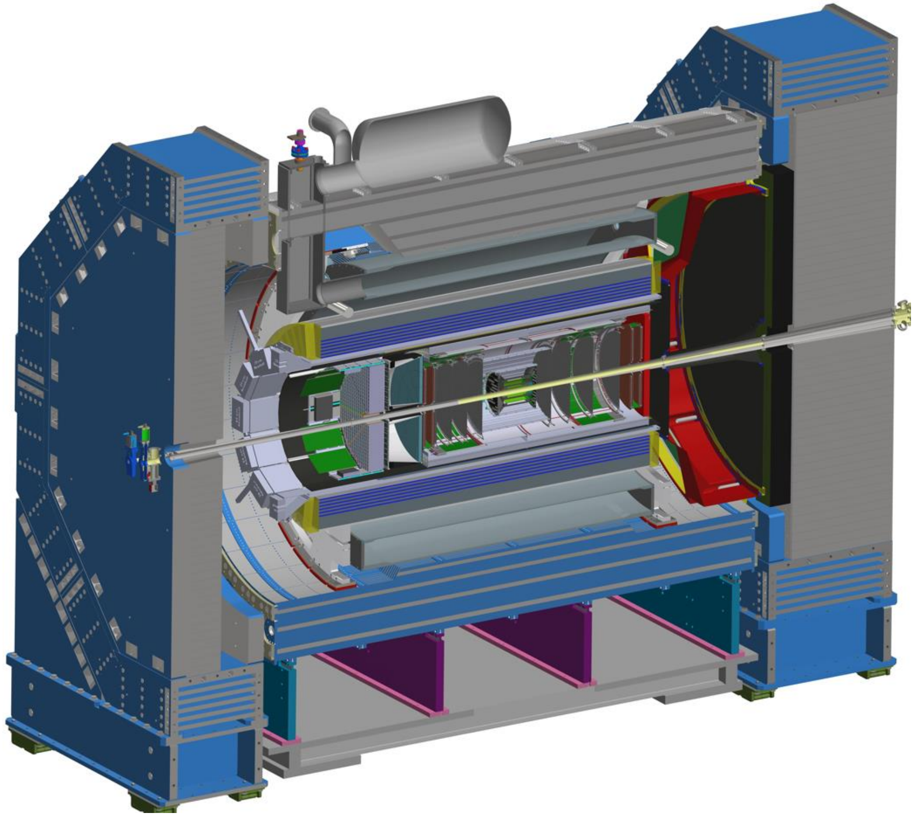
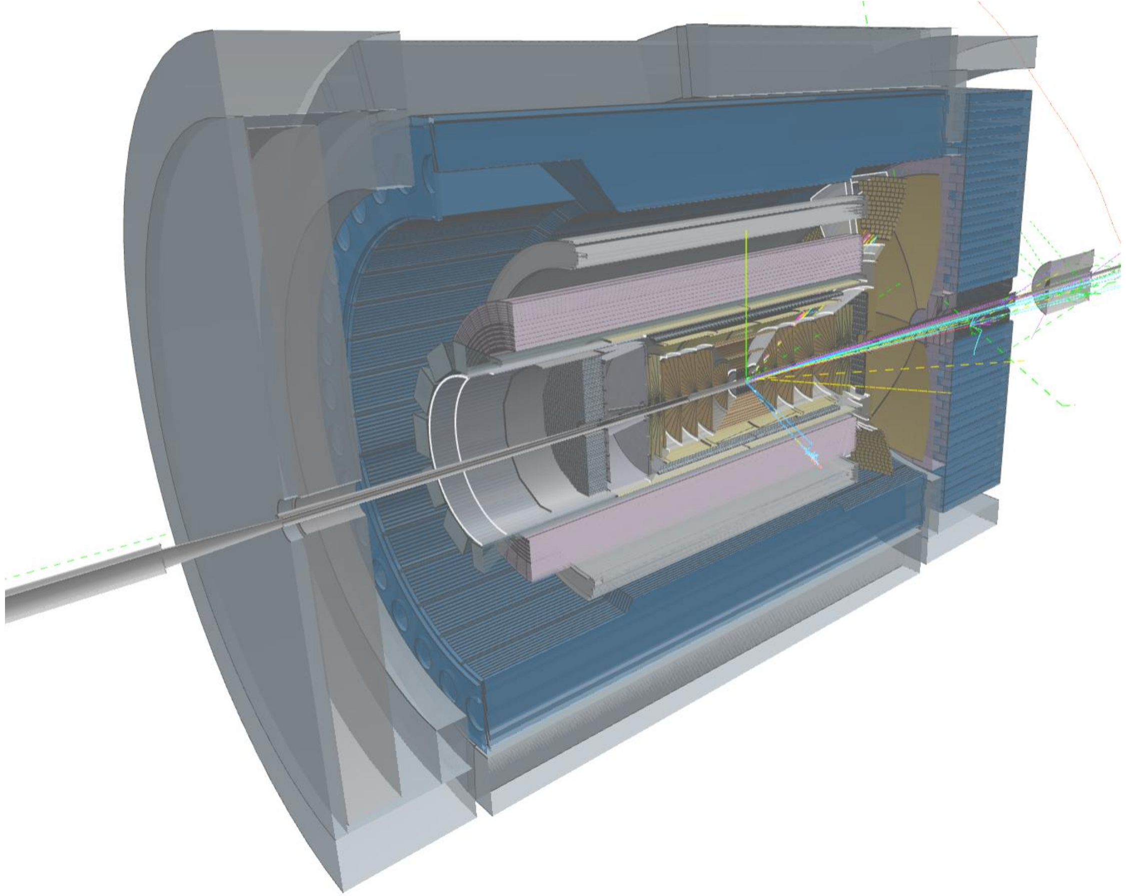
ePIC Software & Computing Report

