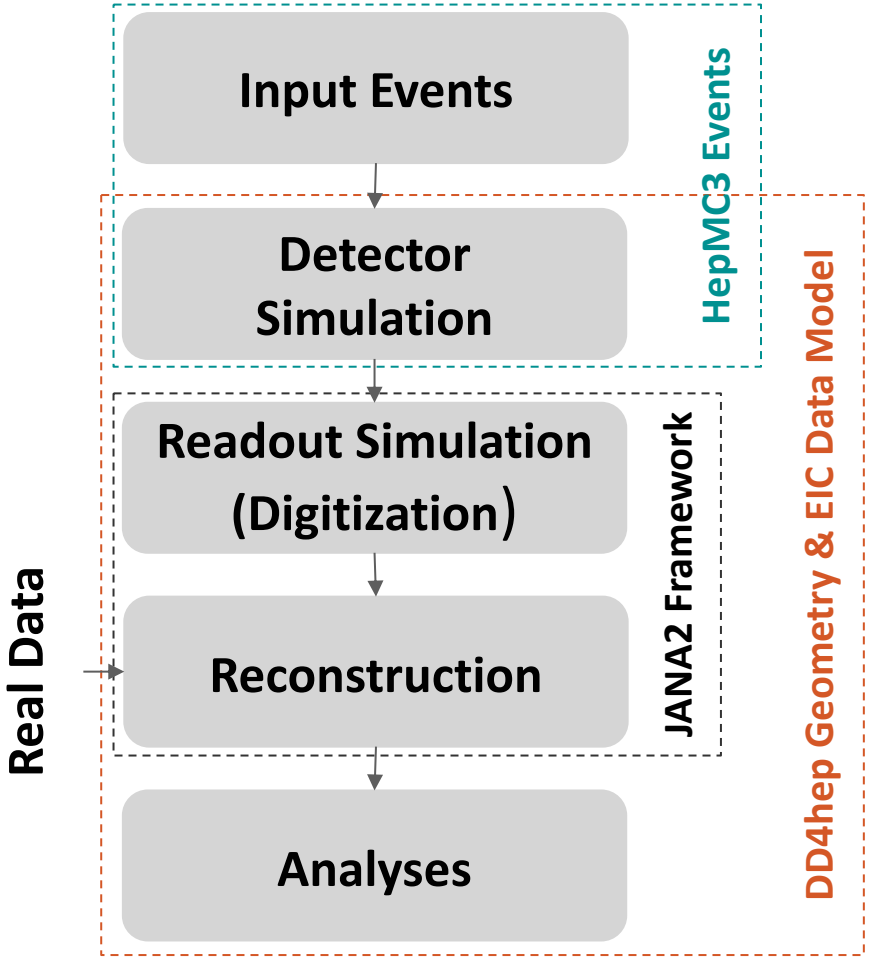
ePIC Software & Computing Report

Markus Diefenthaler (Jefferson Lab) for the ePIC Collaboration

Our Philosophy

- We focus on **modern scientific software & computing practices** to ensure the **long-term success of the EIC scientific program** throughout all CD milestones.
  - Strong emphasis on modular, orthogonal tools.
  - Integration with HTC/HPC, CI workflows, and enable use of standard data science toolkits.
- We **leverage cutting edge sustainable community software** where appropriate, **avoiding the "not invented here" syndrome**.
  - Can build our software on top of a mature, well-supported, and actively developed software stack by using modern community tools, e.g. from CERN, the HPC community, and the data science community.
  - Actively collaborate with external software projects, while externalizing some support burden to external projects.
- We embrace these practices today to avoid starting our journey to EIC with technical debt.
- **We are writing software for the future, not the lowest common denominator of the past!**

ePIC Software Stack: A Modular Simulation, Reconstruction, and Analysis Toolkit



Input events from **MC event generators** or particle guns, with optional physics background merging.

Geant4 simulations with **DD4hep** for geometry description and exchange, output data in the **EIC Data Model** (EDM4hep + EDM4eic, described in Podio).

Algorithms to transform the GEANT4 hits to mimic real detector readout, including background stacking, “pileup”, DAQ frames

Realistic reconstruction algorithms starting from raw detector output (from digitization or real data).

User analyses in plain C++/ROOT or Python/uproot, facilitated by using a flat data model.

Continuous integration for detector and physics benchmarks and monthly production campaigns ensure a production-ready software stack at any time.

ePIC Software & Computing Organization

Spokesperson's Office

Detector | **Software & Computing** | Physics

Guiding Principles:

- *Diversity, Equity, and Inclusion*
- *Statement of Software Principles*
- *Sustainability.*

Software & 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

Deputy Coordinator (Development)
Sylvester Joosten (ANL)

Development Working Groups:

- Physics and Detector Simulation
- Reconstruction
- *Analysis Tools*

Deputy Coordinator (Infrastructure)
Torre Wenous (BNL)

Infrastructure Working Groups:

- Streaming Computing Model
- *Multi-Architecture Computing*
- *Distributed Computing*

