# ePIC/EIC Software & Computing, and AI/ML at BNL

Torre Wenaus

Nuclear and Particle Physics Software (NPPS) Group Leader, BNL Physics Dept

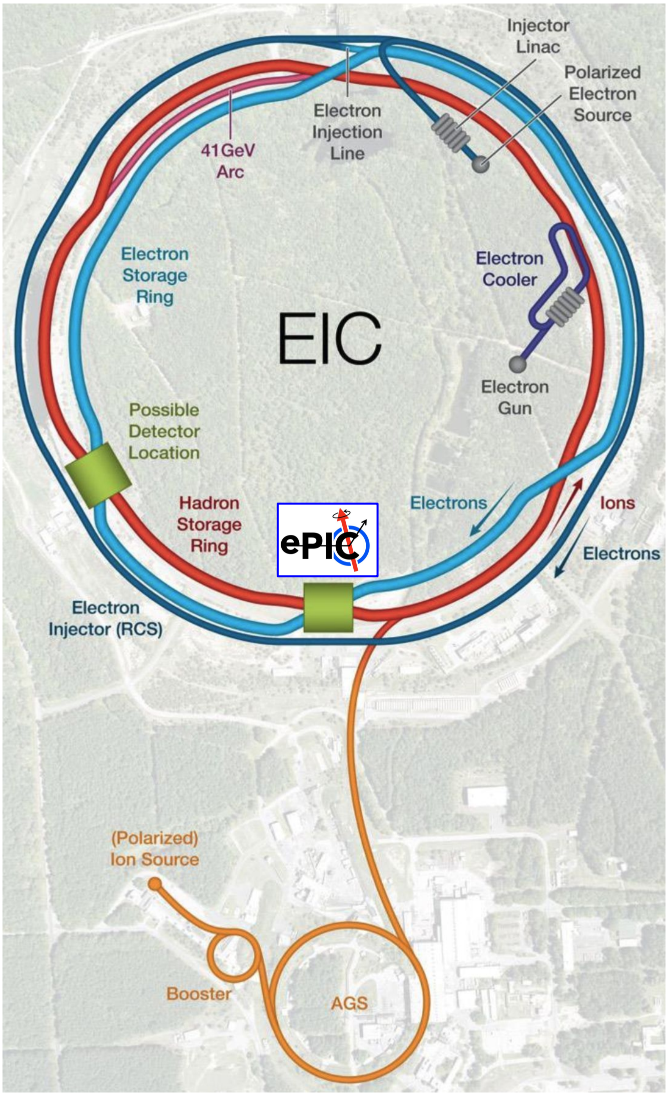
ePIC Deputy S&C Coordinator

ATLAS Distributed Computing Technical Interchange Meeting
Jan 22 2025
Stony Brook University

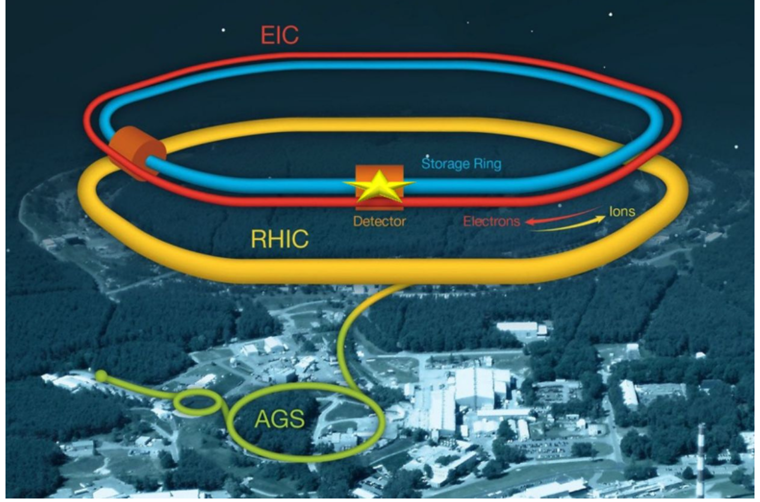
# ePIC/EIC Software & Computing

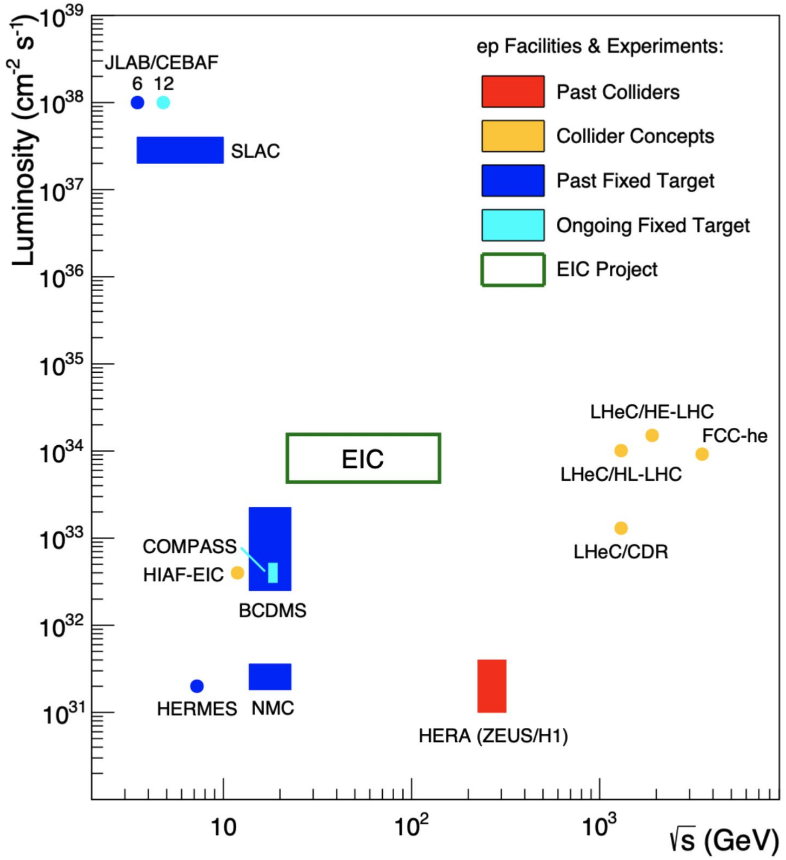
- EIC today has one funded detector, ePIC, so this talk is ePIC S&C
- EIC/ePIC gives high importance to common software; a currently latent EIC-scoped common software activity will be switched back on when needed
- Within the broad scope of ePIC S&C, my focus is on aspects close to ADC
- My S&C deputy S&C coordinator roles include streaming computing model and distributed computing

# The Electron Ion Collider (EIC)



- RHIC’s transformation into the EIC begins after this year’s RHIC run concludes
- **Physics datataking from ~2034**
- It will be the first
  - electron-nucleus collider
  - high-luminosity electron-proton collider
  - e & p spin polarized collider
- Exploring the QCD frontier inside the nucleus
  - Nucleon structure - full 3D spatial and momentum structure, spin structure
  - Origin of nucleon mass
  - Precision study of proton spin
  - Emergent properties of a dense system of gluons
  - **Every collision event has physics interest**
- **Two host labs**, BNL and JLab (Virginia)
  - Close collaboration on all aspects: machine, detector, software and computing
- Together with a **global community** collaborating on the EIC