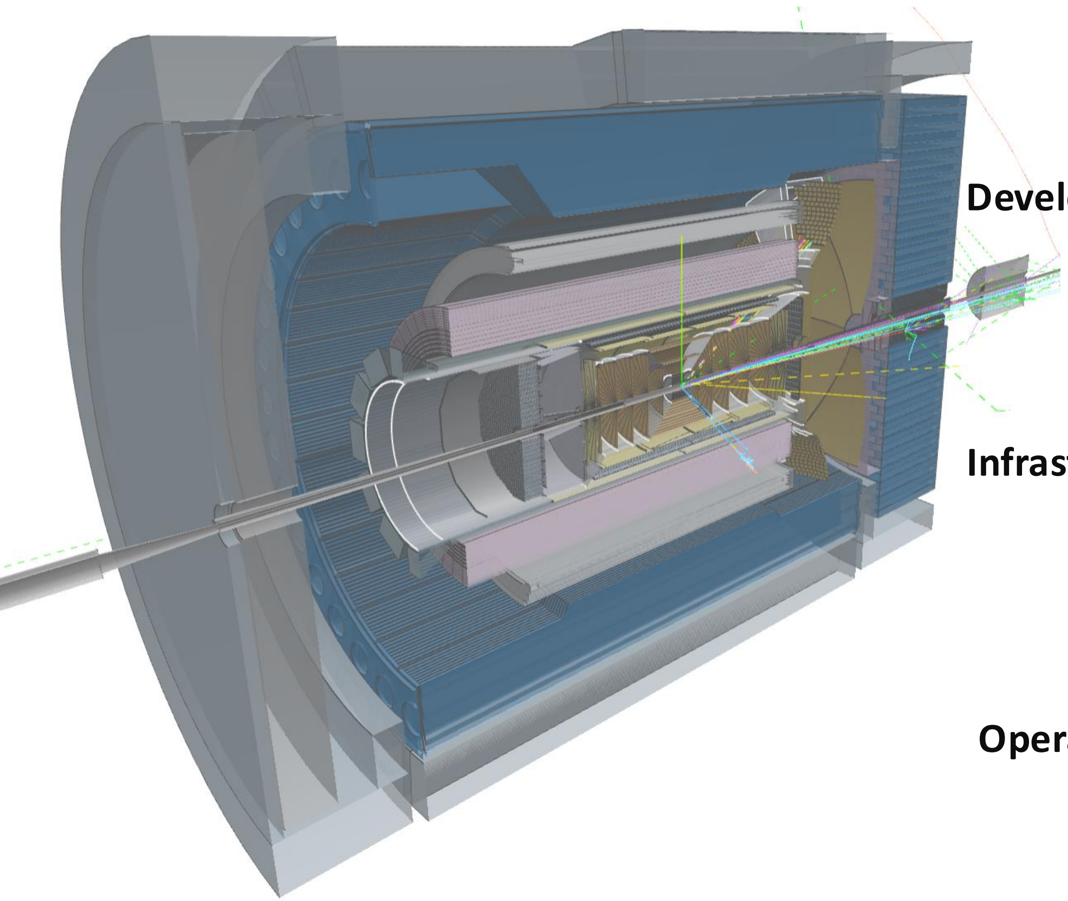




Markus Diefenthaler (Jefferson Lab) on behalf of ePIC Software & Computing

"Software is the Soul of the Detector"

Screenshot from the ePIC Event Display



Coordinators and WG Conveners

Development

Dmitry Kalinkin




Simulation WG: Chao Peng, Sakib Rahman. **Reconstruction WG**: Derek Anderson, Shujie Li.

Infrastructure

Torre Wenaus



Streaming Computing WG: Marco Battaglieri, Taku Gunji, Jeff Landgraf.

Operations

Wouter Deconinck




Production WG: Thomas Britton, Sakib Rahman. **User Learning WG**: Stephen Kay, Holly Szumila-Vance.

