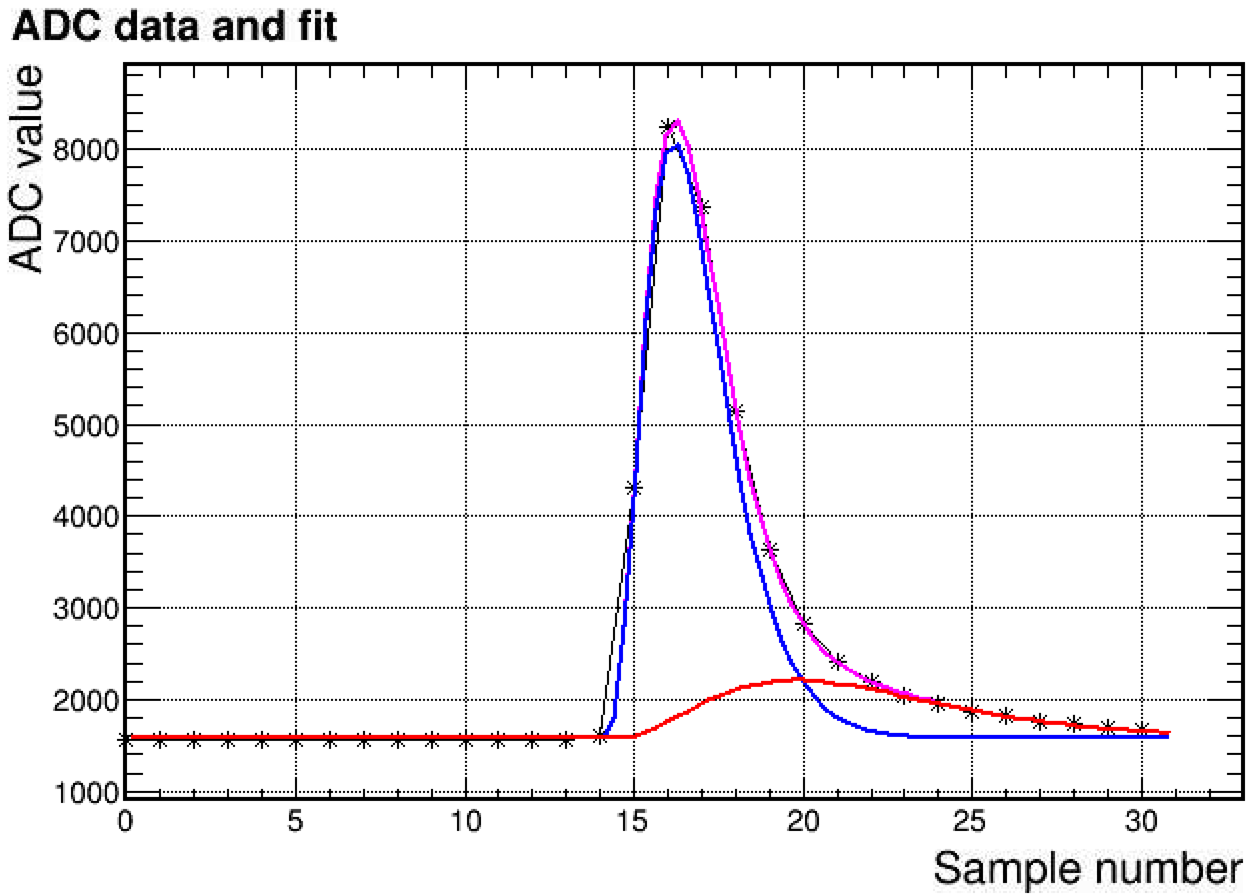
- Maxim Potekhin (NPPS)
- Tim Rinn (sPHENIX)

ADC time series, and reduction based on feature building

- Both sPHENIX and EIC calorimeter will be digitized continuously with FADCs
- An efficient way of storing the information is feature of pulse: amplitude and time of arrival, shape information
- Application of regression with MLP/CNN

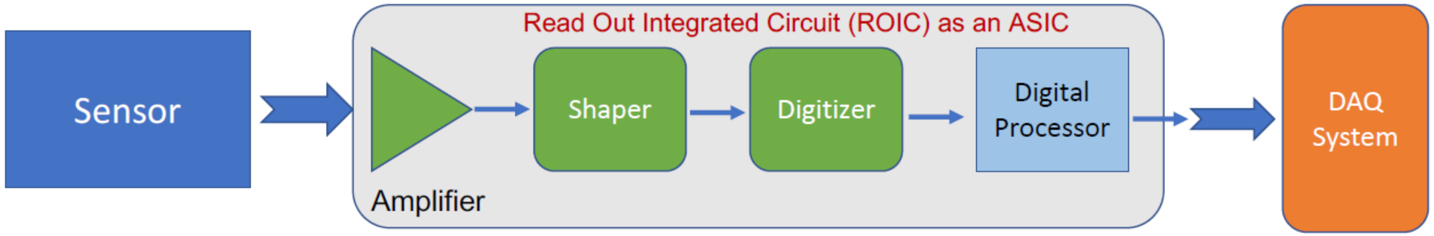
sPHENIX calorimeter test beam data:

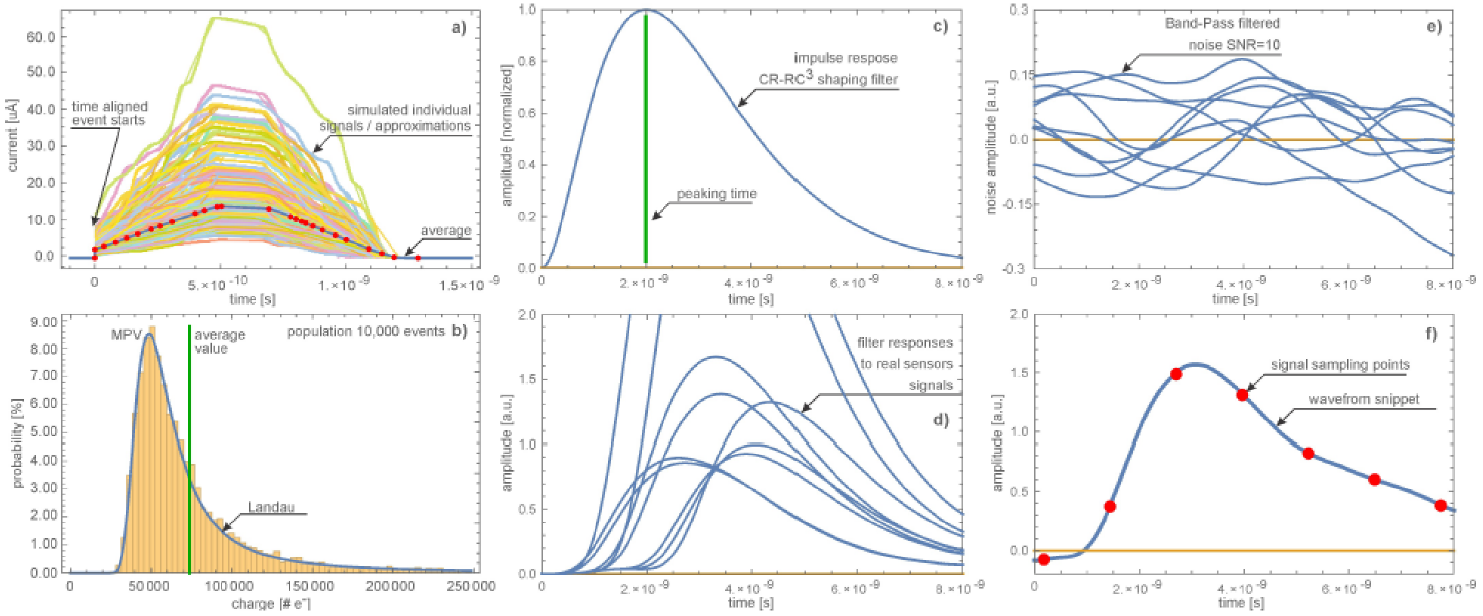
- 2016 data: [10.1109/TNS.2018.2879047](https://doi.org/10.1109/TNS.2018.2879047)
- 2018 data: [10.1109/TNS.2020.3034643](https://doi.org/10.1109/TNS.2020.3034643)



