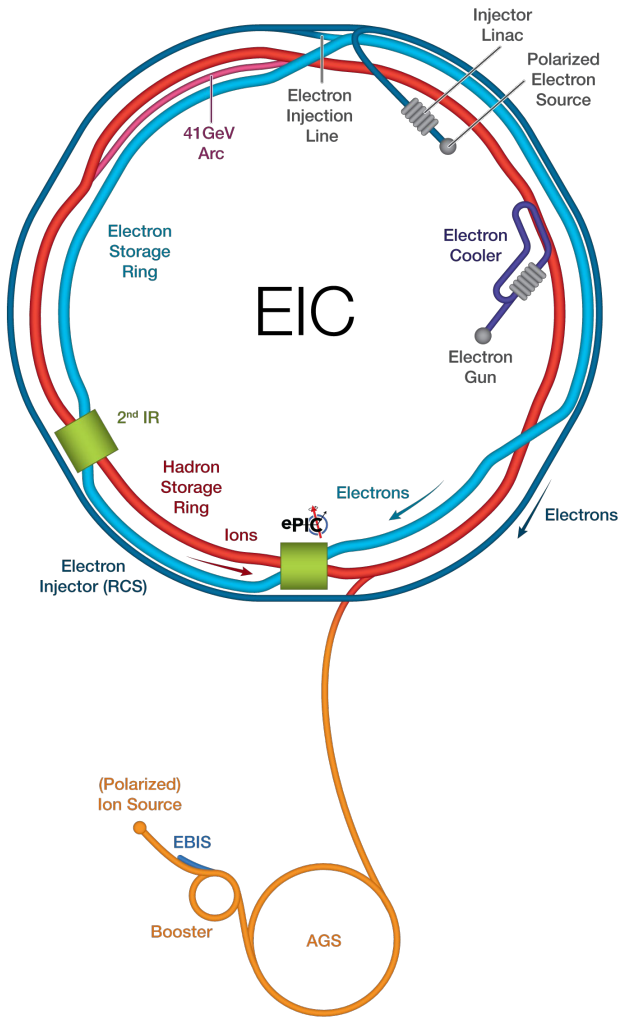
Thomas Britton (Jefferson Lab)
on behalf of Anil Panta and the ePIC collaboration

Integrating Rucio for Advanced Data Management at the ePIC Experiment

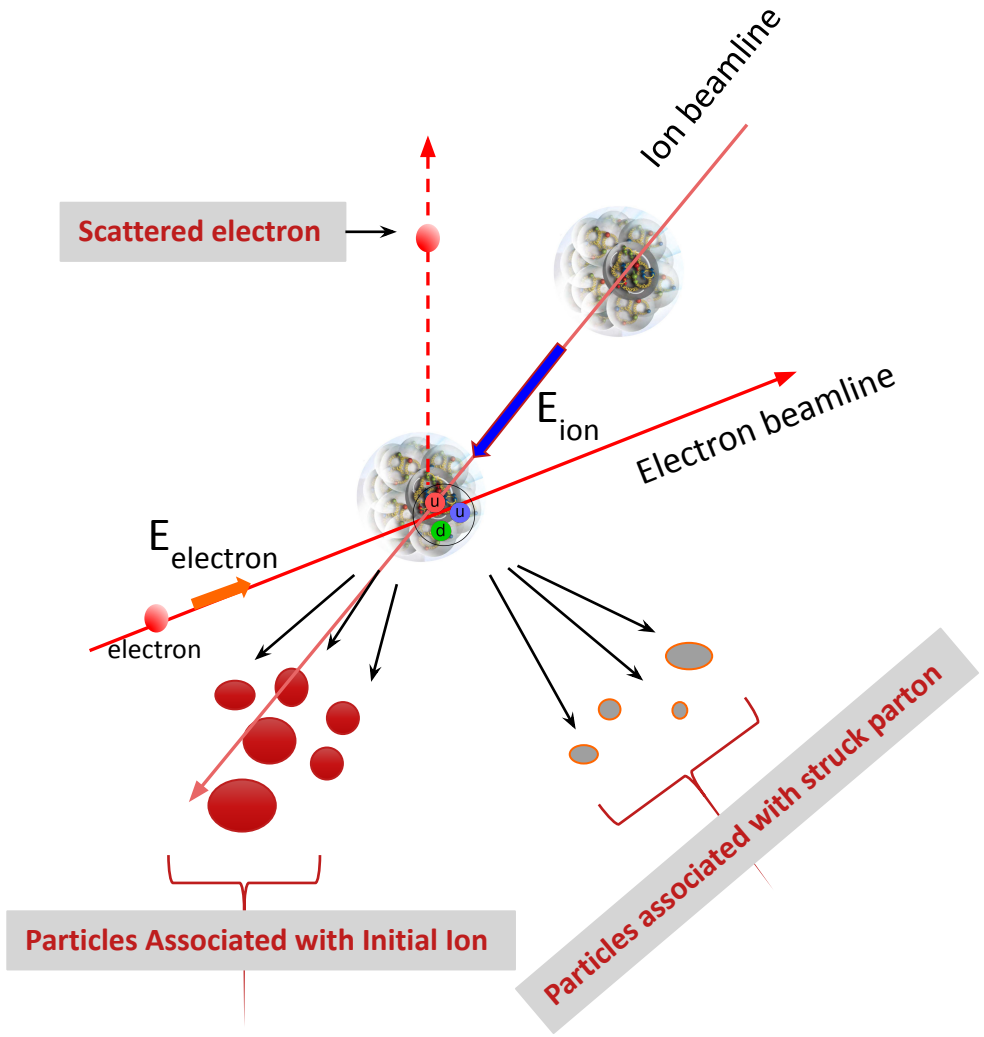
The Electron-Ion Collider (EIC)