Since May: Monthly Simulation Productions

Operations | Development | Infrastructure

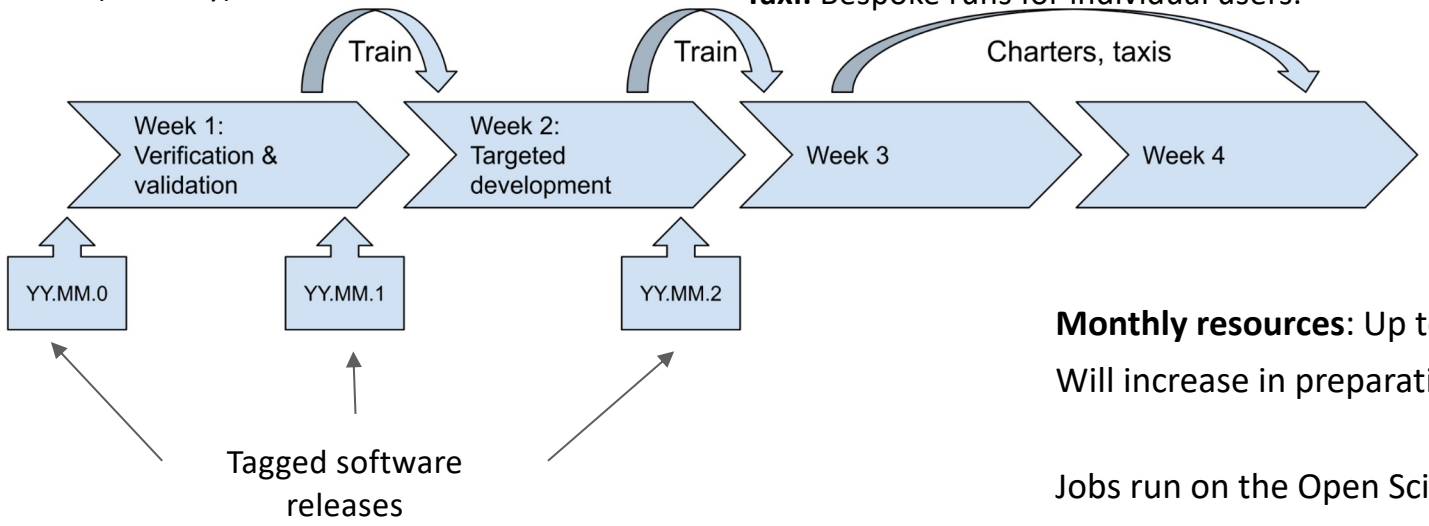
Objectives

1. Achieve **continuous deployment** of the software used for detector and physics simulations.
2. Ensure **regular updates** of simulation productions for detector and physics studies in preparation for the TDR (and all other CD milestones).
3. Implement **timely validation and quality control** for simulation productions on datasets that require substantial time and resources.

Train: Major central campaign at a fixed (monthly) schedule.

Charter: Special interest runs for working groups.

Taxi: Bespoke runs for individual users.



Monthly resources: Up to 10k continuously used cores (about 500 core-years). Will increase in preparation of TDR.

Jobs run on the Open Science Grid (OSG), with output stored at host labs.

Production WG Conveners

Thomas Britton (Jefferson Lab)

Sakib Rahman (U. Manitoba)

ePIC Simulation Campaigns

Details on https://eic.github.io/epic-prod/

Monthly Production Updates:

- Provided at the end of each campaign on the main mailing list of the collaboration.
- Version Format: Year.Month

Live Updates:

- Follow the **firehose** Mattermost channel for real-time information.

Previous Campaign Information:

- Access **reconstructed output files** and **full Geant4 simulation output files** on our campaign history pages.

The list of datasets are displayed in a tree structure with:

- **XrootD server** and **base address** at the top,
- followed by **RECO** for reconstructed output files or **FULL** for full Geant4 simulation files and the **version**.
- Then they are organized by **detector config**, **physics process**, and **beam properties**.

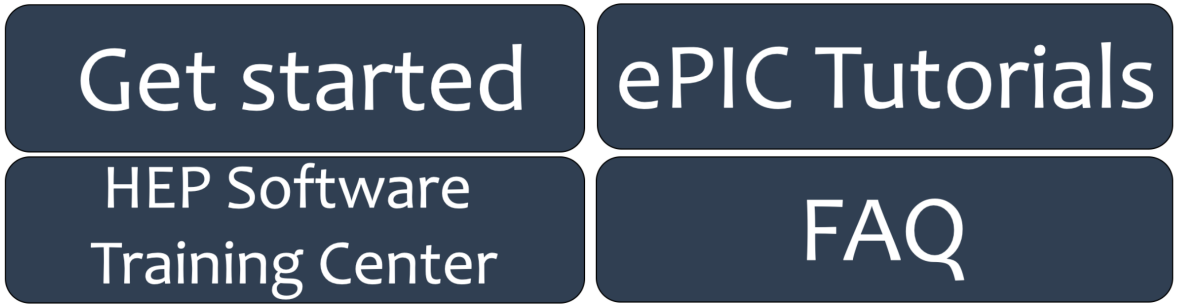
To list the reconstructed output files for a particular dataset, first start eic-shell and execute:

```
xrdfs root://dtn-eic.jlab.org
      XRootD server
ls /work/eic2/EPIC/RECO/23.12.0/epic_craterlake/DIS/NC/10x100/minQ2=10
   Base Address        TYPE and Version  Detector Config      Physics   Beam Properties
```

User Learning

Accomplishments:

- **Landing Page** development and support for new collaboration members:



- Making previous tutorials and documentation more available.
- Created five tutorials for the collaboration meeting with standardized documentation, accessible from the Landing Page.
- Creation of an FAQ with ongoing support. In addition to active **Helpdesk**.

