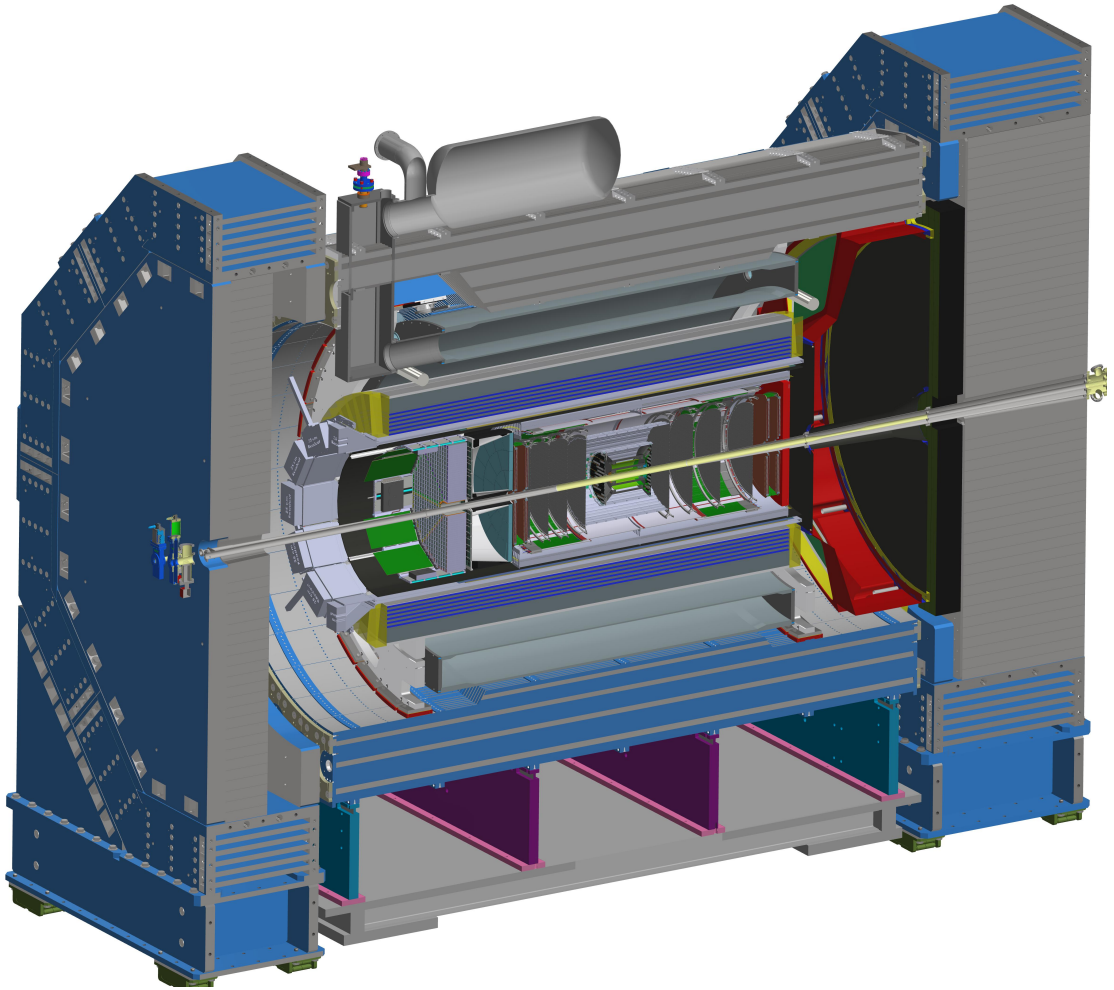
ePIC Software & Computing Report

Markus Diefenthaler (Jefferson Lab)

“Software is the Soul of the Detector”



Great Software for Great Science:

- **Design and Construction:** Integrated and validated simulations are essential for evaluating detector performance and determining physics reach.
- **Operation:** Rapid processing of streamed data using streaming readout, AI, and distributed computing. Autonomous experimentation and control.
- **Physics Analysis:** Software and data enable discovery.

- We **work together**, on a global scale and with other fields, on great software for great science.
- We focus on **modern scientific software & computing practices** to ensure the **long-term success** of the **EIC scientific program**.