Current Priorities

Computing Model and Testbeds

Onboarding and Community Building

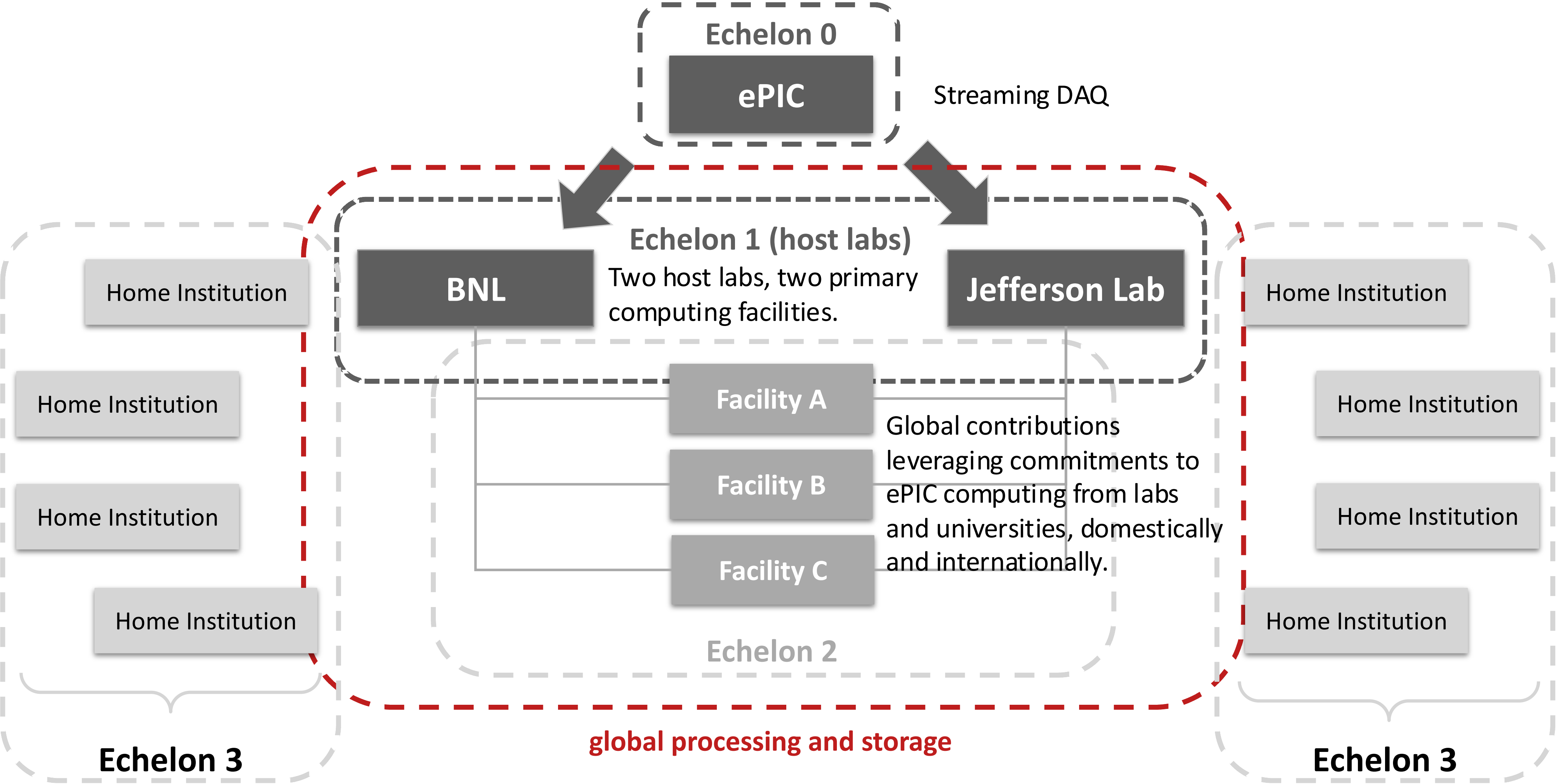
Software and Simulations for the preTDR

Compute-Detector Integration to Maximize and Accelerate Science

- **Maximize Science** **Capture every collision signal, including background.**
  - Event selection using all available detector data for **holistic reconstruction**:
    - **Eliminate trigger bias** and provide accurate estimation of uncertainties during event selection.
  - Streaming background estimates ideal to **reduce background** and related systematic uncertainties.
- **Accelerate Science** **Rapid turnaround of 2-3 weeks for data for physics analyses**.
  - Timeline driven by alignment and calibration.
  - Preliminary information from Detector Subsystem Collaborations indicates that 2-3 weeks are realistic.
- **Technologies** Compute-detector integration using:

Streaming readout for continuous data flow of the full detector information.	**AI** for rapid processing (autonomous alignment, calibration, and validation).	**Heterogeneous computing** for acceleration (CPU, GPU).

The ePIC Streaming Computing Model



Supporting the analysis community *where they are* at their home institutes, primarily via services hosted at Echelon 1 and 2.

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		✓	✓	

Substantial role for Echelon 2 in preliminary resource requirements model

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

Quote from EIC Computing & Software Advisory Committee (ECSAC): The power of distributed computing lies in its flexibility to shift processing between facilities as needed.

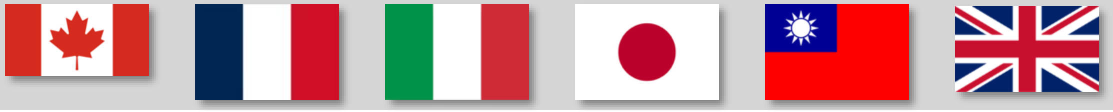
Echelon 1 sites uniquely perform the **low-latency streaming workflows** consuming the data stream from Echelon 0:

- Archiving and monitoring of the streaming data, prompt reconstruction and rapid diagnostics.

- Apart from low-latency, **Echelon 2** sites fully participate in use cases and **accelerate** them:
  - Tentative resource requirements model assumes a **substantial role for Echelon 2**.
  - Capabilities and resource requirements for Echelon 2 sites are developed jointly with the community.

EIC International Computing Organization (EICO) formed:

- Charter being reviewed by EIC RBB.



Computing Resource Needs (EIC Phase I) and Their Implications

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**

O(1M) core-years to process a year of data:

- Optimistic scaling of constant-dollar performance gains would reduce the numbers about 5x:
  - Based on current WLCG measure of 15% per year.
  - But the trend is towards lower gains per year.
- Whatever the gains over time, processing scale is substantial!
- Motivates attention to leveraging distributed and opportunistic resources from the beginning.

~350 PB to store data of one year.

Computing resource needs at a scale of ATLAS and CMS today.

ePIC is compute-intensive experiment; must ensure ePIC is not compute-limited in its science.

ePIC

Streaming Computing: Getting to the Specifics

July 15: Parallel Session

- We have our **streaming computing model document** and an evolving conception of E0-E1 dataflow and workflows, developed in an active **Streaming Computing WG meeting series**.

- Emphasis now is **moving from reports and schematics to the specifics**.
- **Prototyping ideas and tools in testbeds**, guided by requirements.
    - Gathering input on **requirements document for streaming orchestration**:

Recommendation from GRETA experiment: System testing needs orchestration. Prioritize it early.

    - **Testbed plans** are taking concrete shape:

Streaming orchestration: Developing E0-E1 streaming workflows in testbed utilizing Rucio and PanDA.

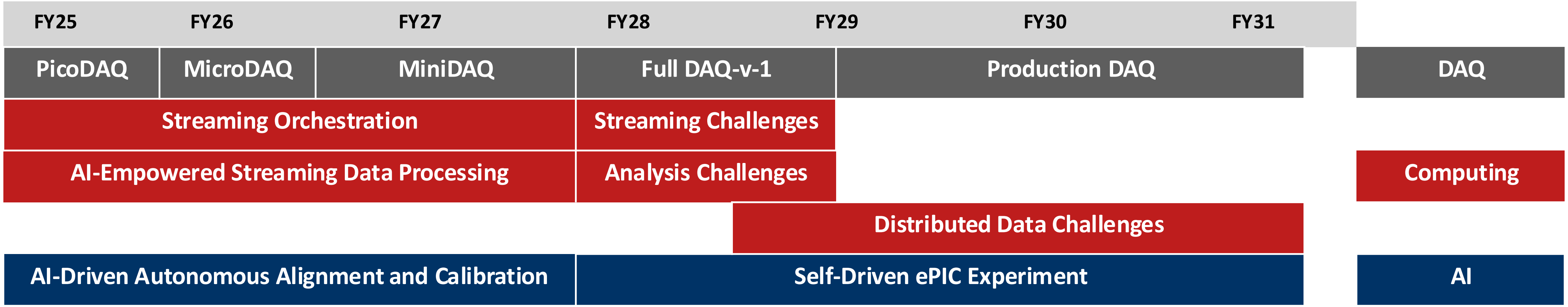
Rapid data processing: Describing and executing complex calibration workflows with their dependencies.

Streaming reconstruction: Raw data stream to collision event identification to reconstruction and analysis.

Streaming analysis: Demonstrate simulation data production streaming to Echelon 2 site.

Ongoing discussions on data handling, storage, and archiving may lead to an additional testbed.

Streaming DAQ and Computing Milestones



- **Compute-Detector Integration**:
  - Joint deliverables between **DAQ** and **computing** to develop integrated systems for detector readout, data processing, and ultimately physics analysis.
  - **Key role of AI**: Empowering data processing and enabling autonomous experimentation and control.