Future plans:

- Tutorials at the ePIC Software & Computing workshop in April, tailored to collaboration needs.
- Hold bi-monthly hybrid tutorials with documentation (recorded and available in remote/asynchronous setting).
- Develop training and best practices for experts giving tutorials.
- Further improve FAQ.

User Learning WG Conveners


Kolja Kauder (BNL)


Holly Szumila-Vance (Jefferson Lab)

Tutorials at Collaboration Meeting

In-person tutorials at the January collaboration meeting, covering five key topics:

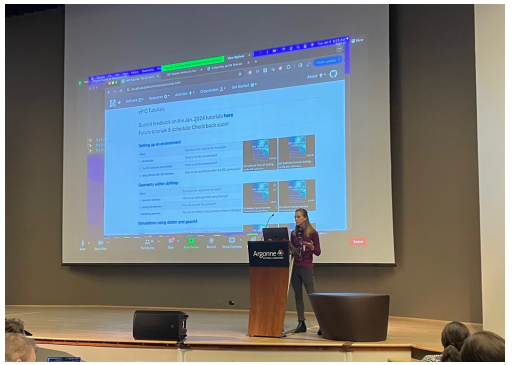
1. **Collaborative Development** (Holly Szumila-Vance)

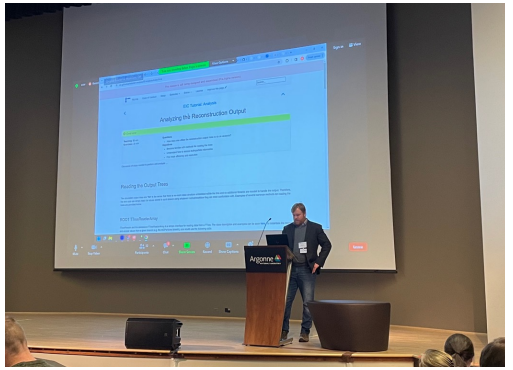
   Eic-shell Easy to get started locally… in only 1 line!

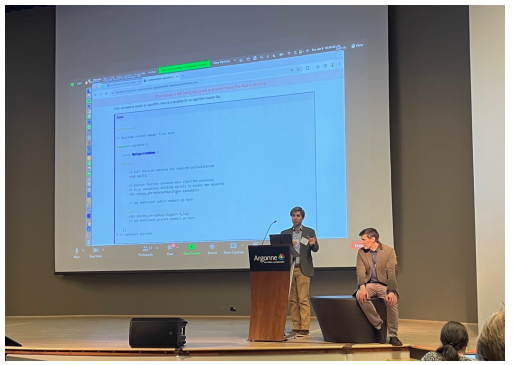
   ```
   curl -L get.epic-eic.org | bash
   ```

   Based on container images, the same images are used for simulation campaigns.

2. **Working with Simulations in Python or ROOT** (Brian Page)
3. **Detector Geometry and Digitization** (Kolja Kauder)
4. **Developing Reconstruction Algorithms** (Nathan Brei and Tyler Kutz)
5. **Developing Benchmarks** (Dmitry Kalinkin and Torri Jeske)

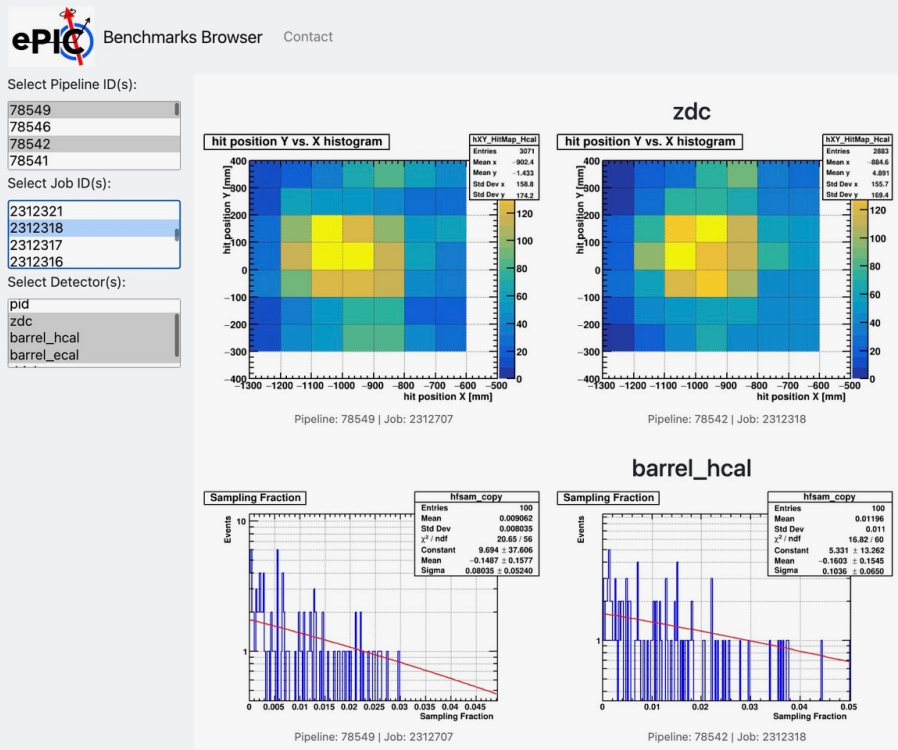






Validation

Benchmark Browser available, will be further refined with user input.