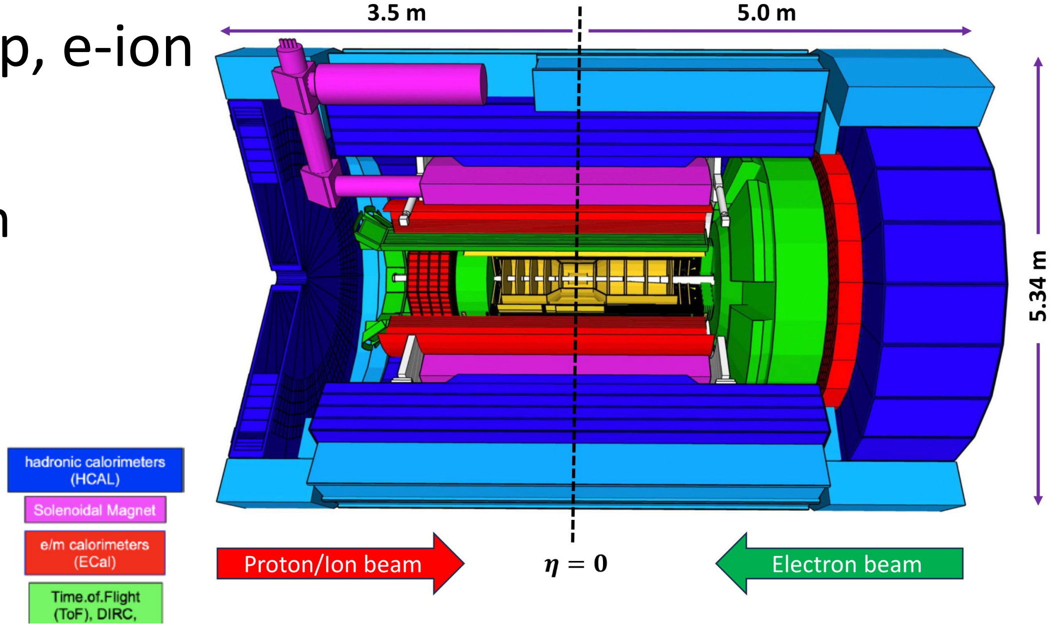


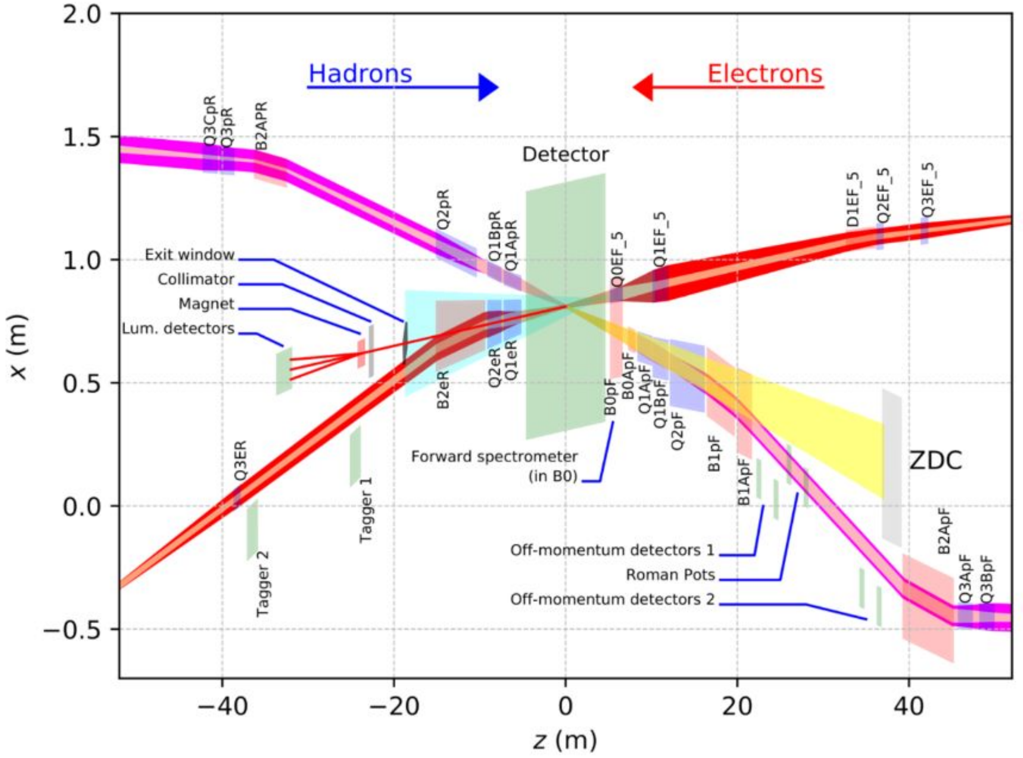


# The ePIC Detector at the EIC

170+ institutions, 25 countries, 500+ participants

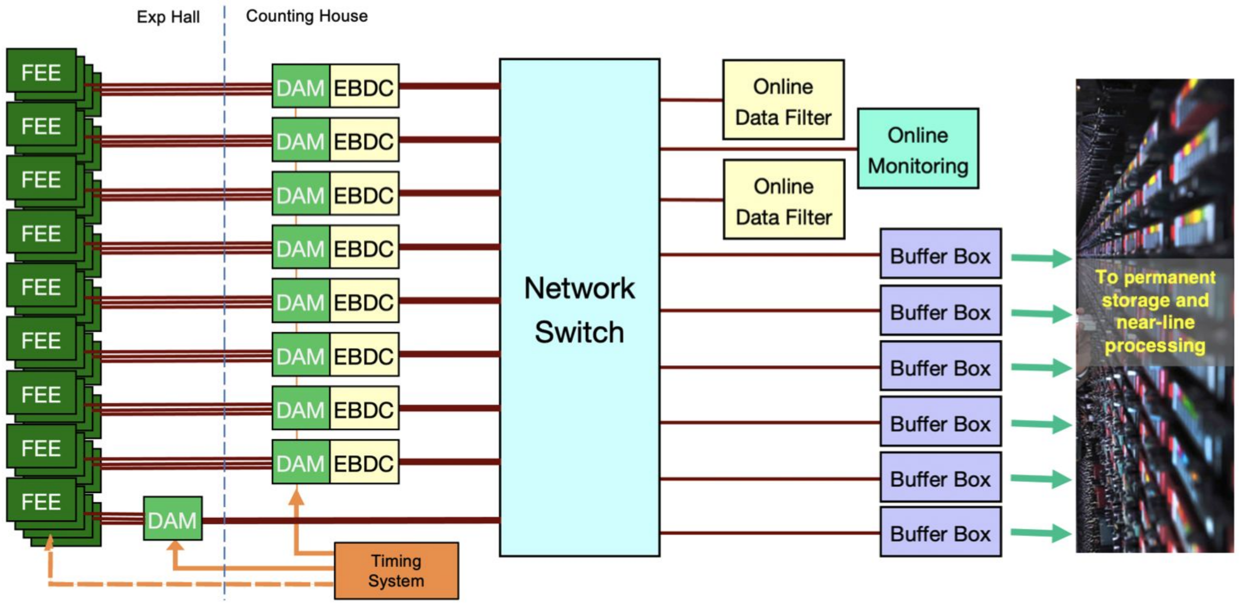
- DOE’s EIC funding includes the ePIC detector to probe e-p, e-ion collisions and the full EIC physics scope
  - including forward/backward detectors extending +/-~45m
- **17+ detector subsystems**
  - charged particle tracking, vertex finding
  - particle identification
  - electromagnetic and hadronic calorimetry
  - precision requirements, many leading edge technologies
- A new 2.8m bore 1.7 tesla superconducting solenoid
- As hermetic a detector as possible
  - **Asymmetric detector systems** optimized for the different particle types and energies in the two directions
- Tightly integrated with the EIC machine and beamline

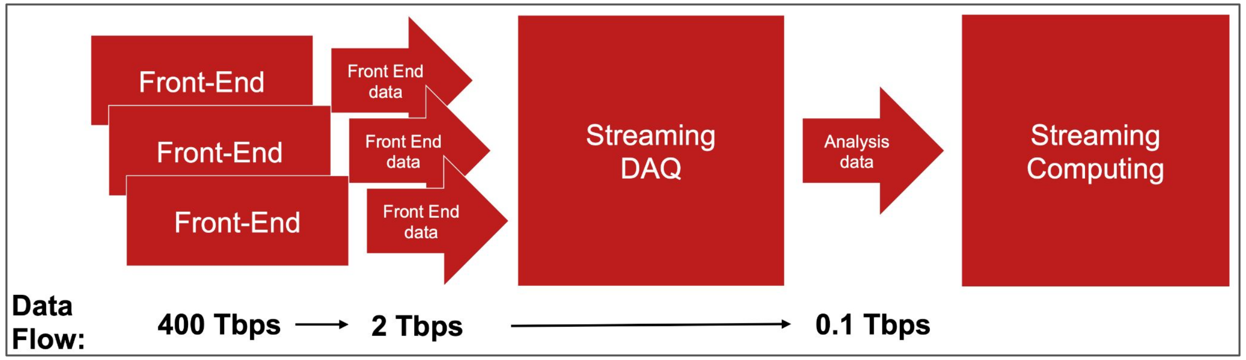




# ePIC Streaming Readout: Maximizing Physics Reach

- EIC luminosity is high, the cross section is not
- **Tractable to read out ~100% of the events**
  - Capture every collision event: they are all of physics interest
    - Backgrounds are substantial, they are in the data stream too
  - A **complete, unbiased event sample**
  - There is substantial noise reduction, compression in DAQ to reduce needed storage
- Data is streamed to prompt reconstruction for **early holistic view of the data**
  - Reconstruct, monitor/diagnose, calibrate, analyze as quickly as possible, O(1min)



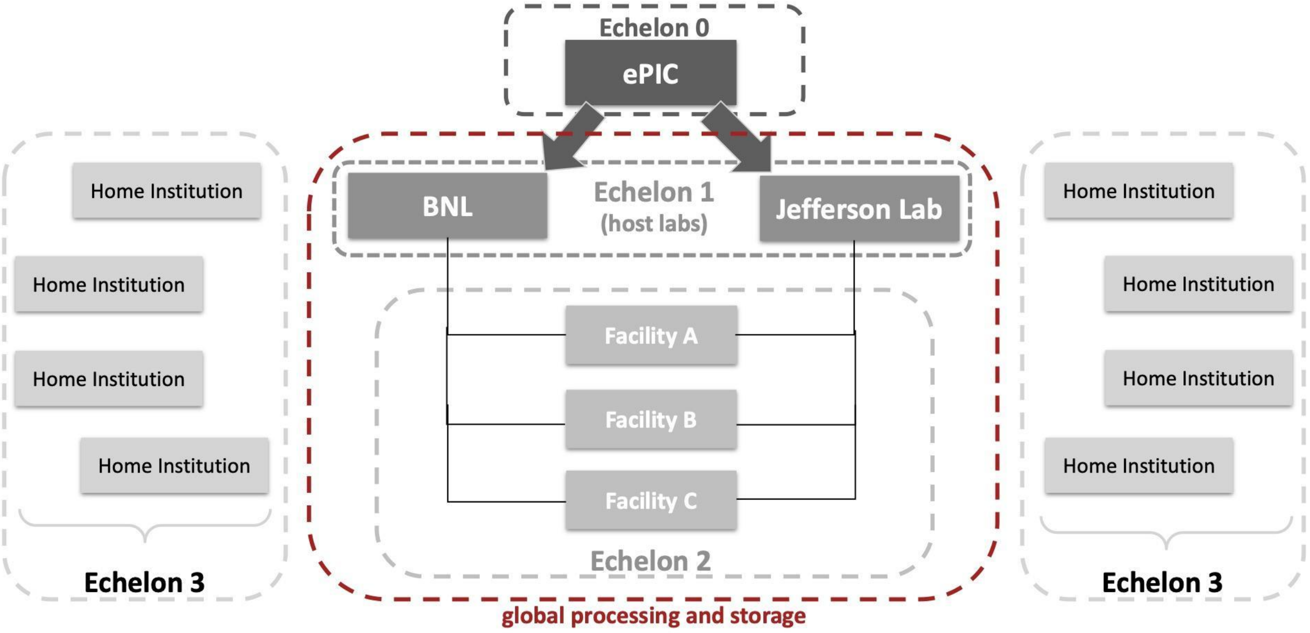


Brookhaven National Laboratory

# The ePIC streaming computing model

ePIC streaming computing model report

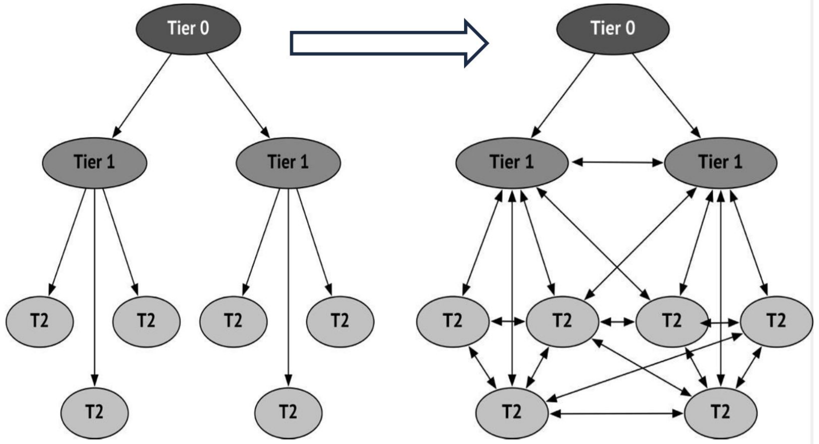
- The two-host-lab organization motivates the ‘butterfly’ model: BNL and JLab are symmetric peers
  - They avoided ‘Tier’ because 0 and 1 levels differ from their LHC meaning
- Globally distributed community and compute
  - Like LHC, distributed computing is essential
  - Pledged and opportunistic: 80% OSG today
- ePIC’s plan is to draw on existing experience and tools from LHC & elsewhere
  - while addressing the unique aspects of its streaming computing model



Lesson from LHC: Mesh outperforms hierarchy

Fully interconnected facilities flexibly and efficiently serve many roles

A hierarchy is *more complex* and less efficient & flexible



**Echelon 0**: ePIC experiment, DAQ system

**Echelon 1**: Two host labs, two primary ePIC computing facilities

**Echelon 2**: Global contributions leveraging commitments to ePIC computing from universities and labs domestically and internationally

**Echelon 3**: Supporting the analysis community where they are at their home institutes, primarily via services hosted at E1s/E2s

# Computing use cases and their Echelon distribution

| Use Case | Echelon 0 | Echelon 1 | Echelon 2 | Echelon 3 |
|---|---|---|---|---|
| Streaming Data Storage and Monitoring | ✓ | ✓ | | |
| Alignment and Calibration | | ✓ | ✓ | |
| Prompt Reconstruction | | ✓ | | |
| First Full Reconstruction | | ✓ | ✓ | |
| Reprocessing | | ✓ | ✓ | |
| Simulation | | ✓ | ✓ | |
| Physics Analysis | | ✓ | ✓ | ✓ |
| AI Modeling and Digital Twin | | ✓ | ✓ | |

Prompt = rapid low-latency processing

Prompt processing of newly acquired data typically begins in seconds, not tens of minutes or longer

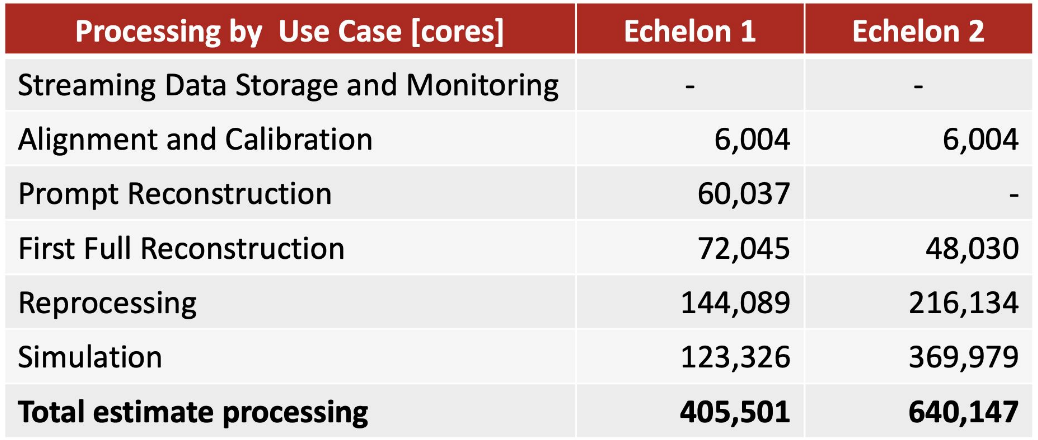
| Assumed Fraction of Use Case Done Outside Echelon 1 | |
|---|---|
| Alignment and Calibration | 50% |
| First Full Reconstruction | 40% |
| Reprocessing | 60% |
| Simulation | 75% |

- **Echelon 1s uniquely perform the low-latency streaming workflows consuming the data stream from Echelon 0**
  - Archiving, monitoring, prompt reconstruction, rapid diagnostics
- There's been discussion over whether Echelon 2s have a role in low-latency streaming processing
  - We shouldn't exclude it in the design, but doing it in year 1,2,... is very unlikely
  - Message from reviewers (e.g. Simone Campana) was don't plan to do this unless you really have to
- **Ensure the E1s have sufficient processing power for the low-latency workflows**
- **Apart from low-latency streaming, Echelon 2s are full participants in the use cases**
  - Resource requirements model assumes a substantial role for Echelon 2