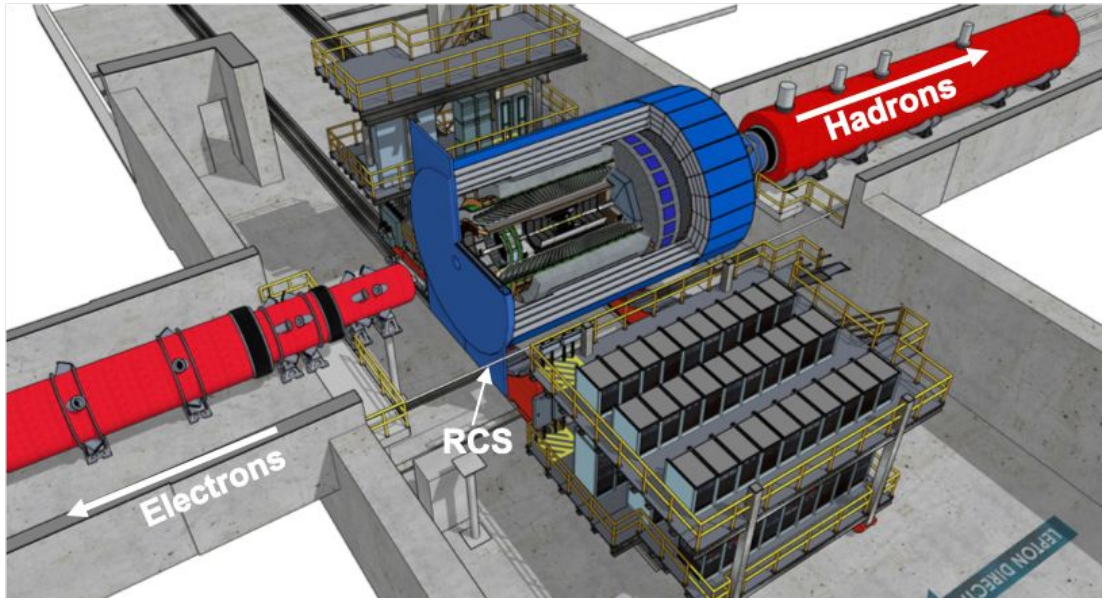
Frontier accelerator facility in the U.S.