Propose **Snakemake for analysis workflow definition**, including detector and physics studies for TDR:

- **Allows to both**:
  - Run workflows locally.
  - Submit batch jobs.
- **Caches intermediate steps** – ideal for rapid development.

Snakemake tutorial available with select examples.

Benchmarks using snakemake will be integrated into validation process of simulation campaigns.

Validation WG Conveners

Torri Jeske (Jefferson Lab)

Dmitry Kalinkin (U. Kentucky)

Community Building

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

12 pages of detailed notes that enabled software progress, pushed the review preparations, and informed our planning.

Topics: Software and simulations for TDR, tutorials; work with HSF and key4hep; MCEG.

Software Readiness for TDR

Physics and Detector Simulation WG Conveners

Kolja Kauder (BNL)

Chao Peng (ANL)

Deadlines that inform our planning:

- **November 2024**: TDR finalization
- **September 2024**: Complete draft of TDR
- **June 2024**: Software release for TDR

Progress and Discussions:

- On January 9, during a parallel session, we discussed the requirements and priorities for software development.
- Results were compiled from four discussion groups who responded to a questionnaire.

Upcoming Plenary Session:

- On January 13, in the plenary session, we will summarize these results and engage in a discussion with the larger collaboration.

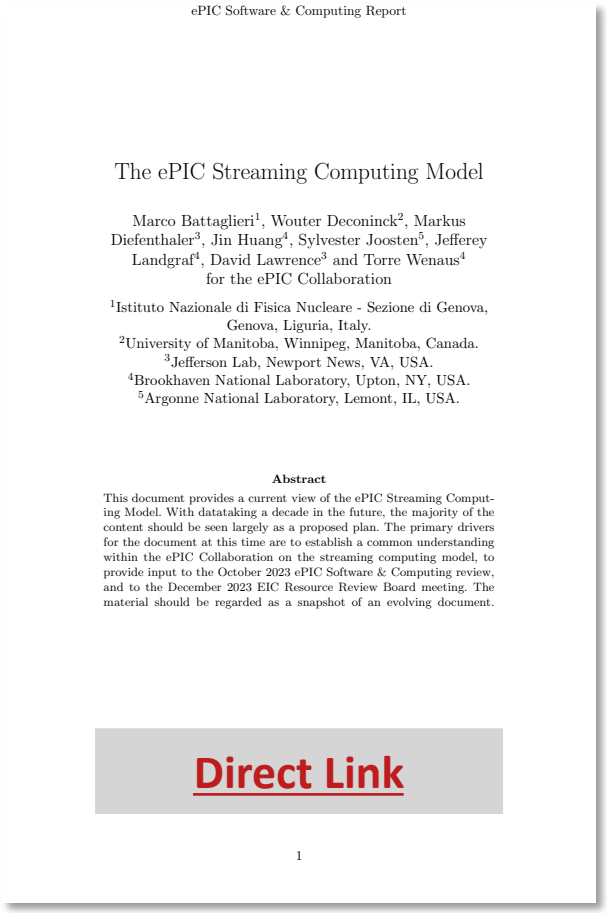
Reconstruction Framework & Algorithms WG Conveners

Derek Anderson (Iowa State)

Shujie Li (LBL)

Streaming Computing Model WG

Operations | Development | **Infrastructure**



ePIC Software & Computing Report

The ePIC Streaming Computing Model

Marco Battaglieri[1], Wouter Deconinck[2], Markus Diefenthaler[3], Jin Huang[4], Sylvester Joosten[5], Jefferey 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.

Abstract

This document provides a current view of the ePIC Streaming Computing Model. With datataking a decade in the future, the majority of the content should be seen largely as a proposed plan. The primary drivers for the document at this time are to establish a common understanding within the ePIC Collaboration on the streaming computing model, to provide input to the October 2023 ePIC Software & Computing review, and to the December 2023 EIC Resource Review Board meeting. The material should be regarded as a snapshot of an evolving document.

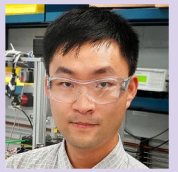
Direct Link

1

- Works in synergy with Electronics and DAQ WG.
- Defined requirements and high-level design for a **computing model** that enables **rapid data processing for physics analyses** in 12 WG meetings since July.
- Started documenting a streaming computing model that can be redefined further with international partners.
- Initial version of the ePIC Streaming Computing Model has been presented in recent ePIC Software & Computing Review.