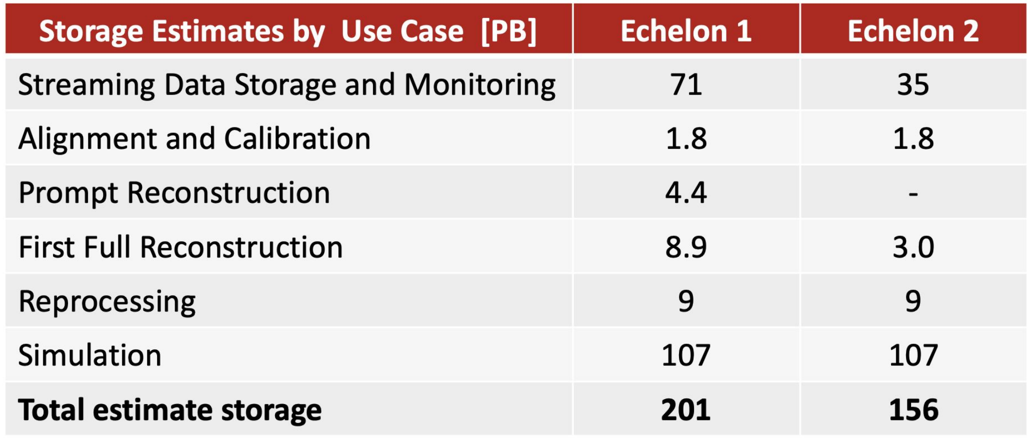
# Computing resource needs and the implications

- See Markus Diefenthaler's talk this week 24-25 for the numbers behind the numbers
- O(1M) core-years to process a year of data, above ATLAS scale today
  - Optimistic constant-dollar performance gains would reduce the numbers about 5x
    - Based on current LHC measure of 15%/yr
    - But the trend is towards lower gains per year
- Whatever the gains over time, the processing scale is substantial
  - Motivates attention to leveraging distributed and opportunistic resources from the beginning
- ~400PB/yr storage scale, also above ATLAS accumulation rate today
  - Archival storage (probably tape) will play a role
  - Motivates attention to data carousel type orchestration

Estimated needs for ePIC Phase I nominal year (circa 2034)



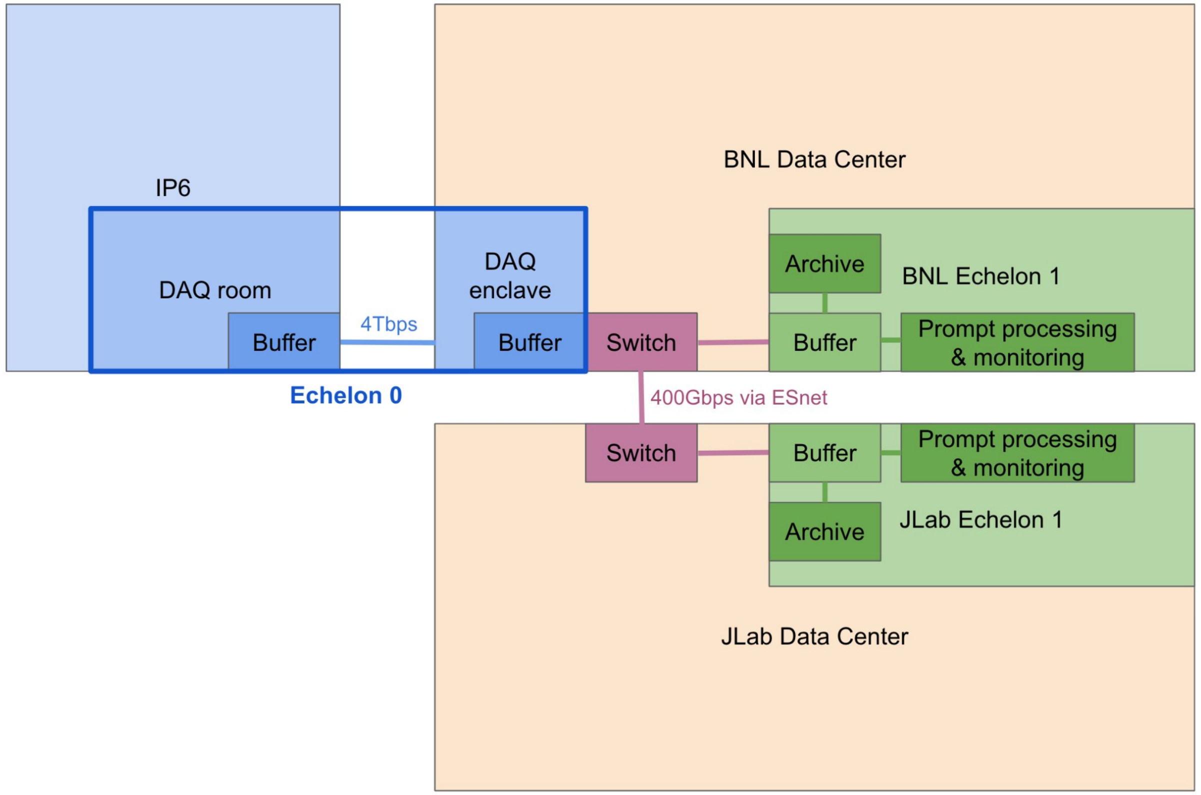
| Processing by Use Case [cores] | Echelon 1 | Echelon 2 |
|---|---|---|
| Streaming Data Storage and Monitoring | - | - |
| Alignment and Calibration | 6,004 | 6,004 |
| Prompt Reconstruction | 60,037 | - |
| First Full Reconstruction | 72,045 | 48,030 |
| Reprocessing | 144,089 | 216,134 |
| Simulation | 123,326 | 369,979 |
| **Total estimate processing** | **405,501** | **640,147** |



| Storage Estimates by Use Case [PB] | Echelon 1 | Echelon 2 |
|---|---|---|
| Streaming Data Storage and Monitoring | 71 | 35 |
| Alignment and Calibration | 1.8 | 1.8 |
| Prompt Reconstruction | 4.4 | - |
| First Full Reconstruction | 8.9 | 3.0 |
| Reprocessing | 9 | 9 |
| Simulation | 107 | 107 |
| **Total estimate storage** | **201** | **156** |

# Echelon 0 (DAQ) to Echelon 1

- We began last fall to work out how data will flow from Echelon 0 to Echelon 1 for prompt consumption
- It's all under discussion; nothing I say here is decided, it's merely my perspective
- DAQ expected to leverage data center resources through a 'DAQ enclave' at SDCC
- Symmetric BNL, JLab Echelon 1 facilities downstream of DAQ are equally capable of performing Echelon 1 workflows
  - The model expresses the desired symmetry
  - Exact E1 roles will be decided by ePIC + facilities



- What's in the data stream sent from DAQ?
  - **Time frames**, each containing all detector data in a time interval, are built in DAQ
  - Time frames are aggregated in **super time frames** (STFs) which are sent out of DAQ to E1
    - Together with a small non-event data component (e.g. slow controls)
  - Data arrives at Echelon 1s in a ~1 week deep disk buffer
  - Workflows consume data from that buffer, first and foremost:
    - Archiving the full stream to tape (at least in early years)
    - Prompt processing, monitoring for a rapid view of data/detector integrity and quality

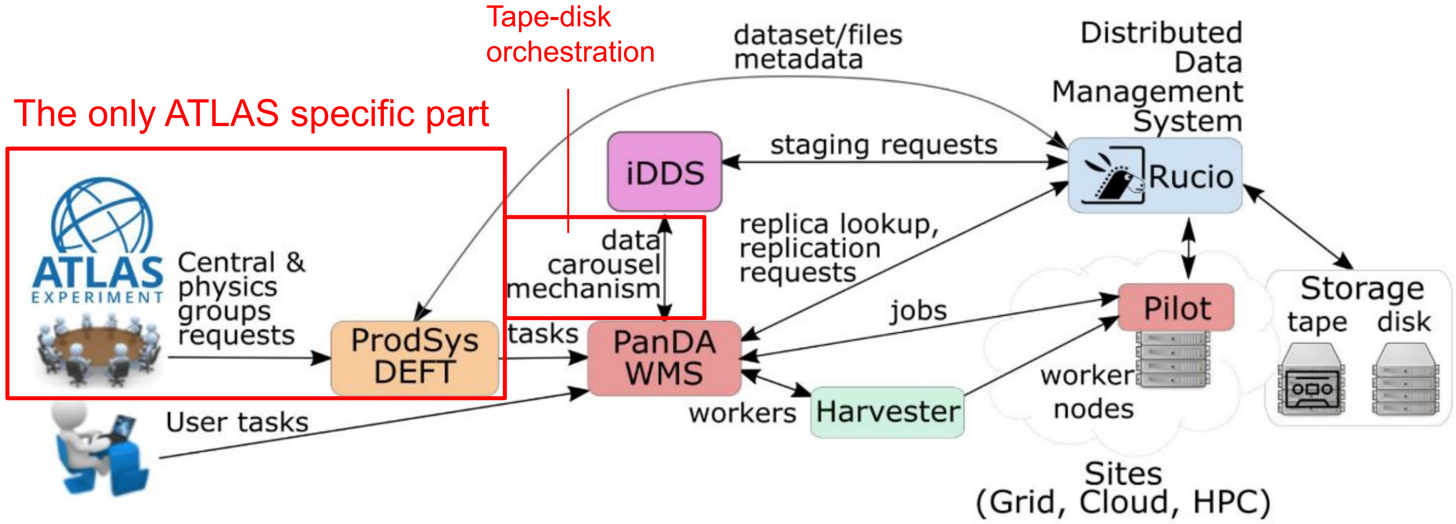
# Time Frames and Super Time Frames

- Each time frame aggregates all detector data within a time window of ~0.6ms
- Super Time Frame (STF) is a contiguous set of ~1000 time frames
  - Within a STF the TFs are time-ordered, as required for reconstruction
  - No overlaps between time frames, so cannot reconstruct the edges
  - Make them large enough that losses from the edges are negligible
- This O(1s), 2GB STF data unit is an appropriate granularity for Echelon 1 processing
  - Short enough for prompt processing
  - Long enough for tractable bookkeeping and file size
- The STFs are *not* (required to be) time-ordered (facilitated by no overlaps)
  - Friendly to distributed computing and parallel orchestration
- **STF is the atomic unit for ePIC data processing**
  - (Such is the present ~4 month old thinking)

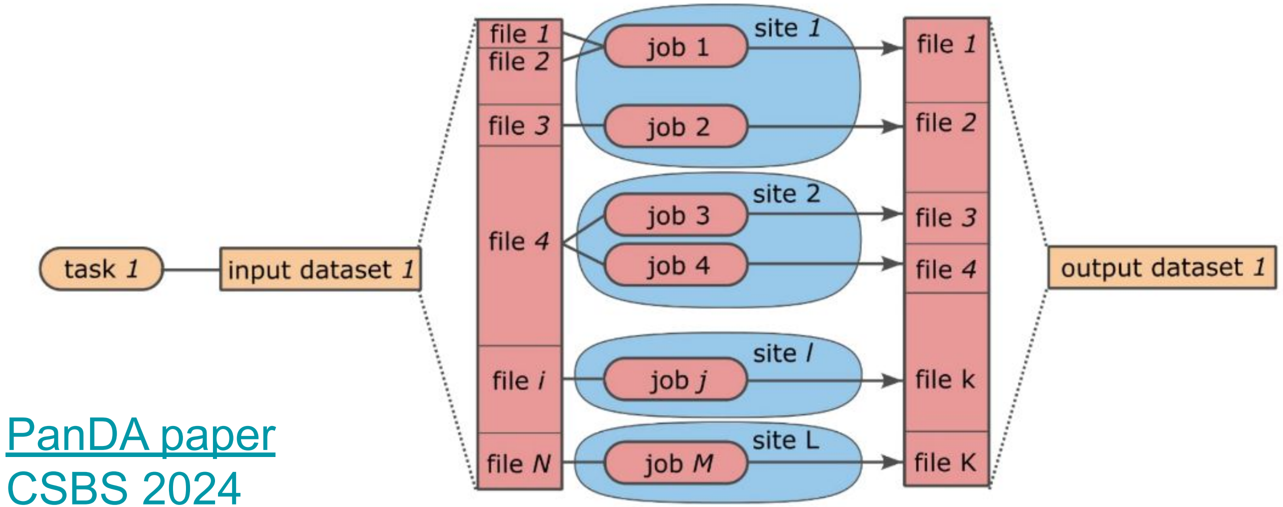
# Streaming orchestration at Echelon 1s

- Each E1 receives and archives 100% of the raw data, packaged identically: each T1 has a full set of super timeframe (STF) replicas
- Prompt processing is the other primary Echelon 1 role
- How will prompt processing be performed? What distributed computing systems will manage the processing at Echelon 1s and downstream?
  - The data management system is decided: Rucio
  - The workflow/workload management system evaluation and decision will take place (we expect) over the next 6mo or so
  - ePIC is very good at doing a considered process of requirements-based surveying, evaluating, and choosing. They've chosen the elements of their software stack this way
    - Key4HEP framework with Jana2 replacing Gaudi; Podio based EDM, etc.
  - Their decision process doesn't demand prototyping, but for PanDA, we do plan to prototype as available effort permits
    - As well as addressing the requirements document ePIC will produce (for which the CMS PanDA eval Q&A is useful input)
  - We are planning for success: demonstrate PanDA + Rucio working together as designed, for the first time outside ATLAS
    - (in Rubin they don't work together; Rubin middleware sits in the middle)