- **World's first collider of**:
  - Polarized electrons and polarized protons,
  - Polarized electrons and light ions (d, ^{3}He),
  - Electrons and heavy ions (up to Uranium).
- **High luminosity (**100 to 1000 times HERA luminosity**)** and **versatile range of center of mass energies** (20 GeV – 140 GeV), **beam polarizations** (longitudinal, transverse, tensor), and **beam species** (e; p ⟶ U).
- The EIC will enable us to embark on a **precision study of the nucleon and the nucleus at the scale of sea quarks and gluons**, over all of the relevant kinematic range.
- The **EIC Yellow Report (**Nucl.Phys.A 1026 (2022) 122447) describes the physics case, the resulting detector requirements, and the evolving detector concepts for the experimental program at the EIC.
- BNL and Jefferson Lab will be the host laboratories for the EIC Experimental Program. Leadership roles in the EIC project are to be shared.

General Purpose Detector for ePIC



Integrated interaction and detector region (+/- 45 m)
to get ~100% acceptance for all final state particles, and measure them with good resolution.



Overall detector requirements:

- Large rapidity (-4 < h < 4) coverage; and far beyond in the far-forward detector regions.
- Large acceptance solenoid of 1.7 T (up to 2 T).
- High control of systematics: luminosity monitor, electron and hadron polarimetry.

Compute-Detector Integration to Maximize Science

Broad ePIC Science Program:

- Plethora of observables, including less distinct topologies where every event is significant.
- High-precision measurements: Reducing systematic uncertainties is very important.

Streaming Readout Capability Due to Moderate Signal Rate:

- **Capture every collision signal**, including background.
- Event selection done 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.

	EIC	RHIC	LHC → HL-LHC
Collision species			
Top x-N C.M. energy	140 GeV	510 GeV	13 TeV
Peak x-N luminosity	10^{34} cm^{-2} s^{-1}	10^{32} cm^{-2} s^{-1}	$10^{34} \rightarrow 10^{35}$ cm^{-2} s^{-1}
x-N cross section	**50 µb**	**40 mb**	**80 mb**
Top collision rate	500 kHz	10 MHz	1-6 GHz
$dN_{ch}/d\eta$	0.1-Few	~3	~6
Charged particle rate	**4M N_{ch}/s**	**60M N_{ch}/s**	**30G+ N_{ch}/s**

Compute-Detector Integration to Accelerate Science

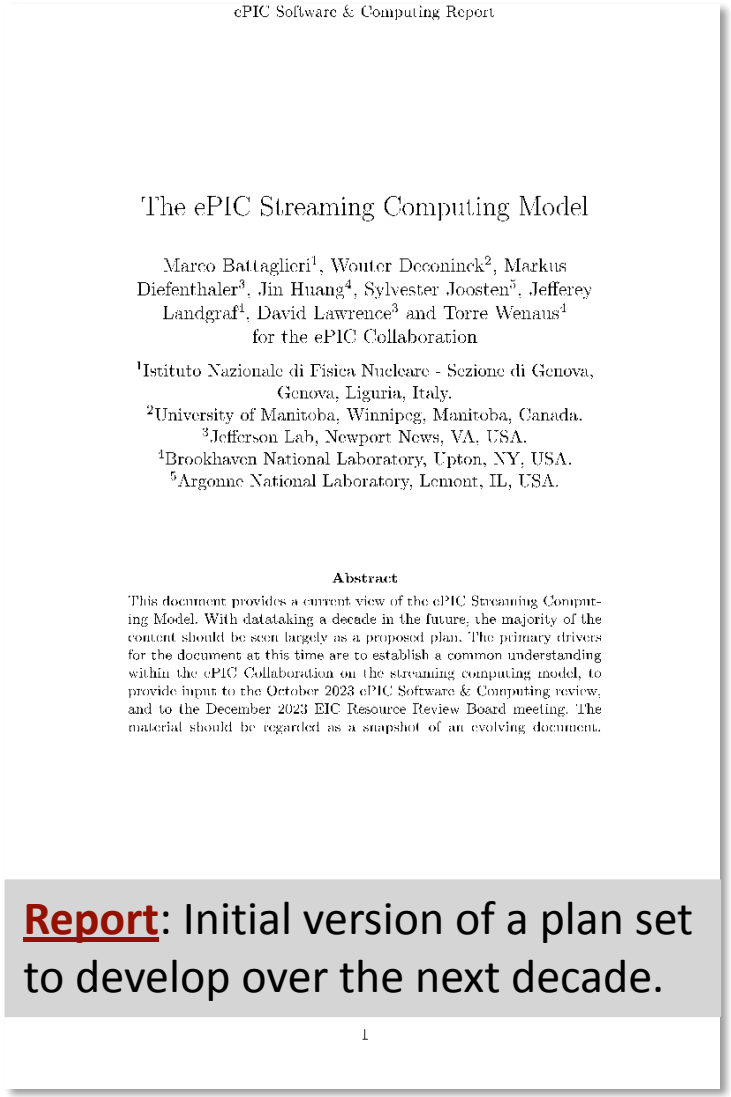
- **Problem** Data for physics analyses and the resulting publications available after O(1year) due to complexity of NP experiments (and their organization).
  - Alignment and calibration of detector as well as reconstruction and validation of events time-consuming.
- **Goal** **Rapid turnaround of 2-3 weeks for data for physics analyses**.
  - calibrations are the limiting reagent.
- **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/ distributed computing for acceleration.