Streaming Computing Model WG Conveners

Marco Battaglieri (INFN Genova)



Jin Huang (BNL)

Jeff Landgraf (BNL)

ePIC Software & Computing Review, Oct 19–20, 2023, Washington, DC

- Convened by and reporting to the host labs: Haiyan Gao (BNL) and David Dean (Jefferson Lab).
- An excellent **review panel**, which will continue as the **EIC Computing and Software Advisory Panel**:
  - Frank Wuerthwein (chair, UCSD), Mohammad Al-Turany (GSI), David Brown (LBNL), Simone Campana (CERN), Pere Mato (CERN), Christoph Paus (MIT), Heidi Schellman (OSU).
- We addressed five charge questions posed in the context of being 10 years away from data:
  - Briefly summarized below with the panel's conclusions; see Indico and the closeout for the full story.
- **Is there a comprehensive and cost effective long term S&C plan?**
  - *Yes. Impressive organization and plan. ePIC should verify software and simulation readiness for the TDR by May 2024. A long term computing needs assessment was not presented; present one in a year.*
- **Are there adequate plans for integrating international partners?**
  - *Yes. The opportunities are significant (Canada, Italy, UK thus far), the collaboration is doing all the right things to leverage them.*
- **Are S&C plans integrated with HEP/NP community developments?**
  - *Yes. ePIC is leveraging tools and services widely adopted and supported in the HEP/NP community. ePIC should contribute to supporting key software. Document dependencies and ePIC contributions.*
- **Are there sufficient S&C resources to deliver the TDR?**
  - *Yes. Computing resources available from OSG (80%) and labs (20%) look sufficient. Software development plan is credible to meet the critical TDR milestones.*
- **Do BNL/JLab joint institute plans integrate sufficiently with experiment S&C?**
  - *Yes. Organization resembles those of the LHC which have worked well. The sharing of responsibilities between experiment and institute, particularly for software, still being discussed. Ongoing evaluation will be part of future reviews.*

Compute-Detector Integration to Accelerate Science

- **Problem** Data for physics analyses and the resulting publications available after O(1year) due to complexity of NP experiments (and their organization).
  - Alignment and calibration of detector as well as reconstruction and validation of events time-consuming.
- **Goal** **Rapid turnaround of 2-3 weeks for data for physics analyses**.
  - Timeline driven by calibrations.
- **Solution** Compute-detector integration using:

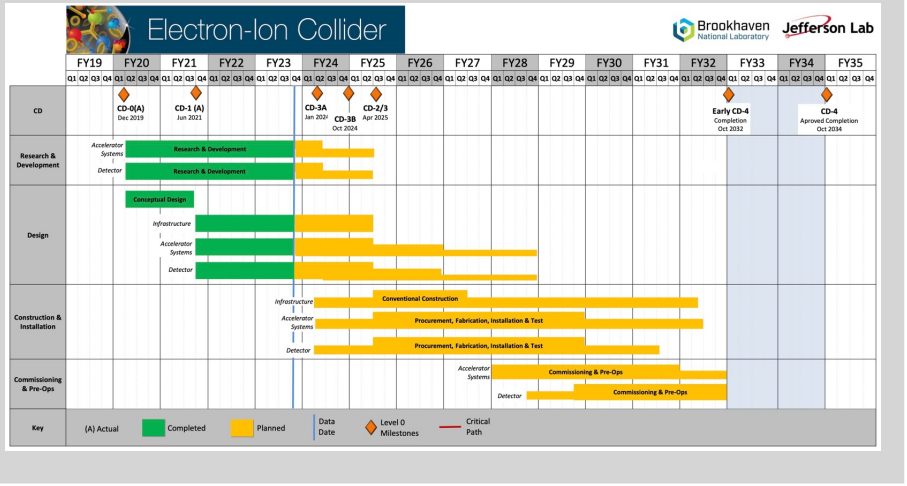
AI for autonomous alignment and calibration as well as reconstruction and validation for rapid processing.