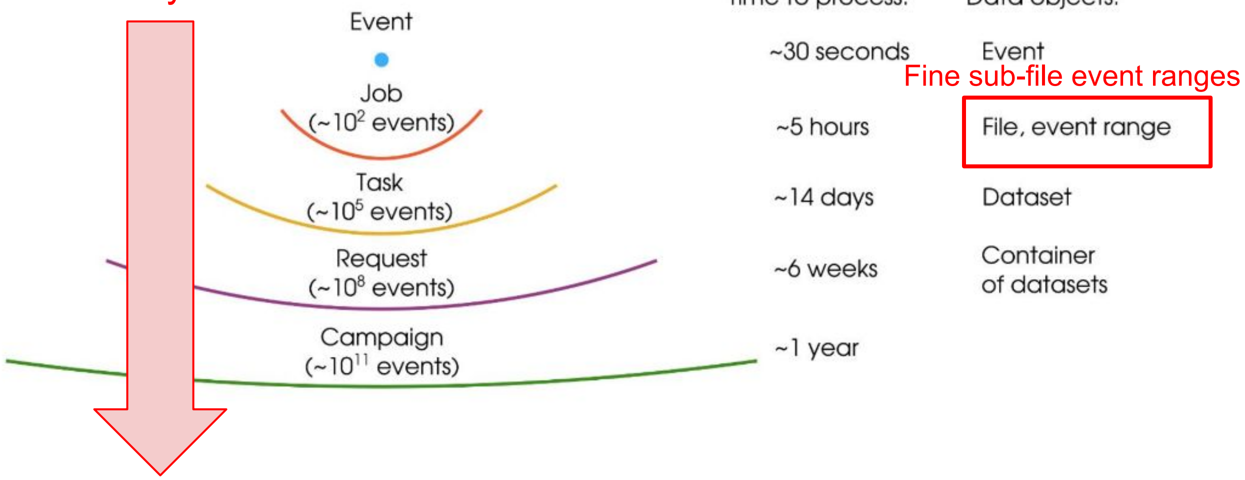
# PanDA is a good fit, including fine grained orchestration



- **DEFT**: Database Engine For Tasks
- **PanDA**: Production ANd Distributed Analysis System
- **Harvester**: resource-facing service between the PanDA and collection of pilots
- **Pilot**: the execution environment on a worker node   Fine grained workflow orchestration
- **iDDS**: Intelligent Data Delivery System
- **Rucio**: Distributed Data Management System



Granularity suited to the use case



PanDA paper CSBS 2024

# Prototyping prompt streaming with Rucio + PanDA

- Rucio is operating at JLab, BNL close to completing an instance for EIC use
  - Rucio used in ePIC production jobs for the first time this week
- PanDA prerequisite has been a PanDA instance at BNL, now in place thanks to Xin, PanDA team, SDCC
  - (Parenthetical: it's through the PanDA@BNL for NP work that we set up OSG + GlideinWMS support, essential to CMS and spotted by them immediately in the last PanDA community meeting, towards their PanDA evaluation)
  - Client #1 for BNL PanDA is our AID2E AI-based detector design optimization project funded by DOE NP
    - ePIC PanDA is riding on its coat tails
- Other workflow management system participants in the evaluation TBD
  - DIRAC(x) is the natural, expected as part of the requirements eval, probably not prototyping?
  - NPPS also works/develops with DIRAC(x), we did the Rucio integration for Belle II; we have the most DIRAC experience in ePIC but we're not going to work two evaluations!
- ePIC streaming computing model WG met last week with two long-time friends of this room
  - Cedric assessed Rucio for the STF-based streaming model, all seems good
  - Dan Van Der Ster assessed object stores as the basis for STF storage; (Ceph) S3 seems good
- For the next weeks: develop a prototyping plan!

# AI/ML at BNL

I will narrow this broad topic to highlights suited to an EIC S&C talk at an ATLAS TIM

Especially as I'm probably out of time right now!

# AI for the EIC - Current and Near Term

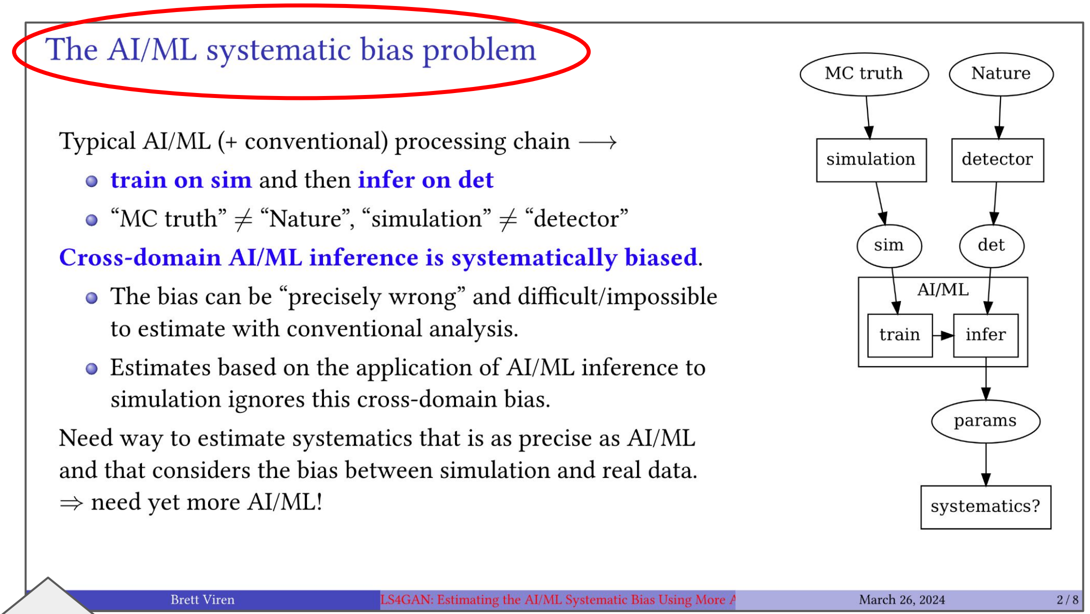
- The 2023 Nuclear Science Advisory Committee Long Range Plan stated that "*EIC could be one of the first large-scale collider-based programs in which AI/ML is integrated from the start.*"
  - We've just left the starting gate and we've begun the work to achieve this!
- AI is already "*a key part of all software and computing working groups in ePIC*" - ePIC Spokesperson John LaJoie at AI4EIC workshop Fall 2023
  - ML-based algorithms supported and **used in the software framework**
  - **Applications** developing in fast calibration, streaming, particle identification, many others
  - Working on **centralized services** for training, model management, workflow integration
  - Monitoring integration of AI methods in monthly **production campaigns as progress metric**
  - **Data and analysis preservation**: key to leveraging rapidly evolving AI approaches

# AI for the EIC - Expectations across the scope

- Accelerator: **proton polarization**, **luminosity optimization**, **bunch merging**
- DAQ: **background/noise reduction**, **zero suppression**, **lossy compression**
- Controls: sub-second corrections/calibrations via AI-guided data prediction
- Conditions: smart conditions monitoring, problem prediction, anomaly detection
- Calibration/QA: fast calibration, **alignment**, real-time anomaly detection
- Operations: custom LLM chatbots, smart service status monitoring, log analysis
- Software: event generation, **fast simulation**, **reconstruction**, **detector design optimization**
- Distributed computing: **workflow/dataflow optimisation**, **resiliency**, **large scale distributed AI services**
- Accelerator-detector-DAQ-processing integration in a '**cognizant AI facility**'
- And throughout analysis
  - Physics object reconstruction, full-event analysis
  - **Understanding and quantifying uncertainties and systematics**
  - In the long term, insights from rapidly evolving AI technologies

**Bold: current or near term BNL activity**

# AI for the EIC Today: Examples with Local Contributions



## The AI/ML systematic bias problem

Typical AI/ML (+ conventional) processing chain ⟶

- **train on sim** and then **infer on det**
- “MC truth” ≠ “Nature”, “simulation” ≠ “detector”

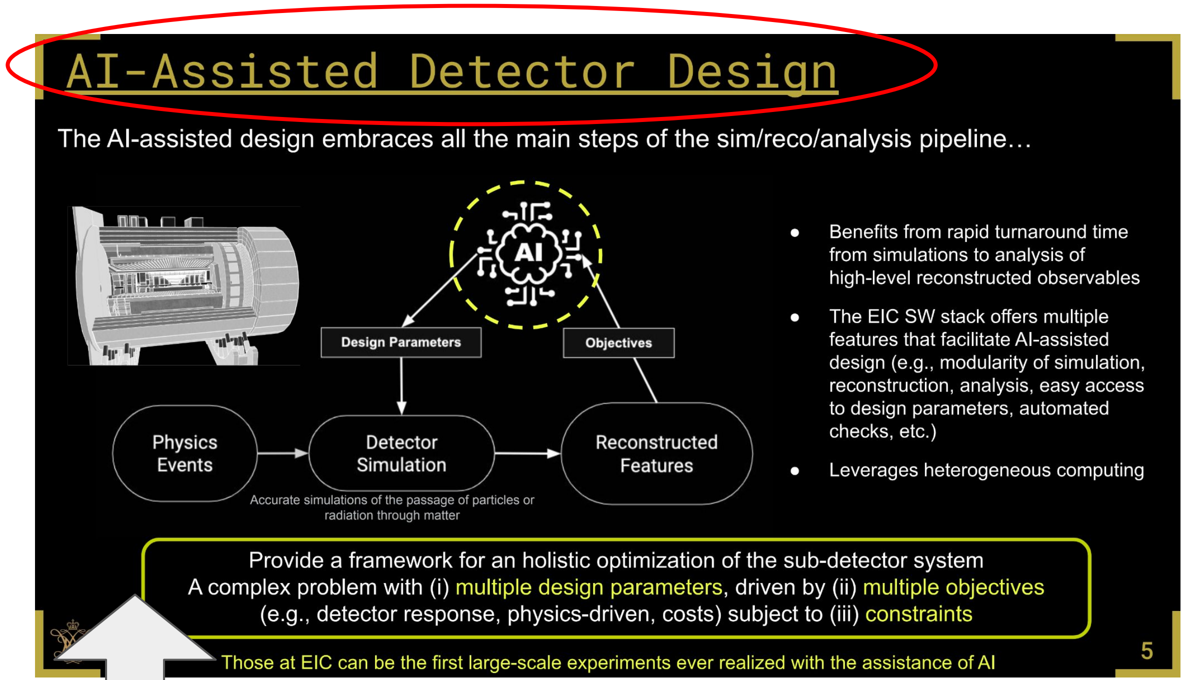
**Cross-domain AI/ML inference is systematically biased.**

- The bias can be “precisely wrong” and difficult/impossible to estimate with conventional analysis.
- Estimates based on the application of AI/ML inference to simulation ignores this cross-domain bias.

Need way to estimate systematics that is as precise as AI/ML and that considers the bias between simulation and real data.
⇒ need yet more AI/ML!

Brett Viren · LS4GAN: Estimating the AI/ML Systematic Bias Using More A · March 26, 2024 · 2 / 8

LS4GAN project, Brett Viren (BNL DUNE)



## AI-Assisted Detector Design

The AI-assisted design embraces all the main steps of the sim/reco/analysis pipeline…

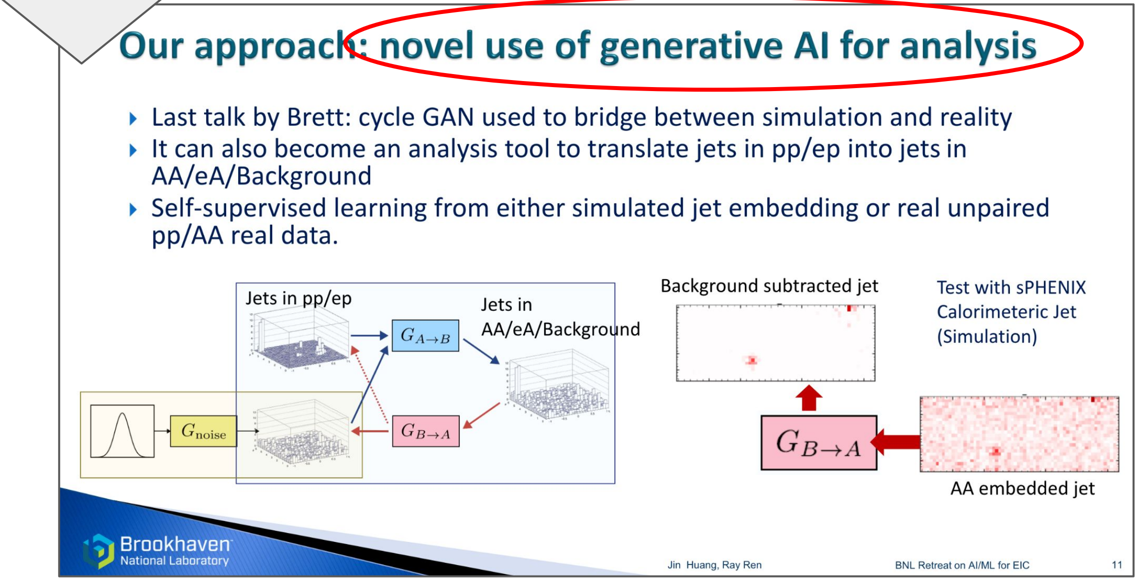
- Benefits from rapid turnaround time from simulations to analysis of high-level reconstructed observables
- The EIC SW stack offers multiple features that facilitate AI-assisted design (e.g., modularity of simulation, reconstruction, analysis, easy access to design parameters, automated checks, etc.)
- Leverages heterogeneous computing

Provide a framework for an holistic optimization of the sub-detector system
A complex problem with (i) multiple design parameters, driven by (ii) multiple objectives (e.g., detector response, physics-driven, costs) subject to (iii) constraints

Those at EIC can be the first large-scale experiments ever realized with the assistance of AI

5

Cristiano Fanelli (CWM EIC)



## Our approach: novel use of generative AI for analysis

- Last talk by Brett: cycle GAN used to bridge between simulation and reality
- It can also become an analysis tool to translate jets in pp/ep into jets in AA/eA/Background
- Self-supervised learning from either simulated jet embedding or real unpaired pp/AA real data.