ePIC Streaming Computing Model



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

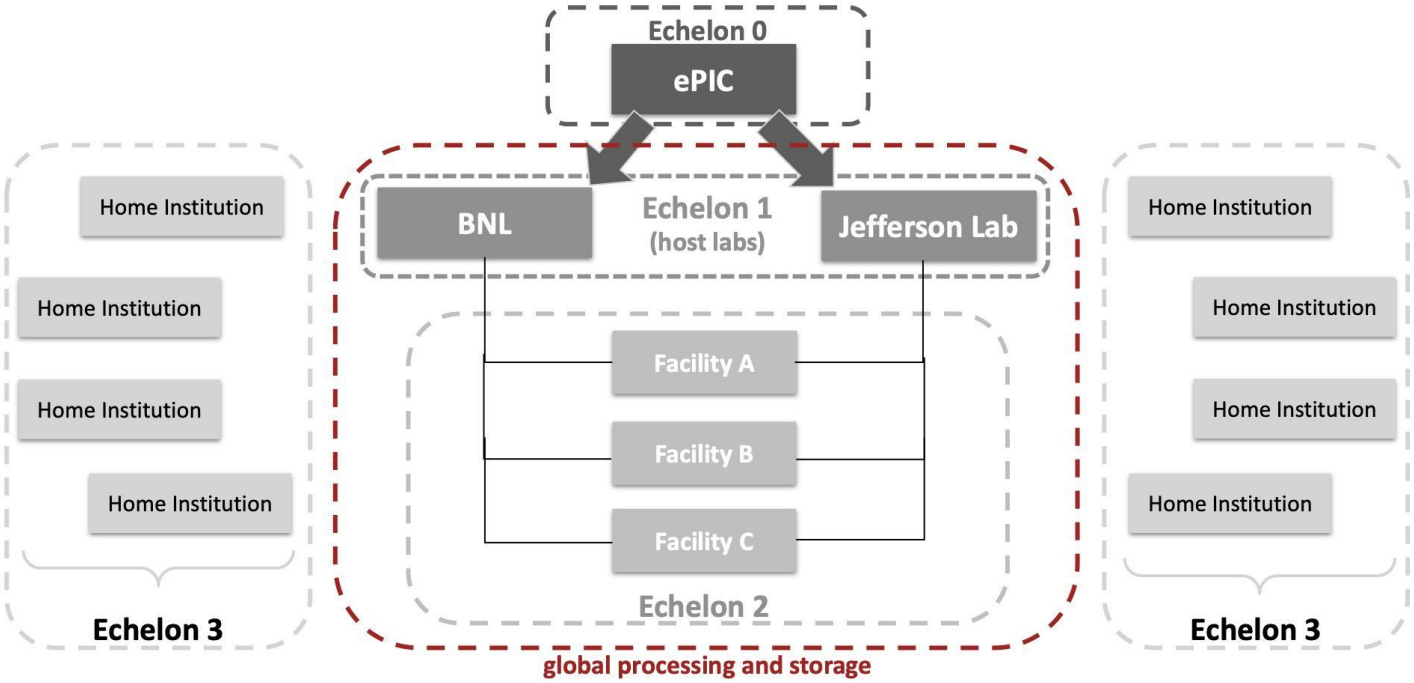
[1]Istituto Nazionale di Fisica Nucleare - Sezione di Genova, Genova, Liguria, Italy.
[2]University of Manitoba, Winnipeg, Manitoba, Canada.
[3]Jefferson Lab, Newport News, VA, USA.
[4]Brookhaven National Laboratory, Upton, NY, USA.
[5]Argonne National Laboratory, Lemont, IL, USA.