Technical Interchange Meeting

ePIC Software & Computing Report

https://doi.org/10.5281/zenodo.14675920

The ePIC Streaming Computing Model
Version 2, Fall 2024

Marco Battaglieri[1], Wouter Deconinck[2], Markus Diefenthaler[3], Jin Huang[4], Sylvester Joosten[5], Dmitry Kalinkin[6], Jeffery Landgraf[4], David Lawrence[3] and Torre Wenaus[4]
for the ePIC Collaboration

[1]Istituto Nazionale di Fisica Nucleare - Sezione di Genova, Genova, Liguria, Italy.
[2]University of Manitoba, Winnipeg, Manitoba, Canada.
[3]Jefferson Lab, Newport News, VA, USA.
[4]Brookhaven National Laboratory, Upton, NY, USA.
[5]Argonne National Laboratory, Lemont, IL, USA.
[6]University of Kentucky, Lexington, KY, USA.

Abstract

This second version of the ePIC Streaming Computing Model Report provides a 2024 view of the computing model, updating the October 2023 report with new material including an early estimate of computing resource requirements; software developments supporting detector and physics studies, the integration of ML, and a robust production activity; the evolving plan for infrastructure, dataflows, and workflows from Echelon 0 to Echelon 1; and a more developed timeline of high-level milestones. This regularly updated report provides a common understanding within the ePIC Collaboration on the streaming computing model, and serves as input to ePIC Software & Computing reviews and to the EIC Resource Review Board. A later version will be submitted for publication to share our work and plans with the community. **New and substantially rewritten material in Version 2 is dark green. The present draft is preliminary and incomplete and is yet to be circulated in ePIC for review.**

1

Purpose:

- Discuss the ePIC (Streaming) Computing Model paper, as presented to the EIC Computing & Software Advisory Committee (ECSAC) in 2023 and 2024,
- Gather feedback and comments from ECSAC.

Outcome:

- Documented ECSAC questions, ePIC responses, and comments made during the meeting (13 pages).
- Meeting helped improve the manuscript by clarifying key points and addressing inconsistencies.
- Also supported preparation for the upcoming review, likely in October.

Key Discussion Points:

- **ECSAC Role**: Scope of the ECSAC review begins with data at the output buffer of E0.
- **Interface Between E0 and E1**: Discussion focused on data aggregation, transfer, and low-level processing.
- **Echelon 2**: Discussed role of E2 in the distributed computing model.
- **Analysis Infrastructure**: Discussed analysis model involving E1, E2, E3 sites.

ePIC

Current Priorities

Computing Model and Testbeds

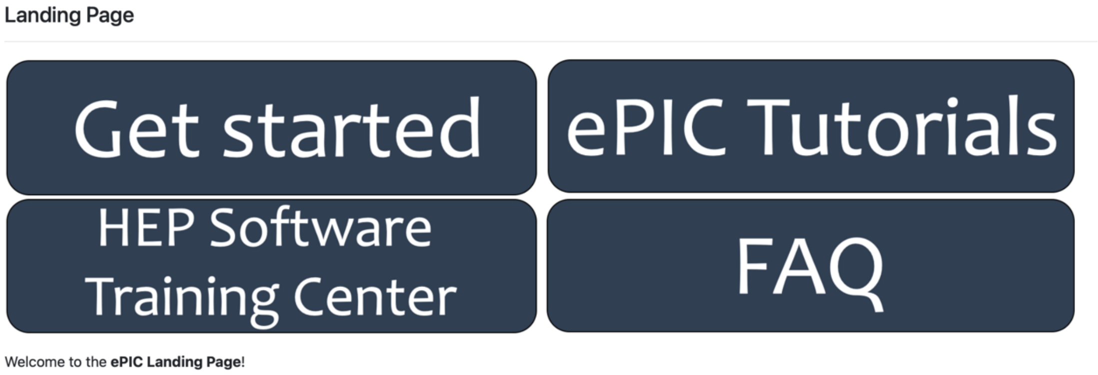
Onboarding and Community Building

Software and Simulations for the preTDR

User Learning

Successful Landing Page for Onboarding

- Many new collaborators successfully onboarded themselves using only the landing page.
- Regarded as a *"game changer"* by Physics Analysis Coordinators:
  - Significantly improved engagement and progress in the Physics WGs.