Brookhaven National Laboratory · Jin Huang, Ray Ren · BNL Retreat on AI/ML for EIC · 11

Jin Huang (BNL eSPHENIX), Ray Ren (BNL CSI)

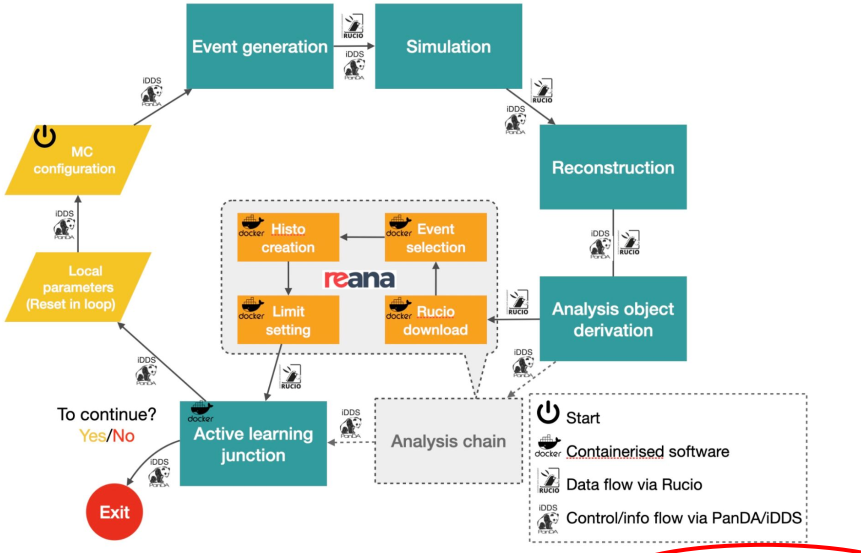
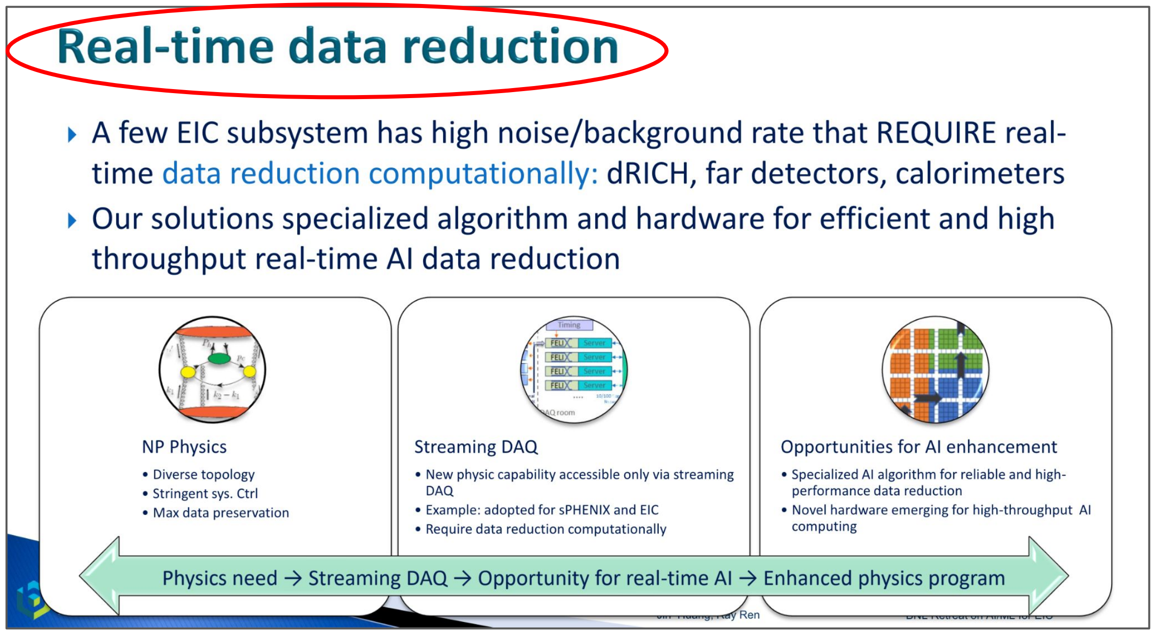


Figure 10: Bayesian optimisation based active learning with PanDA/iDDS and Rucio

Tadashi Maeno, Wen Guan (BNL ATLAS)

# AI for the EIC Today: Local Examples 2



## Real-time data reduction

- A few EIC subsystem has high noise/background rate that REQUIRE real-time data reduction computationally: dRICH, far detectors, calorimeters
- Our solutions specialized algorithm and hardware for efficient and high throughput real-time AI data reduction

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

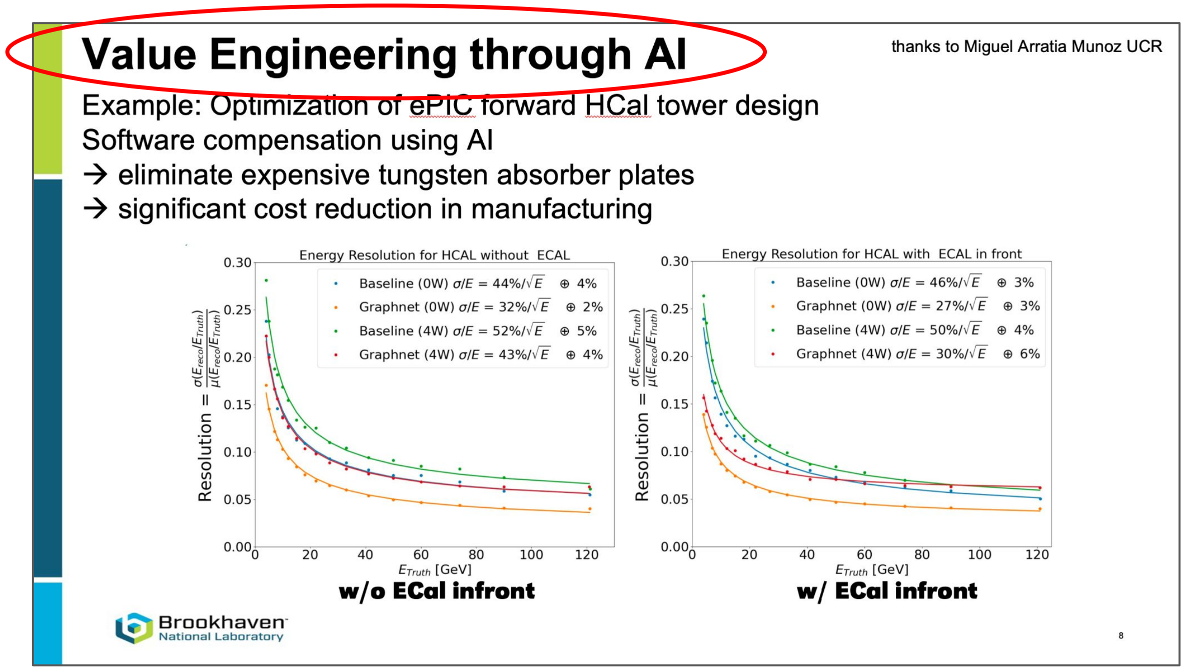
Opportunities for AI enhancement

- 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

Jin Huang, Ray Ren BNL Retreat on AI/ML for EIC

Jin Huang (BNL eSPHENIX)



## Value Engineering through AI

thanks to Miguel Arratia Munoz UCR

Example: Optimization of ePIC forward HCal tower design

Software compensation using AI

→ eliminate expensive tungsten absorber plates

→ significant cost reduction in manufacturing

Energy Resolution for HCAL without ECAL
Baseline (0W) $\sigma/E = 44\%/\sqrt{E} \oplus 4\%$
Graphnet (0W) $\sigma/E = 32\%/\sqrt{E} \oplus 2\%$
Baseline (4W) $\sigma/E = 52\%/\sqrt{E} \oplus 5\%$
Graphnet (4W) $\sigma/E = 43\%/\sqrt{E} \oplus 4\%$
Resolution = $\frac{\sigma(E_{reco}/E_{Truth})}{\mu(E_{reco}/E_{Truth})}$
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0
20
40
60
80
100
120
$E_{Truth}$ [GeV]
Energy Resolution for HCAL with ECAL in front
Baseline (0W) $\sigma/E = 46\%/\sqrt{E} \oplus 3\%$
Graphnet (0W) $\sigma/E = 27\%/\sqrt{E} \oplus 3\%$
Baseline (4W) $\sigma/E = 50\%/\sqrt{E} \oplus 4\%$
Graphnet (4W) $\sigma/E = 30\%/\sqrt{E} \oplus 6\%$
Resolution = $\frac{\sigma(E_{reco}/E_{Truth})}{\mu(E_{reco}/E_{Truth})}$
$E_{Truth}$ [GeV]

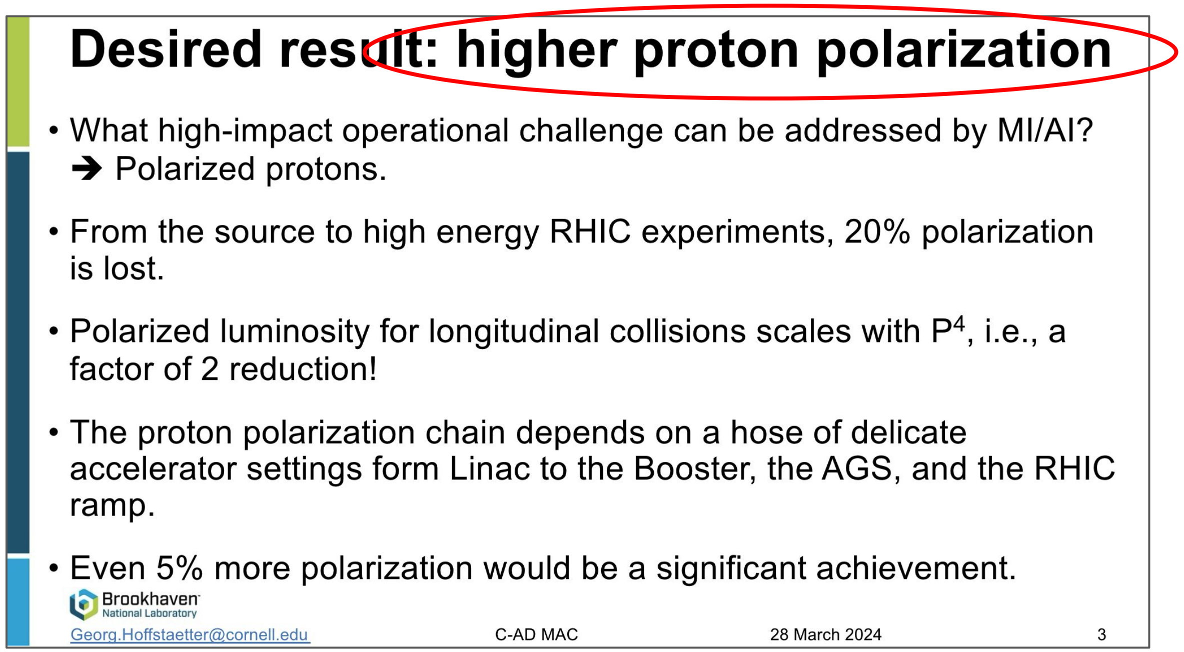
**w/o ECal infront**

**w/ ECal infront**

Brookhaven National Laboratory

8

Elke Aschenauer (BNL EIC)



## Desired result: higher proton polarization

- What high-impact operational challenge can be addressed by MI/AI? ➔ Polarized protons.
- From the source to high energy RHIC experiments, 20% polarization is lost.
- Polarized luminosity for longitudinal collisions scales with $P^4$, i.e., a factor of 2 reduction!
- The proton polarization chain depends on a hose of delicate accelerator settings form Linac to the Booster, the AGS, and the RHIC ramp.
- Even 5% more polarization would be a significant achievement.