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.

1

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



Echelon 0: ePIC experiment.

Echelon 1: Crucial and innovative partnership between host labs.

Echelon 2: Essential global contributions.

Echelon 3: Full support of the analysis community.

ePIC Collaboration

Formed in 2022/2023

ePIC Collaboration Meeting at Jlab in January 2023.

ePIC Collaboration Meeting in Warsaw in July 2023.

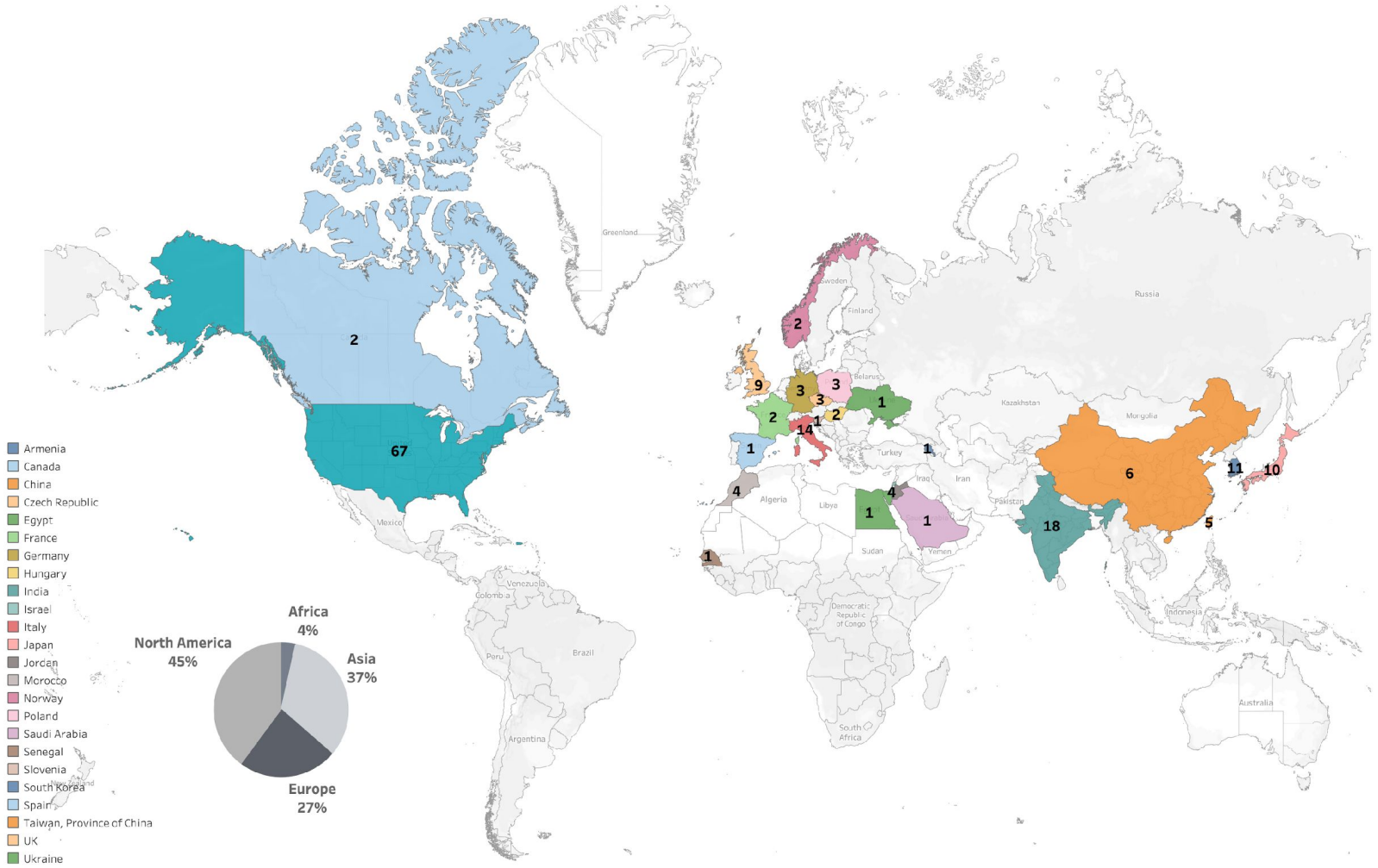
ePIC Collaboration Meeting at ANL in January 2024.