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*

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)*

Current Priorities

Computing Model and Testbeds

Onboarding and Community Building

Software and Simulations for the preTDR

Compute-Detector Integration to Maximize and Accelerate Science

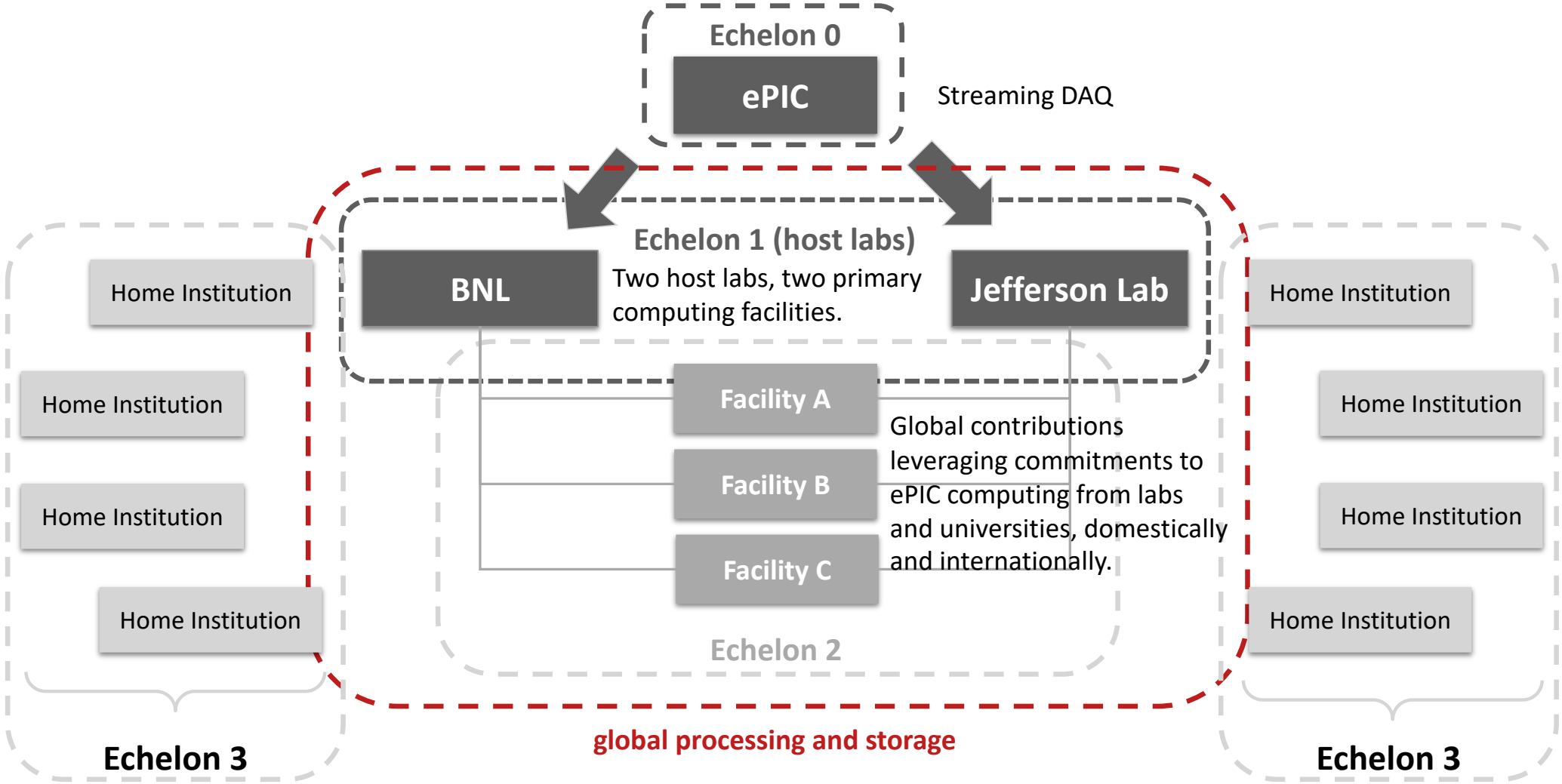
- **Maximize Science** **Capture every collision signal, including background.**
  - Event selection using all available detector data for **holistic reconstruction**:
    - **Eliminate trigger bias** and provide accurate estimation of uncertainties during event selection.
  - Streaming background estimates 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%

- **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 resources developed jointly with the community.
  - Forming EIC International Computing Organization (EICO):
    - Details in Computing Session.
  - The power of distributed computing lies in its flexibility to shift processing between facilities as needed.

