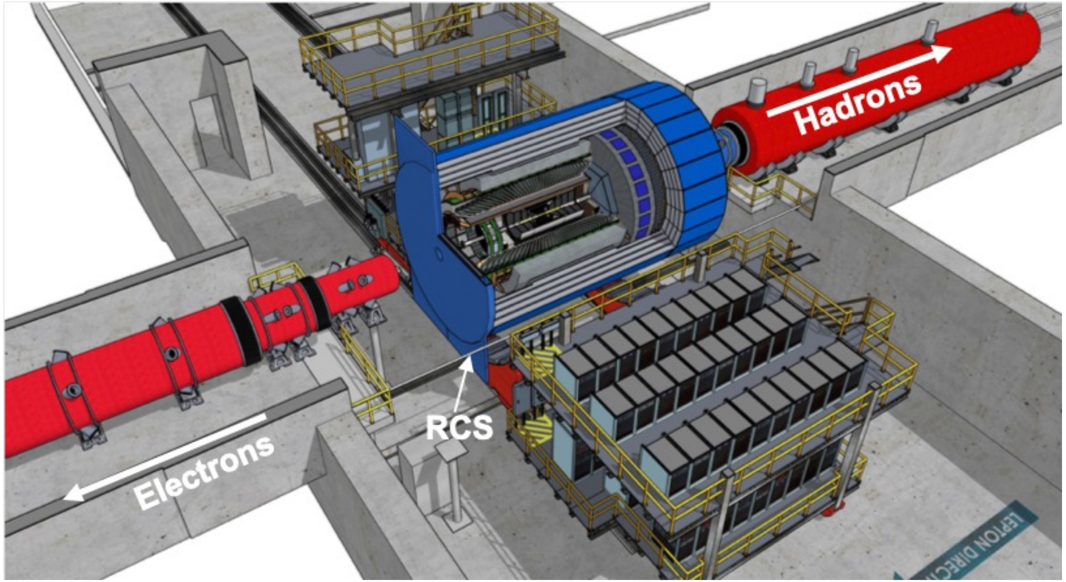
ePIC Software & Computing



"***Software is the soul of the detector.***" (Ian Shipsey, Oxford)

EIC Software: Statement of Principles

EIC SOFTWARE: Statement of Principles

1. **We aim to develop a diverse workforce, while also cultivating an environment of equity and inclusivity as well as a culture of belonging.**

2. **We will have an unprecedented compute-detector integration:**
   - We will have a common software stack for online and offline software, including the processing of streamed data and its time-ordered structure.
   - We aim for autonomous alignment and calibration.
   - We aim for a rapid, near-real-time turnaround of the raw data to online and offline productions.

3. **We will leverage heterogeneous computing:**
   - We will enable distributed workflows on the computing resources of the worldwide EIC community, leveraging not only HTC but also HPC systems.
   - EIC software should be able to run on as many systems as possible, while supporting specific system characteristics, e.g., accelerators such as GPUs, where beneficial.
   - We will have a modular software design with structures robust against changes in the computing environment so that changes in underlying code can be handled without an entire overhaul of the structure.

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.

5. **Our data formats are open, simple and self-descriptive:**
   - We will favor simple flat data structures and formats to encourage collaboration with computer, data, and other scientists outside of NP and HEP.
   - We aim for access to the EIC data to be simple and straightforward.

6. **We will have reproducible software:**
   - Data and analysis preservation will be an integral part of EIC software and the workflows of the community.
   - We aim for fully reproducible analyses that are based on reusable software and are amenable to adjustments and new interpretations.

7. **We will embrace our community:**
   - EIC software will be open source with attribution to its contributors.
   - We will use publicly available productivity tools.
   - EIC software will be accessible by the whole community.
   - We will ensure that mission critical software components are not dependent on the expertise of a single developer, but managed and maintained by a core group.
   - We will not reinvent the wheel but rather aim to build on and extend existing efforts in the wider scientific community.
   - We will support the community with active training and support sessions where experienced software developers and users interact with new users.
   - We will support the careers of scientists who dedicate their time and effort towards software development.

8. **We will provide a production-ready software stack throughout the development:**
   - We will not separate software development from software use and support.
   - We are committed to providing a software stack for EIC science that continuously evolves and can be used to achieve all EIC milestones.
   - We will deploy metrics to evaluate and improve the quality of our software.
   - We aim to continuously evaluate, adapt/develop, validate, and integrate new software, workflow, and computing practices.