Brookhaven National Laboratory

Georg.Hoffstaetter@cornell.edu C-AD MAC 28 March 2024 3

Lucy Lin (BNL C-AD), Georg Hoffstaetter (Cornell)

# Experimental HEP AI/ML at BNL

- As everywhere, expanding activity within all research programs, with HEP-NP collaboration
- Fruitful collaboration with Computational Science Initiative (CSI) is a common denominator
  - Analogy from a BNL HEP physicist: CSI is the 'AI/ML expertise bus' interconnecting and powering our projects with shared expertise
- AI/ML used in most **BNL ATLAS analyses and projects** (Omega, CSI)
  - Improvements of ~50% on HH limit, 1000x on NN+FPGA tracking inference, HH → ɣɣbb yield increase
- **REDWOOD** - Efficiently and resiliently manage very large datasets and complex workflows (NPPS, SDCC, CSI)
  - *Draw on PanDA/ATLAS's deep reservoir of workflow/dataflow data, deliver AI/ML applications in PanDA!*
- AID2E - **AI based scalable distributed detector design optimization** for EIC, NP, HEP (NPPS, CSI)
  - *We deliver 'scalable distributed': leverages and strengthens PanDA's AI/ML workflow capability*
  - *Wen (with support!) has developed a new PanDA/iDDS capability for this, Function as a Service to transform a functional spec for AID2E workflow into a PanDA/iDDS workflow*
  - *Extend to DOE HPCs, in particular Perlmutter, collaborating with CSI, HEP-CCE, ATLAS ADC*
- EIC Simulation Infrastructure project (NPPS)
  - Apply new Geant4 capabilities in integrated fast/full simulation with AI/ML
  - Particular detector focus: Cerenkov detectors (ePIC has 3) with GPU-accelerated optical photon simulation
  - Target EIC while leveraging/benefiting wider NP, HEP
- Using AI to draw **trusted, quantified inferences from real data** (DUNE, sPHENIX, CSI)

# AI for the EIC: Long Term

- EIC with its unbiased data sample encompassing deep emergent physics phenomena is fertile ground for exploring the AI leading edge: how can we use foundation models to increase the physics yield of ePIC
- Timeline is long, being ready requires 10 years of work, apace with the technology
- At an AI4EIC BNL retreat last year we sketched out a possible timeline on two tracks, developing a ‘Large Particle Model’ and a cognizant integrated facility
  - Near term: 1-3 years
    - Complete R&D on using AI to draw trusted, quantified inferences from real data (ie continue the LS4GAN path)
    - Develop a prototype HENP LLM with current data (e.g. ATLAS, RHIC) and first generation feature extraction tools
  - Mid term: 3-5 years
    - Begin work on an integrated accelerator - experiment dataset and its AI instrumentation for a cognizant facility from accelerator to detector to analysis, targeting the EIC
  - Long term: 5-10 years
    - Documented analyses employing the HENP LLM (internal notes and/or peer reviewed publications)
    - Commission and deploy AI for EIC in parallel with machine and detector installation, commissioning and datataking
- The brainstorming informs the proposals we write
  - a BNL-internal one was just accepted on EIC computing with an AI component
  - R&D prototyping is going on for applying foundation model tech, towards FASST readiness (a possible but far from assured DOE AI funding call)

# Conclusion

- ePIC S&C
  - ePIC's streaming computing model presents novel and exciting challenges in the orchestration of fine grained streaming processing at the Echelon 1 centers
  - Distributed computing will extend to Echelon 2 and 3, with much LHC similarity
  - Workflow management must serve a wide range of workflows from streaming orchestration at E1s, to batch production at E1/E2, to complex calibration workflows, to opportunistic processing
  - Putting systems to the test, including Rucio+PanDA, begins this year
- AI/ML at BNL
  - Rich programs in both HEP and NP
- EIC aims to be the first large-scale collider-based program in which AI is integrated from the start
  - The ePIC experiment sees AI as already a key tool throughout
  - AI capable framework, applications developing throughout online, offline, analysis
- The EIC machine, detector(s), readout and analysis are conducive to AI approaches
  - Integrated facility from accelerator to detector to readout to analysis: 'cognizant' via AI
  - Streaming readout from front-end electronics to reconstruction/analysis in real time
    - Fast AI for on-the-fly data reduction, calibration, monitoring, reconstruction
  - A complete unbiased data sample dense with emergent physics phenomena
    - Will AI technologies 10 years from now be capable of deepening our physics insights?
    - Being ready to address the question requires R&D from today in a fast-moving field
- AI/ML at BNL benefits from strong collaborations among lab communities
  - RHIC, EIC, HEP experiments (ATLAS, DUNE, …), Computational Science Initiative (CSI)
- Drawing on support from DOE NP, HEP, ASCR as well as experiments

# More info

- ePIC collaboration meeting this week is public
  - Markus Diefenthaler's S&C report
- EIC website
- EIC/ePIC github (ePIC + common software)
- ePIC Streaming Computing Model Report (currently V2)
- ePIC streaming computing model WG meeting notes
- Science Requirements and Detector Concepts for the Electron-Ion Collider: EIC Yellow Report
- AID2E report at the DOE PI meeting Dec 2024
- AI4EIC 2023 Workshop (an ongoing series)

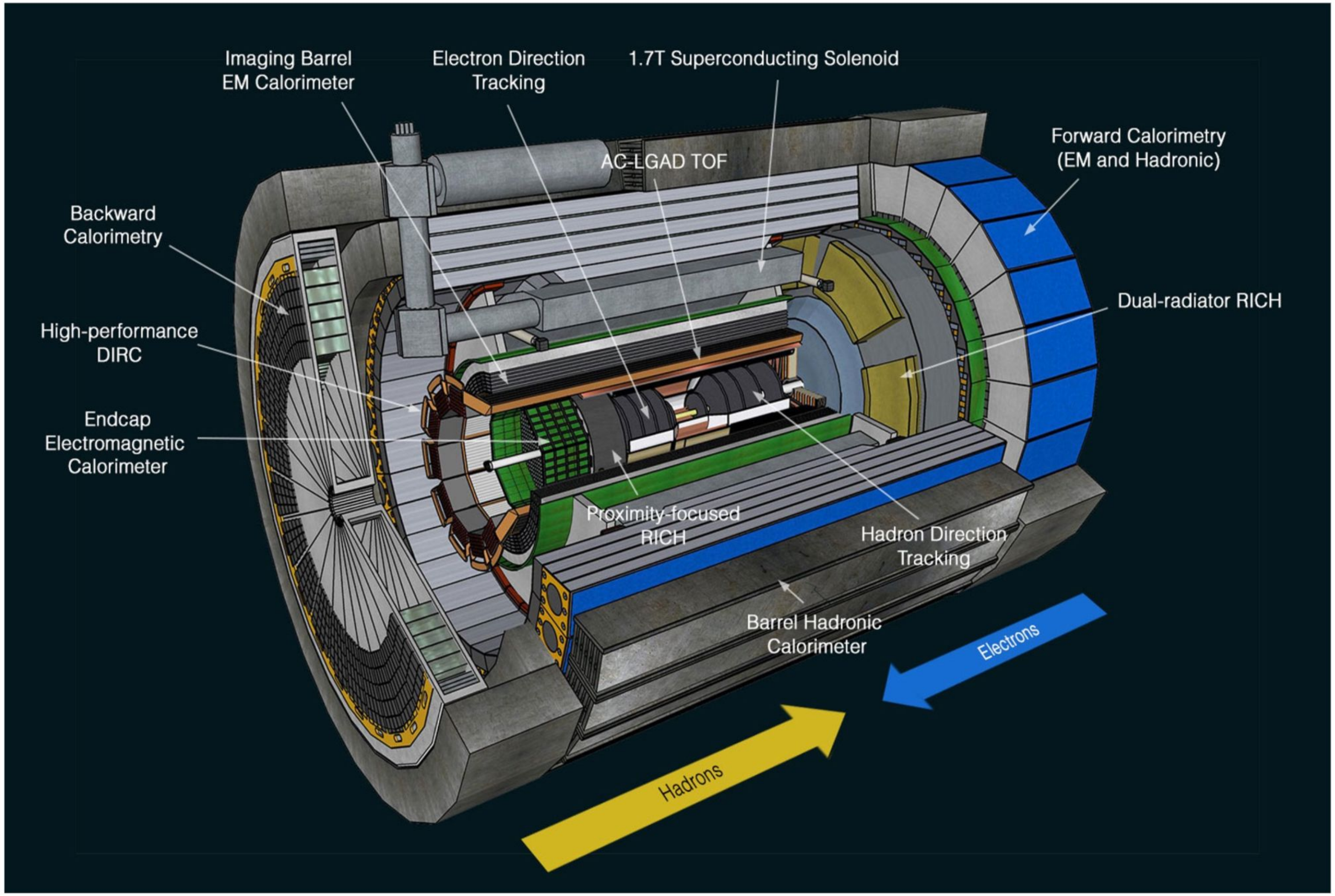
# Thank you

Thank you to ePIC S&C and HEP/NP AI folks, these and others!

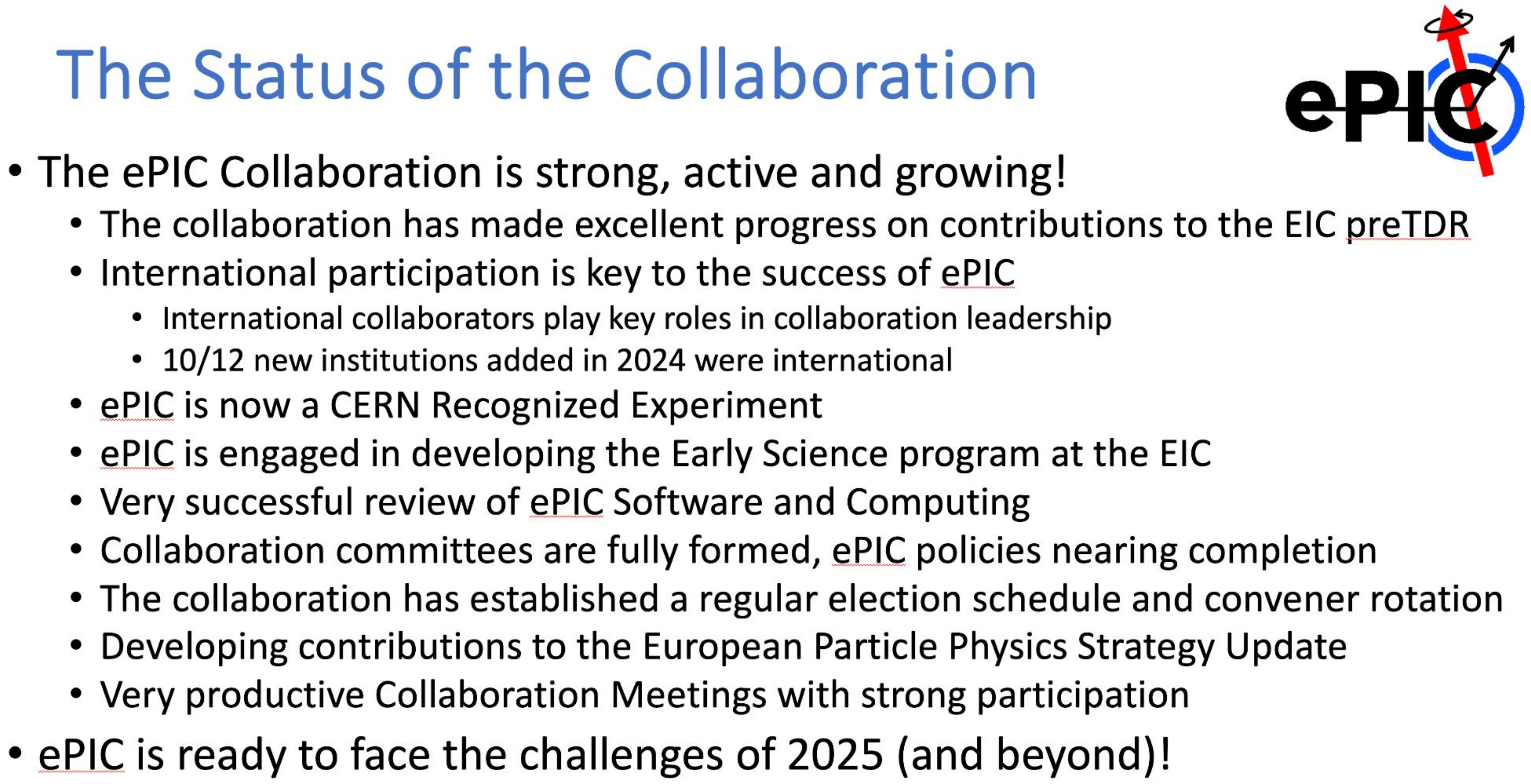
Thank you to Dmitry Arkhipkin, Elke Aschenauer, Kevin Brown, Wouter Deconinck, Markus Diefenthaler, Jamie Dunlop, Haiyan Gao, Yuan Gao, Wen Guan, Xiaofeng Gu, Adolfy Hoisie, Georg Hoffstaetter, Jin Huang, Alexander Jentsch, Dmitry Kalinkin, Kolja Kauder, Alexei Klimentov, John Lajoie, Jeff Landgraf, Meifeng Lin, Lucy Lin, Tadashi Maeno, Hong Ma, Jennefer Maldonado, Shigeki Misawa, Sergei Nagaitsev, Frank Rathmann, Ray Ren, John de Stefano, Brett Viren, Xin Zhao, Shinjae Yoo, my BNL NPPS and SDCC colleagues, and others whose work and ideas have contributed to what I've presented