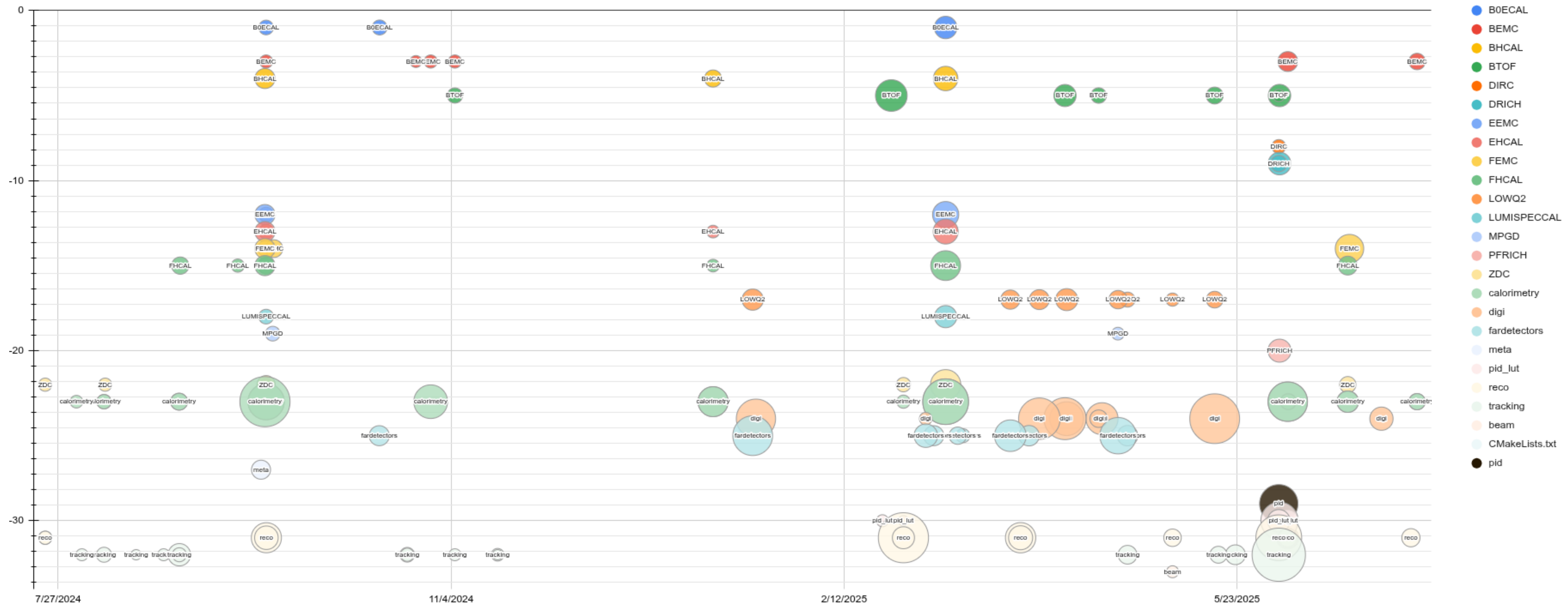


Any member of the collaboration can **directly contribute by submitting change requests**.

Tutorial Series Continues

- Understanding the Simulation Output (Shujie Li): How is the information in the simulation output generated (algorithms, data model)?
- Analysis and Working with the Simulation Output (Stephen Kay): How to work with the simulation output for detector and physics studies?
- Getting Started with a Physics Analysis (Alex Jentsch): How does one get started with a physics analysis based on the simulation output?
- Inclusive Kinematics Reconstruction (Stephen Maple): How can one obtain the most accurate information about inclusive DIS kinematics?

DSC Contributions to the ePIC Software Stack



Contributions to the ePIC Software Stack vary across DSCs:

- For some, this may indicate that they have reached a sufficient level of implementation for now. For others, it may point to challenges that are limiting their ability to contribute.
- We use this kind of metadata analysis to gain early insights and to identify where additional support or investigation may be needed (More on slide 19).

Community Building in

The HSF-India project aims to join networks in India to networks in the U.S. and Europe in order to build international research software collaborations.

HSF-India/ePIC Meeting

- **Five introductory talks covering** the EIC and the science it will enable, a theory perspective on spin physics, EIC/ePIC in India, research software collaborations, and contributing to ePIC.
- Several long-format tutorials by:
    - Chandradoy Chatterjee (INFN Trieste)
    - Stephen Kay (University of York)
    - Charlotte Van Hulse (University of Alcalá)
- **Meeting Survey**: Students found the workshop content very useful and appropriately paced.
- Planning a hackathon-style follow-up later this year to encourage students to initiate their own projects.

44 registered participants, including 33 master's and doctoral students
Students from: Central University of Karnataka, Central University of Tamil Nadu, IIT Bombay, IIT Indore, IIT Madras, IIT Mandi, Malaviya National Institute of Technology, Ramaiah University of Applied Sciences

ePIC

Discoverable Software

EIC SOFTWARE:
Statement of Principles

4. We will aim for user-centered design:

- We will enable scientists of all levels worldwide to actively participate in the science program of the EIC, keeping the barriers low for smaller teams.
- EIC software will run on the systems used by the community, easily.
- We aim for a modular development paradigm for algorithms and tools without the need for users to interface with the entire software environment.

- **Part of User-Centered Design**: How easy is it to **find**, **understand**, and **use** our software for detector and physics studies?
- As an initial step, we aim to assess the following:
  - How **discoverable** is our software?
  - What actions can we take to improve its **discoverability**?
    - **We need to set priorities**.
- To address these questions:
  - Ongoing **survey** among **Early Career** scientists:
    - Shared during Early Career Workshop on July 11
    - Excerpt shown on the right.
  - **Focus group discussions** after the survey.

Where do you prefer to access documentation?

- [ ] Dedicated Website
- [ ] In-code comments and docstrings
- [ ] README files in repositories
- [ ] Wiki or collaborative documentation platform
- [ ] Dowloadable PDFs/manuals
- [ ] Other: ____

What types of documentation are most useful to you?

- [ ] Step-by-step examples or tutorials
- [ ] Reference documentation
- [ ] Conceptual overviews
- [ ] Troubleshooting guides
- [ ] FAQs
- [ ] Other: ____

On a scale of 1 to 10, how helpful are code examples in learning the software?

	1	2	3	4	5	6	7	8	9	10	
Not helpful at all	○	○	○	○	○	○	○	○	○	○	Very helpful

Current Priorities

Computing Model
and Testbeds

Onboarding and Community Building

Software and Simulations for the preTDR

Priorities for 2025

- **Improved Communication**
  - **Biweekly Software News**: Regular reports highlighting major changes to the software stack, notable code merges, and updates from the WGs.
  - **Meeting Notes**: Summaries of WG meetings, including outcomes and next steps, are provided to enable asynchronous participation.
- **Updated WG Charges and Priorities** (right and slides 25–29)
  - Reflect outcomes of the collaboration meeting in Frascati, Italy.
  - Example priorities include background integration in simulations and resolving discrepancies between engineering and simulation designs.
- **Collaboration with Physics WGs**
  - **Ongoing coordination** on **simulation targets**, **reconstruction**, and **analysis tools**.
  - Support for analysis tutorials.
- **Coordination with DSCs** (slide 19)
- **Collaboration with EICUG**
  - ePIC Software used for Detector II.
  - MC4EIC 2025.

July 15: Ilkka Helenius will report on MC4EIC.

Physics and Detector Simulation

- **Charge**:
  - Development of accurate MC simulations using a suite of physics and background generators and detector simulation based on Geant4 and DD4hep
- **Priorities for 2025**:
  - Continue to support the **detector design** and integration with services.
  - Collaborate with the EIC Project to evaluate the **differences between the engineering and simulation designs**, and lead discussions with the DSCs on how to address these differences.
  - Continue to support the development of **background modeling** and implement its timing structure in physics and detector simulations, together with the Background TF.
  - Enable **simulation of streaming readout** by providing the option to switch between streaming data and event data modes.
  - Coordinate the **development of digitization and noise models** with the DSCs and the Electronics and DAQ WG.