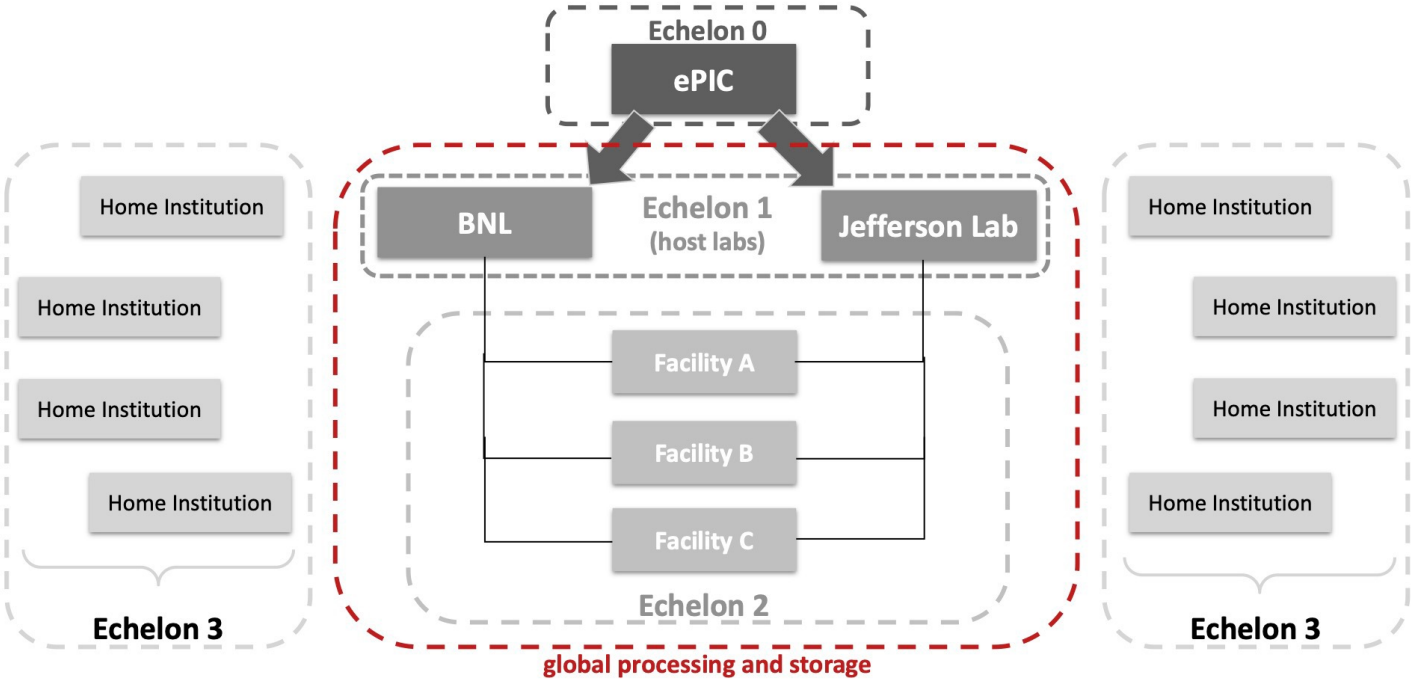
Streaming readout for continuous data flow of the full detector information.

Heterogeneous computing for acceleration.

Current View of the Computing Model

In a decade: Start of the ePIC operations





Initial version of a plan set to develop over the next decade.

Echelon 2: Global ePIC Computing

ePIC is an international collaboration and so is its computing:

- **Echelon 2** includes **global resources contributed by collaborating institutions**.
- Achieving scientific goals relies on effectively using Echelon 2's resources.
- Design of computing model aims for **effective integration and management**.
- **EIC Computing & Software Joint Institute (ECSJI)** by host labs oversees complex computing fabric of the EIC.

International computing contributions are essential.

International Contributions:

- **From the review close out**:

  "There are clearly very significant opportunities in in-kind computing infrastructure contributions."

- Canada, Italy, and the United Kingdom are engaged as a proof of concept in this context.
  - Integration of resources from international partners into Open Science Grid foreseen on timescale of TDR.
- Computing centers of these countries were already included in large-scale simulation efforts for the EIC.

Current Estimate for Compute Resources

Streaming DAQ sends data in **1ms time frames**.

Each time frame corresponds to 10MB of data.

Based on our current detector readout design and when running at peak luminosity and in standard operating conditions.
40% of data bunch crossing related, 60% background.

In a year, we will record 15.5 billion frames.

Assuming a 50% up-time for 6 months.

Number of expected events (assuming a 50% up-time for 6 months):

- The event rate at peak luminosity is 500kHz, which gives roughly 4 x 10^{12} events.
  - Lower at start of operations, where the luminosity will be lower (but relatively speaking background rate is expected to be higher).
- The expected number of physics events of interest for one year of running at peak luminosity is ~ 10^{10}.
  - The actual physics events is only a very small fraction of the total physics bunch crossings.

Number of simulation events:

- We expect to simulate 10x events for each event of interest, yielding O(10^{11}) simulated events.

  While considerable (~ 60k core years on today's hardware), this should be a realistic target in a decade.

Core-seconds for simulation and reconstruction (on a typical modern machine):

- Our current simulations including background take ~17s for simulation and ~ 2s for reconstruction, per event.
  - Simulation and reconstruction on event level only.
- **Unknown**: How much this will change once changing to streaming data processing?
- **Priority target for TDR**: Prototype of event reconstruction from realistic frames.