The "Statement of Principles" represent guiding principles for EIC Software. They have been endorsed by the international EIC community. For a list of endorses, see LINK.

EIC Software is:

1. Diverse
2. Integrative
3. Heterogeneous
4. User-centered
5. Accessible
6. Reproducible
7. Collaborative
8. Agile

ePIC Software Stack

Our software design is based on **lessons learned in the worldwide NP and HEP community** and a **decision-making process** involving the whole community. We will continue to work with the worldwide NP and HEP community.

Modular Simulation, Reconstruction, and Analysis Toolkit using tools from the NP-HEP community

MC Event Generators → Detector Simulations in Geant4 → Readout Simulation (Digitization) → Reconstruction in JANA2 → Physics Analyses

Edm4eic data model based on edm4hep and podio.
Geometry Description and Detector Interface using DD4hep.

Continuous Integration for Detector and Physics Benchmarks and Reproducibility

We are providing a production-ready software stack throughout the development:

- **Milestone**: Software enabled first large-scale simulation campaign for ePIC.

We have a good foundation to meet the near-term and long-term software needs for ePIC.

Software and Computing Coordinator (Markus)

+ Deputy Coordinator **Operations (Wouter)**

+ Deputy Coordinator **Development (Sylvester)**

+ Deputy Coordinator **Infrastructure (Torre)**

Guiding Principles: DE&I, Software Principles, Sustainability

Operation WGs:

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

Development WGs (CI):

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

Infrastructure WGs:

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

Cross-cutting WG:

- Data and Analysis Preservation

How to engage with the Software & Computing effort

- **Indico category**: https://indico.bnl.gov/category/410/

User Learning	15 events	
Weekly Meetings	48 events	
Working Group Meetings	195 events	
Workshops	1 event	

Main meeting to follow for updates
Wednesdays at 11:00 a.m. (EDT)

- **Mattermost for communication**: https://chat.epic-eic.org/

  New instance!
  Click here for invite!

  - **Helpdesk on Mattermost**: https://chat.epic-eic.org/main/channels/helpdesk

- **Mailing list for announcements**: eic-projdet-compsw-l@lists.bnl.gov

Production WG

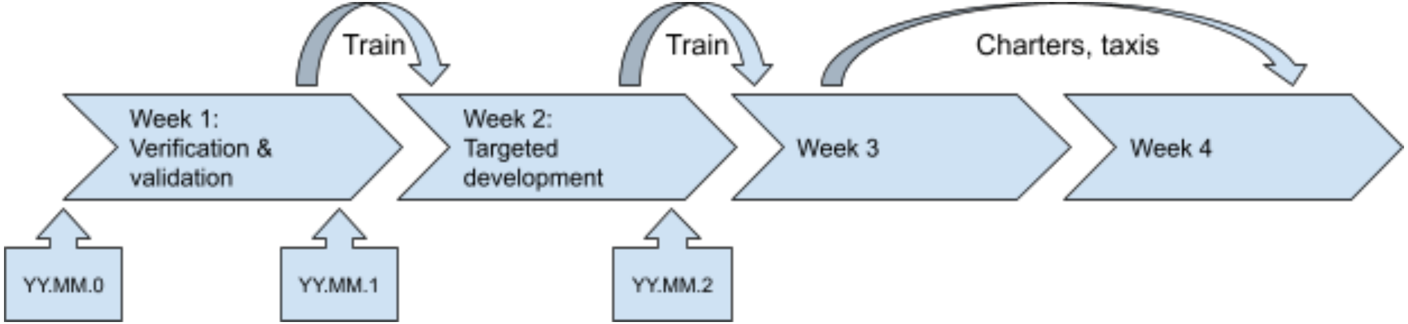
- **Conveners**: Thomas Britton and Sakib Rahman
- **Charge**: Responsible for the coordination and production of simulation campaigns based on priorities from the Technical and Analysis Coordinators. Develop automated production workflows that scale with the needs of the collaboration.
- **Priorities for 2023**:
  - Implement and document our Simulation Production Strategy, together with Validation WG.
  - Survey current production resources and identify potential future resources.
  - Inform us when the Distributed Computing WG needs to start.

Update by Thomas Britton (JLab) in the Software & Computing Session

New Simulation Campaign Strategy Implemented

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, as well as for geometry and algorithm development.
3. Implement **timely validation and quality control** for simulation production on data sets that require significant time and resources.