# Supplementary

# ePIC Central Detector


Imaging Barrel EM Calorimeter
Electron Direction Tracking
1.7T Superconducting Solenoid
AC-LGAD TOF
Forward Calorimetry (EM and Hadronic)
Backward Calorimetry
Dual-radiator RICH
High-performance DIRC
Endcap Electromagnetic Calorimeter
Proximity-focused RICH
Hadron Direction Tracking
Barrel Hadronic Calorimeter
Electrons
Hadrons

# From the collaboration meeting underway

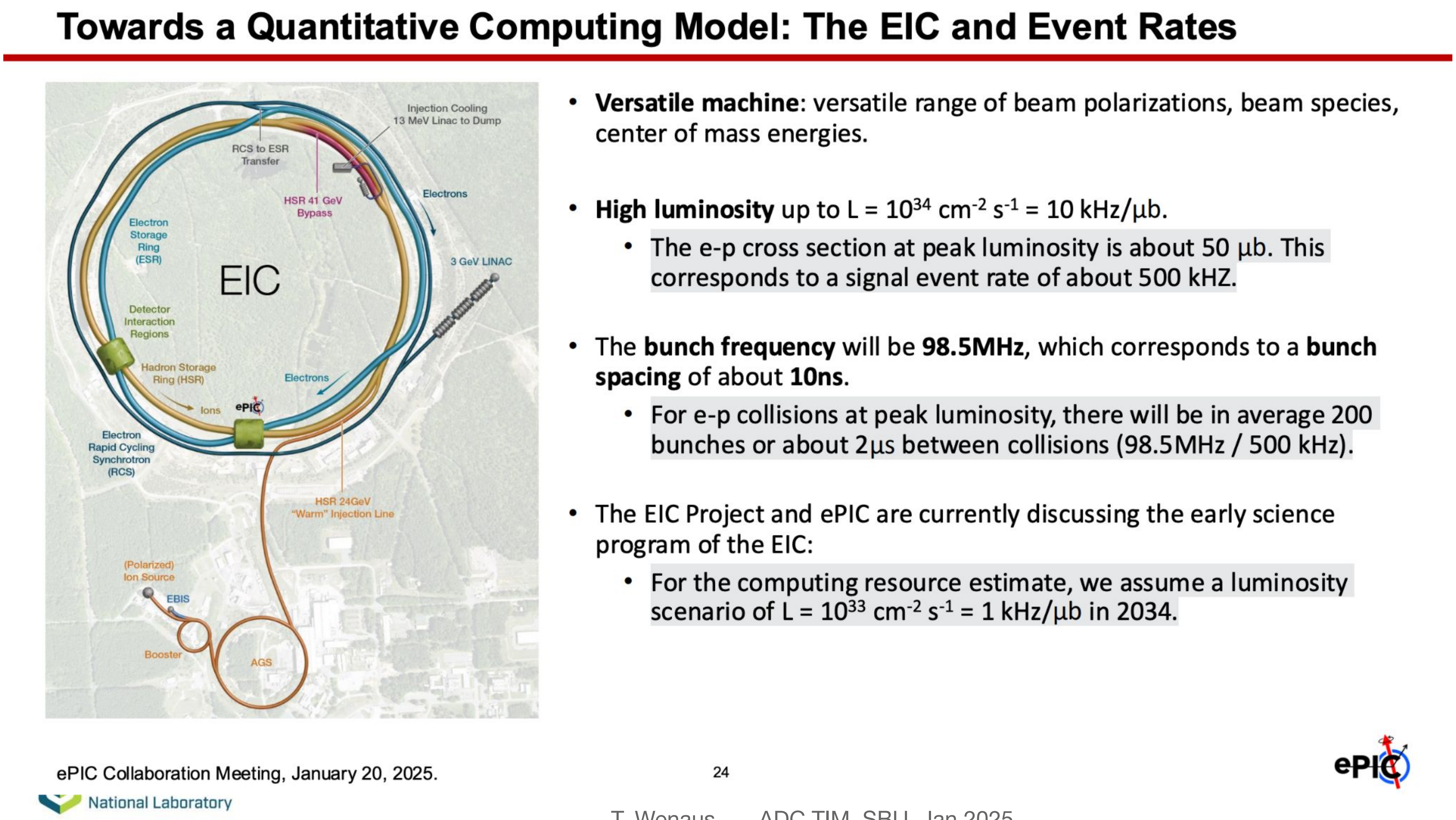


## The Status of the Collaboration

- The ePIC Collaboration is strong, active and growing!
  - The collaboration has made excellent progress on contributions to the EIC preTDR
  - International participation is key to the success of ePIC
    - International collaborators play key roles in collaboration leadership
    - 10/12 new institutions added in 2024 were international
  - ePIC is now a CERN Recognized Experiment
  - ePIC is engaged in developing the Early Science program at the EIC
  - Very successful review of ePIC Software and Computing
  - Collaboration committees are fully formed, ePIC policies nearing completion
  - The collaboration has established a regular election schedule and convener rotation
  - Developing contributions to the European Particle Physics Strategy Update
  - Very productive Collaboration Meetings with strong participation
- ePIC is ready to face the challenges of 2025 (and beyond)!

John Lajoie (ORNL), ePIC Spokesperson

# Basis of ePIC computing requirements estimates



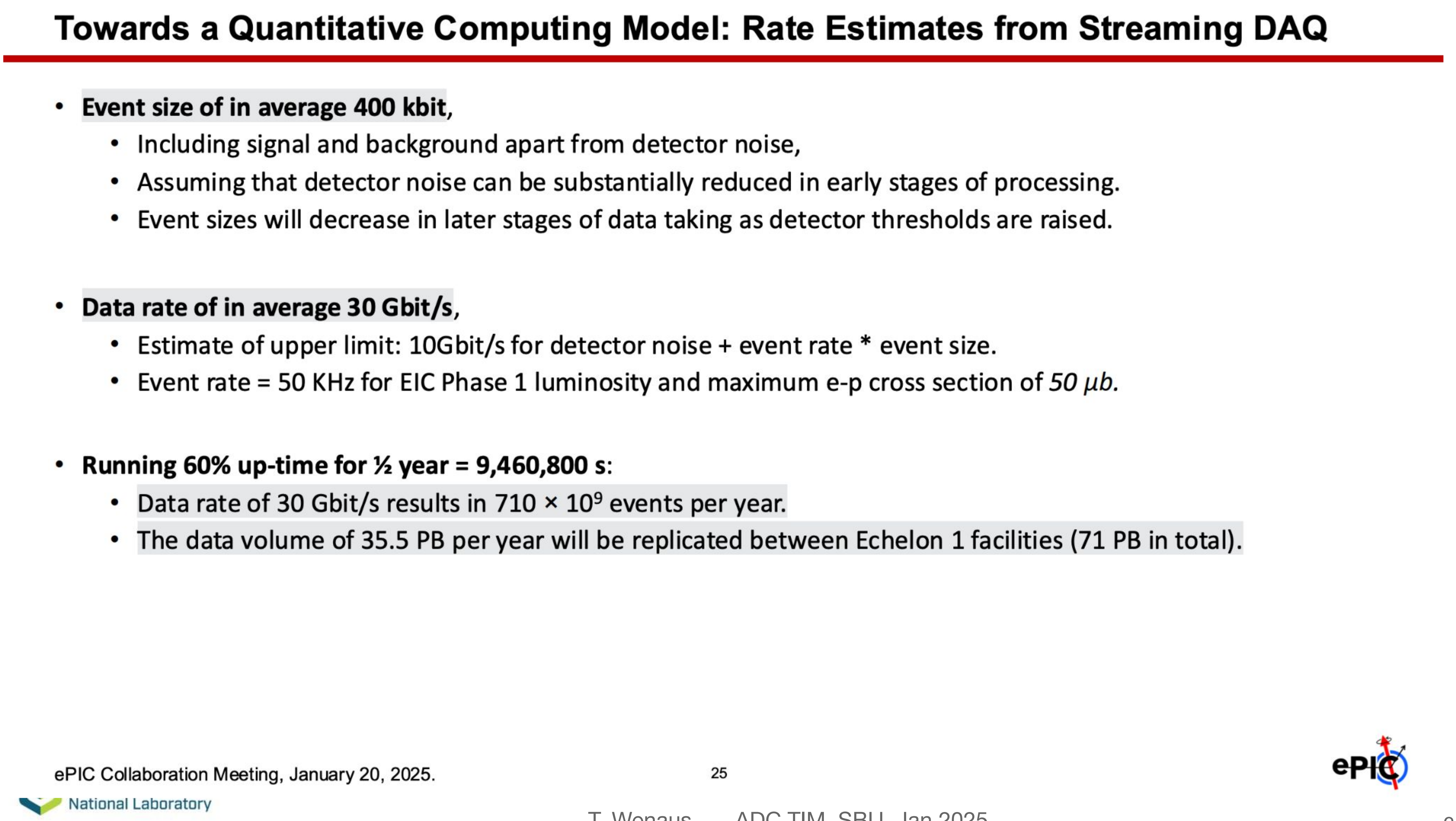
## Towards a Quantitative Computing Model: The EIC and Event Rates

- **Versatile machine:** versatile range of beam polarizations, beam species, center of mass energies.
- **High luminosity** up to L = $10^{34}$ $cm^{-2}$ $s^{-1}$ = 10 kHz/µb.
  - The e-p cross section at peak luminosity is about 50 µb. This corresponds to a signal event rate of about 500 kHZ.
- The **bunch frequency** will be **98.5MHz**, which corresponds to a **bunch spacing** of about **10ns**.
  - For e-p collisions at peak luminosity, there will be in average 200 bunches or about 2µs between collisions (98.5MHz / 500 kHz).
- The EIC Project and ePIC are currently discussing the early science program of the EIC:
  - For the computing resource estimate, we assume a luminosity scenario of L = $10^{33}$ $cm^{-2}$ $s^{-1}$ = 1 kHz/µb in 2034.

ePIC Collaboration Meeting, January 20, 2025. 24

National Laboratory

# Basis of ePIC computing requirements estimates



## Towards a Quantitative Computing Model: Rate Estimates from Streaming DAQ

- **Event size of in average 400 kbit**,
  - Including signal and background apart from detector noise,
  - Assuming that detector noise can be substantially reduced in early stages of processing.
  - Event sizes will decrease in later stages of data taking as detector thresholds are raised.

- **Data rate of in average 30 Gbit/s**,
  - Estimate of upper limit: 10Gbit/s for detector noise + event rate * event size.
  - Event rate = 50 KHz for EIC Phase 1 luminosity and maximum e-p cross section of *50 $\mu b$.*

- **Running 60% up-time for ½ year = 9,460,800 s**:
  - Data rate of 30 Gbit/s results in 710 × $10^9$ events per year.
  - The data volume of 35.5 PB per year will be replicated between Echelon 1 facilities (71 PB in total).