Realizing the ePIC Streaming Computing Model

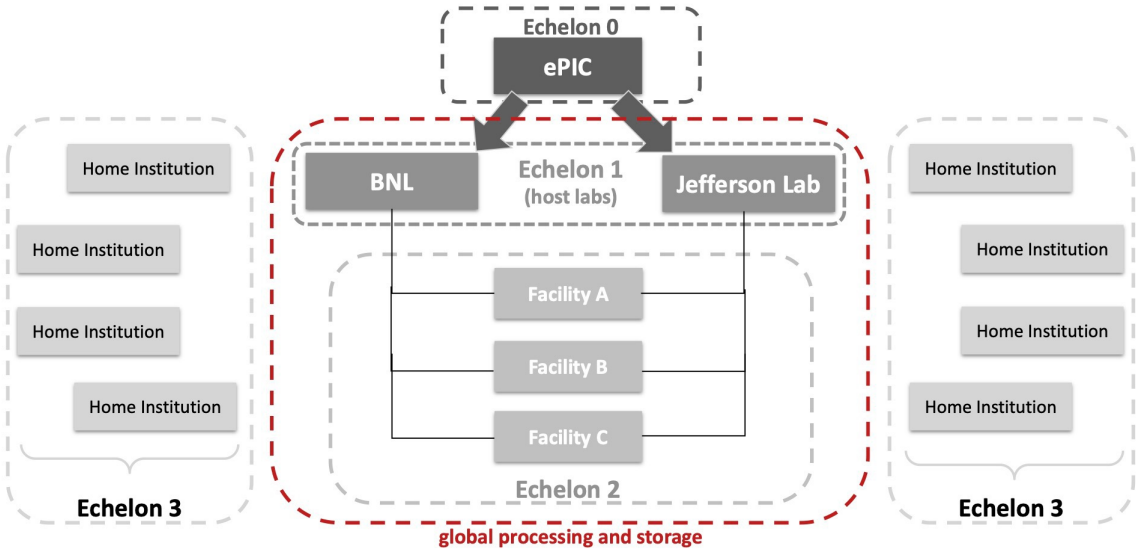
Four Tiers:

Echelon 0: ePIC Experiment

Echelon 1: Crucial and innovative partnership between host labs.

Echelon 2: Global processing and data facilities, includes High-Performance (HPC) and High-Throughput Computing (HTC) resources.

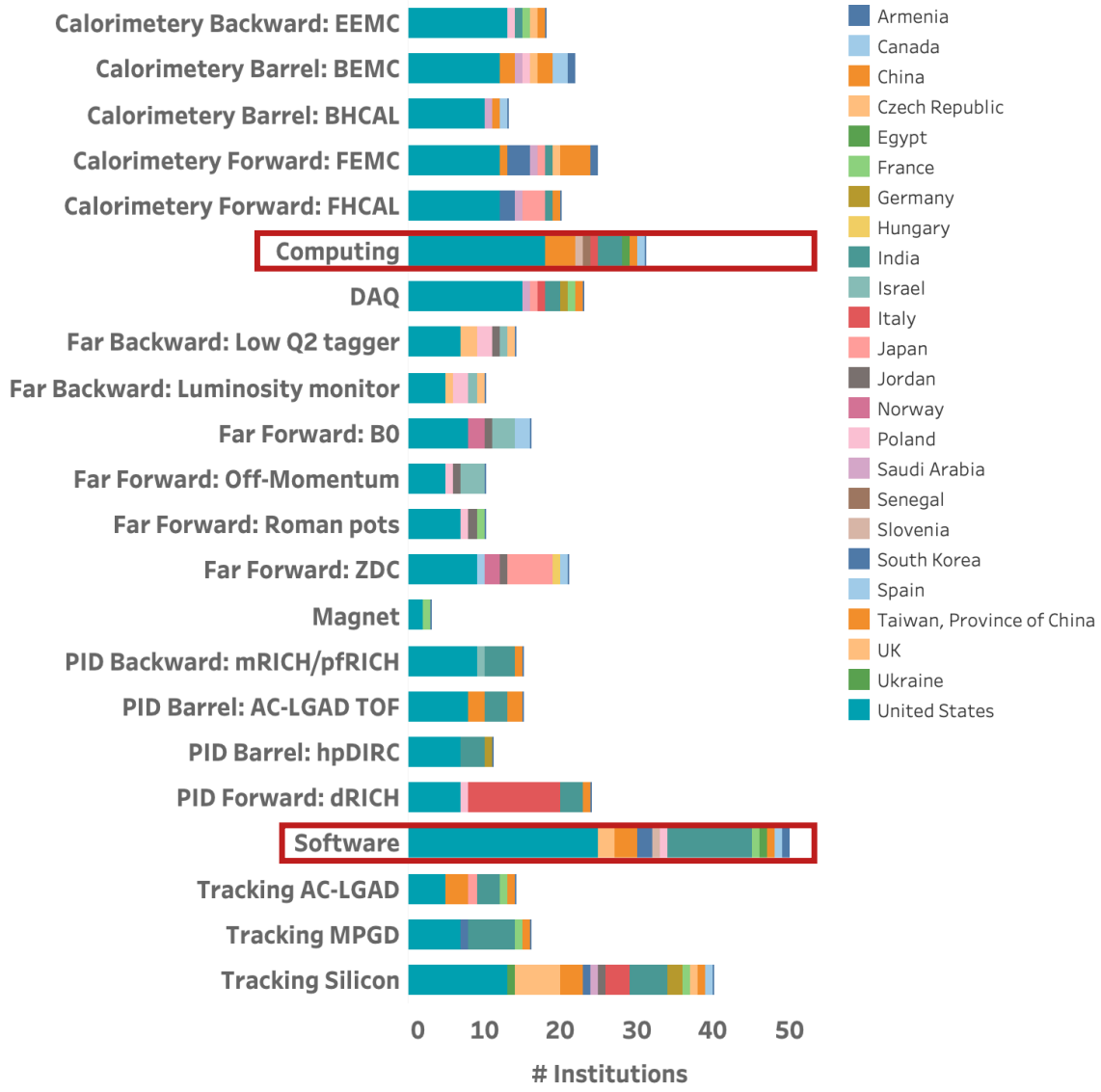
Echelon 3: Full support of the analysis community at the home institutions.