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

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





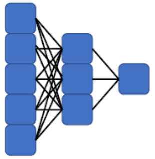
Network selection

[LDRD 21-023, JINST in press]

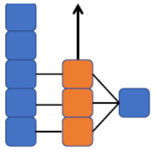
- 18 sample at 2GHz,
- 16 bit input /output
- Tested MLP and CNN
- Various sizes

5-FC layer

18x8, (8x8)^3, (8x1)



- 3-Conv1D + 2-FC layer
- Conv^3, flatten, x16, (16x1)

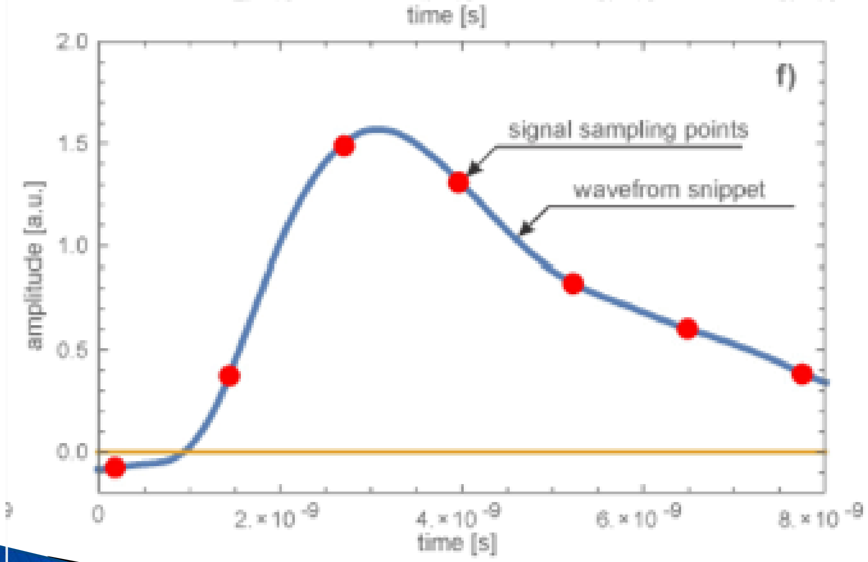


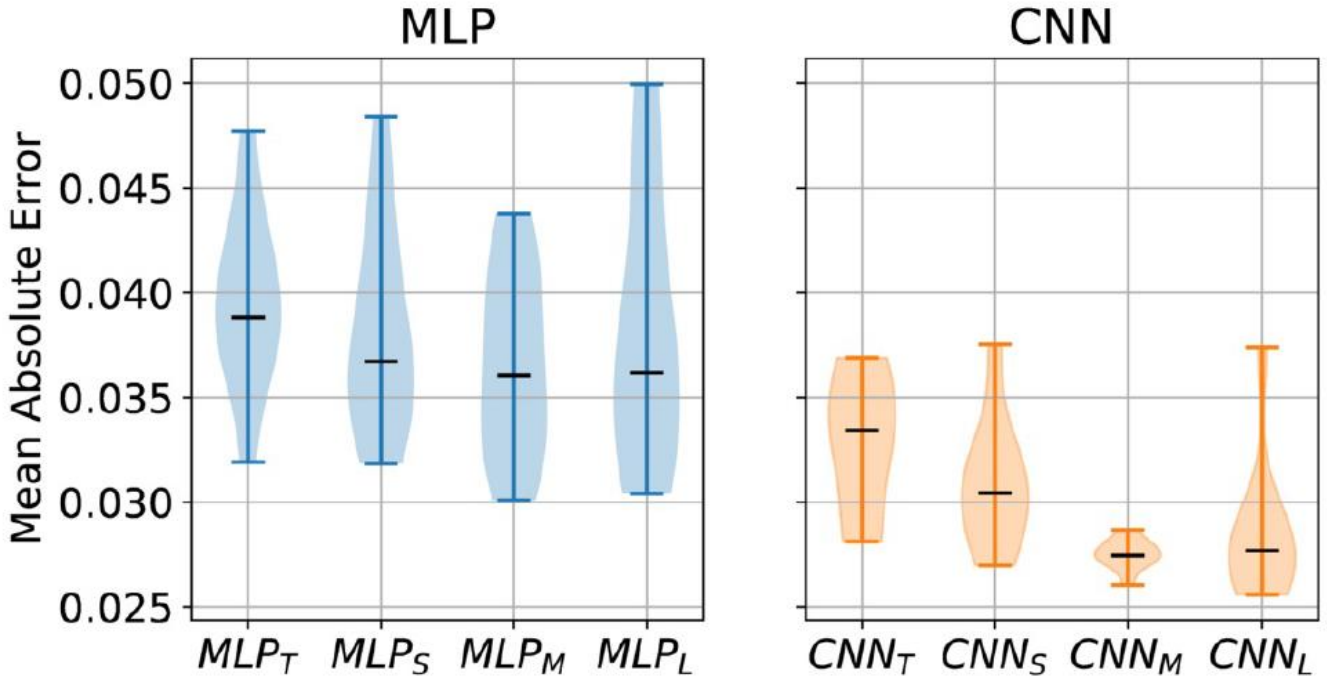
(a) Model Configurations of MLP

MLP	Config.	# Param.
Tiny (T)	8-8-8-8	377
Small (S)	16-16-16-16	1137
Medium (M)	32-32-32-32	3809
Large (L)	52-52-52-52	9309

(b) Model Configurations of CNN

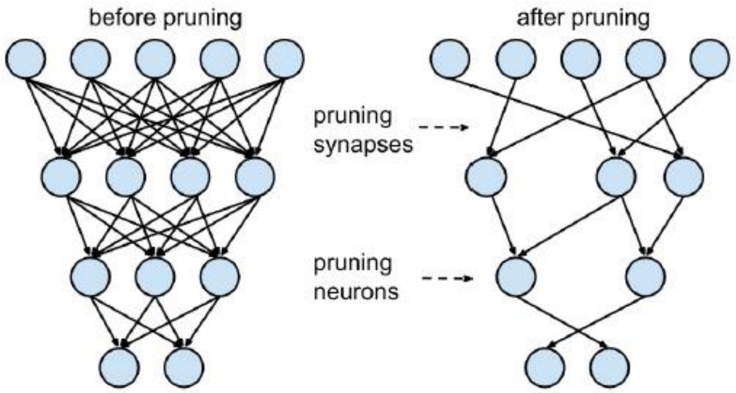
CNN	Config.	# Param.
Tiny (T)	2-2-2-16	453
Small (S)	3-3-3-32	1289
Medium (M)	5-5-5-64	4149
Large (L)	6-6-6-128	9725