Weekly meeting in the first working week: Summary of changes, identification of missed targets, and prioritization of sprint goals.

Weekly meeting in second working week: Overview of the verification and validation, and of the first *train* data sets.

Production WG: Open Tasks

1. **Demonstrate writing to XRootD remotes in DDSim:** We currently already read input events from XRootD, but writing to a remote XRootD file is not tested yet. This may work, or it may require adding some Python support.
2. **Demonstrated reading from and writing to XRootD remotes in EICrecon:** We currently run EICrecon only on local files, and reading from or writing to remote XRootD files is not tested yet. This may require some work within the podio and/or EICrecon source code.
3. **Modify the DDSim simulation utility to allow multithreaded simulations with Geant4:** Although both DD4hep and Geant4 can run in multithreaded mode, this functionality is not present in the DDSim utility that is used for our simulations. Modifying DDSim to run multithreaded simulations (without changing the command line interface) may require some Python and C++ programming, but there are example Python code segments for multithreaded running available.
4. **Improve performance of DIRC simulations:** The DIRC optical photon simulations are currently a performance bottleneck which prevents them from being enabled. An approximate optical photon propagation model within geant4 could be implemented to radically increase the speed of simulation, or we could try to work within the current geant4 physics model but increase minimum step sizes. This will require some experience with geant4 physics modeling, or a willingness to dive in.
5. **Split the digitization from reconstruction in production simulations:** The current simulation production model runs the digitization as a step of the reconstruction. We would like to move to a model where digitization is a separate process, with intermediate digitized output files, which is then followed by a reconstruction process. This modularization would allow other digitization approaches, and would remove the need for a reproducible pseudo-random number generator in the reconstruction code base.

Validation WG

- **Conveners**: Torri Jeske and Dmitry Kalinkin
- **Charge**: Responsible for the validation of the simulations via a suite of detector and physics performance plots. Develop autonomous checks and verification of the validation plots.
- **Priorities for 2023**
  - Implement and document our Simulation Production Strategy, together with Production WG.
  - Develop and maintain a collection of plots that showcase the performance of the ePIC detector, its physics reach, and enable comparison to a baseline or previous simulation campaigns.
  - Drive the development of unit tests for the ePIC software, together with the Development WGs.

Will be also covered by Thomas Britton (JLab)

Physics and Detector Benchmarks

- Discussion on validation before the release (e.g., CI and unit tests) and after the release (e.g., benchmarking and validation).
- Benchmarks to showcase the measurement capabilities of the ePIC detector in a specific design:
    - Verify our ability to carry out the EIC Science Program as outlined in the NAS report.
    - Compare between detector designs, e.g., our reach in low *t*, is a driver for the design.

Before the release, it was decided to:
- Prioritize software development topics and use the meeting slot for release management discussions.
- Emphasize the important role of conducting thorough code reviews for validation but also sustainability.
- Encourage volunteers to actively participate in code reviews and offer an "apprenticeship" program for new collaboration members interested in software development.
- Welcome suggestions for potential candidates for the "apprenticeship" program.

After the release, the following actions will be taken:
- Involvement of DSCs in detector benchmarks to ensure comprehensive validation.
    - Consideration of tracking performance plots as an example to start with and show to the DSCs.
    - Show examples from physics benchmarks we have received.
- Organize dedicated discussion on infrastructure for benchmarking and validation, utilizing eicweb and resources available at the host labs.