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.

Streaming DAQ and Computing Milestones

Streaming DAQ Release Schedule:

PicoDAQ **FY26Q1**
- Readout test setups

MicroDAQ: **FY26Q4**
- Readout detector data in test stand using engineering articles

MiniDAQ: **FY28Q1**
- Readout detector data using full hardware and timing chain

Full DAQ-v1: **FY29Q2**
- Full functionality DAQ ready for full system integration & testing

Production DAQ: **FY31Q3**
- Ready for cosmics

Streaming Computing Milestones:

Start development of streaming orchestration, including workflow and workload management system tool.
Start streaming and processing streamed data between BNL, Jefferson, DRAC Canada, and other sites.
Support of test-beam measurements, using variety of electronics and DAQ setups:

Establish autonomous alignment and calibration workflows.

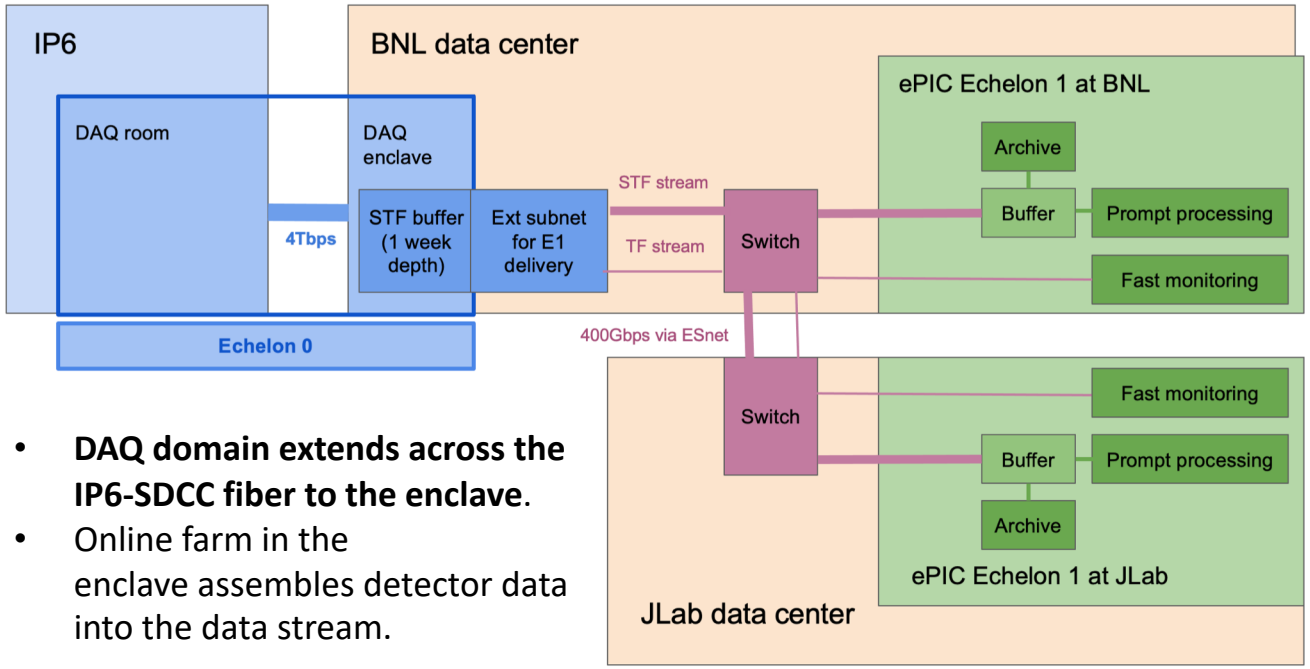
Analysis challenges exercising end-to-end workflows from (simulated) raw data.

Streaming challenges exercising the streaming workflows from DAQ through reconstruction, and the Echelon 0 and Echelon 1 computing and connectivity.
Analysis challenges exercising autonomous alignment and calibrations.