Reconstruction Framework and Algorithms

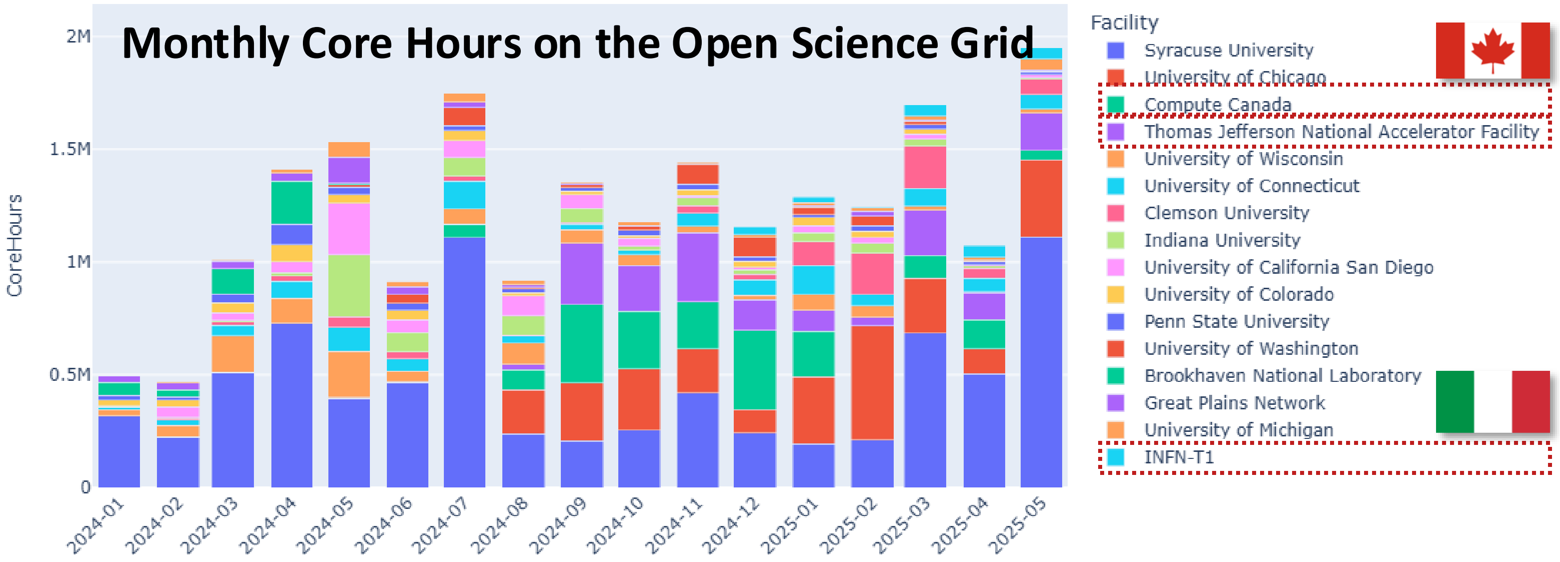
- **Charge**:
  - Development of a holistic and modular reconstruction for the integrated ePIC detector.
- **Priorities for 2025**:
  - Drive the **development of the reconstruction framework to meet ePIC needs**, e.g., on modularity or streaming data processing.
  - Host collaboration-wide discussions on all aspects of reconstruction, driving the **work toward holistic reconstruction**.
  - Enable reconstruction algorithms to **handle physics events with background**.
  - Collaborate with **PWGs** on **shared reconstruction priorities**, which currently include:
    - Secondary vertexing
    - Hadron identification
    - Particle flow algorithms for jet reconstruction
    - Event kinematics
  - Integrate continued development of **web-based event display** in reconstruction efforts.

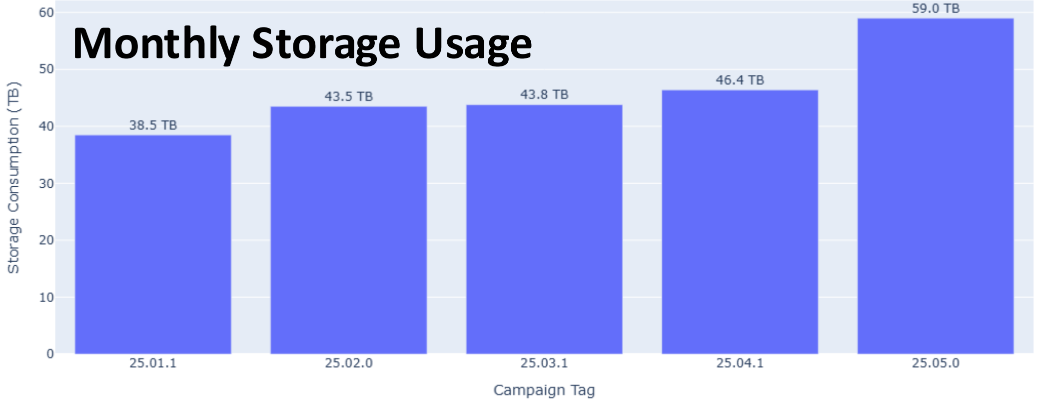
The specific charges and priorities for each WG are provided on slides 25 to 29.

Simulation Campaigns

Production Liaisons Wanted

- We provide **simulation productions including background** tailored to the needs of the collaboration, as defined by the DSCs and PWGs.
- Simulation campaigns are conducted monthly, based on the software release for the corresponding month (e.g., 25.05 for May 2025).
- These simulations serve as the standard for detector and physics studies for the preTDR and also the Early Science Program.
- In the past year, monthly simulation campaigns consumed approximately 15 million core hours on the Open Science Grid (OSG), generating over 500 TB of simulation data.





We are capable of integrating new detector geometry and algorithms within a month, processing millions of events needed to assess scientific impact. See Sakib Rahman's announcement of the June campaign (25.06) for latest details.