Validation WG: Open Tasks

1. **Implementing unit tests for existing reconstruction algorithms** EICrecon inherited several algorithms for reconstruction for calorimetry (central, far backward, far forward) and central tracking. Those came from the Juggler without unit tests. Volunteers with basic C++-related skills are needed to go through the algorithms, identify test cases and implement tests within Catch2 framework (some examples are provided). A volunteer with advanced C++ skill is needed to implement support for testing of algorithms that rely on DD4hep and ACTS geometries (suggested implementation involves improving EICrecon service abstraction and providing mock geometry objects). Achieving those objectives would significantly advance our ability to develop experiment's simulation and reconstruction software forward with confidence.
2. **Consolidation of detector benchmarks** The current plan is to continue evolving detector benchmarks based on ATHENA collaboration's original efforts. Original benchmarks were implemented as self contained analyses consisting of generating a small detector-specific simulation sample and analyzing it for key performance observables. Experts with knowledge of detector subsystems are needed to identify benchmarks missing or in need of an update. We intend to use benchmarking step to also extract detector calibrations and deliver those continuously in a machine-readable format. A volunteer would be needed to provide a first implementation for calorimeters or for ML based reconstruction that is intended to be used for the far backward. Detector expert is needed to survey existing detector simulations supporting the ongoing beam tests, the software needs to be published and references made available.
3. **Development of tooling for presenting CI metrics** We have metrics generated on several CI systems (GitHub Actions, Git Lab CI on eicweb) for several repositories (eic/epic, eic/EICrecon) that would need to be scraped for data or modified to deliver it actively. We are looking to find a way to track, compare and visualize those multidimensional metrics (n-dimensional histograms and profiles) along with relevant metadata (references to software versions - git commits, references to datasets - campaign dataset IDs). A volunteer with understanding of Internet protocols and Web development is needed to implement a prototype. Ideally, that would work on top of a public computing infrastructure (GitHub Pages, GitHub Actions).
4. **Consolidation of Physics benchmarks** The current plan is to continue evolving detector benchmarks based on ATHENA collaboration's original efforts. Original benchmarks were implemented as self contained analyses consisting of generating a small physics process -specific simulation sample and analyzing it for key performance observables. We are looking to extend the scope of those analyses to incorporate processing of simulation campaign input. A volunteer is needed to implement a pipeline for processing large multi-file datasets (map-reduce) into the existing analyses. Ideally, that extended system would work on top of a public computing infrastructure (OSG). Another volunteer with broad understanding of EIC physics is needed to coordinate implementation of missing physics analyses to produce summaries of ePIC detector reach and comparing it to physics goals outlined in the Yellow Report. These goals are to be achieved in close cooperation with analysis coordination and User Learning WG.
5. **Development of defensive checks for reconstruction and geometry** We are looking to further improve our development processes by implementing even more automation to catch bugs and configuration mistakes. Volunteer is needed to work on EICrecon parameter system to make it robust against configuration mistakes. A first step would involve fixing the existing UX issues of the JANA2 configuration system: silent ignoring of unexpected parameters, inability to handle whitespaces in values. We are looking to implement a consistency check to ensure compatibility between the simulation and reconstruction geometries. A basic checksumming ability was implemented in DD4hep, and can be adapted and improved to be used in the cross check. We are looking to ensure parameter consistency by propagating values along the reconstruction chain. An experienced volunteer is needed to implement storing and use of the values in the metadata: beam parameters are to be stored in HepMC3 metadata at all the afterburner/generator facilities to be propagated by ddsim into the PODIO metadata, some additional fields to be added here from ddsim configuration will need to be identified, then at EICrecon that metadata is to be propagated into the user-consumable output with addition of all configuration parameters and readout specs of all detectors (for standalone data analysis without loading of the DDhep geometry). Another volunteer is needed to complete the work on Unit System. We are looking to implement explicit units ("X * GeV") for the EICrecon parameters. We need to investigate a refactoring that would establish a consistent way to handle units systems in EICrecon (there are different numerical values used in EDM4hep and DD4hep, possibly ACTS). Possible directions for solutions are wrapped values or implementing invariancy checks (running against software compiled with different unit constants).