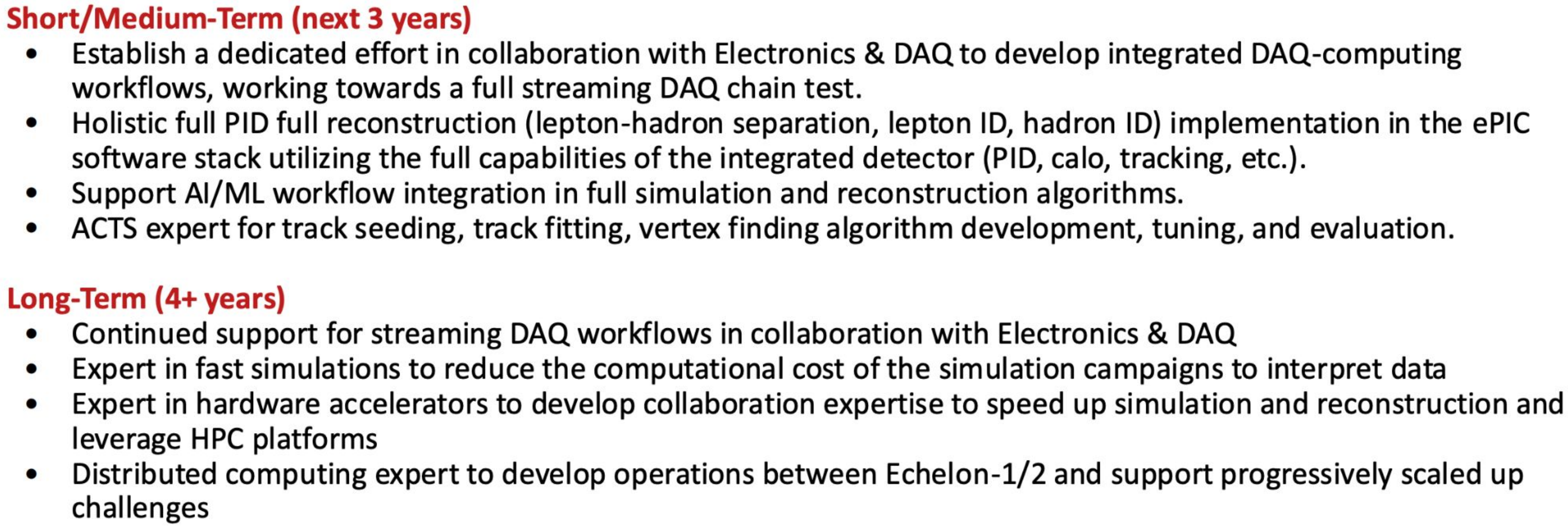
ePIC Collaboration Meeting, January 20, 2025.

25

National Laboratory

# ePIC S&C priorities

As reported to our reviewers last fall (“if you had more effort how would you use it?”)



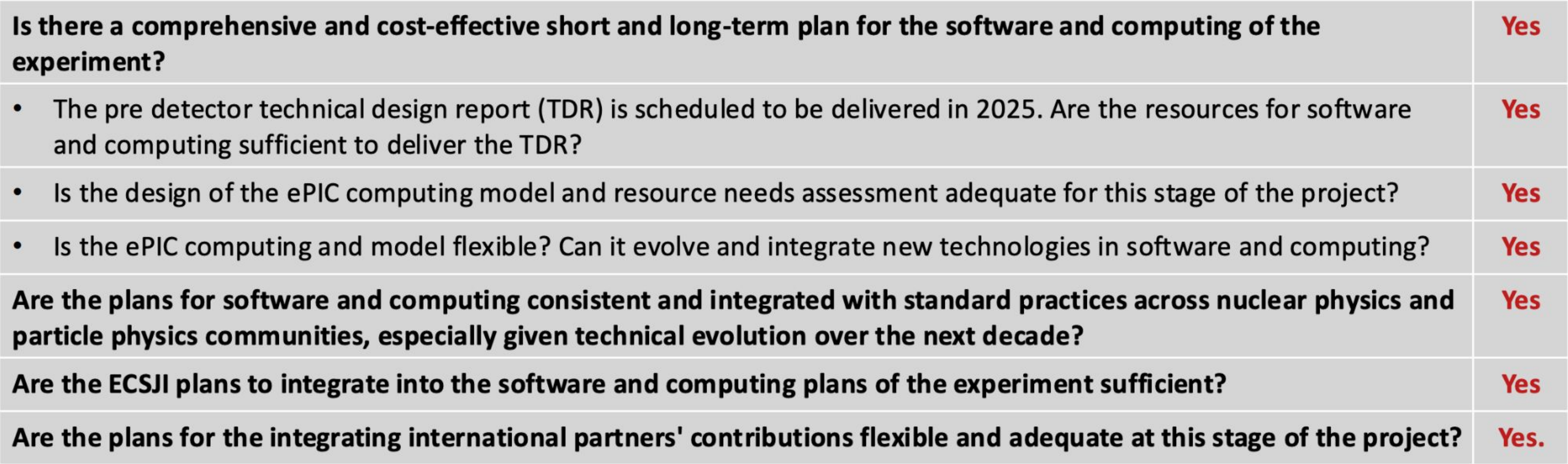
**Short/Medium-Term (next 3 years)**

- Establish a dedicated effort in collaboration with Electronics & DAQ to develop integrated DAQ-computing workflows, working towards a full streaming DAQ chain test.
- Holistic full PID full reconstruction (lepton-hadron separation, lepton ID, hadron ID) implementation in the ePIC software stack utilizing the full capabilities of the integrated detector (PID, calo, tracking, etc.).
- Support AI/ML workflow integration in full simulation and reconstruction algorithms.
- ACTS expert for track seeding, track fitting, vertex finding algorithm development, tuning, and evaluation.

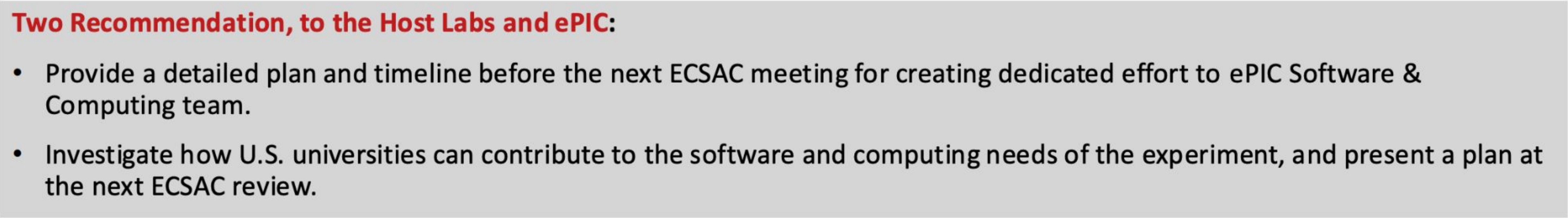
**Long-Term (4+ years)**

- Continued support for streaming DAQ workflows in collaboration with Electronics & DAQ
- Expert in fast simulations to reduce the computational cost of the simulation campaigns to interpret data
- Expert in hardware accelerators to develop collaboration expertise to speed up simulation and reconstruction and leverage HPC platforms
- Distributed computing expert to develop operations between Echelon-1/2 and support progressively scaled up challenges

# Outcome of Fall 2024 ePIC S&C review



| | |
|---|---|
| **Is there a comprehensive and cost-effective short and long-term plan for the software and computing of the experiment?** | Yes |
| • The pre detector technical design report (TDR) is scheduled to be delivered in 2025. Are the resources for software and computing sufficient to deliver the TDR? | Yes |
| • Is the design of the ePIC computing model and resource needs assessment adequate for this stage of the project? | Yes |
| • Is the ePIC computing and model flexible? Can it evolve and integrate new technologies in software and computing? | Yes |
| **Are the plans for software and computing consistent and integrated with standard practices across nuclear physics and particle physics communities, especially given technical evolution over the next decade?** | Yes |
| **Are the ECSJI plans to integrate into the software and computing plans of the experiment sufficient?** | Yes |
| **Are the plans for the integrating international partners' contributions flexible and adequate at this stage of the project?** | Yes. |



**Two Recommendation, to the Host Labs and ePIC:**

- Provide a detailed plan and timeline before the next ECSAC meeting for creating dedicated effort to ePIC Software & Computing team.
- Investigate how U.S. universities can contribute to the software and computing needs of the experiment, and present a plan at the next ECSAC review.

BNL ‘blue sky AI vision’ slide from last Spring for a DOE info call

# AI model challenge in Accelerator based Nuclear and Particle Physics

**Critical Opportunity:**

- Experiments at future accelerators such as the Electron Ion Collider (EIC) will employ data streaming systems that preserve virtually all the data such that a **fully unbiased study** of the data sample, together with accelerator data, can be made.
- These massive datasets, rich in the complex physics embedded within, are an ideal basis to draw on the **techniques of the AI LLM revolution to bring a transformative change** in deriving scientific insights from experimental HENP data.
- **We have the opportunity to build a fully cognizant facility from accelerator to detector and analysis.**
- **BNL is ideal for this work as the only US laboratory hosting multiple user facilities across different science domains, and is home to the only collider in the US**
  - **RHIC, US ATLAS and the EIC comprise some of the largest HENP datasets available today and in the future.**

**Expected Impacts:**

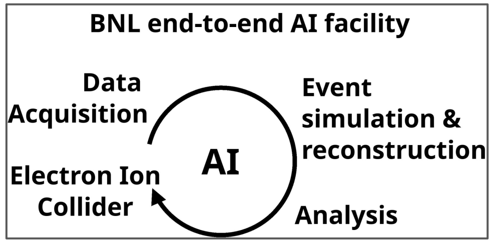
- Experimental HENP complexes such as the EIC are multi billion dollar enterprises that warrant application of the most sophisticated techniques to **maximize their discovery potential, teasing out the greatest possible science return**
- In economic impact,
  - AI-driven efficiency improvements will yield the same data in much less time, **reducing operations and energy costs**.
  - Can reduce the compute and storage intensive demands of HENP analysis, **saving compute and energy costs**.
    - Cost saving example: Improving the signal to background in recorded EIC data by 10% through the use of AI (intelligent DAQ) would lead to a $300k/year saving for archiving media.
  - Techniques developed for accelerators should be **readily transferable to medical and industrial applications**
- An ideal training ground for the nation’s AI workforce, ensuring **continued US leadership in this critical new technology**

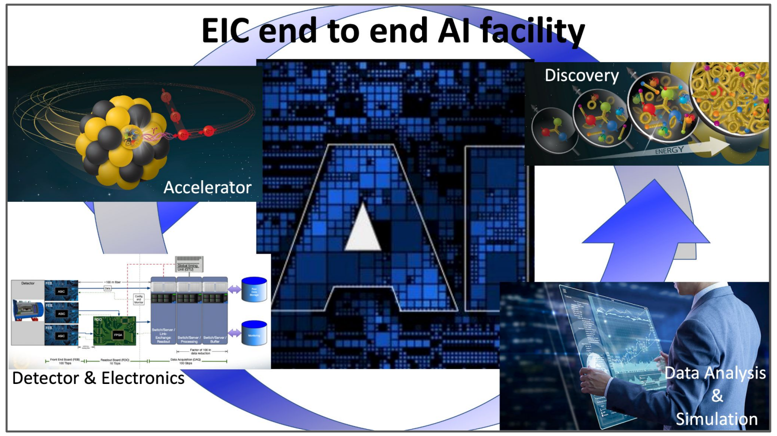
**Required R&D:**

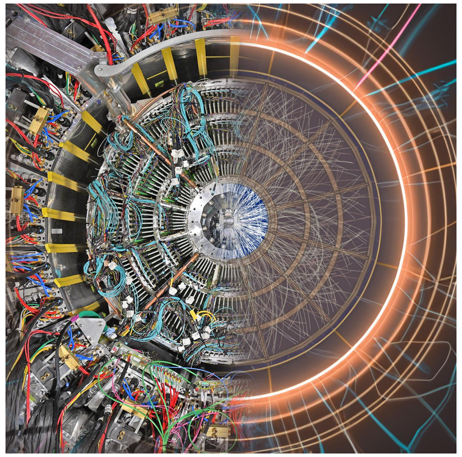
- AI dataset integration of accelerator, experiment and calibration/QA systems, e.g. for real-time optimization of luminosity and background; **optimization in both directions between the machine and the experiment**
- Create a **HENP LLM** (HENP data elements replacing ‘words’), and develop training and feature extraction techniques
- Develop techniques to **enhance trustworthiness in building and using AI**

**Timeline:**

- Near term: 1-3 years
  - Complete R&D on using AI to draw trusted, quantified inferences from real data ***[ie continue the LS4GAN path]***
  - Develop a **prototype HENP LLM with current data** (e.g. ATLAS, RHIC) and first generation feature extraction tools
- Mid term: 3-5 years
  - Begin work on an integrated accelerator - experiment dataset and its AI instrumentation for a cognizant facility from accelerator to detector to analysis, targeting the EIC
- Long term: 5-10 years
  - Documented analyses employing the HENP LLM (internal notes and/or peer reviewed publications)
  - Commission and deploy AI for EIC in parallel with machine and detector installation, commissioning and datataking







From detector to AI in sPHENIX