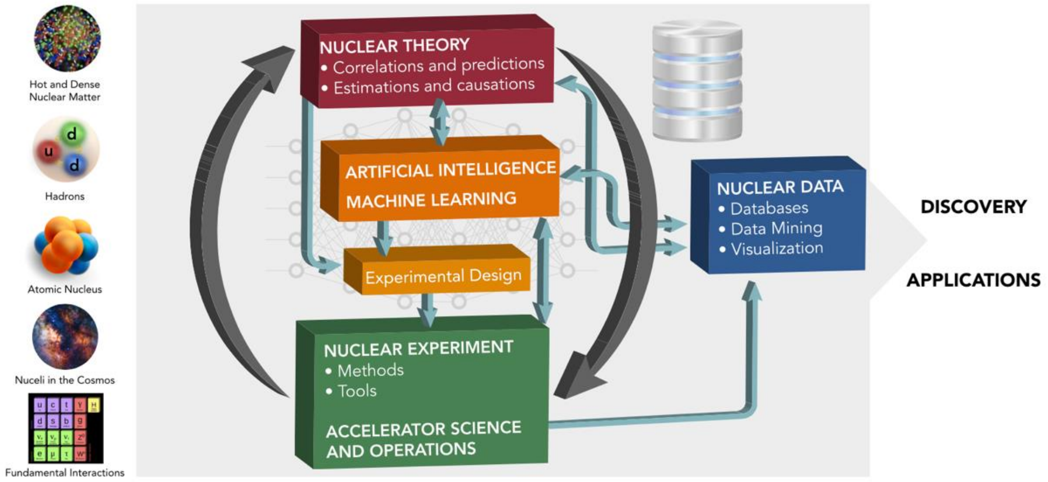
Software & Simulation Priorities Across Detector Systems

- **Coordination with DSCs**:
  - Requested **updates on software priorities** to facilitate the integration of DSCS simulation efforts into ePIC software and simulation campaigns.
    - **Example**: Improved shared development by aligning four DSCs under a common digitization work plan.
  - **Almost all** detector groups submitted priority lists.
- **Common Themes Across DSCs**:
  - Need for improved geometry modeling, often linked to mechanical design.
  - Integration of realistic detector effects (digitization, noise, optical photons).
  - AI/ML integration for reconstruction.
  - Benchmarking and validation tools requested widely.
- **Challenges Identified**:
  - Workforce limitations and tool access barriers (e.g., CAD tools, database infrastructure).
- **Next Steps & Coordination Needs**:
  - Streamline simulation realism needs across DSCs (digitization, noise).
  - Continue support on reconstruction tools, ML integration, and validation workflows.
  - Support of test-beam measurements.

July 16: Parallel Session and Hackathon

Thank you to the DSCs for the detailed input. To complete the picture as much as possible, we have awaited responses from most DSCs. We will share the detailed responses by next week at the latest.

A Word on AI



Numerous AI prototypes and R&D efforts in ePIC span the full workflow—from low-level data reduction to reconstruction, classification, and analysis.

- Monthly **simulation campaigns incorporate AI techniques**, provided to the whole collaboration:
  - Currently focused on reconstruction (EEEMCal, low Q2 tagger, roman pots, event kinematics).
- AI workflows are **supported** through containerized environments including **JAX**, **PyTorch**, and **TensorFlow**, **ONNX** for inference, and the **ePIC data model**:
  - The data model enables data-driven API design, promoting modularity through standardized interfaces across simulation, reconstruction, and analysis tasks.
- Simulation data is accessible from both Python and ROOT, ensuring interoperability across analysis and data science environments.
- Actively building connections to AI efforts in the wider scientific community.

Summary

Computing Model and Testbeds

- **Compute-detector integration** using **streaming readout**, **AI**, and **distributed computing**.
- ECSAC feedback on **Streaming Computing Model paper**; next step circulation within ePIC.
- **Testbeds launched** for streaming orchestration, rapid processing, reconstruction, and analysis.
- ePIC will be **compute-intensive experiment**; **substantial role for Echelon 2** foreseen.
- EIC Intl. Computing Organization (**EICO**) formed; draft charter being reviewed by EIC RRB.

Onboarding and Community Building

- Successful **landing page** for onboarding new collaboration members.
- **Tutorial series** on a rolling schedule, adapted to the evolving needs of the collaboration.
- **Community building** through in-person software weeks, including recent **HSF-India/ePIC meeting**.
- Using Early Career feedback to assess and improve **software discoverability**.

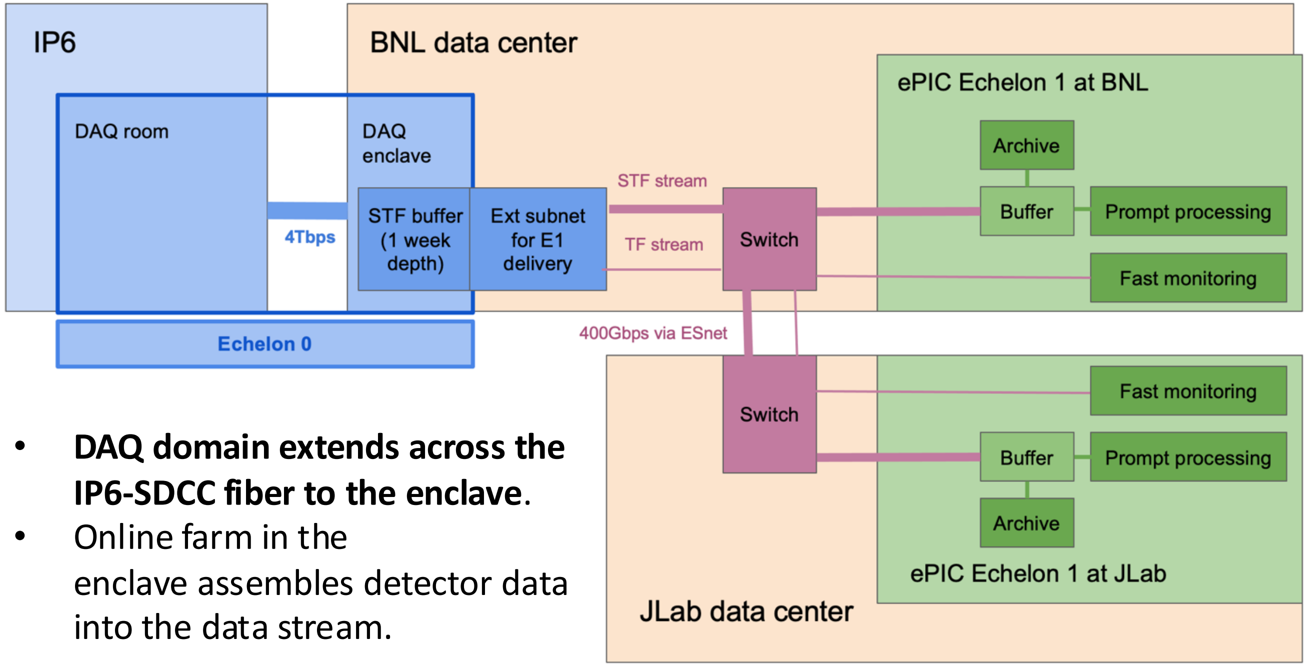
Software and Simulations for the preTDR

- **2025 priorities** emerged from discussions at the Frascati collaboration meeting.
- Simulation, Reconstruction, Streaming Computing Model, Production, and User Learning **WGs are delivering** on 2025 priorities.
- **Coordination with Physics WGs** on simulation targets, reconstruction, and analysis tools.
- **DSCs** provided valuable **input on software and simulation priorities**; common needs include improved geometry modeling, realistic detector effects.
- Progress communicated via **meeting notes** and **biweekly Software News**.