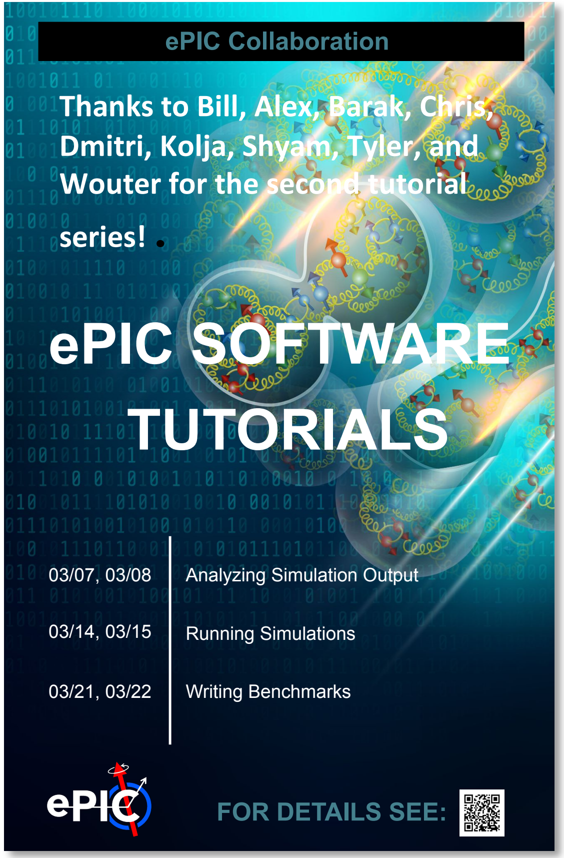
User Learning WG

- **Conveners**: Kolja Kauder and Holly Szumila-Vance
- **Charge**: 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 2023**:
  - Develop https://eic.github.io into the centralized documentation hub for ePIC Software, compiling all relevant documentation and curating a landing page that enables the collaboration to get started on software use and development.
  - Establish a regular, predictable schedule of training that includes introductory and intermediate materials. Incorporate relevant materials from the HSF Training WG.
  - Restart soon the help desk office hours, with a staffing schedule that distributes this workload.

User Learning WG: Open Tasks

1. **Develop model analysis macros in Python (uproot/awkward) and using ROOT RDataFrame, demonstrating best (or good enough) practices:** Analyzers are mostly left on their own with the simulation production data files, with little guidance. This leads to low code analysis quality which is typically not shared.
2. **Develop a code review training document:** A lot of our code quality depends on the pull request review process. We do not have a set of best (or good enough) practices for code review to help onboard new reviewers. Collecting a set of resources and combining them in a short document could help us get more code reviewers.
3. **Develop a “Day 1” checklist for new students and researchers:** While we are focusing a lot on computing tutorials, we risk overlooking other actions that we want our users to take: register for mailing lists, mattermost, etc. A periodically-reviewed checklist that all advisers can point their new researchers to would be useful.

User Learning



- Maintain a list of regularly updated reference tutorials:
  - Setting up your environment.
  - Geometry development with DD4hep.
  - Detector simulations using ddsim and Geant4.
  - Reconstruction Algorithms in JANA2
  - EIC Analysis Bootcamp
- Collected feedback from collaboration on next tutorial topics. Excellent input on FAQs.
- Actively working on landing page for onboarding new members.

Physics and Detector Simulation WG

- **Convener**: Kolja Kauder and Chao Peng
- **Charge**: Development of accurate MC simulations using a suite of physics and background generators and detector simulation based on Geant4 and DD4hep.
- **Priorities for 2023**:
  - Support the detector design and integration with services.
  - Support the needs for eA simulations.
  - Support the development of background modeling and its implementation in physics and detector simulations, together with the Background Task Force.
  - Embrace modularity and separate the simulation of the detector readout (digitization) from the reconstruction.
  - Embrace streaming readout in simulation, using it as the default data format and enabling the capability to stream the continuous readout of the detector.

Update by Kolja Kauder (BNL) in the Software & Computing Session

Physics and Detector: Open Tasks

1. **Tune simulation parameters (Geant4 region limits) and benchmark them:**
   Example: beamline components and subdetectors in far-forward region. We need to improve the performance while maintaining a high precision.
2. **Build effective models for service structure:**
   Validation through material scan (comparing to the original CAD model)
3. **Benchmark the effects/performance of the service structure:**
   Study the impact on the detector performance and the service structure's own simulation performance
4. **Implement timing information in digitization for subdetectors:** Machinery is ready and most of the subdetectors may have it. Need to coordinate with the DSC to see if any special implementation is needed.
5. **Test of different Geant4 physics lists:** Some physics lists (LHEP) is known to be fast for shower simulation with a cost of precision. Need a thorough test of different physics lists for the resource-consuming components (e.g., beamlines in FF) to achieve a good performance with a high precision.

Reconstruction Framework and Algorithms WG

- **Conveners**: Derek Anderson and Shujie Li
- **Charge**: Development of a holistic and modular reconstruction for the integrated ePIC detector.
- **Priorities for 2023**:
  - Enforcing modularity for clear separation between the development of reconstruction algorithms and the development of the framework and its services.
  - Embrace algorithmic development that utilizes the holistic information from detector components or the entire detector.
  - Integrate far-forward and far-backward detectors in reconstruction.
  - Implement a web-based event display.

Update by Derek Anderson (ISU) in the Software & Computing Session

Reconstruction WG: Open Tasks

1. **[Particle Flow]** Validation of existing cluster-splitting
2. **[Particle Flow]** Validation of existing MC-cluster associations
3. **[Particle Flow]** Development of cluster/track visualizers
4. **[Particle Flow]** Extending track-cluster associations to HCal
5. **[Particle Flow]** Implementation of PF algorithm + factories
6. **[Jets]** Enabling user-configured jet finding parameters
7. **[Jets]** Enabling proper PODIO jet-constituent associations

Software Progress

Monthly tagged releases since April 2023!