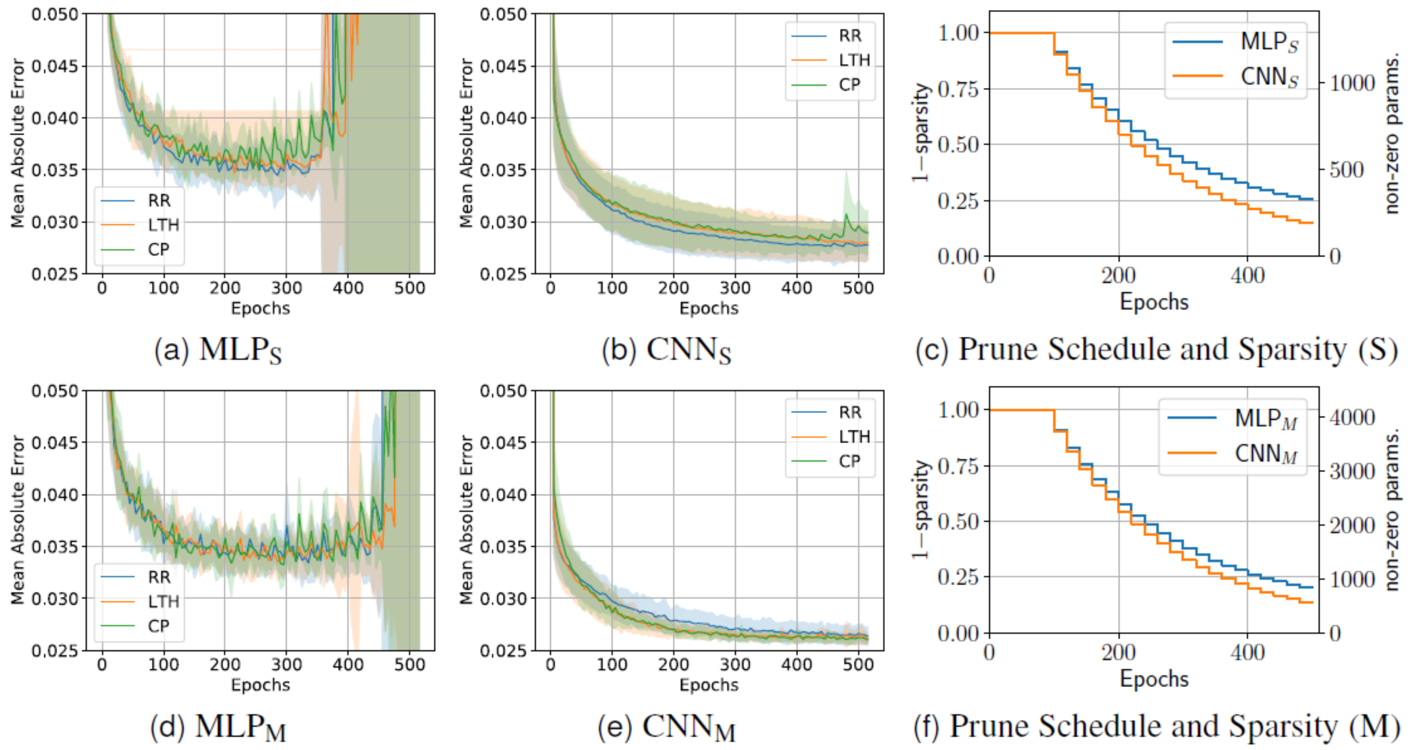




Network pruning

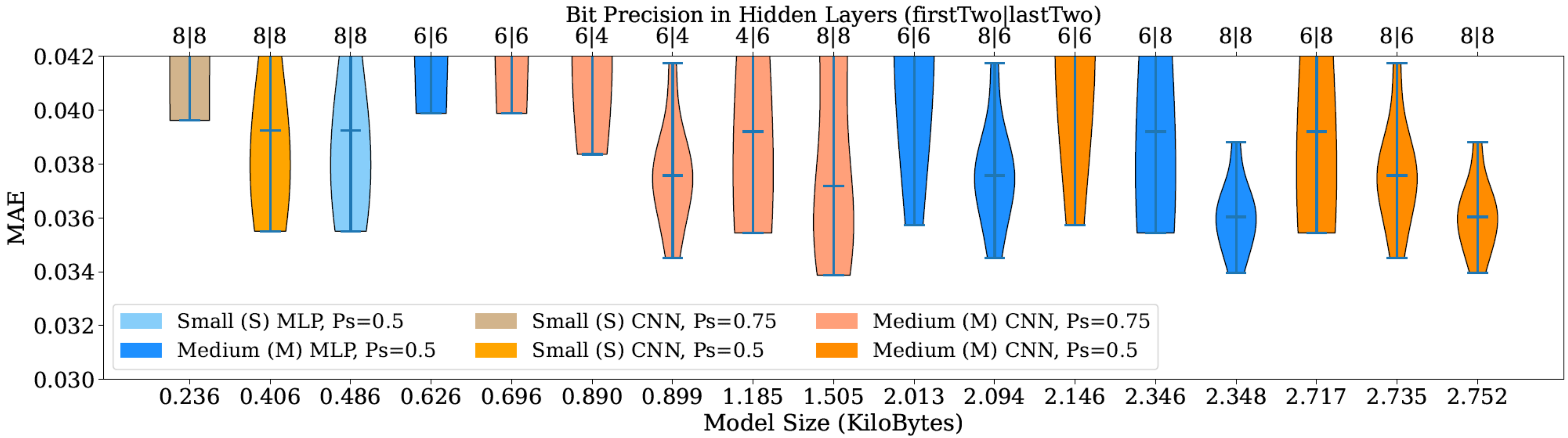
[LDRD 21-023, JINST in press]





Pruning + Variable Bit Quantization-aware Training

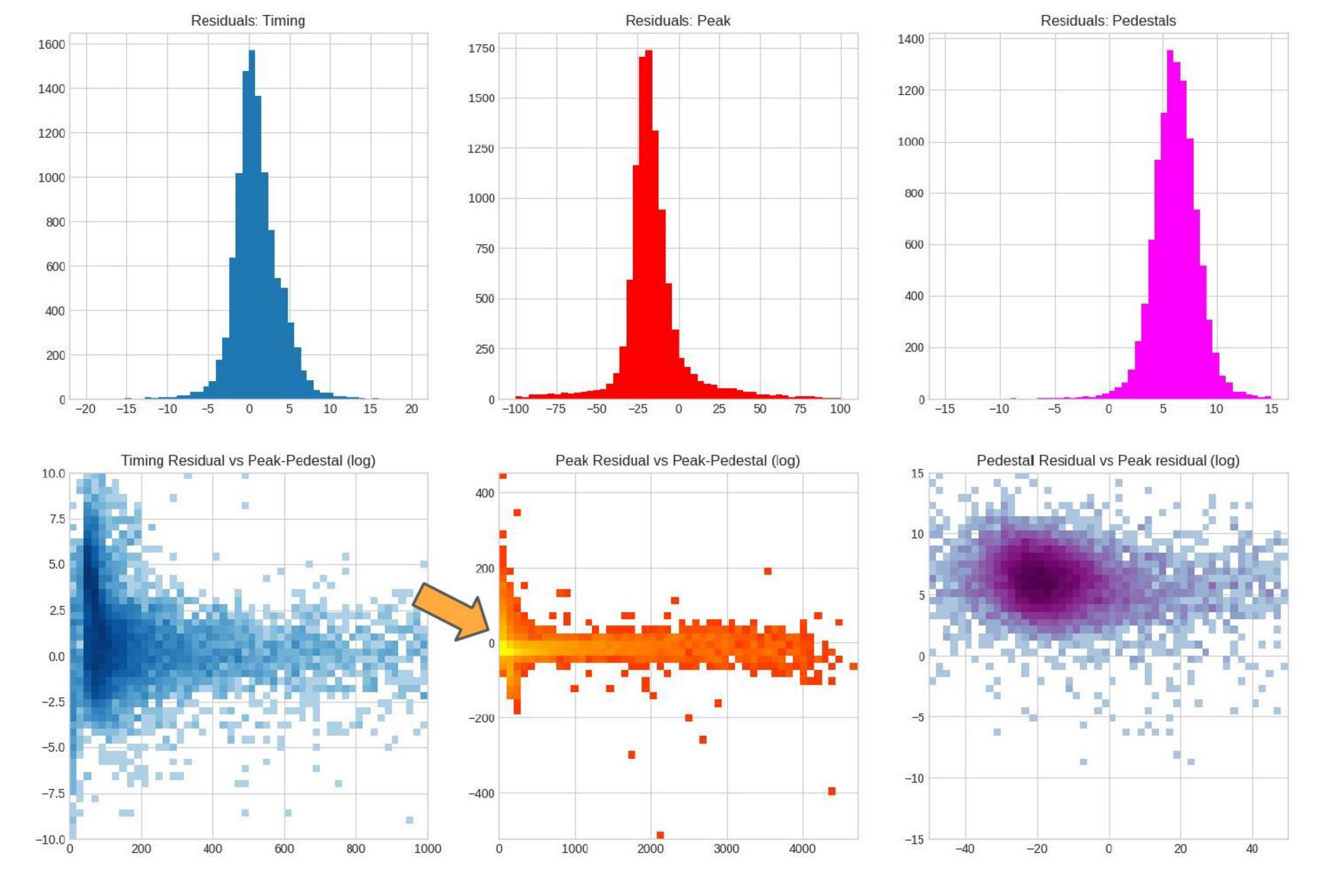
- Highly pruned (sparsity=0.75) CNN with 8bit internal precision appears strike good performance (smaller error) and small model size