ePIC Members
> 850

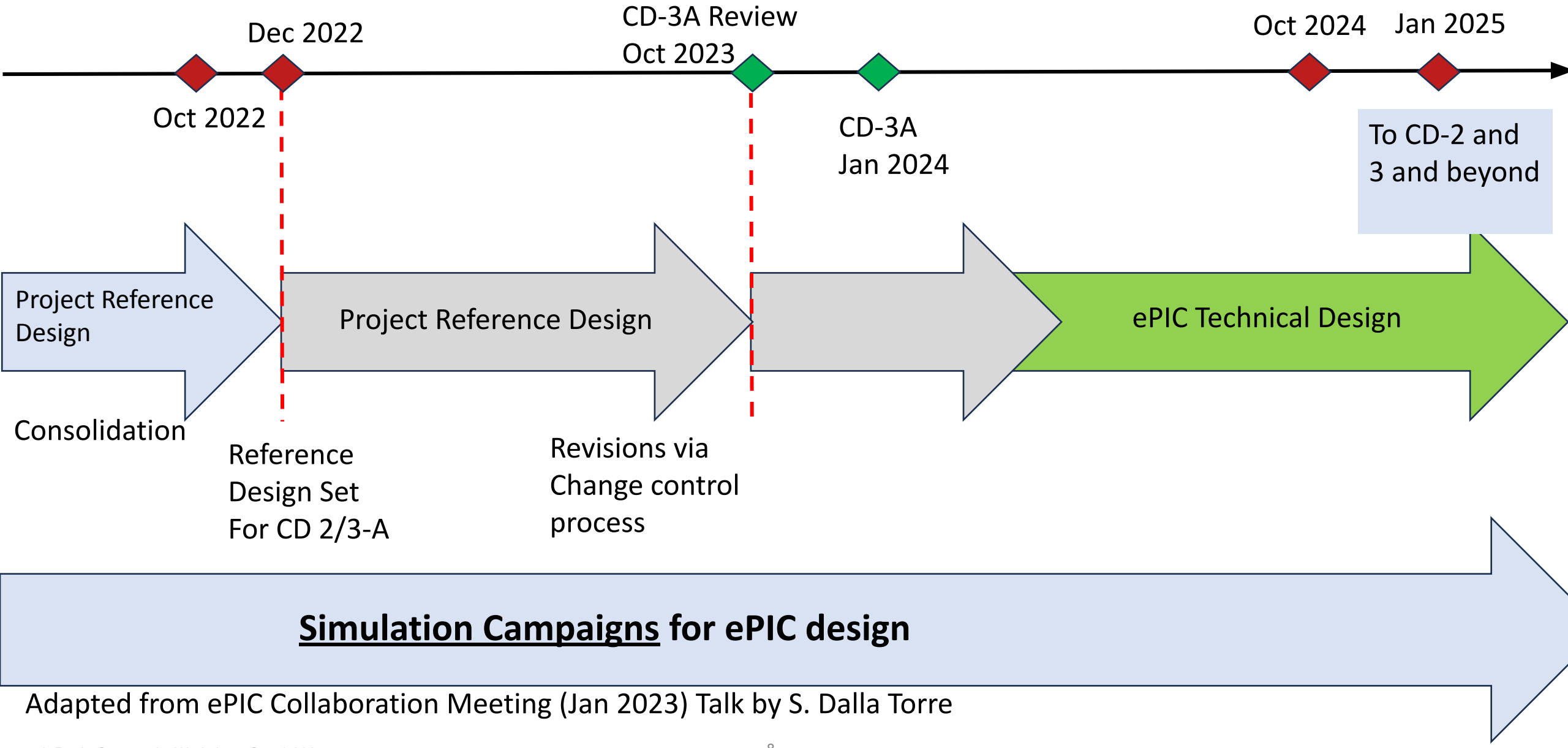
ePIC Institutions
173

ePIC Countries
25

ePIC World Regions
4



ePIC Detector Consolidation and Optimization Process



Adapted from ePIC Collaboration Meeting (Jan 2023) Talk by S. Dalla Torre

Simulation Production

- Major central campaigns (**Trains**)
  - Run without user intervention
  - Runs on time
    - Within our control
  - Pre-established “flavors”
- Special interest runs (**Charters**)
  - Working Group level requests
  - Needed for design/development