Next Steps:

- Provide for each use case detailed **estimates on the compute resources**; obtain an idea of how resources are provided by host labs (Echelon 1) and international partners (Echelon 2).
- Alternately, detail the needs **to fully support** future EIC science data taking and analysis in this model. Requires coordination with the BNL+JLab EIC Computing and Software Joint Institute, as well as R&D in collaboration with theory as well as computer and data scientists.
- **Advance distributed computing capabilities** and start **integrating resources from international partners.**
    - Will start integrating resources from international partners on the Open Science Grid (OSG).
- Design the system for both batch and **dynamic processing** to ensure resilience against technology evolution.
- **Streaming challenges** exercising the streaming workflows from DAQ through offline reconstruction.
- Analysis challenges exercising **autonomous alignment** and **calibrations**.

Collaboration Interest in Software & Computing



The interest does not yet reflect actual, active participation in ePIC Software & Computing.

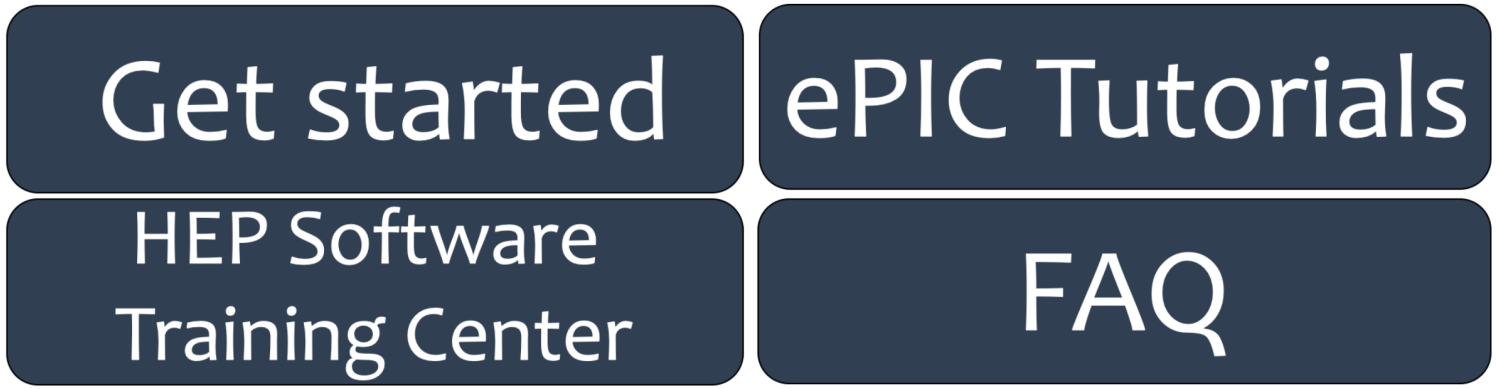
Key priorities for the TDR include conducting **simulation campaigns**, **validating our software and simulation outputs**, and developing **reconstruction algorithms**.

Efforts on these key priorities will benefit substantially from an increase of the workforce.

Busy TDR preparation period is ideal time to get actively involved in ePIC Software & Computing!

How To Get Involved

Landing Page https://eic.github.io/documentation/landingpage.html



Weekly Updates on ePIC Software & Computing:

- Join our weekly meetings every Wednesday at 11:00 a.m. EST.
- **Indico category**: **https://indico.bnl.gov/category/435/**

Active Involvement in ePIC Software & Computing:

- Reach out to Coordinators and WG Conveners.
- Tomorrow's plenary discussion on software and simulation readiness for the TDR: Gain insights into tasks for involvement.

Summary

- The **ePIC software stack** is a **modern** and **modular toolkit** for simulation, reconstruction, and analysis.
- **Streaming Computing Model** integral part of ePIC:
  - Rapid turnaround of 2-3 weeks for data for physics analyses.
- **Successful ePIC Software & Computing Review** on October 20–22:

Is there a comprehensive and cost effective long term S&C plan?	Yes.
Are there adequate plans for integrating international partners?	Yes.
Are S&C plans integrated with HEP/NP community developments?	Yes.
Are there sufficient S&C resources to deliver the TDR?	Yes.
Do BNL/JLab joint institute plans integrate sufficiently with experiment S&C?	Yes.

- **High level milestones** ensures that the agile development process is continuously confronted with real world exercising of the software and the developing realization of the computing model:
  - Priority always given to meeting near-term needs:
    - ePIC leverages monthly production campaigns, CI-driven benchmarks, and timeline-based prioritization to ensure timely completion of the simulation studies for the TDR.
    - **January 13: Plenary discussion on software and simulation readiness for TDR!**
  - Longer range timeline progressively exercising the streaming computing model to deliver for the needs of the CD process, for specific applications, e.g. test beams, for scaling and capability challenges, and ultimately for the phases of datataking.