Data challenges exercising scaling and capability tests, including exercising the functional roles of the Echelon 2 sites.

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.

Streaming Computing: Getting to the Specifics

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

ePIC Computing & Software Reviews

EIC Computing and Software Advisory Committee (ECSAC)

advise host laboratories on the progress and status of computing and software for the ePIC collaboration.

Mohammad Al-Turany (GSI)

Simone Campana (chair, CERN)

Pere Mato (CERN)

Christoph Pauss (MIT)

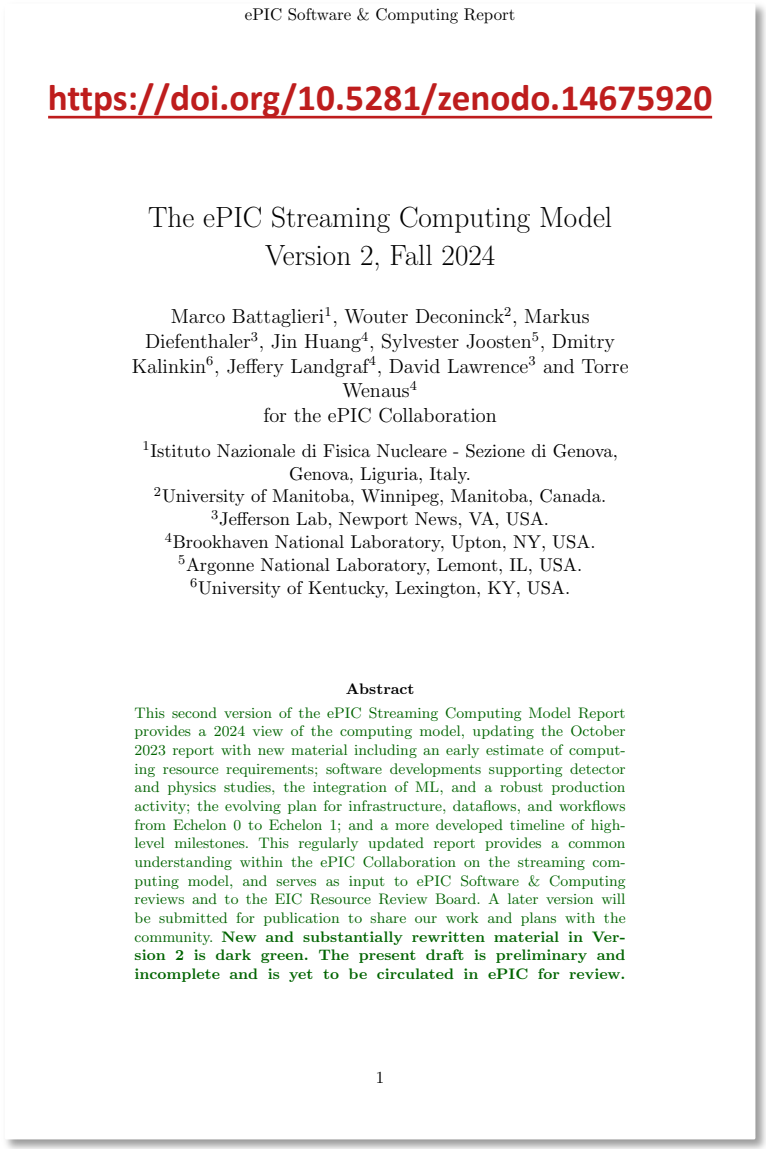
Verena Martinez Outschoorn (UMass Amherst)

Frank Würthwein (UCSD)

Annual reviews with a **charge reflective of** the **EIC schedule**, the **stage of the ePIC experiment**, and **impending deadlines**.

- **2023**: Assessed the adequacy and flexibility of our planning, including the integration of international contributions.
- **2024**: Resource estimates for streaming computing model.
- **2025**: In preparation.

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 at the ECSAC Reviews 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.

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?

Community Building

Regular meetings to drive forward priority targets and provide an avenue for new collaboration members to engage.

Summary: Status and plans for software and simulations; preparations for first ePIC Software & Computing review.