On sPHENIX test beam data

- Study by Maxim Potekhin, data from Tim Rinn
- sPHENIX Test beam data
- MLP in Tensor flow
- Orders of magnitude speed up comparing to current iterative fit algorithm

Maxim Potekhin [sPHENIX software meeting, Jan 18, 2022]

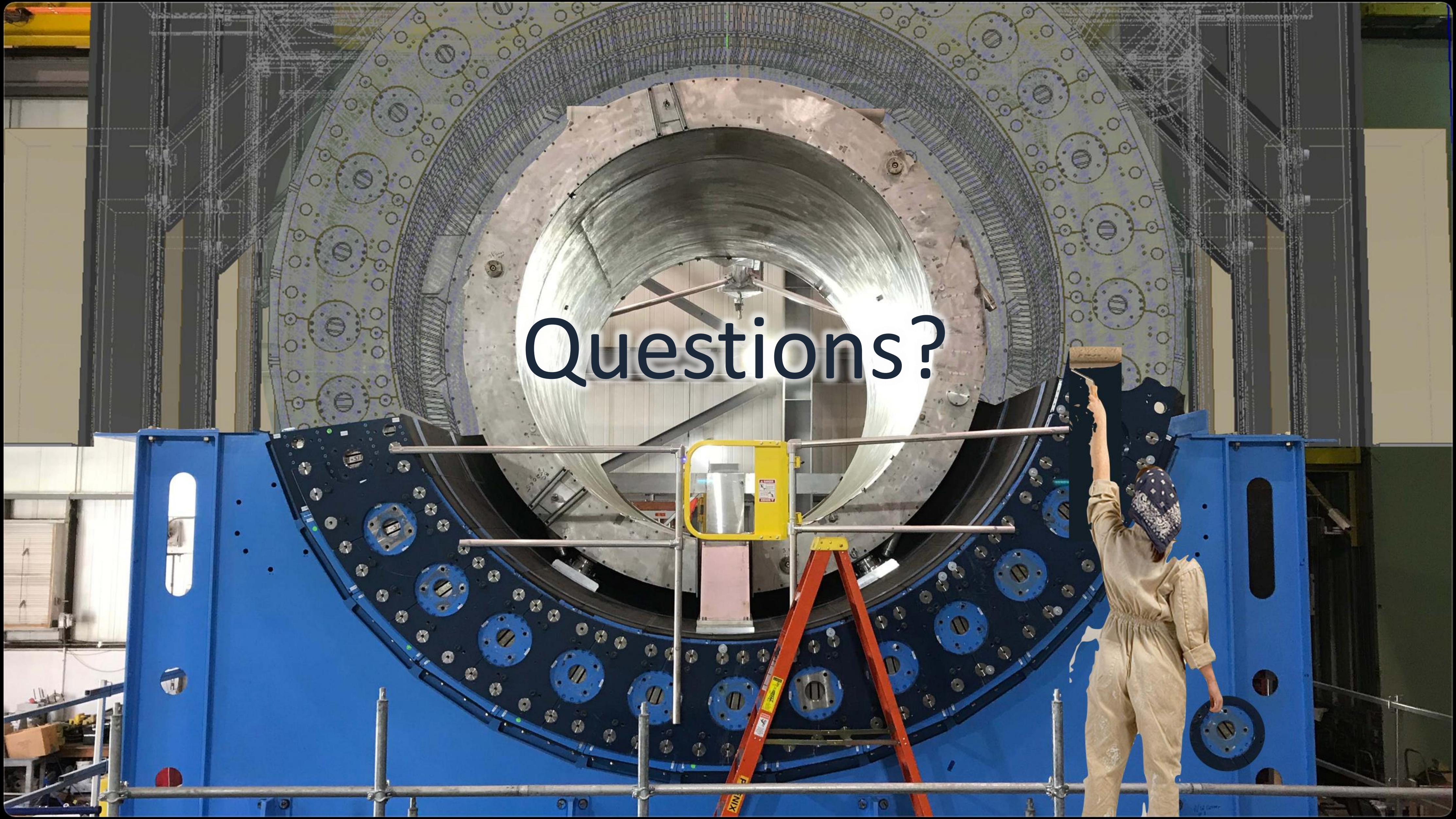


Possible sPHENIX and EIC applications

- Improving calorimetric trigger energy threshold reconstruction (require FPGA application)
- Preserve tower energy for below-threshold towers for jet background estimation
- Realtime data reduction via wavelet feature extraction for calorimeter, LGAD, PID detectors
- Fast online tower energy reconstruction for monitoring and abnormally detection
- Fast offline reconstruction

Summary

- Unique real time challenge at RHIC and EIC calls for AI applications
  - Both RHIC and EIC has much lower collision signal data rate comparing to LHC
  - But background is important and can be dominating
  - We DO NOT want to drop any event for systematic uncertainty control and physics interests
  - Key research is reducing data sufficiently in real time to fit into storage, where AI comes in
- Selected opportunities for Real-time AI highlighted here
  - Lossy compression of data, noise filtering (LDRD 19-028)
  - Feature extraction: Energy time extraction from ADC time-series (LDRD 21-023, NPPS)
  - Feature extraction: tracking, vertexing, HF signal selection (Seminar Feb 1st D.T. Yu)
- Still in exploration stages but aim to have a deployment at sPHENIX with unique physics capability gain. Your inputs welcomed