- Bespoke (**Taxis**)
  - Individual user needs (think specific analyses)

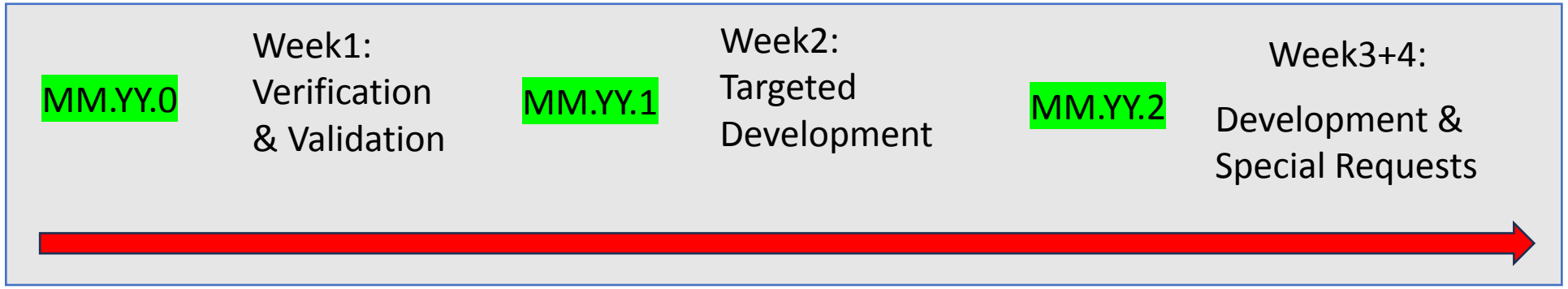
Jefferson Lab

Simulation Production

- Monthly cadence of production
  - Run on the **Open Science Grid**

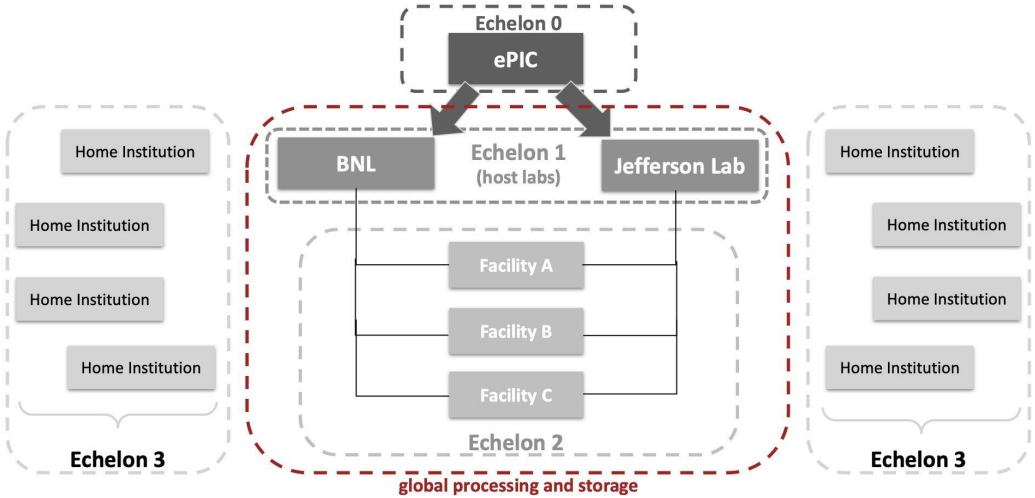
- 1 default detector configuration but multiple test configurations are possible based on demand
- Benchmarked core year estimates for different campaigns for default config:
  - MM.YY.X ~ 100 coreyears
- Total of ~15-20 TB a month.