Updates in July Release:

- **Detector Simulation**: *service integration procedure (Denali), low Q2 tagger, new tracker configuration (Craterlake), magnetic fields in FF, updates to imaging calorimeter geometry and LFHCAL*
- **Streaming Readout Simulation**: *Time information propagation to simulate streaming*
- **Reconstruction**: New targeted task forces (Electron Finder, Hadron Identification, Jet Reconstruction, Realistic Seeding, TOF): *Foundational work to enable progress towards August-September*

Updates in May Release:

- **Detector Simulation**: *Improved realism of ePIC detector geometry (default to Brycecanyon, hpDIRC improvements, updates to silicon tracker, ZDC baseline, forward electron beampipe)*
- **Reconstruction**: *added EcalLumiSpecCal reconstruction, cluster splitting for the central detector, ACTS iterative vertex finder*

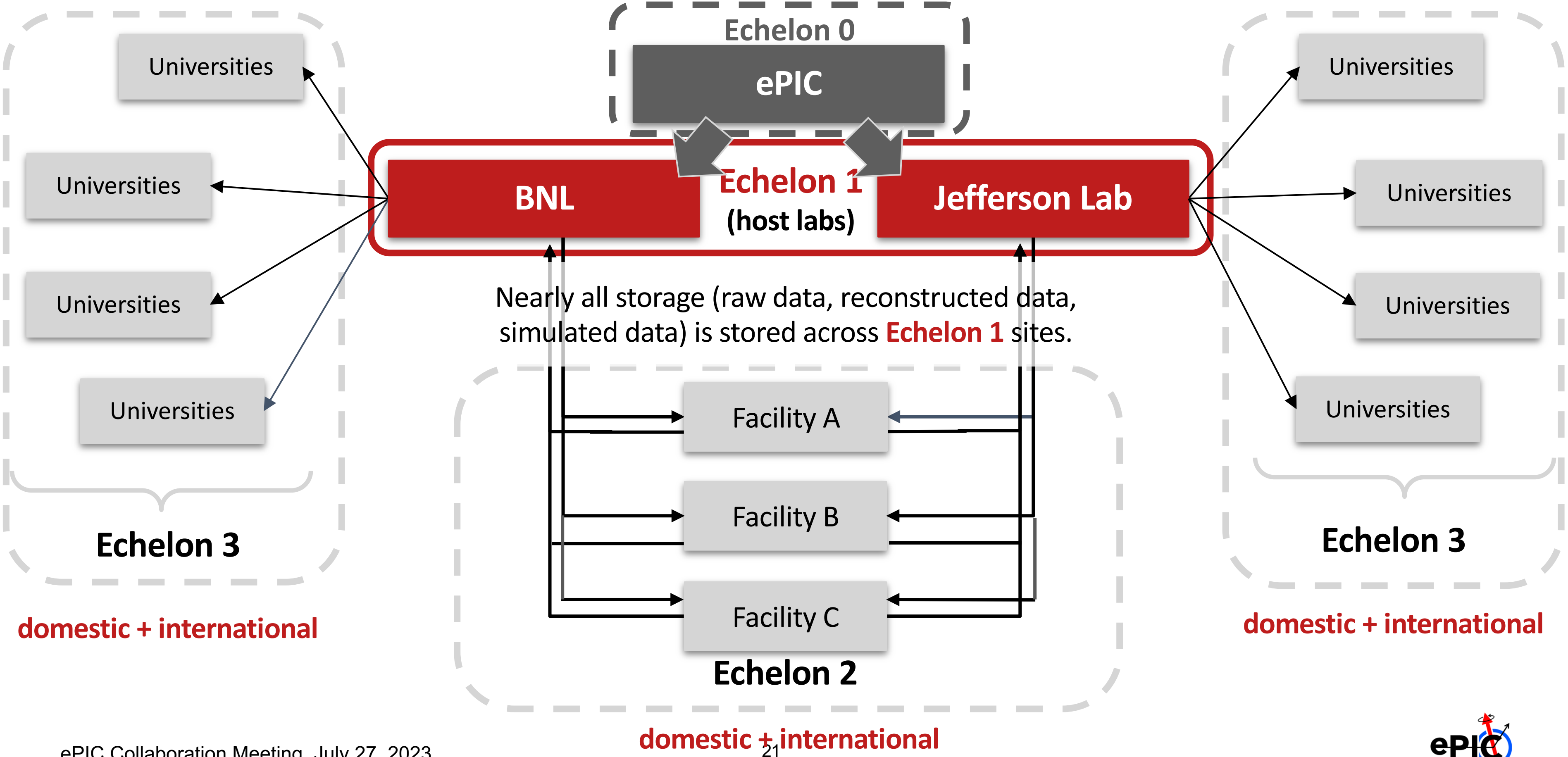
Updates in April Release (includes work for many months):