Remix credit: Dave Morrison

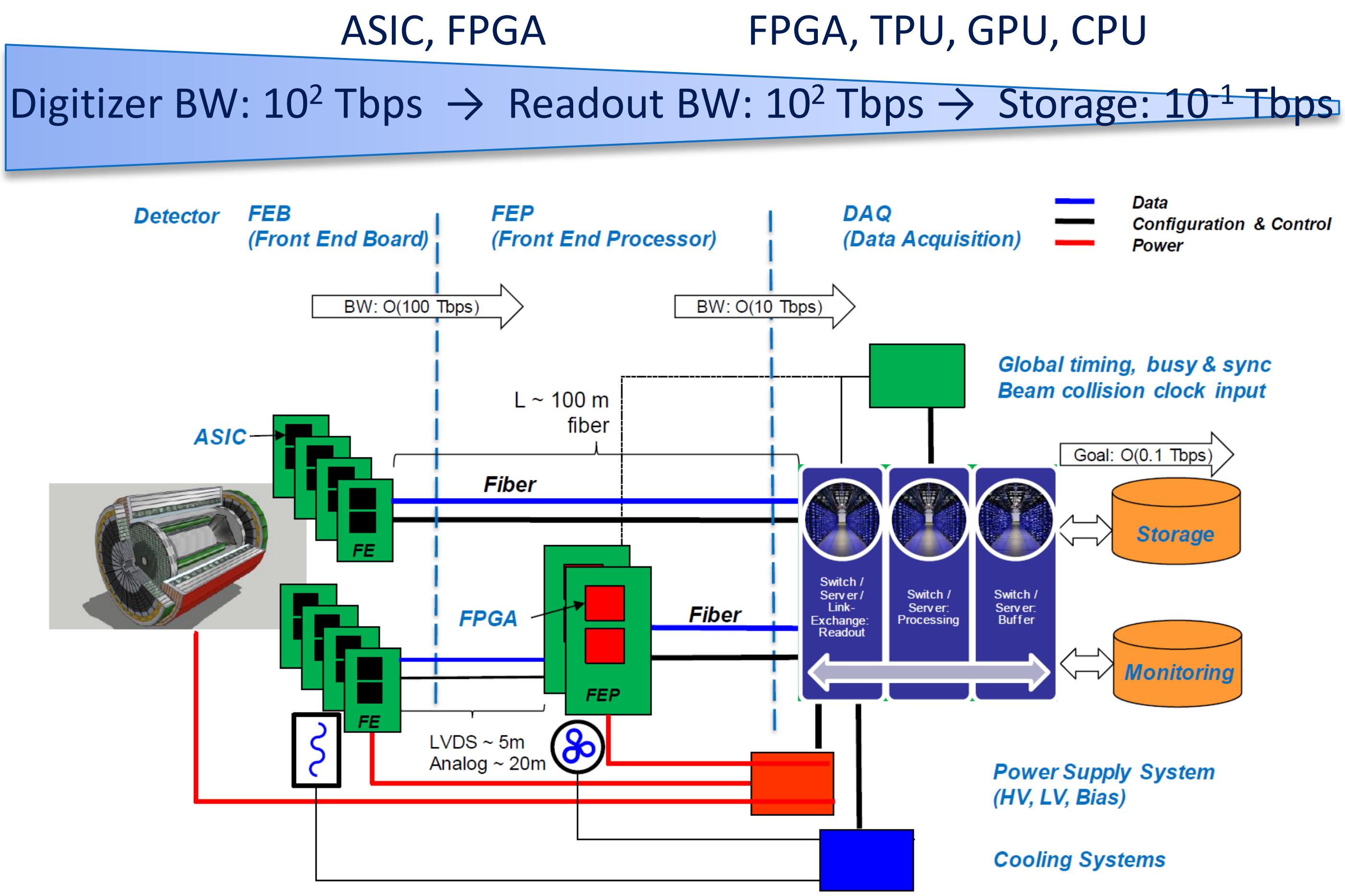
Extra information

Online computing for streaming data – trigger throttling

- At the beginning of the EIC operation, background & noise rate could be unpredictable and high
- A contingency method: throttling streaming data with triggering
  - Immediately reduce streaming data by orders of magnitudes
  - Widely used hardware producing trigger, fix latency or HLT (Aaji's talk)
  - Has physics loss, added systematic uncertainty for hardware trigger efficiency
- Can utilize ML to produce more complex triggering on FPGA
  - PID trigger, e.g. ref: S. Furletov @ streaming workshop VIII [link]
  - Tracking-event topology trigger: D. Yu @ AI4EIC workshop [link]

Real-time computing for streaming data pipeline

- Despite low signal rate, the raw data rate can be filled with noises and background
  - Need low background & low noise detector & electronics design
- An essential job of EIC real-time computing: reliable streaming data reduction to fit permanent storage (next topics)
- And more traditional roles for online/offline server farm:
  - Online monitoring/fault det.
  - Calibration
  - Production → Initial analysis pass

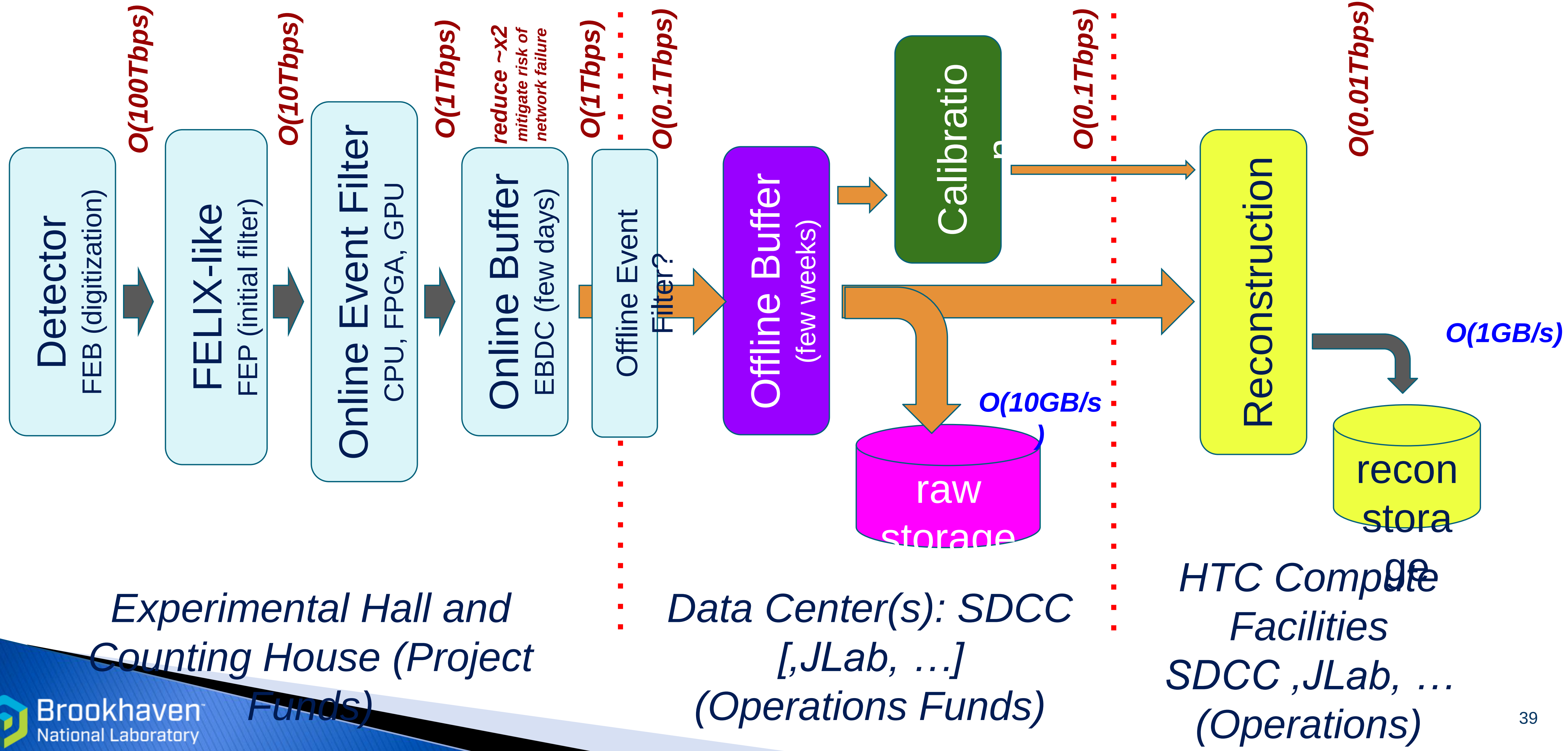


[EIC CDR]