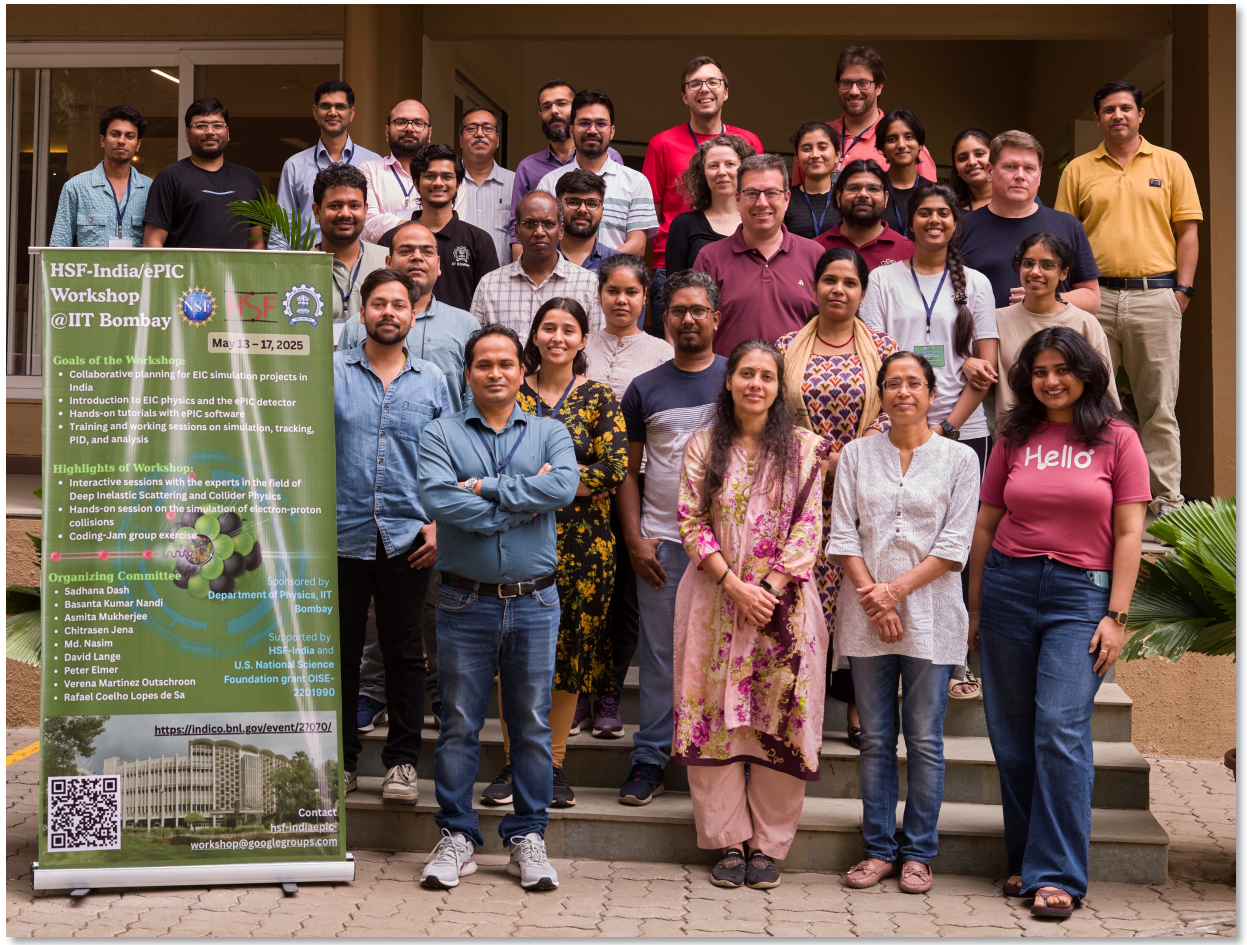
Summary: Status and plans for software and simulations, including TDR priorities, tutorials; streaming computing; software projects with HEP.

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

Current Priorities

Computing Model and Testbeds

Onboarding and Community Building

Software and Simulations for the preTDR

Priorities for 2025

- **Improved Communication**
  - **(Almost) Weekly 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 23–27)
  - 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**.
- **Coordination with DSCs**
  - Requested **updates on software priorities** to facilitate the integration of DSCS simulation efforts into ePIC software and simulation campaigns.
  - Coordination **improved shared development**, e.g., by aligning four DSCs under a common digitization work plan.

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.