- **Detector Simulation**: *Too many to list*
- **Reconstruction**: *Too many to list*

Streaming Computing Model WG

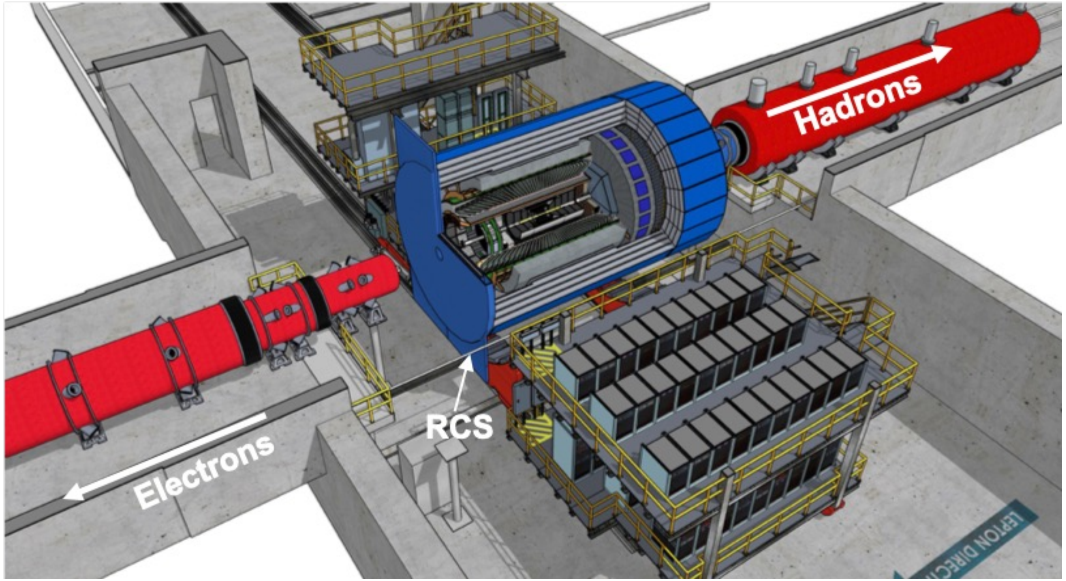
- **Convener**: Marco Battaglieri, Jin Huang, Jeff Landgraf (interim)
- **Charge**: Development of the computing model for the compute-detector integration using streaming readout, AI/ML, and multi-architecture computing (CPU, GPU, …) with a specific focus on the data flows after the FEE layer.
- **Priorities for 2023**:
  - Establish a collaborative dialogue with the Electronics and DAQ WG.
  - Define the requirements and high-level design for a computing model that enables rapid processing for the data for physics analyses, while leveraging external compute resources, including those provided by international partners.
  - Coordinate activities on prototyping streaming computing systems, together with the Physics and Detector Simulation and Reconstruction Framework and Algorithms WGs.
  - Document a streaming computing model that can be redefined further with international partners.

Distributed Computing Model



Optimize Physics Reach

Integrated interaction and detector region (+/- 40 m)
Get ~100% acceptance for all final state particles, and measure them with good resolution. All particles count!