Blured boundary with offline computing

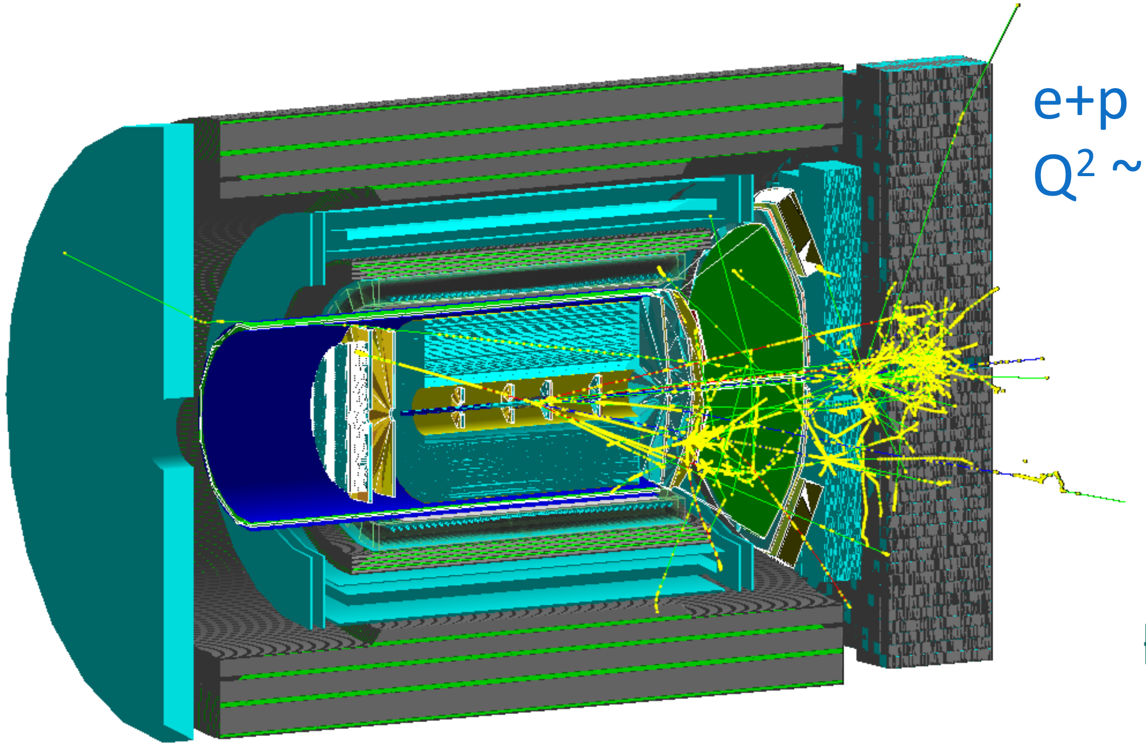
Countesy: David Lawrence
ECCE computing model [link]

See also: last talk M. Battaglieri

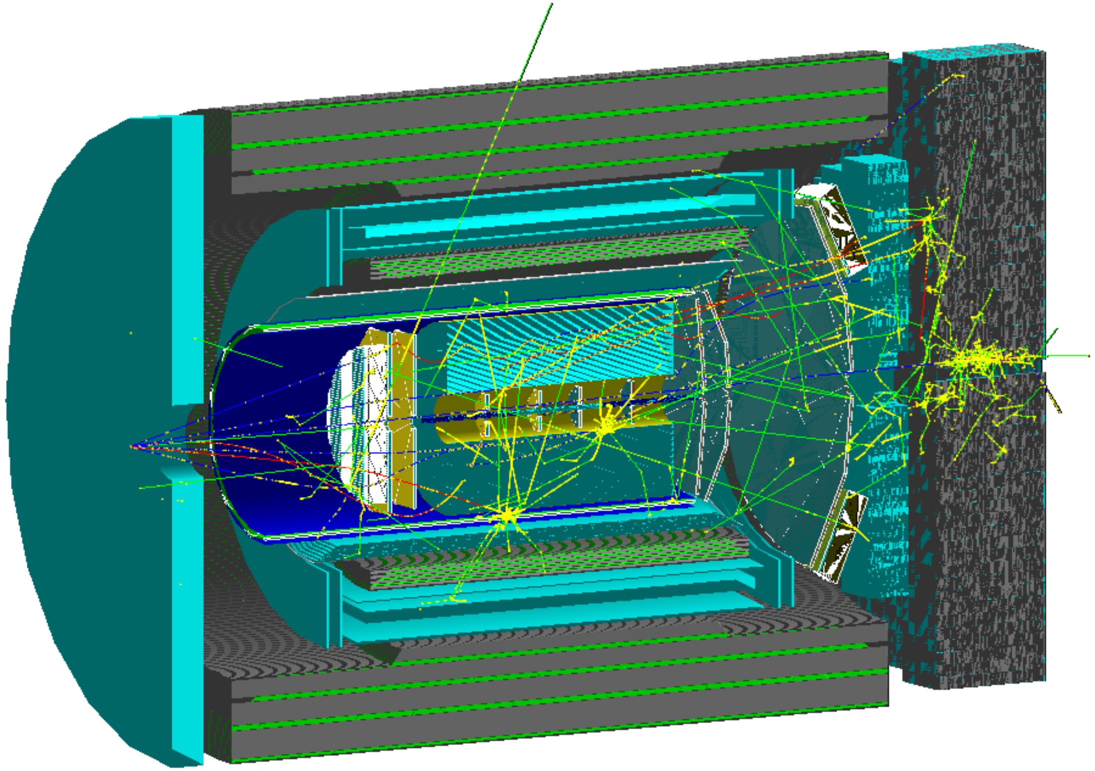


EIC DAQ in Fun4All-EIC simulation

Refs: EIC CDR, sPH-cQCD-2018-001: https://indico.bnl.gov/event/5283/



e+p DIS 18+275 GeV/c
$Q^2 \sim 100\ (\mathrm{GeV}/c)^2$



Beam gas event
p + p(gas), 275 GeV/c
at z=-4 m

Data Rate

MAPS silicon tracker

Raw data: 16-24 bit / MAPS hit (3-layer ALPIDE model)

TPC

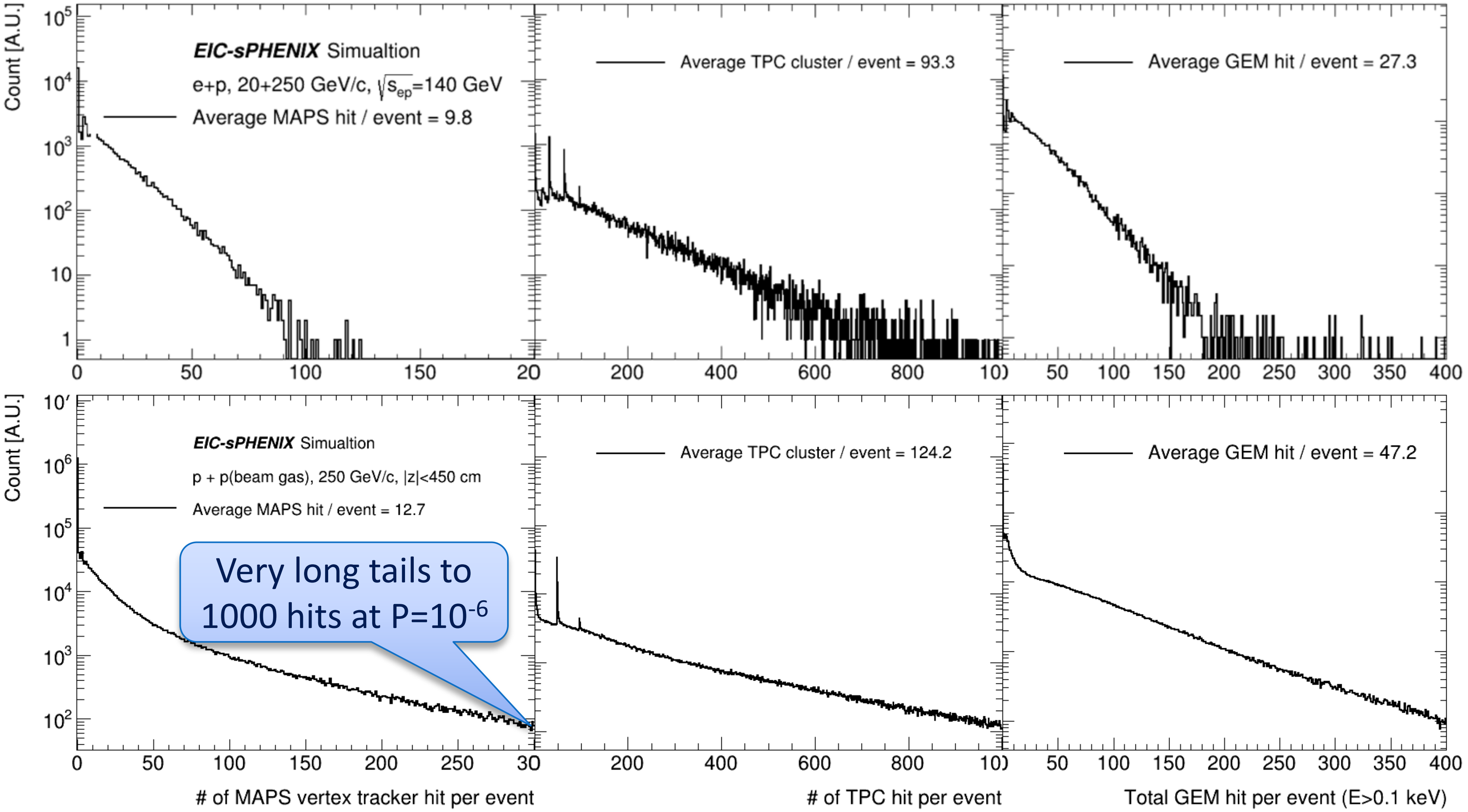
Raw data: 3x5 10 bit / TPC hit + headers (60 bits)