Monthly Campaigns Strategy



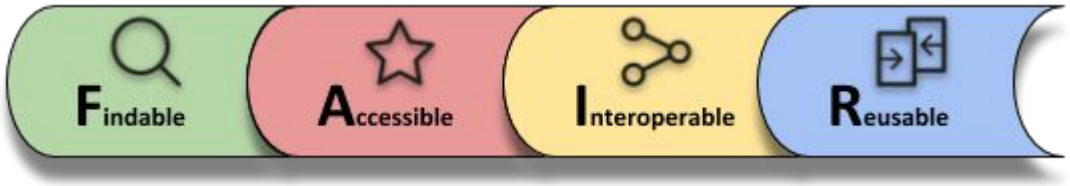
The Problem

- How do we get all of this data into the hands of collaborators world wide?!
  - Many underlying storage elements
    - many different data sets
  - ePIC desires a transparent method for users
  - Have need for both data cataloging and access/transfer



Introducing Rucio

- Why Rucio?
  - **Scalable**: can handle the growing data management needs of ePIC
  - **Policy Driven**: Different data/users may require different management policies. We can spend time figuring out how we want to manage the data not how to technically manage the data
  - **Aligned principles**



Roadmap

- Stand up test instance of Rucio at JLab
- Coordinate a similar setup at BNL
- Successfully initiate bidirectional transfers
- Develop policies for managing ePIC data
- Migrate older datasets to Rucio
- Roll out to production

Roadmap

- ✓ **Stand up instance of Rucio and FTS3 at Jefferson Lab**
  - Rucio, FTS, XrootD production and test instances deployed

- ✓ **Coordinate storage element setup at Brookhaven National Laboratory**
  - XrootD instance up and running!

- ✓ **Successfully initiate bidirectional transfers with Rucio/FTS3**
  - x509 authentication
- ✓ **Develop policies for managing ePIC data**
  - Initial draft under review
  - Rucio Policy package created as a pip package.
  - Created customized Rucio server, daemon container with the policy package.
    - Will deploy once the policy is finalized.
- **Migrate older datasets to Rucio**
- **Roll out to production**

Ongoing Work: CILOGON TOKEN

- We are currently testing token authentication with CILogon.
  - CILogon provides a WLCG-profile token.

- Rucio with CILogon:
  - For Client authentication - Rucio uses Code Flow.
    - For Code Flow to work with CILogon: we need https://github.com/rucio/rucio/issues/6630 resolved to start full testing
      - In user authentication the IAM audience parameter needs to be made optional as CILogon does not have such a parameter
        - removing parameter confirmed working by JLab and a member of DUNE
  - For transfers- Rucio uses Client Credential Flow
    - **CILogon doesn't provide Client Credential Flow.**
    - We could use **User Token Exchange** flow with CILogon…
      - Has this been disabled? (a discussion has been started at https://github.com/rucio/rucio/discussions/7108)

Conclusion

- **ePIC is the EIC's first experiment**
  - comprised of over **850 members** (and growing), from over 170 institutions, across 25 countries

- ePIC has a monthly cadence of simulations
  - **~20TB of data per month** on top of individual analyses and work (and eventually real data!)

- It is challenging to manage all this data and make it easily accessible internationally in a way that is scalable and transparent to the users

- The solution is **Rucio**!
  - Efforts are underway to deploy Rucio at ePIC
    - Test instances working
    - Currently working out the issues with authentication
    - Need to migrate old data

Backups

JLAB TAPE with XrootD

- **GOAL: JLAB TAPE as a grid storage element (SE) endpoint**

- Work has already been done to integrate JLAB tape with the Grid storage workflow
  - All integration complete.
    - *Authentication is still x509. (we will revisit token implement)

- Rucio/FTS3 test to archive, bring-online and transfer from/to TAPE is working.

- This will be necessary when we start data taking.