Lessons learned in the NHEP community:

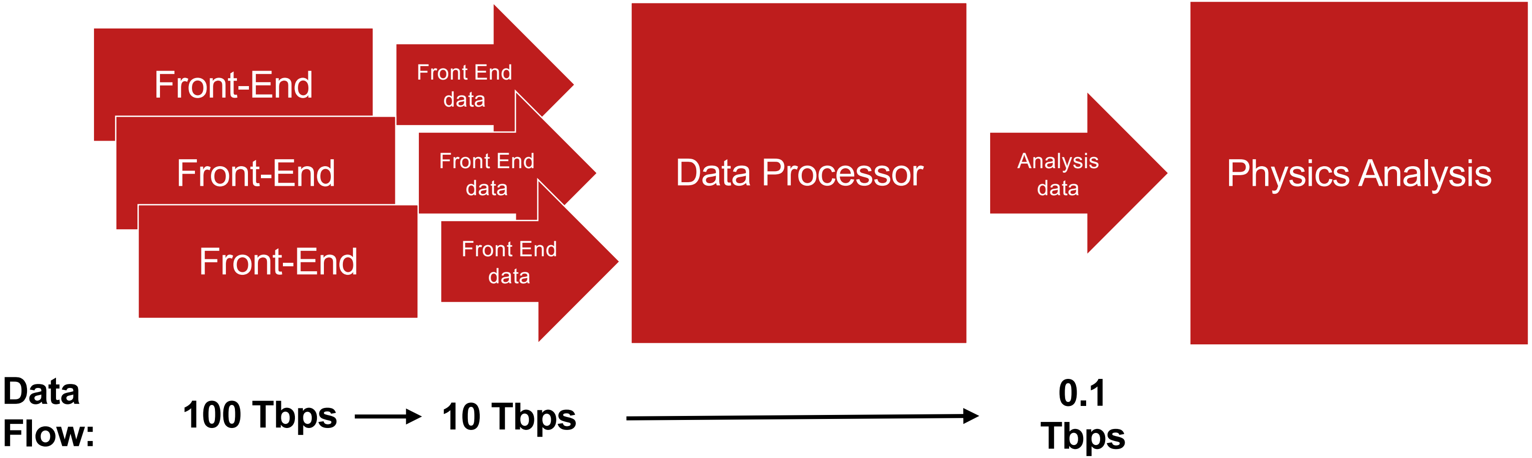
Software & Computing are an integral part of the experiment.

Develop computing model hand in hand with the detector.

Compute-Detector Integration
Extend integrated interaction and detector region into detector readout (electronics), data acquisition, data processing and reconstruction, and physics analysis.

Compute-Detector Integration to Maximize 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 data for physics analyses.
- **Solution** Compute-detector integration using:
  - AI/ML for autonomous alignment and calibration as well as reconstruction in near real time,
  - Streaming readout for continuous data flow and heterogeneous computing for acceleration.



Computing Model Discussion

Context:

- ePIC has been asked to present an **update on the computing model at the EIC RRB** in December, describing how international partners can contribute to computing for the EIC.
- The ePIC Streaming Computing Model WG and SCCs will guide the **discussion on the computing model**.
- In our discussions, we would like to the involve the ePIC collaboration at large as well as the host labs and other computing experts worldwide.
- Prior to the EIC RRB in December, the host labs will organize a review of the computing model. The review will be likely on October 19–20.
- We aim to publish the ePIC computing model and have a draft ready for the review.

Kickoff the Discussions:

- We have discussed unique requirements of **computing models that feature streaming**.
- We have discussed the **status of the ePIC Computing Model**.
- On August 8, we will discuss the event and data sizes from Echelon 0 (ePIC Detector)
- Next, we will discuss the workflows for data productions, physics analysis, and simulation.
- We will hear **perspectives from the international community**.
- **Focus: EIC RRB in December and the role of international partners in computing.**

Essential to gather feedback from the collaboration at large.

Mark Your Calendars!

ePIC AI Town Hall Meeting

- **Date and time**: Wednesday, August 16, at 11:00 a.m. (EDT)
- Showcase AI/ML projects and inspire work on AI/ML in ePIC.
- Insight into various AI/ML activities within the collaboration.

ePIC Software & Computing Meeting at UIC

- September 20–22
- https://indico.bnl.gov/event/20159/
- **Thanks to Olga for hosting us!**

Summary

- We have a **modular simulation, reconstruction, and analysis toolkit** for the development of the our detector and science program.
- The toolkit is based the '**Statement of Software Principles**' and a **decision-making process involving the whole community**.
- The collaboration has approved our WG structure, and we have defined for priorities for 2023.
- We have implemented an **simulation campaign strategy**, enabling collaborative release management with well-defined timelines and continuously deployment of our software.
- **We have a good foundation to meet the near-term and long-term software needs for ePIC.**

- A priority for the next months is the **development of the ePIC Computing Model**.
- We have had first successes with onboarding new members, but to ensure a sustainable ePIC Software & Computing effort, we **need additional workforce and support**.