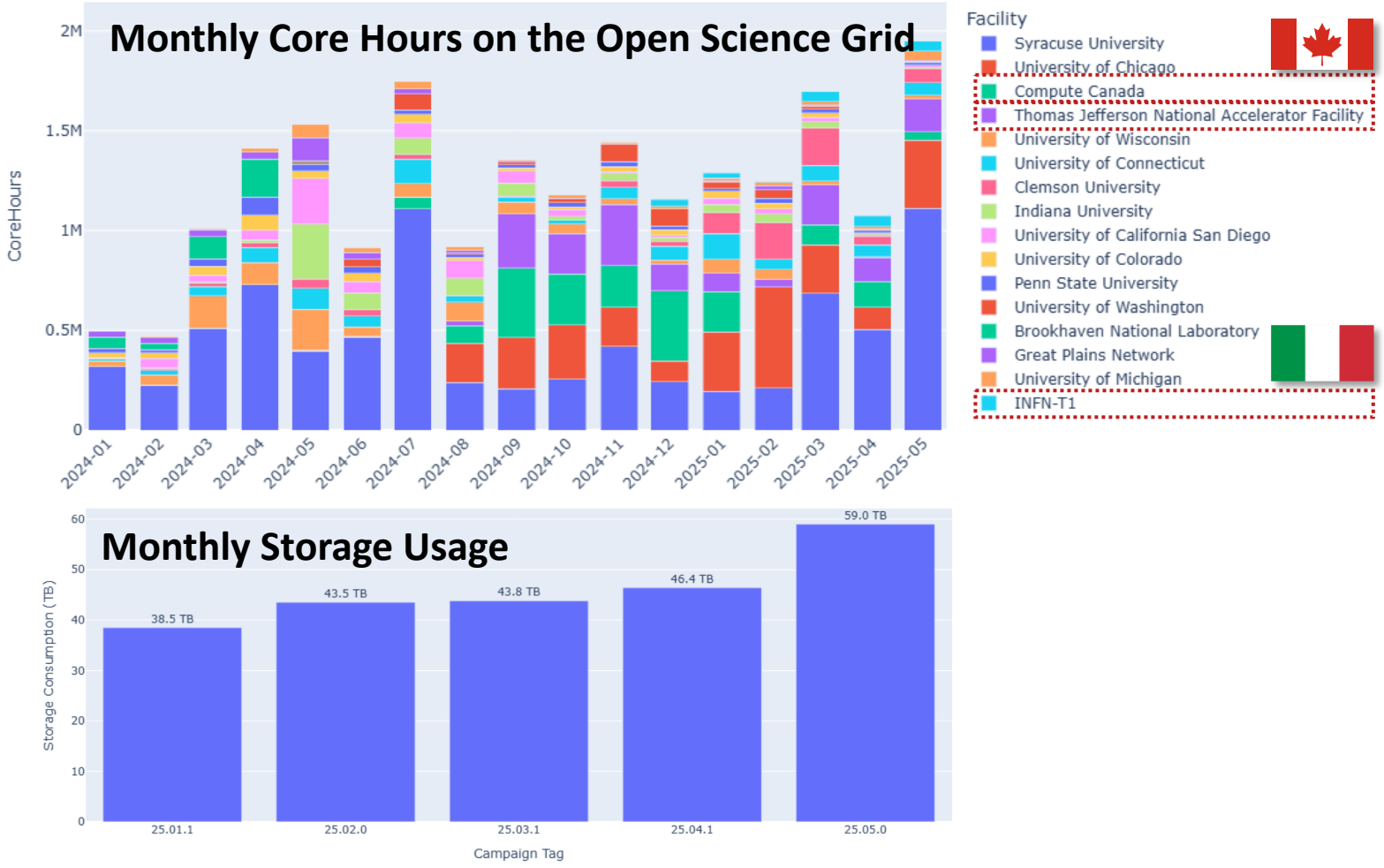
Reconstruction Framework and Algorithms

- **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 23 to 27.

Simulation Campaigns

- We provide simulation productions 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.

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 International Computing Organization (**EICO**) formed; discussion in Computing Session.

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**.

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.
- Updates from DSCs on software priorities have led to **improved shared development**.
- Progress communicated via **meeting notes** and (almost) **weekly Software News**.

Backup

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.