Forward/backward GEM

Raw data: 3x5 10 bit / GEM hit + headers (60 bits)

e+p, Pythia6 Q2>0

p+p(gas) Pythia8



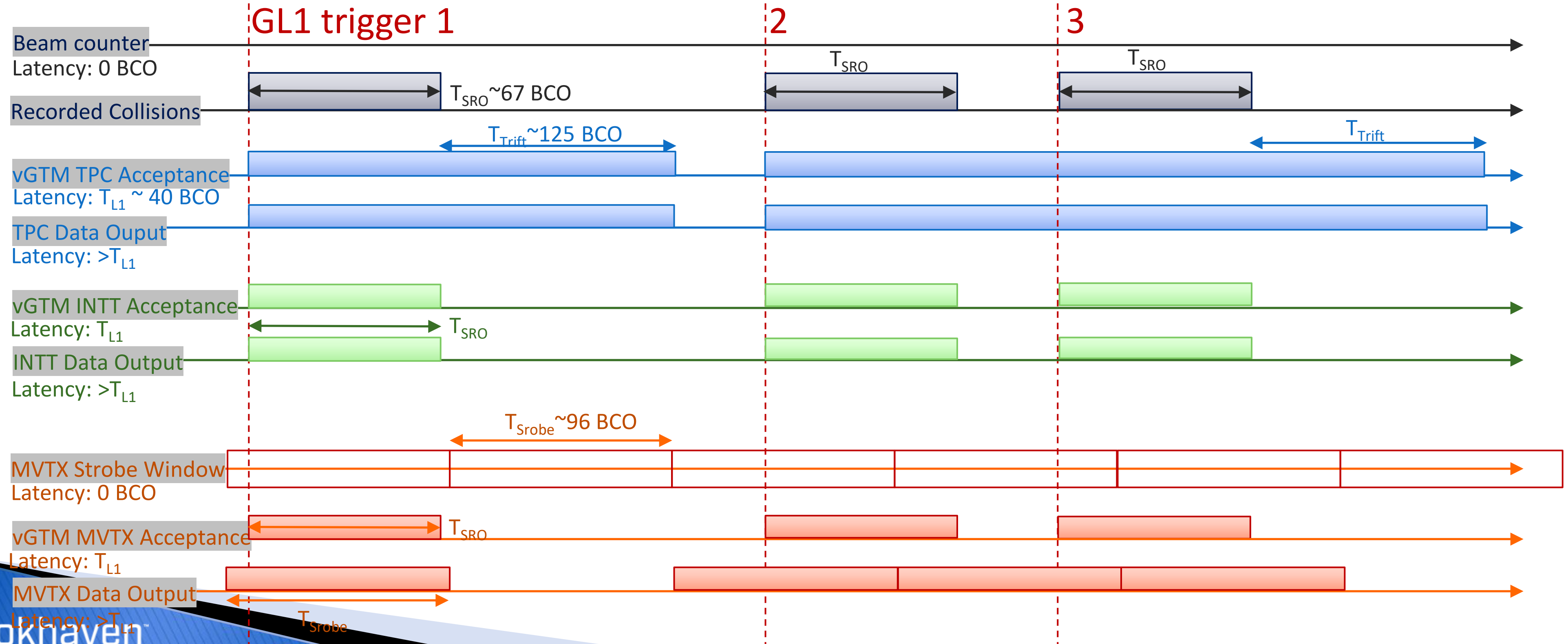
Refs: ECI CDR, sPH-cQCD-2018-001: https://indico.bnl.gov/event/5283/

SRO-Mode1-Simple [Recommended]

Simply prolong L1-Acceptance signal to each subsystem, from 1 BCO to T_{SRO}~67 beam crossings (~7us or 10% SRO data)

→ x500 increase of hard-to-trigger p+p sample

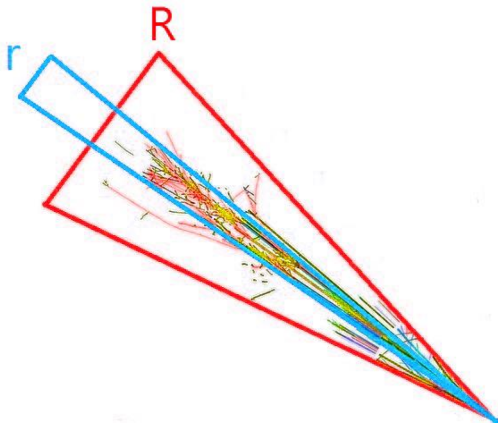
→ at cost only 50% increase in data vol. (by piggy back on long TPC readout window of 13us)