Backup

Echelon 0 – Echelon 1 Data Flow and Processing

- EIC Echelon 0 + 1 half-day mini-workshop at BNL on April 2 focused on planning, particularly around data streaming from Echelon 0 to 1 and testbed activities. It emphasized collaboration between computing facilities at BNL and Jefferson Lab and ePIC.



- **DAQ domain extends across the IP6-SDCC fiber to the enclave.**
- Online farm in the enclave assembles detector data into the data stream.

Data stream consists of:

Time Frame (TF):
- 0.6ms of detector data,
- Used for low-latency use case, e.g., monitoring.

Super Time Frame (STF):
- Contiguous blocks of ~1000 TFs,
- Atomic unit for raw data processing after the DAQ.

- **Echelon 1 Sites are symmetric peers**:
  - Handle archiving, prompt processing, monitoring.
  - Maintain redundant raw data copies.
  - Leverage ePIC distributed computing capabilities supporting the E0/1/2/3 streaming computing model.

ePIC Software & Computing Organization

Spokesperson's Office

Detector | **Software & Computing** | Physics

Guiding Principles:

- *Statement of Software Principles*
- *Sustainability*

Software and Computing Coordinator
Markus Diefenthaler (Jefferson Lab)

Cross-cutting Working Group:

- *Data and Analysis Preservation (not yet activated)*

Deputy Coordinator (Operations)
Wouter Deconinck (U. Manitoba)

Operation Working Groups:

- Production
- User Learning
- *Validation (paused)*

Deputy Coordinator (Development)
Dmitry Kalinkin (U Kentucky)

Development Working Groups:

- Physics and Detector Simulation
- Reconstruction Framework and Algorithms
- *Analysis Tools (not yet activated)*

Deputy Coordinator (Infrastructure)
Torre Wenaus (BNL)

Infrastructure Working Groups:

- Streaming Computing Model
- *Heterogeneous Computing (not yet)*
- *Distributed Computing (not yet activated)*

Development Priorities

Chao Peng (ANL), Sakib Rahman (BNL)

Physics and Detector Simulation

- **Charge**:
  - Development of accurate MC simulations using a suite of physics and background generators and detector simulation based on Geant4 and DD4hep
- **Priorities for 2025**:
  - Continue to support the **detector design** and integration with services.
  - Collaborate with the EIC Project to evaluate the **differences between the engineering and simulation designs**, and lead discussions with the DSCs on how to address these differences.
  - Continue to support the development of **background modeling** and implement its timing structure in physics and detector simulations, together with the Background TF.
  - Enable **simulation of streaming readout** by providing the option to switch between streaming data and event data modes.
  - Coordinate the **development of digitization and noise models** with the DSCs and the Electronics and DAQ WG.

Development Priorities

Derek Anderson (ISU), Shujie Li (LBNL)

Reconstruction Framework and Algorithms

- **Charge**:
  - Development of a holistic and modular reconstruction for the integrated ePIC detector.
- **Priorities for 2025**:
  - Drive the **development of the reconstruction framework to meet ePIC needs**, e.g., on modularity or streaming data processing.
  - Host collaboration-wide discussions on all aspects of reconstruction, driving the **work toward holistic reconstruction**.
  - Enable reconstruction algorithms to **handle physics events with background**.
  - Collaborate with **PWGs** on **shared reconstruction priorities**, which currently include:
    - Secondary vertexing
    - Hadron identification
    - Particle flow algorithms for jet reconstruction
    - Event kinematics
  - Integrate continued development of **web-based event display** in reconstruction efforts.

Infrastructure Priorities

Marco Battaglieri (INFN Genoa), Taku Gunji (Tokyo), Jeff Landgraf (BNL)

Streaming Computing Model

- **Charge**:
  - Development of the computing model for the compute-detector integration using streaming readout, AI/ML, and heterogeneous computing, in collaboration with the Electronics and DAQ WG.
- **Priorities for 2025**:
  - Define **requirements for streaming orchestration** and **set up corresponding testbeds**:
    - Develop a testbed for event reconstruction from streamed data in EICrecon, separating signal from background events and demonstrating how we will reconstruct physics events.
    - Establish an initial testbed for super time frame building and processing, and deliver a corresponding requirements document.
  - Document alignment and calibration workflows jointly with the DSCs and identify **requirements for autonomous alignment and calibration**.
  - **Publish the ePIC Streaming Computing Model report**, and the related section in the (pre)TDR.

Operation Priorities

Thomas Britton (JLab), Sakib Rahman (BNL)

Production

- **Charge**:
  - Responsible for the coordination and production of simulation campaigns based on priorities from the Technical and Analysis Coordinators.
  - Develop automated production workflows that scale with the needs of the collaboration.
- **Priorities for 2025**:
  - **Automation Priorities**:
    - Improve the exposure and organization of monitoring so that no one needs to be an OSG expert to track progress, thereby enabling more individuals to participate in operating the monitoring.
    - Explore workflow and workload management tools.
  - **Simulation Campaign Priorities**:
    - Roll out Rucio to the collaboration as the default method for finding and accessing simulation productions.
    - Establish liaisons with DSCs and PWGs to actively participate in the simulation campaigns.

Operation Priorities

Stephen Kay (York), Holly Szumila-Vance (FIU)

User Learning

- **Charge**:
  - Responsible for onboarding via a landing page for new collaboration members and additional appropriate mechanisms.
  - Responsible for support via documentation, help desk, and training.
  - Ensure that software is discoverable (easy to use with only minimal instructions) and simulated data and metadata is findable.
- **Priorities for 2025**:
  - New initiative: **Roadmap towards discoverable software**.
  - **Revised and frequently updated FAQs**.
  - **Rolling schedule of software tutorials** that incorporates updated versions of existing tutorials, new material, and relevant resources from the HSF Training WG.