Core physics programs

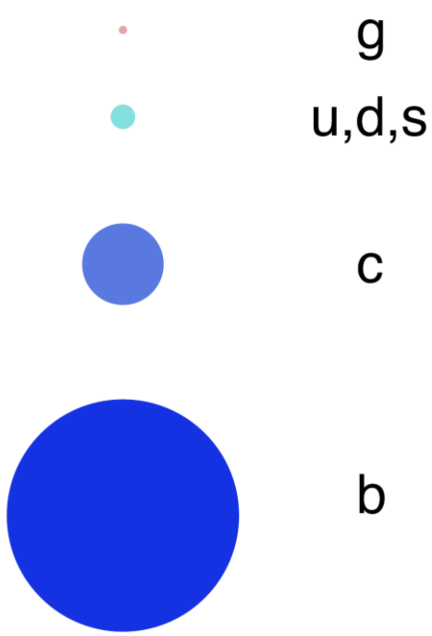
Jet cor. & substructure

Vary momentum/angular size of probe



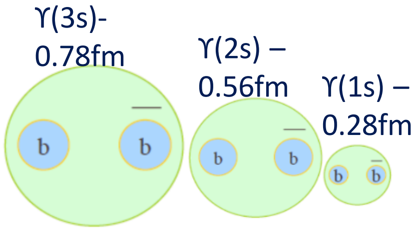
Parton energy loss

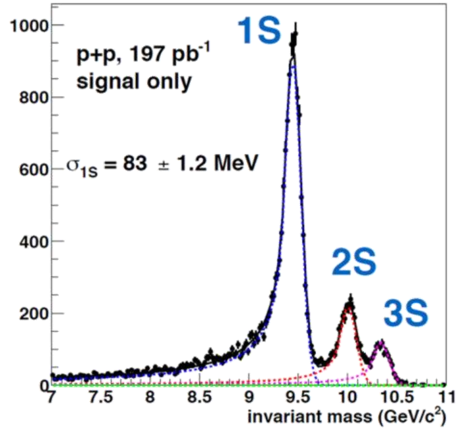
Vary mass/momentum of probe



Upsilon spectroscopy

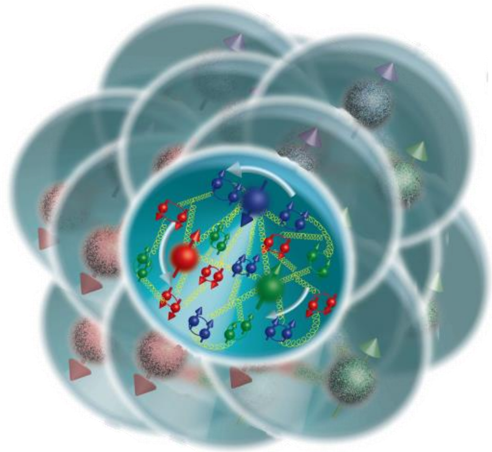
Vary size of the probe

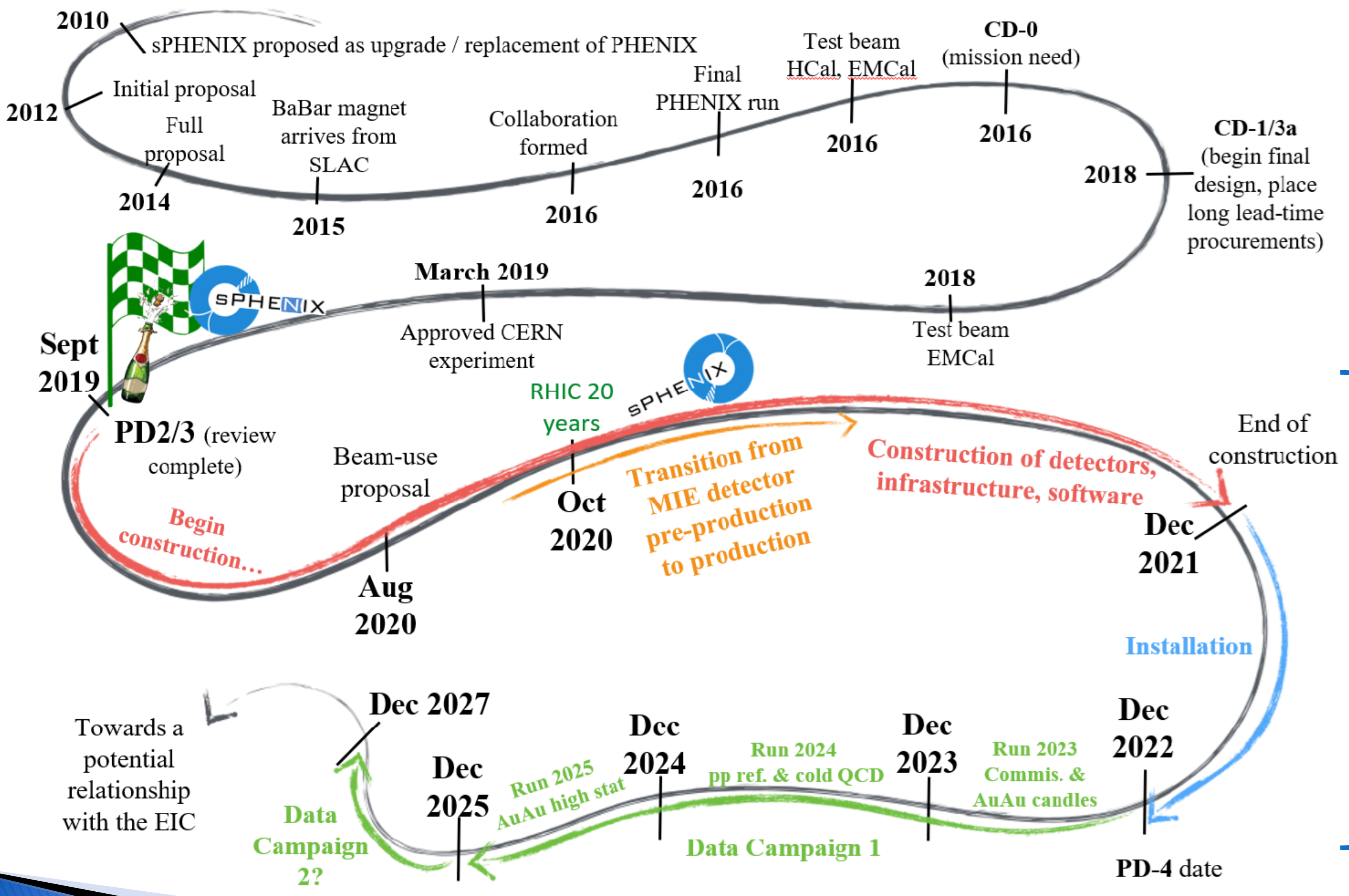




Cold QCD

Vary temperature of QCD matter





~1 years !
From now to first data


sPHENIX magnet installation
RHIC IP8 Hall, Oct 7, 2021
← 13/32 sPHENIX hadronic
calorimeter sectors installed

EIC x-sec : further quantification [Courtesy E. Aschenauer]

- Inelastic e+p scattering x-sec:
  - For a luminosity of 10^{34} $cm^{-2}s^{-1}$ 50ub corresponds to **500 kHz**
- Elastic e+p cross-section:
  - For EIC central barrel, elastic cross section is **small** comparing to the inclusive QCD processes
- Beam gas interaction:
  - Beam proton – beam gas fix target inelastic interactions. The pp elastic cross section is smaller (~7 mb)
  - For a vacuum of 10^{-9} mbar in the detector volume (10m) this gives a rate of **14 kHz**

Beam [GeV]	HERA	5 x 50	10 x 100	18 x 275
$Q^2>10^{-9}$ GeV	65.6	**29.9**	**41.4**	**54.3 ub**
$Q^2>1$ GeV	1.29	0.45	0.65	0.94 ub

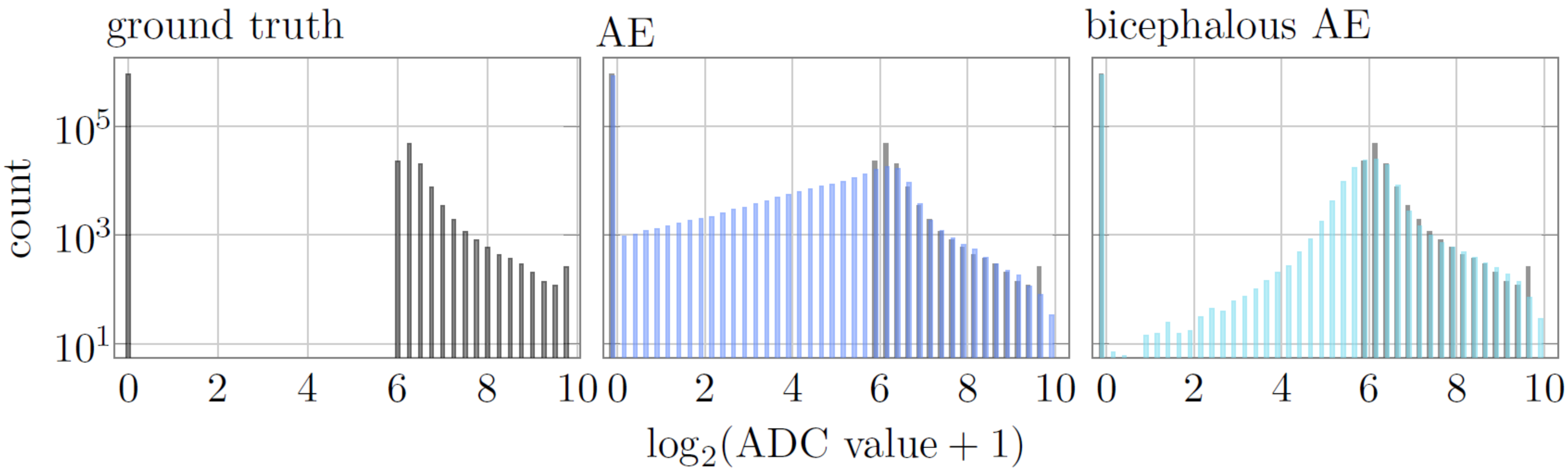
Beam [GeV]	HERA	5 x 50	10 x 100	18 x 275
σ [y_{Exp} >-4]	5 pb	**5 ub**	**0.7 ub**	**0.06 ub**
σ [y_{Exp} >-6]	11 ub	420 ub	100 ub	29 ub

E_p:	50 GeV	100 GeV	275 GeV	920 GeV
	38.4 mb	**38.4 mb**	**39.4 mb**	41.8 mb

Results I: AE v.s. Bicephalous AE

Compression ratio is 1 : 27

(1 : 3 for SAMPA ASIC for this busiest event)



Result III. Ablation Study